JOURNÉES OUVERTES
DE BIOLOGIE
INFORMATIQUE
& MATHÉMATIQUES

28 > 30
JUIN
ENS LYON

JOBIM 2016

En présence de :
Stein Aerts (Belgique)
Isabelle Callebaut (France)
Robert Gentleman (USA)
Louis Lambrecht (France)
Jeroen Raes (Belgique)
Alexandros Stamatakis (Allemagne)

Génomique des populations, Ecologie,
Biologie des systèmes et réseaux moléculaires
Phylogénie, Evolution
Statistique & Algorithmique pour la biologie à haut débit

2016

Préface

C'est à Lyon qu'a lieu la dix-septième édition des Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM). Par ailleurs, c'est la deuxième fois dans son histoire que cette conférence est organisée dans la capitale des Gaules. Depuis la première édition, organisée à Montpellier en 2000, JOBIM se veut un lieu de partage et d'échange pour la communauté francophone des bioinformaticiens et, comme chaque année, ces journées sont placées sous l'égide de la Société Française de Bioinformatique (SFBI).

Cette année, quatre évènements satellites sont associés à la conférence : la journée des Jeunes Bioinformaticiens de France (JeBiF), le workshop Statistique, Informatique et Modélisation pour l'Oncologie (SIMOnco), le séminaire du Groupe de Travail en Génomique Comparative (GTGC) et la réunion du groupe de travail sur la Biologie Systémique Symbolique (BioSS).

JOBIM est également l'occasion de la tenue d'assemblées générales d'organismes impliqués dans la structuration et l'animation de la recherche en bioinformatique. Cette année, JOBIM accueille ainsi les assemblées du Groupement de Recherche en Bioinformatique Moléculaire (GdR BiM), de l'Institut Français de Bioinformatique (IFB) et de la SFBI.

Une des caractéristiques historiques de JOBIM est la diversité des thématiques qui y sont abordées. Cette diversité se reflète tout d'abord dans la liste des conférenciers invités qui sont, dans l'ordre des interventions : Robert Gentleman, Alexandros Stamatakis, Louis Lambrecht, Jeroen Raes, Stein Aerts et Isabelle Callebaut. Pour la première fois dans l'histoire de ces journées, nous avons décidé de mettre en place un mode de soumission unique pour les communications orales et affichées. Notre intention était de permettre à un maximum de personnes d'avoir la possibilité de présenter leurs travaux. Le comité de programme a ainsi sélectionné 48 communications, 7 démonstrations et 130 posters. Nous en profitons pour remercier chaleureusement les relecteurs pour leur travail d'évaluation.

Pour la première fois également, les actes de la conférence sont disponibles sous forme électronique uniquement. Le lecteur y retrouvera les communications sur les thèmes habituels de la conférence, regroupés en dix grandes catégories : statistique des données à haut débit, phylogénie et évolution, analyse de séquences nucléiques et protéiques, génomique et génétique des populations, métagénomique et métabarcoding, études d'association, biologie des systèmes, bioinformatique structurale, bioinformatique pour la santé et données de cellules uniques.

Ce congrès n'aurait pu avoir lieu sans le soutien de nos partenaires institutionnels, scientifiques et industriels. La page qui leur est consacrée sur le site de la conférence [*], montre à quel point ils sont nombreux. Nous voudrions remercier tout particulièrement les principaux donateurs institutionnels qui sont le CNRS, l'ENS de Lyon, l'IFB, l'INRA, et la Région Auvergne Rhône-Alpes.

Il nous reste à vous souhaiter un bon colloque et de nombreux échanges, que vous revivrez en feuilletant (électroniquement) ce volume. Celui-ci restera, avec les précédents, la trace de la recherche bioinformatique française dans les années 2010.

Guy Perrière et Franck Picard, pour le comité de programme de JOBIM 2016.

*. http://jobim2016.sciencesconf.org/resource/sponsors

Présentations orales

Session Statistique des données à haut débit
mardi 28 09h30
Amphi Mérieux

Adventures in mRNA splicing

Robert Gentleman [†1]

[1] 23andMe – 23andMe, Mountain View, Californie, États-Unis

Typical short-read sequencing provides extensive detail about local mRNA splicing events. In this talk I will highlight recent work with L. Goldstein [‡] on mRNA spicing detection in cancer. Splicing is also interesting in the study of human genetics. Rather than rare somatic mutations, polymorphisms arise in the population that can either create or degrade donor and acceptor sites. Evidence of these and how they affect our interpretation of Genome Wide Association Studies (GWAS) will also be presented (work with B. Alipanahi at 23andMe).

†. Intervenant Invité
‡. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156132

Supervised multivariate analysis for biological data integration, dimension reduction and feature selection with mixOmics

Florian Rohart *[1]

[1] The University of Queensland Diamantina Institute (UQDI) – Translational Research Institute, QLD 4102, Australia/Australie

Session Statistique des données à haut débit
mardi 28 11h00
Amphi Mérieux

The statistical analysis of 'omics data ('omics referring to the analysis of data acquired from new biological technologies ending in -omics), such as genomics for the study of genomes, transcriptomics for the study of transcriptomes or proteomics for the study of proteins, is a challenge because of the high dimensionality of the data and the presence of highly correlated features (e.g. genes, transcripts). Exploration and analysis of these data are usually performed on each 'omic data independently through unsupervised or supervised multivariate statistical methodologies. The former includes methods such as Principal Component Analysis (PCA) while the latter includes Partial Least Square - Discriminant Analysis (PLS-DA, Barker and Rayens 2003). Although integration of multiple type of 'omics data can give more insight into complex biological systems, such statistical analysis remains a challenge due to the heterogeneity of the 'omics data. Since 2009, we have developed the R package mixOmics (`http://mixomics.org/`, and available on CRAN) which provides data analysts with an extensive range of multivariate statistical methodologies to analyse, visualise and extract relevant information from the integration of large data. Depending on the analytical question that is asked, one or several methodologies can be applied. The package is devoted to 'omics data analysis but can also be applied in several other fields where data integration is required.

In this presentation, I will present the latest improvements developed for mixOmics, with a specific focus on supervised analysis and feature selection for homogeneous (mixMINT) and heterogeneous (mixDIABLO) data integration. Those latest improvements are variants of the Partial Least Square - Discriminant Analysis approach, which we first describe.

PLS-DA is an extension of Partial Least Square to classification frameworks through a dummy matrix transformation approach. Briefly, a dummy matrix is used to explicit which sample belongs to which class. PLS was developed by Wold (1966) and initially applied as an econometric technique before becoming popular in chemometrics. Applications of PLS has increased over recent years as this multivariate dimension reduction method can handle large and highly correlated data. PLS-DA is a multivariate iterative method that constructs H successive artificial components that are linear combination of the original variables; the vector of weights in the linear combination is called loading vector. Resulting from the PLS-DA algorithm, each sample is assigned H scores that are calculated as the projection of this sample on each of the H PLS-components. The sample's scores are the new coordinates of a sample in a small subspace spanned by the PLS-components. These scores are used in sample plot representation.

Le Cao et al. (2011) developed a sparse extension based on L1 penalisation (sPLS-DA). sPLS-DA identifies relevant biomarkers (e.g. genes) on each PLS-component and classifies the samples based on this biomarker panel; this further improves classification accuracy and interpretability of the results. sPLS-DA has been a core function of the mixOmics package for several years and has been applied in several studies from our collaborators and elsewhere. Recently, we showed in Shah et al (2015) that a a combination of proteins selected with sPLS-DA led to a higher diagnostic

*. Intervenant

power than considering each protein individually. We also concentrated our efforts in developing user-friendly web interfaces using Shiny.

We described below our latest improvements for data integration.

Combining homogeneous data, e.g. combining multiple transcriptomics study that focus on the same biological question, is made complicated by the so-called batch-effects. Batch-effects are usually due to different commercial platforms being used for each study, different protocols, etc. This technical and unwanted variation in the data is especially evident on a Principal Component Analysis sample plot where the samples cluster by study instead of by biological phenotype. Several methodologies exist to accommodate for batch-effects but most of them are limited because they can not be included in an unbiased prediction framework (as all samples are used to remove the unwanted variation).

mixMINT is a new framework that we developed to perform homogeneous integration in a classification context. MINT is a multivariate integrative method that integrates independent studies while selecting the most relevant and discriminant variables to classify samples and predict the class of new samples. In MINT, the PLS-components of each study are constraint to be built on the same loading vectors in X and Y (called global loading vectors). We apply this novel method to two case study that include independent test sets and we show that MINT outperforms its concurrent in terms of accuracy (highest classification on a learning set), reproducibility (highest classification on a test set) and computational time. Implemented in mixOmics, sample plots representation of the combined studies are available to assess the overall classification accuracy while study-specific outputs are available to benchmark a study against others, which can serve as a quality control step. A first study involves 15 independent studies for a total of 342 samples and 13,313 genes, and aims to discriminate three human cell types. A second study involves four cohorts of breast cancer patients for a total of 3,068 samples and 15,755 genes, and aims to discriminate the 4 subtypes Basal, HER2, Luminal A and Luminal B. I will present the results of the second study, which highlights a new subset of genes that best explains the cancer subtypes.

DIABLO (Data Integration Analysis for Biomarker discovery using a Latent variable approach for Omics studies) is a classification framework that we recently developed to perform heterogeneous integration, e.g. combining different type of 'omics data with matching samples in order to classify samples (e.g. breast cancer subtypes). DIABLO is the first methodology that identifies a discriminative subset of correlated variables that are predictive of multiple groups of subjects using multiple omics datasets. DIABLO is a multivariate classification method based on the Sparse Generalized Canonical Correlation Analysis (sgcca, Tenenhaus 2014) approach, which extends the PLS framework to the analysis of more than two datasets. DIABLO substantially improved sGCCA for a supervised framework and enhances prediction. The implementation of DIABLO in mixOmics provides users with several graphical outputs that depicts the correlation structure among datasets. This includes sample plot representation of each Omics data, heatmap and circos plot that highlights the correlation between features across datasets.

We apply DIABLO to integrate heterogeneous 'omics breast cancer data and I will show that DIABLO improves the classification accuracy compared to independent omics analysis.

Reference

Barker, M. and Rayens, W. (2003). "Partial least squares for discrimination". *Journal of chemometrics*, 17(3):166–173.

AK Shah, K-A. Lê Cao, E Choi, D Chen, B Gautier, D Nancarrow, DC Whiteman, NA Saunders, AP Barbour, V Joshi and MM Hill (2015). "Serum glycoprotein biomarker discovery and qualification pipeline reveals novel diagnostic biomarkers for esophageal adenocarcinoma". *Mol Cell Proteomics*, 14(11):3023-39.

A. Tenenhaus, C. Phillipe, V. Guillemot, K-A. Lê Cao, J. Grill, V. Frouin (2014). "Variable selection for generalized canonical correlation analysis". *Biostatistics*, 15(3):569-83.

H. Wold (1966). "Estimation of principal components and related models by iterative least squares". *J. Multivar. Anal.*, 391-420.

Mots clefs : multivariate, dimension reduction, integration, grande dimension, R, mixOmics

Eigen-Epistasis for detecting Gene-Gene interactions in GWAS

Virginie Stanislas [*] [†1], Christophe Ambroise[1], Cyril Dalmasso[1]

[1] Laboratoire de Mathématiques et Modélisation d'Évry (LaMME) – ENSIIE, CNRS : UMR8071, Université Paris-Est Créteil Val-de-Marne (UPEC), Université Paris V - Paris Descartes, Institut national de la recherche agronomique (INRA), Université d'Évry-Val d'Essonne – 23 bvd de France, F-91 037 Évry, France

Session Statistique des données à haut débit
mardi 28 11h20
Amphi Mérieux

Genome Wide Association Studies (GWAS) aim at finding genetic markers associated with a phenotype of interest. Typically, hundreds of thousands of single nucleotide polymorphism (SNP) are studied for a limited number of individuals using high-density genotyping arrays. The association between each SNP and the phenotype is usually tested by single marker approaches. Multiple markers may also be considered but are typically selected with simple forward selection methods. GWAS represent a powerful tool to investigate the genetic architecture of complex diseases and have shown success in identifying hundreds of variants. However, they have explained only a small part of the phenotypic variations expected from classical family studies. Indeed complex diseases may also partly result from complex genetic structures such as multiple interactions between markers, known as epistasis.

In past years, numerous methods have been proposed for studying epistasis and have been reported in various reviews [Niel et al., 2015; Wei et al., 2014; Steen, 2012]. They vary in terms of data analysis (genome-wide or filtering) and statistical methodology (Bayesian, frequentist, machine learning or data mining). Most of them focus on single-locus interactions, but considering interactions at gene level may offer many advantages. Firstly, as genes are the functional unit of genome, results can be more biologically interpretable. Furthermore, genetic effects can be more easily detected when SNP effects are aggregated together. Finally, gene based analysis simplifies the multiple testing problem by decreasing the number of variables. In the past few years several gene-gene methods have been proposed, they rely on a summarizing step to obtain information at the gene level and a modeling phase to represent interactions. Most of these methods resume gene information using Principal Component Analysis (PCA) and represent the interaction between two genes by the product of the first components [Zhang et al., 2008; Li et al., 2009; He et al., 2011], some proposition have also been done using Partial Least Square analysis (PLS) that extract components that summarize both the information among SNPs in a gene and the correlation between SNPs and the outcome of interest [Wang et al., 2009].

Directly modeling all gene-gene interactions would be inefficient due to computational challenge and lack of power. For the most recent methods, filters or penalized models are used to make the method applicable to a large number of genes. Penalized regression methods as LASSO allow to select a subset of important predictors from a large number of potential ones. These methods operate by shrinking the size of the coefficients. The coefficients of predictors with little or no apparent effect are pushed on a trait down toward zero, allowing to reduce the effective degrees of freedom and in many cases to performs model selection. Some gene-gene detection methods using PCA and penalized regression have been developed as the propositions of D'Angelo et al. [2009] and Wang et al. [2014].

Here we propose a Group LASSO based method [Yuan and Lin, 2006] that takes into account the group structure of each gene in order to detect epistasis at the gene level. We introduce the Gene-Gene Eigen Epistasis (G-GEE) as a new approach to compute the gene-gene interaction

*. Intervenant
†. Corresponding author : virginie.stanislas@math.cnrs.fr

part of the model. The method first compute interaction variables for each gene pair by finding its Eigen-epistasis Component defined as the linear combination of Gene SNPs having the highest correlation with the phenotype. The selection of the significant effects results from a group LASSO penalized regression method combined to an adaptive ridge approach [Bécu et al., 2015] allowing to control the False Discovery Rate.

We conduct a simulation study to compare G-GEE with Principal Component Analysis (PCA) and Partial and Least squares (PLS). We simulated genotypes using a multivariate random vector with a block diagonal correlation matrix adapted from the model used in Wu et al. [2009] with an extension to control the minor allele frequency (MAF) of each SNP. Phenotypes are generated from two different linear models varying in the manner of construct interaction effects. The strength of the association and the difficulty of the problem are calibrated by the coefficient of determination R2 which is used in the generation of the error term. Causal effects are chosen considering two different simulation settings, a first setting where the main effects and interaction effects involve the same genes, a second setting where interaction genes are different from main effect genes. We then compare the power of the three different methods to detect interaction effects under various R2 values. Overall the G-GEE method performs well to detect interactions in all tested settings. The power of the PLS vary depending of the considered setting, favorable contexts appear only when the related main effect are also present. When the simulated main and interaction effects do not concern the same genes, the detection performances of the PLS approach drastically collapses. PCA detects the main effects but encounter problem with the interactions effects which are often considered as main effects.

We then apply the G-GEE approach in a genome-wide association study on ankylosing spondylitis. Ankylosing spondylitis is a common form of inflammatory arthritis predominantly affecting the spine and pelvis. Genetic factors contribute for more than 90 % to the susceptibility risk to the disease. Human leukocyte antigen (HLA) class I molecule HLA B27, belonging to the Major Histocompatibility Complex (MHC) region, was the first genetic risk factor identified as associated with ankylosing and remains the most important risk locus for this pathology but recent studies suggests that other genetic factors are involved [Tsui et al., 2014; Reveille et al., 2010]. Here, we focus our analysis on a list of 29 genes previously identified as having a main effect in GWAS, we compare the results of the three different methods and demonstrate the power of the G-GEE approach by detecting new gene-gene interactions.

References

Bécu JM, Grandvalet Y, Ambroise C, Dalmasso C. 2015. Beyond Support in Two-Stage Variable Selection. *ArXiv* 150507281 Stat.

D'Angelo GM, Rao D, Gu CC. 2009. Combining least absolute shrinkage and selection operator (lasso) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BMC Proc* 3(Suppl 7):S62.

He J, Wang K, Edmondson AC, Rader DJ, Li C, Li M. 2011. Gene-based interaction analysis by incor- porating external linkage disequilibrium information. *Eur J Hum Genet* 19:164–172.

Li J, Tang R, Biernacka JM, de Andrade M. 2009. Identification of gene-gene interaction using principal components. *BMC Proc* 3:S78.

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Front. Genet*, page 285.

Reveille JD, Sims AM, Danoy P, Evans DM, Leo P, Pointon JJ, Jin R, Zhou X, Bradbury LA, Appleton LH, Davis JC, Diekman L, Doan T, Dowling A, Duan R, Duncan EL, Farrar C, Hadler J, Harvey D, Karaderi T, Mogg R, Pomeroy E, Pryce K, Taylor J, Savage L, Deloukas P, Kumanduri V, Peltonen L, Ring SM, Whittaker P, Glazov E, Thomas GP, Maksymowych WP, Inman RD, Ward

MM, Stone MA, Weisman MH, Wordsworth BP, Brown MA. 2010. Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nat Genet* 42:123–127.

Steen KV. 2012. Travelling the world of gene-gene interactions. *Brief Bioinformatics* 13:1–19.

Tsui F, Tsui HW, Akram A, Haroon N, Inman R. 2014. The genetic basis of ankylosing spondylitis: new insights into disease pathogenesis. *Appl Clin Genet* 7:105-115.

Wang T, Ho G, Ye K, Strickler H, Elston RC. 2009. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. *Genet Epidemiol* 33:6–15.

Wang X, Zhang D, Tzeng JY. 2014b. Pathway-Guided Identification of Gene-Gene Interactions. *Annals of Human Genetics* 78:478–491.

Wei WH, Hemani G, Haley CS. 2014. Detecting epistasis in human complex traits. *Nat Rev Genet* 15:722–733.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67.

Zhang F, Wagener D. 2008. An approach to incorporate linkage disequilibrium structure into genomic association analysis. *Journal of Genetics and Genomics* 35:381–385.

Mots clefs : Genome wide association study, Gene-gene interactions, Epistasis, Group Lasso

Computational identification of genomic features that influence 3D chromatin domain formation

Raphaël Mourad *[1]

[1] Laboratoire de biologie moléculaire eucaryote du CNRS (LBME) – CNRS : UMR5099, Université Paul Sabatier (UPS) - Toulouse III – Bâtiment IBCG 118 route de Narbonne, F-31 062 Toulouse Cedex 4, France

Session Statistique des données à haut débit
mardi 28 11h40
Amphi Mérieux

Introduction

Recent advances in long-range Hi-C contact mapping have revealed the importance of the 3D structure of chromosomes in gene expression. A current challenge is to identify the key molecular drivers of this 3D structure. Several genomic features such as architectural proteins and functional elements were shown to be enriched at topological domain borders using classical enrichment tests. Here we propose multiple logistic regression to identify those genomic features that positively or negatively influence domain border establishment or maintenance. The model is flexible and can account for statistical interactions among multiple genomic features. Using both simulated and real data, we show that our model outperforms enrichment test and non-parametric models such as random forests for the identification of genomic features that influence domain borders.

Results

The proposed multiple logistic regression models the influences of p genomic features $X=\{X_1,\ldots,X_p\}$ such as DNA-binding proteins on the variable Y that indicates if the genomic bin belongs to a border (Y=1) or not (Y=0) (see supplementary data). The set $\beta=\{\beta_1,\ldots,\beta_p\}$ denotes slope parameters, one parameter for each genomic feature. The model can easily accommodate interaction terms between genomic features (for instance, by multiplying two genomic features). By default, model likelihood is maximized by iteratively reweighted least squares to estimate unbiaised parameters. However, when there are a large number of correlated genomic features in the model, L1-regularization is used instead to reduce instability in parameter estimation.

We illustrate the proposed model with two different scenarios (see supplementary data). In the first scenario, protein A positively influences 3D domain borders, while protein B colocalizes to protein A. In this scenario, enrichment test will estimate that the parameter associated with protein A $\beta A>0$ and the parameter associated with protein B $\beta B>0$. In other words, both proteins A and B are enriched at 3D domain borders. Multiple logistic regression will instead estimate that parameters $\beta A>0$ and $\beta B=0$. This means that protein A positively influences 3D domain borders, while protein B does not. This is because multiple logistic regression can discard spurious associations (here between protein B and 3D domain borders). One would argue that enrichment test can also be used to discard the spurious association if the enrichment of protein B when protein A is absent is tested instead. However such conditional enrichment test becomes intractable when more than 3 proteins colocalize to domain borders, whereas multiple logistic regression is not limited by the numbers of proteins to analyze within the same model. In the second scenario, the co-occurrence of proteins A and B influences 3D domain borders, but not the proteins alone. Enrichment test will find that each protein alone is enriched at 3D domain borders ($\beta A>0$ and $\beta B>0$) as well as their interaction ($\beta AB>0$). The proposed model will instead find that only the interaction between proteins A and B influences 3D domain borders ($\beta A=0$, $\beta A=0$ and $\beta AB>0$).

*. Intervenant

We compared multiple logistic regression (MLR) with enrichment test (ET) and random forests (RF) using real data in human. For this purpose, we analyzed new 3D domains detected from recent high resolution Hi-C data at 1 kb for GM12878 cells for which 69 ChIP-seq data were available [5]. Multiple lines of evidence indicate that CTCF and cohesin serve as mediators of long-range contacts [4]. However several proteins also colocalize or interact with CTCF, including Yin Yang 1 (YY1), Kaiso, MYC-associated zing-finger protein (MAZ), jun-D proto-oncogene (JUND) and ZNF143 [1]. In addition, recent work has demonstrated the spatial clustering of Polycomb repressive complex proteins [7]. Using the large number of available proteins in GM12878 cells, we could compare MLR with ET and RF to identify known or suspected architectural proteins CTCF, cohesin, YY1, Kaiso, MAZ, JUND, ZNF143 and EZH2. For this purpose, we computed receiver operating characteristic (ROC) curves using Wald's statistics for ET, beta parameters for MLR, and variable importances for RF. We carried out computations at the very high resolution of 1 kb. ROC curves revealed that MLR clearly outperformed ET and RF to identify architectural proteins (AUCET = 0.613, AUCRF = 0.558, AUCMLR = 0.827; see supplementary data). Lower performance of ET was likely due to its inability to account for correlations among the proteins (average correlation = 0.19). Regarding RF, its low performance could be explained by its well-known inefficiency with sparse data (at 1 kb, there are 99.4 % of zeros in the data matrix X). At a lower resolution of 40 kb (88.5 % of zeros), RF performed much better (AUCRF = 0.746) but still lower than MLR (AUCMLR = 0.815).

Discussion

Here, we describe a multiple logistic regression (MLR) to assess the roles of genomic features such as DNA-binding proteins and functional elements on TAD border establishment/maintenance. Based on conditional independence, such regression model can identify genomic features that impact TAD borders, unlike enrichment test (ET) and non-parametric models. In addition, we show that our model outperforms enrichment test and non-parametric model for the identification of genomic features that influence domain borders. Using recent experimental Hi-C and ChIP-seq data, the proposed model can identify genomic features that are most influential with respect to TAD borders at a very high resolution of 1 kb in both Drosophila and human. The proposed model could thus guide the biologists for the design of most critical Hi-C experiments aiming at unraveling the key molecular determinants of higher-order chromatin organization.

Enrichment test shows slight differences of enrichments among architectural proteins. This could suggest that domain borders are determined by the number and levels of all proteins present at the border rather than the presence of specific proteins [8]. However MLR instead reveals that only some architectural proteins influence the presence of 3D domain borders. Moreover, MLR retrieves both positive and negative contributions among most influential proteins, depending on contexts such as co-occurence. From these novel results, we propose a biological model for 3D domain border establishment or maintenance (see supplementary data). In this model, three kinds of proteins are distinguished: positive drivers (βMLR> 0), negative drivers (βMLR< 0), and proteins that are enriched or depleted at borders but are not drivers (βET> 0 or βET< 0, and βMLR$=0$). Positive drivers favor attraction between domain borders leading to the formation of 3D domains. CTCF and cohesin are well-studied positive drivers in mammals [5]. By contrast little is known about negative drivers of 3D domain borders that could favor repulsion between specific chromatin regions [6]. Repulsion phenomenon could be the result of allosteric effects of loops in chromatin [2]. Negative drivers could also regulate disassembly of protein complex that mediate long-range contacts [3].

References

[1] Caelin Cubenas-Potts and Victor G. Corces. Architectural proteins, transcription, and the three-dimensional organization of the genome. *FEBS Letters*, 589(20PartA):2923–2930, 2015.

[2] Boryana Doyle, Geoffrey Fudenberg, Maxim Imakaev, and Leonid A. Mirny. Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS Computational Biology*, 10(10):e1003867+, October 2014.

[3] Andrew F. Neuwald, L. Aravind, John L. Spouge, and Eugene V. Koonin. AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Research*, 9(1):27–43, 1999.

[4] Jennifer E. Phillips-Cremins, Michael E. G. Sauria, Amartya Sanyal, Tatiana I. Gerasimova, Bryan R. Lajoie, Joshua S. K. Bell, Chin-Tong Ong, Tracy A. Hookway, Changying Guo, Yuhua Sun, Michael J. Bland, William Wagstaff, Stephen Dalton, Todd C. McDevitt, Ranjan Sen, Job Dekker, James Taylor, and Victor G. Corces. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, 153(6):1281–1295, June 2013.

[5] Suhas S. P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander, and Erez L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, February 2015.

[6] Saeed Saberi, Pau Farre, Olivier Cuvier, and Eldon Emberly. Probing long-range interactions by extracting free energies from genome-wide chromosome conformation capture data. *BMC Bioinformatics*, 16:171, May 2015.

[7] Stefan Schoenfelder, Robert Sugar, Andrew Dimond, Biola-Maria Javierre, Harry Armstrong, Borbala Mifsud, Emilia Dimitrova, Louise Matheson, Filipe Tavares-Cadete, Mayra Furlan-Magaril, Anne Segonds-Pichon, Wiktor Jurkowski, Steven W. Wingett, Kristina Tabbada, Simon Andrews, Bram Herman, Emily LeProust, Cameron S. Osborne, Haruhiko Koseki, Peter Fraser, Nicholas M. Luscombe, and Sarah Elderkin. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nature Genetics*, 47(10):1179–1186, August 2015.

[8] Kevin Van Bortle, Michael H. Nichols, Li Li, Chin-Tong Ong, Naomi Takenaka, Zhaohui S. Qin, and Victor G. Corces. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biology*, 15(5):R82+, June 2014.

Mots clefs : Chromatin, HiC, ChIPseq, Biostatistics, Logistic regression

Session Phylogénie
Évolution
mardi 28 13h30
Amphi Mérieux

Challenges & Problems in Phylogenetic Inference and Bioinformatics

Alexandros Stamatakis [†1]

[1] Heidelberg Institute for Theoretical Studies – The Exelixis Lab, Heidelberg, Allemagne

The fields of phylogenetic inference and Bioinformatics in general are currently facing two major challenges. Firstly, the scalability challenge that is driven by the molecular data avalanche and secondly the software quality challenge. In the first part of the talk, I will provide an overview of current and future approaches for tackling the scalability challenge in large-scale phylogenetic inference. In the second part of the talk, I will provide a critical review of software quality issues in Bioinformatics. More specifically, I will provide four examples for what can go wrong: (i) a code quality analysis of popular software tools for evolutionary biology, (ii) a major bug in several tree visualization tools that leads to incorrect display and hence interpretation of branch support values, (iii) a bug in our own analysis pipeline for a Science paper, and (iv) a mathematical error in Gotoh's pair-wise sequence alignment algorithm that has been propagated to implementations as well as bioinformatics teaching slides and text books. I will conclude with suggestions on how to improve software quality in Bioinformatics.

†. Intervenant Invité

Détection de signatures moléculaires de convergences évolutives

Olivier Chabrol [*1,2], Gilles Didier [†1], Manuela Royer Carenzi [‡1], Pierre Pontarotti [§1,2]

[1] Institut de Mathématiques de Marseille (I2M) – École Centrale de Marseille, CNRS : UMR7373, Aix-Marseille Université – Centre de Mathématiques et Informatique, Château Gombert, Campus de Luminy, France

[2] Évolution Biologique & Modélisation – CNRS : UMR7353, Université de Provence - Aix-Marseille I – Aix-Marseille Université, CNRS UMR7353, équipe Évolution Biologique et Modélisation, Campus Saint-Charles, 3 place Victor Hugo, F-13 331 MARSEILLE, France

Session Phylogénie 1
mardi 28 14h40
Amphi Mérieux

Nous proposons une nouvelle approche pour la détection de protéines potentiellement impliquées dans des apparitions indépendantes d'un même caractère chez des espèces évolutivement distantes. Pour ce faire, nous définissons une quantité appelée *indice de convergence*, que l'on calcule pour chaque site d'un alignement de protéines orthologues.

Au moyen de données simulées, nous montrons que cet indice détecte mieux les sites sous convergence que l'approche standard, basée sur la reconstruction ancestrale. Les protéines potentiellement impliquées dans l'apparition du caractère sont détectées si elles contiennent un nombre significatif de sites ayant un indice de convergence élévé.

Nous appliquons enfin notre approche sur un jeu de données biologiques relatif à l'apparition de l'écholocation chez les mammifères.

Mots clefs : évolution, convergence, données moléculaire

*. Intervenant
†. Corresponding author : gilles.didier@univ-amu.fr
‡. Corresponding author : manuela.carenzi@univ-amu.fr
§. Corresponding author : pierre.pontarotti@univ-amu.fr

Lifemap : exploring the entire tree of life

Damien M. de Vienne [*1]

[1] LBBE, CNRS – PRES Université de Lyon – France

Session Phylogénie 1
mardi 28 15h00
Amphi Mérieux

The decrease of DNA sequencing costs [1] associated with improved phylogenetic and phylogenomic methods for reconstructing phylogenetic trees [2,3,4] helped resolve, in the recent years, many parts of the Tree of Life. The aggregation of these portions, using taxonomy for unresolved parts and to name monophyletic groups, produces a phylogenetic classification scheme describing the evolutionary relationships among the species under consideration [5]. The NCBI taxonomy is one of these classifications that has the advantage of offering a good compromise between the number of represented species ($\approx$ 1.4 million as of April 2016) and the reliability of the relationships between them. Unsurprisingly, NCBI is the most widely used source of taxonomic information for biologists.

Despite huge ongoing efforts to assemble the Tree of Life [6,7], no tool exists today to explore it entirely in an interactive manner. A parallel has been drawn earlier between what has been achieved in cartography with the development of Google maps or OpenStreetMaps [OSM,8], and what should be done for exploring the tree of life [9]. The logic behind the use of the cartographic paradigm for visualizing a taxonomy is straightforward: like geographic entities (countries, regions, cities…), taxonomic levels have names (kingdoms, families, classes, orders, genera…) and are nested within each other. Consequently, it is possible to propose a visualization where zooming-in increases the level of detail displayed while hiding upper-level information. This idea was first materialized in a tool called OneZoom [9] that allows to visualize, on demand, large trees by zooming and panning. This tool is convenient for trees with a few tens of thousands of tips, but suffers from some limitations that make it inappropriate for visualizing the complete Tree of Life. For instance, the fractal representation it uses prevents the presence of multifurcations in the trees (one node connected to more than two descendants) which is very frequent in the Tree of Life in its present state (83 % of the nodes are not binary), and the amount of RAM that the computer would need to handle a dataset of more than a million species is unlikely to be available on a personal computer.

I developed Lifemap, a tool largely inspired by the technology developed for cartography, that is free of the limitations described above. Its approach differs from that of OneZoom (9) both in the representation of the tree and in the way images are displayed and interacted with on the screen. This allows a fast and smooth exploration of the biggest tree ever proposed on a single page for exploration.

Lifemap comes in two versions that differ by the tree that is displayed and the information that is associated to tips and nodes. The general public version is based on a reduced NCBI taxonomy obtained by removing non-identified clades and all taxa below the species level. When clicking on nodes or tips, a short description and a picture are displayed. The *expert* version displays the whole NCBI taxonomy, and is updated every week. When clicking on a node, the user can (i) get additional information about the current taxa (taxid, number of species), (ii) reach the NCBI web page corresponding to the node, or (iii) download the corresponding subtree in Newick Extended Format (parenthetic) for further analysis. Both versions give the possibility to compute, visualize and explore "paths" in the Tree of Life. This is done by choosing a source and a destination taxa. The path is computed instantly and highlighted on the tree. The Most Recent Common Ancestor (MRCA) is indicated with a marker, and the list of taxa encountered in the route from the source to the destination is returned.

*. Intervenant

Lifemap may become a useful source of information for the general public interested in evolution and biodiversity, but also for education and research in various fields related to ecology, evolutionary biology and genomics. Lifemap is available online at `http://lifemap.univ-lyon1.fr/` (general public version) and `http://lifemap-ncbi.univ-lyon1.fr/` (expert version). The general public version is also available as a mobile App for Android phones and tablets on the Play Store (Lifemap – Tree of Life), and should soon be available for iOS devices and Windows Phones.

References

[1] NHGRI Genome Sequencing Program (GSP). `https://www.genome.gov/27541954/dna-sequencing-costs/`

[2] Guindon, S. & Gascuel, O. *Syst. Biol.* 52, 696-704 (2003).

[3] Stamatakis, A. *Bioinf.* 22, 2688-2690 (2006).

[4] Ronquist, F. & Huelsenbeck, J. *Bioinf.* 19, 1572–1574 (2003).

[5] Federhen, S. *The NCBI Handbook*, McEntyre & Ostell, Bethesda, Maryland, USA, 2003.

[6] Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J. *et al. Nat. Microbiol.* 16048 (2016).

[7] Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R. *et al. PNAS.* 112, 12764-12769 (2015).

[8] Haklay, M. & Weber, P. *IEEE Pervas. Comp.* 7, 12-18 (2008).

[9] Rosindell, J & Harmon, L. J. *PLoS Biol.* 10, e1001406 (2012).

Mots clefs : Arbre de la vie, Évolution, Visualisation, MRCA, Tree of Life

Nucleotide, gene and genome evolution : a score to bind them all

Wandrille Duchemin [*†1], Éric Tannier [*1,2], Vincent Daubin [*1]

Session Phylogénie 1
mardi 28 15h20
Amphi Mérieux

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] BEAGLE (Insa Lyon / INRIA Grenoble Rhône-Alpes / UCBL) – INRIA, Institut National des Sciences Appliquées [INSA] - Lyon, Université Claude Bernard - Lyon I (UCBL) – Antenne INRIA Lyon la Doua, Bâtiment CEI-1, 66 Boulevard Niels Bohr, F-69 603 Villeurbanne, France

The history of genomes is usually studied through a pipeline of independent, successive steps, each answering a specific question by optimizing a specific metric. For instance sequences homology must first be assessed in order to delineate gene families. The sequences of each gene from a family are then aligned together. This alignement is used to generate an history of the gene family, for instance a unique phylogenetic tree (a "gene tree") or a distribution of phylogenetic trees if one wants to account for the uncertainties of the reconstruction. Using a species tree, it is then possible to associate each node of a gene tree with a species (ie. a node in the species tree) and one or several evolutionary events (for instance: speciation, duplication or horizontal gene transfer). This process is called a reconciliation. It is also possible through this method to detect the loss of a gene copy in a lineage of the species tree. The result of such an association is termed a reconciled gene tree. The reconciled tree of different gene families are then combined with relationships information such as gene-order, interactions, co-regulation or co-expression that links some extant gene copies together. Finally, ancestral relationships can be inferred along with some information about the history of these relationships.

Each of these steps represents a complex task and each has a consequent body of literature attached to it. As each step is performed independently, it is possible that the optimization of a previous step sets the following ones in a sub-optimal space of solutions. For instance, the gene tree with the maximal likelihood might require numerous evolutionary events of transfer and loss to be reconciled with the species tree, while a gene tree only slightly less likely with respect to the alignement could allow some simpler reconciliation scenarios. Furthermore, as this pipeline is done for each gene family separately, choices are made that might downplay the amount of coevolution that can be expected from genes evolving in the same species and with sometimes related function or neighbouring positions on a chromosome [Liang2010]. Some methods jointly infer the gene tree topology and the gene tree reconciliation, yielding a more comprehensive view of the gene history [Szöllősi2013b, Scornavacca2014], but they still consider each gene family independently from each other. For instance, two genes both undergoing a duplication event in the same species are usually seen as two separate duplication events (and thus costs twice the cost of a duplication in a parsimonious framework). Yet, if this two genes happen to be neighbours in that species, then it is likely that only one duplication event occurred which encompassed both genes.

Here we will focus on integrating three levels of inference in order to allow for the co-evolution of genes: gene tree reconstruction, gene tree reconciliation and adjacency tree building. In this work, we define a gene as a block of nucleotides that cannot be broken into sub-genes or undergo internal rearrangement. In practice, it can be a protein domain, a entire coding sequence, or any segment of a chromosome. We define adjacencies as binary relationships between genes. Adjacencies might represent diverse notions of relatedness such as co-function, co-expression

*. Intervenant
†. Corresponding author : wandrille.duchemin@univ-lyon1.fr

or co-occurrence on a chromosome. The history of a group of homologous adjacencies can be summed up in an adjacency tree [Bérard2012].

We propose a score that integrates gene tree reconstruction, reconciliation and adjacency history building into a parsimonious framework, but also incorporates a notion of coevents: events regrouping several gene copies. By introducing this notion, we account for the coevolution between neighbouring genes. We are able to propose solutions that may not be optimal when gene families are considered as independently evolving but yield more coherent global scenarios of evolution when all gene families are allowed to co-evolve.

Methodologically, several algorithms are assembled to provide the different component of the proposed score. By using conditional clades probability [Höhna2012], we can estimate the score of a single gene tree from a posterior distribution of trees. We can also compute the score of a gene tree reconciliations, including events of duplication, loss and lateral gene transfer possibly from extinct or unsampled lineages of the species tree [Szöllősi2013a]. This is typically done using the parsimonious reconciliation algorithm of [Doyon2010] as implemented in TERA [Scornavacca2014]. The score of gene adjacency histories are computed using the algorithm of DeCoLT [Patterson2013], which computes the parsimonious adjacency histories given reconciled gene tree and extant adjacencies. Coevents are computed from the results of the DeCoLT algorithm and used to correct a weighted sum of the scores yielded by the different algorithms in order to obtain the global score. Optimizing such a score through an exhaustive exploration of the space of solution would be intractable, as the spaces of all gene trees topologies, all gene trees reconciliation, and adjacency history of all gene families would have to be combined.

We propose an exploration strategy based on heuristic and local moves. The idea is that the initial solution has been obtained by optimizing each element (gene family, for instance) without taking any notion of coevent into account. Thus, to get a better global score, coevents whose global score correction compensate the loss of local optimality must be proposed.

This approach can be used at a variety of scale, e.g. by considering protein domains as unit to reconstruct the history of modular proteins, or by considering genes to reconstruct the history of chromosomes and metabolic networks. Other than providing more coherent evolutionary scenarios, coevents can prove a useful tool to study the dynamics of genomic events of duplication, loss and transfer. They could be used to study their size (not being limited at the size of a gene) or give a better estimate of the frequency of these events (because they can effectively detect several neighbouring gene undergoing the same event as one event rather than several).

References

[Doyon2010] J.-P. Doyon, C. Scornavacca, V. Ranwez, V. Berry (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: *Proceedings of the 2010 International Conference on Comparative Genomics, RECOMB-CG'10*, Springer-Verlag, Berlin,Heidelberg, 93-108

[Liang2010] Z. Liang, M. Xu, M. Teng, et al. (2010) Coevolution is a short-distance force at the protein interaction level and correlates with the modular organization of protein networks. *FEBS Letters* 19:4237-4240

[Bérard2012] S. Bérard, C. Gallien, B. Boussau, et al. (2012) Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics* 28(18):i382-i388

[Höhna2012] S. Höhna, A. Drummond (2012) Guided tree topology proposals for Bayesian phylogenetic inference. *Systematic Biology* 61:1–11.

[Patterson2013] M. Patterson, G. J. Szöllősi, V. Daubin, É. Tannier (2013) Lateral gene transfer, rearrangement, reconciliation. *BMC bioinformatics* 14(Suppl 15):S4

[Szöllősi2013a] G. J. Szöllősi, É. Tannier, N. Lartillot, V. Daubin (2013) Lateral Gene Transfer from the Dead. *Syst Biol* 62:386-397

[Szöllősi2013b] G. J. Szöllősi, W. Rosikiewicz, B. Boussau, et al. (2013) Efficient exploration of the space of reconciled gene trees. *Systematic Biology* 6:901-912

[Scornavacca2014] C. Scornavacca, E. Jacox, G. J. Szöllősi (2014) Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics* 6:841-848

Mots clefs : phylogeny, reconciliation, adjacency, coevent

Robustness of the parsimonious reconciliation method in cophylogeny

Laura Urbini [*1,2], Blerina Sinaimeri [†1,2], Catherine Matias[3], Marie-France Sagot [‡1,2]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 VILLEURBANNE Cedex, France
[2] INRIA Grenoble Rhône-Alpes – INRIA – France
[3] Laboratoire de Probabilités et Modèles Aléatoires – CNRS : UMR7599 – France

Session Phylogénie 1
mardi 28 15h40
Amphi Mérieux

Almost every organism in the biosphere is involved in a so-called symbiotic interaction with other biological species, that is, in an interaction which is close and often long term. Such interactions (one speaks also of symbiosis) can involve two or more species and be of different types, ranging from mutualism (when both species benefit) to parasitism (when one benefits to the detriment of the other). Understanding symbiosis in general is therefore important in many different areas of biology. As symbiotic interactions may continue over very long periods of time, the species involved can affect each other's evolution. This is known as coevolution. Studying the joint evolutionary history of species engaged in a symbiotic interaction enables in particular to better understand the long-term dynamics of such interactions. This is the subject of cophylogeny.

The currently most used method in cophylogenetic studies is the so-called *phylogenetic tree reconciliation*. In this model, we are given the phylogenetic tree of the hosts H, the one of the symbionts S, and a mapping ϕ from the leaves of S to the leaves of H indicating the known symbiotic relationships among present-day organisms. In general, the common evolutionary history of the hosts and of their symbionts is explained through four main macro-evolutionary events that are assumed to be recovered by the tree reconciliation: cospeciation, duplication, host-switch and loss. A reconciliation is a function λ which is an extension of the mapping ϕ between leaves to a mapping that includes all internal nodes and that can be constructed using the four types of events above. An optimal reconciliation is usually defined in a parsimonious way: a cost is associated to each event and a solution of minimum total cost is searched for. If timing information (i.e. the order in which the speciation events occurred in the host phylogeny) is not known, as is usually the case, the problem is NP-hard. A way to deal with this is to allow for solutions that may be biologically unfeasible, that is for solutions where some of the switches induce a contradictory time ordering for the internal nodes of the host tree. In this case, the problem can be solved in polynomial time. In most situations, among the many optimal solutions, some are time-feasible.

However, an important issue in this model is that it makes strong assumptions on the input data which may not be verified in practice. We examine two cases where this situation happens.

The first is related to a limitation in the currently available methods for tree reconciliation where the association ϕ of the leaves is for now, to the best of our knowledge, required to be a function. A leaf s of the symbiont tree can therefore be mapped to at most one leaf of the host tree. This is clearly not realistic as a single symbiont species can infect more than one host. We henceforth use the term *multiple association* to refer to this phenomenon. For each present-day symbiont involved in a multiple association, one is currently forced to choose a single one. Clearly, this may have an influence on the solutions obtained.

*. Intervenant
†. Corresponding author : blerina.sinaimeri@inria.fr
‡. Corresponding author : marie-france.sagot@inria.fr

The second case addresses a different type of problem related to the phylogenetic trees of hosts and symbionts. These indeed are assumed to be correct, which may not be the case already for the hosts even though these are in general eukaryotes for which relatively accurate trees can be inferred, and can become really problematic for the symbionts which most often are prokaryotes and can recombine among them. We do not address the problem of recombination in this work, but another one that may also have an influence in the tree reconciliation. This is the problem of correctly rooting a phylogenetic tree. Many phylogenetic tree reconstruction algorithms in fact produce unrooted trees. The outgroup method is the most widely used in phylogenetic studies but a correct indication of the root position strongly depends on the availability of a proper outgroup. A wrong rooting of the trees given as input may lead to an incorrect output.

The aim of this study is, in the two cases, to explore the robustness of the parsimonious tree reconciliation method under "editing" (multiple associations) or "small perturbations" of the input (rooting problem). Notice that the first case is in general due to the fact that we are not able for now to handle multiple associations, although there could also be errors present in the association of the leaves that is given as input. The editing or perturbations we will be considering involve, respectively: (a) making different choices of single symbiont-host leaf mapping in the presence of multiple associations, and (b) re-rooting of the symbiont tree. In both studies, we explore the influence of six cost vectors that are commonly used in the literature. We thus considered the following cost vectors $c = < c\_cosp, c\_dupl, c\_switch, c\_loss > C$ where $C = \{ < -1,1,1,1 >, < 0,1,1,1 >, < 0,1,2,1 >, < 0,2,3,1 >, < 1,1,1,1 >, < 1,1,3,1 > \}$.

To quantify the differences between two outputs of the method we introduced a mesure to compare sets of tree reconciliations which may be of independent interest. We tested the parsimonious reconciliation method both on real and simulated datasets. The final objective is to arrive at a better understanding of the relationship between the input and output of a parsimonious tree reconciliation method, and therefore at an evaluation of the confidence we can have in the output.

In the case of the multiple associations problem, we observed that the choice of leaf associations may have a strong impact on the variability of the reconciliation output. Although such impact appears not so important on the cost of the optimum solution, probably due to the relatively small size of the input trees (the smaller is composed of a pair of host and symbiont trees which have each 8 leaves. The bigger is composed of a host tree which has 34 leaves and a symbiont tree which has 42 leaves). The difference becomes more consequent when we refine the analysis by comparing, not the overall cost, but instead the patterns observed in the optimal solutions with the dissimilarity measure.

We were able to do the analysis on the choice of leaf associations only for the real biological datasets because we are currently not capable of simulating the coevolution of symbionts and hosts following the phylogenetic tree of the latter and allowing for an association of the symbionts to multiple hosts. This is an interesting and we believe important open problem in the literature on reconciliations which we are currently trying to address.

In the re-rooting problem, we are also interested in the case described by *Gorecki et al.* who showed the existence of a certain property in models such as the Duplication-Loss for the gene/species tree reconciliation. Such property, which the authors called the *plateau property*, states that if we assign to each edge of the parasite tree a value indicating the cost of an optimal reconciliation when considering the parasite tree rooted in that edge, the edges with minimum value form a connected subtree in the parasite tree, hence the name of plateau. Furthermore, the edge values in any path from a plateau towards a leaf are monotonically increasing. Here, for both biological and simulated datasets, we count the number of plateaux (i.e. subtrees where rootings lead to minimal optimum cost), and we further keep track whether the original root belongs to a plateau.

We were able to show that allowing for host switches invalidates the plateau property that had been previously observed (and actually also mathematically proved) in the cases where such

events were not considered. Again here, the number of plateaux observed is small for the real datasets (this number is indeed 2).

Moreover, such increase from 1 to 2 does not concern all pairs of datasets and of cost vectors, even though for all, except one of the cost vectors tested, there is always a biological dataset for which 2 plateaux are observed. We hypothesised that for the real datasets, this might indicate that the original root is not at its correct position. It would be interesting in future to try to validate this hypothesis. If it were proved to be true, an interesting, but hard open problem would be to be able to use as input for a cophylogeny study unrooted trees instead of rooted one, or even directly the sequences that were originally used to infer the host and symbiont trees. In this case, we would then have to, at a same time, infer the trees and their optimal reconciliation.

Clearly, the effect in terms of number of plateaux depends on the presence of host switches since this number was proved to be always one when switches are not allowed. Perhaps the most interesting open problem now is whether there is a relation between the number of plateaux observed as well as the level of dissimilarity among the patterns obtained on one hand, and the number of host switches in the optimal solutions on the other hand. Actually the relation may be more subtle, and be related not to the number of switches but to the distance involved in a switch, where by distance of a switch we mean the evolutionary distance between the two hosts involved in it. This could be measured in terms of the number of branches or in terms of the sum of the branch lengths, that is of estimated evolutionary time.

Mots clefs : cophylogeny, parsimony, event, based methods, robustness, measure for tree reconciliation comparison

An illustration of the consequence to take into account the local score length for statistical significance test on biological sequence analysis and result on the position of the local score realization

Session Statistique
mardi 28 14h40
Salle place de l'école

Agnès Lagnoux[1], Sabine Mercier [*] [†1], Pierre Vallois[2]

[1] Institut de Mathématiques de Toulouse (IMT) – PRES Université de Toulouse – France
[2] Institut Elie Cartan – Université de Lorraine – France

Introduction and motivation

Biological sequence data bases have been created since the 80's and they are still growing. In practice a biological sequence is considered as a succession of letters which belong to a finite set $A_1, \ldots, A_k$. For the case of the DNA, the letters of interest are A, C, G and T. A score is a function s that gives a real number to any letter A_i, an hydrophobic score for example. The local score for sequence analysis quantifies the degree of presence of a physico-chemical property locally in the sequence.
$H_n := \max\{0 \leqslant i \leqslant j \leqslant n\}(X_i + \cdots + X_j)$ where $X_0 = 0$ and $X_k = s(A_k)$ for $k \geqslant 1$. The local score can be rewritten as the maximum of the Lindley process (U_k)
$H_n = \max\{0 \leqslant k \leqslant n\} U_k$
where $(U_k)_{\{k \geqslant 0\}}$ is defined recursively by $U_0 := 0$ and $U_k := \max((U_{\{k-1\}} + X_k), 0)$ for $k \geqslant 1$. The Lindley process can be described as a succession of excursions above 0 and the local score is the high of the highest excursion.

It is usually supposed that the (X_i) are independent and identically distributed (i.i.d.). A crucial question is to establish the distribution or the p-value of this local score in order to distinguish ordinary regions or sequences from really interesting ones. The 90's have seen the most famous mathematical result on the probability of the local score, the Gumble approximation of Karlin et al. when the average score is non positive $(E[X] < 0)$. But there are now other results for this distribution and we will present them (Mercier and Daudin 2001, Cellier et al. 2003, Étienne and Vallois 2003).

The length L_n of the sequence region that realizes the local score has always been a subject of interest since its definition in the 80's but few results exist. In 2014, a new approach is proposed using Brownian motion theory. For this work, a slightly different local score H_n^* is defined on adaptively truncated sequences: The usual local score H_n is defined on every excursion including the final incomplete one whereas H_n^* and its realization length L_n^* are defined on the complete excursions only. The figure illustrates the different notation. An approximation is established for the probability of the pair (H_n^*, L_n^*) (see Chabriac et al. 2014).

We study the accuracy of the different results on the local score distribution and we illustrate what does it change to take into account the length of the local score when testing if the biological sequences are significant are not.

*. Intervenant
†. Corresponding author : mercier@univ-tlse2.fr

Studying the difference between H_n^* and H_n allows us to highlight a non intuitive result on the position of realization of the local score. We have investigate both simulated sequences with different length and different average score and real sequences of SCOP files (Structural Classification of Proteins). Results are quite surprising.

What does it change to take into account the length

On a set of 10^5 i.i.d. simulated sequences and using a Monte Carlo approach, we compare probabilities of the pair $P(H_n \leqslant h; L_n \geqslant \ell)$ and the usual p-values $P(H_n \geqslant h)$ for the observed values of h and ℓ. We observe that the probabilities can be strongly different. We get around 30 % sequences that have a p-value for the pair that is significant: $P(H_n \geqslant h; L_n \leqslant \ell) < \alpha$ for a given threshold α ; whereas these sequences are not significant using the usual p-value of the local score only: $P(H_n \geqslant h) > \alpha$.

But more important it is to see that the most interesting sequence order is modified. For the 606 sequences of a SCOP file (CF scop2dom 20140205aa, `http://scop2.mrc-lmb.cam.ac.uk/downloads/`), using the hydrophobic scale of Kyte and Doolittle, the local score and its length are calculated. The p-value of the local score is calculated using the exact method (see Daudin et Mercier 2001). An ordered list of the most significant sequences is deduced. An empirical p-value for the pair (H_n, L_n) is proposed by a Monte Carlo approach for good candidate sequences and the deduced order list is compared to the first one. We can see for example that the sequence at position 192 on the 606 sequences using the local score p-value is at the position 10 of the most significant sequences using the pair distribution (see the table in supplementary document).

An approximation when $E[X] = 0$ for the distribution of the pair (H_n^*, L_n^*) is deduced from Brownian theory. Using the classical Kolmogorov and Smirnov goodness-of-fit test we study the accuracy of this result and the usual ones. The result on the pair distribution becomes accurate for very long sequences only (upper than 10^4). This still must be improved. But we can also notice that using such a statistical test presents also the approximation of Karlin et al. as not very accurate too. Statistical tests focusing on tail distribution only should be developed.

We also study specificity and sensitivity of the different methods. For this, we simulate a set of sequences composed with a first set of sequences under an i.i.d. model of a given distribution D (H_0 hypothesis, the true negative sequences), and a second one. For this second set, we simulate sequences as for H_0 but for each sequence we exchange a segment of this sequence with a segment simulated with an other distribution than the one in H_0. The second set corresponds to true positive sequences (H_1 hypothesis). ROC curves (false positive rate versus true positive rate) shows that taking into account the length of the local score improves largely the quality of the analysis.

Position of the local score realization

The probability that the maximum of the reflected Brownian Motion is achieved in a complete excursion is established (see Lagnoux et al. 2015). An approximation of the probability that the local score H_n equals H_n^* is deduced when $E[X] = 0$. That gives information when the usual local score is achieved in the final part of the sequence, the last incomplete excursion. It is surprising to see that even for a growing sequence length that gives more chance to realize the local score "inside" the sequence, the percentage is constant: we have $P(H_n = H_n^*)$ equals around 1/3.

This is illustrated in different settings for i.i.d. simulated sequences and also for real sequences of SCOP files (Structural Classification of Proteins). For the 606 sequences of CF scop2dom 20140205aa SCOP File and with an hydrophobic scale, we get $E[X] = -0.23$ and 64 % of the sequences realize their usual local score in the final incomplete excursion.

Conclusion

There are still many interests in studying the distribution of the local score. Taking into account the length of the local score realization change the order of the most significant sequences and the quality of the significance test. The result on the probability for the pair (local score - realization length) are accurate for very long sequences and it must be improved. We observed that for real context, there many sequences that realized their local score in the final incomplete excursion of their corresponding Lindley sequence.

References

Cellier D., Charlot, F. and Mercier, S. An improved approximation for assessing the statistical significance of molecular sequence features (2003), *Jour. Appl. Prob.* (juin 2003), Vol. 40, 427-441.

Chabriac, C., Lagnoux, A., Mercier, S. and Vallois, P. Elements related to the largest complete excursion of a reflected Brownian motion stopped at a fixed time. Application to local score. *Stochastic Processes and their Applications*, 124(12), 2014.

Étienne, M.-P. and Vallois, P. Approximation of the distribution of the supremum of a centered random walk. Application to the local score. *Methodology and Computing in Applied Probability*, 6:255–275, 2004.

Karlin, S. and Altschul, S.-F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *PNAS*, 87:2264–2268, 1990.

Lagnoux, A. and Mercier, S. and Vallois, P. Probability that the maximum of the reflected Brownian motion over a finite interval [0,t] is achieved by its last zero before t. *Electronic Communications in Probability*, 20(62):1-9, 2015.

Lagnoux, A. and Mercier, S. and Vallois, P. Statistical significance for sequence analysis. Illustration of new results on length and position of the local score. Submitted.

Mercier, S. and Daudin, J.J. Exact distribution for the local score of one i.i.d. random sequence. *Jour. Comp. Biol*, 8(4):373–380, 2001.

Mots clefs : Score local, signification statistique, analyse des données, longueur

Revealing the genetic basis for host-pathogen interaction using machine learning

Mélodie Sammarro[1,2], Anja Friedrich[2], Jonathan Marshall[1,2], Patrick Biggs[2], Nigel French[2], Matthieu Vignes [*1]

[1] Institute of Fundamental Sciences, Massey University (IFS) – Private Bag 11222, Palmerston North 4442, Nouvelle-Zélande

[2] Institute of Veterinary, Animal & Biomedical Science, Massey University (IVABS) – Private Bag 11 222, Palmerston North, 4442, Nouvelle-Zélande

Session Statistique
mardi 28 15h00
Salle place de l'école

Introduction

Background *Campylobacter* is a bacterium causing intestinal infections that are generally mild, but can be fatal among very young children, elderly and immunosuppressed individuals. The bacteria normally inhabits the intestinal tract of warm-blooded animals such as poultry and cattle, and are frequently detected in foods derived from these animals. *Campylobacter* causes the most common foodborne illness in New Zealand. Studies have found that New Zealand has the highest rates of *Campylobacter* food poisoning in the developed world – three times higher than England and Wales and ten times higher than the US. Some 75,000 New Zealanders contract the illness every year and around 500 require hospital treatment. Currently, there are 17 species and 6 subspecies assigned to the genus Campylobacter, of which the most frequently reported in human diseases are *C. jejuni* and *C. coli*. Our study focuses on fifteen strains, which belong to four lineages identified by sequence typing. Several environmental factors are known to affect the behaviour of *C. jejuni* in its host. For example, the temperature or medium richness in different carbon sources imply different growth responses, which are the phenotypes we want to predict.

By examining the phenotype of hosts associated with different parasite strains, it appears that some host-pathogen associations do produce an interaction (e.g. host disease status, pathway modification), whilst others don't, or to a lesser extent.
Association trait prediction based upon genetic characteristics of pathogen strains could help prevent human infection in many cases without overindulging in unnecessary prevention. Humans are indeed at the end of the food chain and pathogens can infect wild animal, livestock and domestic animals before humans. Additionally, studying the genetic variation (e.g. patterns or rare genetic variants) informs us on its origin and on very fundamental molecular processes. The grail would be to identify which genes are involved in a disease outcome. More refined predictions may be concerned with pathway modification or with differential responses in different environments.

Research question We want to predict a phenotypic response which characterises the behaviour of a strain in a given environment (temperature and available nutriment) using environmental factors and genome data (GWAS approach).
We propose an interpretable tool for genetic association studies to gain a deeper understanding of the mechanisms at hand.

Data, methods and (very) preliminary results

Data The Phenotype MicroArray system is able to simultaneously capture a large number of phenotypes by recording an organism respiration over time (48h at 15 min intervals in our

*. Intervenant

setting) on distinct substrates (or nutriments). The phenotypic reaction of different strains of Campylobacter jejuni is recorded on sets of 96-well microtiter plates (flat plates with multiple wells used as small test tubes). We then end up with 96 graphics for each strain and each temperature (38 or 42°C to mimic different hosts condition, e.g. poultry, cow or human) containing multiple curves depending on the number of replicates (technical or biological) of the strain. Figure 1 shows an example of such phenotypic measures for one strain and one temperature conditions. We only exploit 91 of the wells since some of them are false positives: they produce a response on the negative control.
Typical features of interest of such curves include: the lag (i.e. the delay before an increase in the respiration curve is actually observed), the maximum slope (i.e. the maximum increase in respiration which corresponds to the slope of a smoothed representation of the curve), maximum recorded value, and area under the curve (AUC) of the respiration curve.
Notice that each well is characterised by a specific nutriment. This nutriment is a metabolite and can hence be part of the metabolic pathways of an organim (host or pathogen) and not of another.

Number of markers (genotyped loci): 2,362. Note however that some gene clusters are missing in one or several of the 15 strains, while others are represented several times. For example 1,261 gene clusters have exactly 15 gene copies, one for each strain, 232 gene clusters are present in only one strain and one gene cluster has 53 copies of the gene scattered accross the 15 strains.

Data analysis

Challenge 1: Carry out a clustering of curves on the data in order to differentiate the possible responses. Despite the large number of response profiles, we ended up with a reasonable number of representative curves comprised between 6 and 10 using k-means for time series data with various model selection criterions (e.g. silhouette, BIC, Calinski-Harabasz). We will keep ten clusters to characterise with a finer grain the different responses. An example of such a clustering with 6 clusters is visible in Figure 2 for one well.

Challenge 1 bis: estimate parameters for each well accounting for biological/technical replicates: we have at our disposal a large amount of information which allows us to study associated variation in the response at different levels. So transforming the response from a dynamical respiratory 48h-curve into four parameters. We will keep three of them as the AUC appears to be highly dependent on A (maximum recorded value), lambda (lag) and mu (the maximum increase in respiration which corresponds to the slope of a smoothed representation of the curve). A combination of these parameters will allow us to charactrise highly/medium/low levels of response.

The web-based software Phenolink (Bayjanov et al. 2012) was used to perform a genome-wide association study, combining the phenotypic profiles and whole genome sequences of the isolates that were identified as genes of relevance. Phenolink uses a random forest approach as its underlying algorithm to determine the importance of the genes in the observed phenotype (Friedrich 2014). To investigate further the complex interactions between the response variable and the covariates, regression trees were applied to the data.

REEMtrees (Random effects-Expectation Maximization trees) are a combination of mixed effects models for longitudinal data with the flexibility of tree-based estimation methods. The underlying algorithm is based on recursive partitioning, in order to minimise the variability in each node. For example, using our data, it has been shown that the loss of the lineage ST-42 (ruminant associated strain) ability to utilise L-Glutamine at 42 degrees Celsius could potentially be associated with the difference between poultry and ruminant core body temperatures. This study was the first one to test the ability of a wide variety of *C. jejuni* isolates from different hosts and STs to respire in 96 substrates as sole carbon sources.

A linear mixed effects model was used to examine possible relationships between temperature, insertion of the ykgC gene and ST and the utilisation of the 15 examined isolates. The first split in the tree divided the wells that are utilised by all isolates from the ones that are only utilised by a

subset. Further down the tree, the algorithm started to split on STs, indicating a specific phenotypic difference between the lineages. As an example, we know that these methods associated the ST-42 lineage with the utilisation of L-glutamine and the ST-474 lineage with the utilisation of citric acid.

We will extend the analysis in the framework of probabilistic graphical models, namely Bayesian networks (Scutari and Denis 2014). A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). These graphical structures are used to represent knowledge about an uncertain domain. In particular, each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables.

Our goal here is to define the probability distribution of the association between the genotype space (G, the variables for prediction) and the phenotypic space (P, the response). We also want to capture the impact of the temperature (T) on the phenotypic responses.

In a first phase, we will apply Bayesian networks constraining all genes to have an independent effect on the phenotypic response. In a second phase, we will relax the independence of the genes and allow genes to interact with each other in the form of gene regulations. We will in this latter scenario, have to deal with the high-dimensionality of the data.

We will also be interested in integrating the metabolite characteristic information, specific to each well in the gene network information. Indeed, genes would encore enzymes which certainly play a role in relevant metabolic pathways. The absence/presence of genes or specific alleles might have a measurable impact and be identified as key regulators in some environments.

References

* Bayjanov, J.R., Molenaar, D., Tzeneva, V., Siezen, R.J. and van Hijum, S.A.F.T. PhenoLink - a web-tool for linking phenotype to -omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains. *BMC Genomics*, 13:170, 2012

* Scutari, M. and Denis, J.-B. Bayesian Networks with Examples in R. *Chapman et Hall*, 2014.

* Friedrich, A. *Campylobacter jejuni* microevolution and phenotype:genotype relationships, PhD thesis, Massey University, 2014.

Mots clefs : host, pathogen, gene marker data, phenotypic response, Bayesian networks

Latent block model for metagenomic data

Julie Aubert [*1], Sophie Schbath[2], Stéphane Robin[1]

[1] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay – AgroParisTech, Institut National de la Recherche Agronomique - INRA, Université Paris-Saclay – 16 rue Claude Bernard, F-75 231 Paris Cedex 05, France

[2] Institut National de Recherche Agronomique - Centre de Jouy-en-Josas (MaIAGE) – Institut national de la recherche agronomique (INRA) – France

Session Statistique
mardi 28 15h20
Salle place de l'école

Introduction

Count matrices are widely used in numerous fields and namely in ecology. Metagenomics which studies microbial communities directly from environmental samples, provides abundance matrices where rows correspond to bacteria and columns to biological samples. One major goal is to find associations between bacterial communities and biological samples, and biclustering is one way for that. Indeed, biclustering aims at simultaneously dividing a data matrix into several homogenous subsets of rows and columns. This technique has been applied in various fields such as collaborative filtering (De Castro et al. 2007) or gene expression data (Madeira et al. 2004). It has proved its usefulness in discovering local patterns in which a subset of genes exhibits a similar pattern over a subset of samples. Two different strategies may be used to this aim: model-based or algorithmic methods. We will consider here the latent block model framework introduced by (Govaert et Nadif 2010), with an adaptation to overdispersed data to better fit metagenomics data.

Latent Block Model (LBM)

The LBM assumes that unobserved labels exist for features and samples, respectively. These two latent variables are supposed to be independent and to follow multinomial distributions. We assume that conditionally on these unknown labels the observed data are independent and follow some parametric distribution.

Poisson-gamma LBM

We focus on the negative binomial distribution which is the reference law for data from next-generation technology (Anders et al. 2010, Robinson et al. 2012, White et al. 2009, Lindner et al. 2015). We use the Poisson-Gamma parametrization of the negative-binomial distribution so that overdispersion is accounted for by a third hidden layer corresponding to unobserved heterogeneity. This model based on multiplicative heterogeneity increases variability in the Poisson mean but in expectation leaves the Poisson mean unchanged. This leads to increased probability of occurrences of low and high counts.

In order to take into account the specificities of the data, we introduce row and column effects. The row effects may be interpreted as a specific bacteria effect while the column effect as a sampling effort effect. As samples may correspond to replicates of a same condition we would like to cluster together, we introduce a new subscript for replicate hierarchized in a specific biological condition.

Inference

The aim of this inference is to estimate both the hidden variables and the parameters. Since latent variables are not independent conditionally on observed variables, the classical maximum

*. Intervenant

likelihood inference is intractable. Govaert and Nadif (2010) and Keribin et al. (2014) suggest inference algorithms based on a variational approach (Wainwright et Jordan 2008) as an alternative. This approach consists in maximizing a lower bound of the log-likelihood of the observed data using a variational version of the Expectation-Maximization algorithm alterning two steps.

Variational E-step :
We consider the class of factorisable distributions and look for the best approximation of the conditional distributions of the latent variables given the data in term of Kullback-Leibler divergence in this class (mean-fields approximation). An optimal solution for the approximate distribution of each hidden variable is given as a function of moments evaluated with respect to the distribution of all the other variables. This leads to closed-form interdependent variational update equations which must be solved iteratively.

Variation M-step :
Keeping the approximate conditional distribution fixed, we maximize the lower bound with respect to the mixing parameters and the distribution parameters, and update the parameter values.

Model selection

In practice the number of co-clusters is unknown and should be chosen by an adequate selection model procedure. We propose a model selection criterion based on the integrated classifical likelihood (ICL) to select the couple of number of groups.

Classification

We use a tractable approximation of the Maximum A Posteriory rule to classify the rows and the columns.

Applications

Different studies have shown that microbial communities living in animals (human included), in or around plants have a significant impact on health and disease of their host and on various services, such as adaptation under stressing environment. We apply our model on three real metagenomic datasets to study the interactions between the structure of microorganisms communities and biological samples they come from. The first one is a well-known human gut microbiota dataset (Arumugam et al. 2011). The second one is the MetaRhizo dataset which aims to study the plant-microbial communities interactions in the rhizosphere, the region of soil directly influenced by root secretions and associated soil microorganisms. The third one concerns the phyllosphere bacteria and fungi communities of oaks infected by the pathogen *Erysiphe alphitoides*. (Jakuschkin et al. 2016).

Mots clefs : Clustering, Mixture models, Count data, Negative binomial distribution, Metagenomics

Statistical modelling of expression patterns in hybrid species

Anthime Schrefheere[1], Matthew Campbell[1], Jennifer Tate[1], Murray Cox[1], Matthieu Vignes [*1]

[1] Institute of Fundamental Sciences, Massey University (IFS) – Private Bag 11222, Palmerston North 4442, Nouvelle-Zélande

Session Statistique
mardi 28 15h40
Salle place de l'école

Background and Introduction

Hybridisaton is the instantaneous formation of a new species from the merger of two different parental species. It is surprisingly common in plants, fungi and even animals. In this work, we focus on allopolyploidy: the global merger of the genomes of two parental sources from two different species. In this case, even surviving the 'genome shock', that is the set of molecular conflicts arising from having two copies of two different genomes, is poorly understood (McClintock 1984). We are primarily interested in a subset of this broader question: how patterns of gene expression in the allopolyploid hybrid relate to the newly formed genome and what are its relationships to parental transcriptomes. Our research question is summarised in Figure 1.

Our study is based on *Tragopogon miscellus*, which result of the interbreeding of *Tragopogon dubius* and *Tragopogon pratensis*. *T. miscellus* is of particular interest because it has formed reciprocally from *T. dubius* and *T. pratensis*. The reciprocally formed populations of *T. miscellus* differ in a simple flower phenotype: when *T. dubius* is the maternal progenitor, *T. miscellus* has long ligules, while when *T. pratensis* is the maternal parent, the ligules are short, see Figure 2 and (Soltis et al. 2004). The reciprocally formed populations cannot interbreed. We have here the opportunity to determine the effect of the sexual origin of each species on gene expression patterns to explain the observed difference in phenotypes.

Data

The data we use to determine gene expression comes from RNA sequencing of the flowers at a fixed physiological stage. All plants were cultivated in the same environmental conditions. We use 2 samples (technical replicates) of each factor (sex, see paragraph above and natural vs. synthetic, we coined generation, see next paragraph) combination and 2 for *T. dubius* and *T. pratensis*. Our study was restricted to genes, which were identified as responsible for flower growth and shape in related species (Chapman et al. 2008, Kim et al. 2008, Chapman et al. 2012).

In addition to parental gene expression, we have gene expression data for a direct interbreeding between *T. dubius* and *T. pratensis* to produce a new plant coined *synthetic*. Synthetic individual are tetraploid with exactly two copies of each chromosome coming from *T. dubius* and two copies coming from *T. pratensis* (2n=24). We also have transcriptome data on *natural* plants, which are also tetraplodoid (2n=24). In these individuals, we are very likely to have from zero to four copies of each parental sequence at each gene. We observe the resulting combined gene expression. Notice that a loss of some gene clusters can also occur (Buggs et a. 2012), but we can't distinguish it from very low levels of gene transcription.

By comparing gene expression patterns in synthetic versus natural plants, we can theoretically observe how the copy number affects the gene expression and the phenotype. We termed the

*. Intervenant

factor taking values natural or synthetic *generation* since synthetic individuals are obtained after the first hybridisation event, while natural plants arose after an undetermined but large (≈ 50, Soltis et al. 2004) number of generations since the 1900's.

Research question In this framework, the goal of the present study focuses on determining the pattern of gene expression in the hybrid in regards of the expression patterns of the parents. We model parental and hybrid genomes as complex systems. More specifically, we will be interested in the genetic regulation (i) controlled by parental sex and generation and (ii) driving an observed phenotype: the ligule length.

Methods

All analyses are performed after data normalisation (with the 'DESeq2' R package).

In a first phase, we explored individual gene expression distributions: globabally via heatmaps and locally, looking at typical expression data for some representative genes.

In a second phase, we aim at reconstructing gene regulatory network driving flower size and shape (sl/ll). The rationale for doing that is that genes which were considered are certainly very related to each other: possibly (i) some genes are directly affected by the generation and/or the parental sex factors, (ii) many of them regulate others and in turn (iii) some regulate the observed phenotype.

We answer this research question using Bayesian networks (Scutari and Denis 2014). A Bayesian network is made of a Directed Acyclic Graph (DAG) and of conditional probability distributions of node (gene expression) given the status of its parents in the DAG. We require that the DAG obeys some structural constraints which translate simple biological assumptions:

- the Sex of Parent 1 can't depend of any other node and acts on gene expressions only.
- Any gene expression is allowed to affect other gene expressions and/or the phenotype.
- The generation factor directly affects the number of copies steming from Parent 1. This number of copy has a direct affect on gene expressions. Genes with a close location on chromosome may be linked so that they are inherited together from a parent.
- Any gene expression and no other variable can modify the phenotype. The phenotype does not affect gene expression.

We present an illustration of such rules for our graphical modelling in Figure 3. BN analyses were done using the R package 'bnlearn'.

Preliminary results

The kernel density estimation helped the interpretation of differences in gene expression. Figures 4 to 7 show the expression of 13 genes selected as representative of detected patterns. It seems to be possible to identify two groups of gene expression pattern:

- One phenotype *T. miscellus* follows the gene expression of one parent and the other phenotype follows the gene expression of the other parent: in Figure 4, the sl phenotype expressions are closer to the expression of the maternal *T. pratensis*. On the opposite, in Figure 5, the expressions of individuals with a ll phenotype are similar to the maternal *T. dubius*.
- One phenotype *T. miscellus* follows the gene expression of the parents and the other phenotype has a different expression pattern: in Figure 6, we see that sl plant gene expressions are similar to that of parents. In Figure 7, the expression of ll phenotype individuals are more similar to either parent expressions.

It seems that the phenotype of *T. miscellus* results by different association rules of the parents genes expression. For some genes especially for the phenotype long liguled, we can observe also that the gene expressions never follow the gene expressions of both parents or of the other phenotype (see Figure 6). The corresponding gene expressions are either under- or over-expressed

as compared to the other group. For example, we can imagine that the under- or over-expression of some of these genes results in a one phenotype and an average level of expression (like that of the parents) results in the other phenotype. We aim at capturing such complex patterns in a global graphical model, namely Bayesian networks.

Among the tested network reconstruction algorithms, the Hill-climbing algorithm (model selection via BIC) was the only method which allowed us to observe paths from SexP1 to Phenotype via some genes. In a first stage, we investigated the DAG reconstruction without interactions between genes. 28 genes don't seem to have any interaction with SexP1; SexP1 seems to have an effect on 29 genes. Only gene CDM_37 has an effect on the phenotype. In a second stage, despite the low sample size, we relaxed the constraint, and genes could affect the expression of other genes. In the obtained DAG (see Figure 8), only one path (highligted in red) links SexP1 to Phenotype. This pathway is composed by 12 genes, some which were not deteted among the 29 genes above. They are characterised by specific expression like those studied above (see Figure 4 to 7). Again, CDM_37 is the key gene link to the phenotype node.

Future Work

Our work will now be to consolidate they analyses we just started here. We will also consider that each gene expression is the superposition of the translation of one (to simplify) copy inherited from one parent and another copy inherited from the other. Using sequence reads, the HyLiTE software (Duchemin et al. 2015) allows to decipher the expression origin in each gene. More precisely, it is possible to determine the parental origin of the genetic expression in the hybrid relying on sequence read similarity. Hence the expression of each gene in the hybrid can be decomposed into a contribution by each parent.

Xg, the expression of gene g would now be considered as a couple Xg=(Xg(P1), Xg(P2)). We could now consider that the action is driven by one parental copy of the gene, the other, or a combined action of both. And in turn, they may impact only one copy of other genes, the other or both. This would make the modelling slighlty more complex.

Later work might be concerned with data augmentation, to include the expression of other genes (some key genes in the process might not have been selected with the available prior knowledge) or to consider the use of latent variables. Moreover, retrieving other expression for different samples might be necessary to avoid the high-dimension curse we are facing.

References

McClintock, B. The significance of responses of the genome to challenge. *Science*, 226:792-801, 1984.

Soltis, D.E., Soltis, P.S., Pires, J.C., Kovarik, A., Tate, J.A. and Mavrodiev, E. Recent and recurrent polyploidy in Tragopogon (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biological Journal of the Linnean Society*, 82:485-510, 2004.

Chapman, M.A., Leebens-Mack, J.H. and Burke, J.M. Positive selection and expression divergence following gene duplication in the sunflower CYCLOIDEA gene family. *Molecular Biology and Evolution*, 25:1260-1273, 2008.

Kim, M., Cui, M.L., Cubas, P., Gillies, A., Lee, K., Chapman, M.A., Abbott, R.J., Coen, E. Regulatory genes control a key morphological and ecological trait transferred between species. *Science*, 322:1116-1119, 2008.

Chapman, M.A., Tang, S., Draeger, D., Nambeesan, S., Shaffer, H., Barb, J.G., Knapp, S.J., Burke J.M. Genetic analysis of floral symmetry in Van Gogh's sunflowers reveals independent recruitment of CYCLOIDEA genes in the Asteraceae. *PLoS Genetics*, 8:e1002628, 2012.

Buggs, R.J., Chamala, S., Wu, W., Tate, J.A., Schnable, P.S., Soltis, D.E., Soltis, P.S., Barbazuk, W.B.. Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Current Biology*, 22(3):248-252, 2012.

Scutari, M. and Denis, J.-B. Bayesian Networks with Examples in R, *Chapman & Hall*, 2014.

Duchemin, W., Dupont, P.-Y., Campbell, M.A., Ganley, A.R.D. and Cox, M.P. HyLiTE: accurate and flexible analysis of gene expression in hybrid and allopolyploid species. BMC Bioinformatics, 16:8, 2015.

Mots clefs : hybridisation, allopolyploid, RNAseq gene expression, differential expression, gene regulatory network, Bayesian network

Large scale analysis of amyloidogenic regions in proteins from evolutionary diverse organisms

Étienne Villain [*1,2], François Richard [†1,2], Philippe Fort[2], Andrey Kajava [‡1,2]

Session séquences protéiques
mardi 28 14h40
Salle des thèses

[1] Institut de Biologie Computationnelle (IBC) – Université de Montpellier – 95 rue de la Galéra, F-34 095 Montpellier, France
[2] Centre de Recherche en Biologie cellulaire de Montpellier (CRBM) – CNRS : UMR5237, Université Montpellier II - Sciences et techniques, Université Montpellier I – 1919 route de Mende, F-34 293 Montpellier Cedex 5, France

Introduction

Various normally innocuous and soluble proteins have a potential to polymerize to form insoluble amyloid fibrils. Amyloid fibrils are the subject of special interest mainly due to their link to a broad range of human diseases (amyloidoses), which include, but are not limited to, type II diabetes, rheumatoid arthritis, and perhaps most importantly, debilitating neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, Huntington's disease and infectious prion diseases [Pepys, 2006]. It has been also shown that in some organisms amyloid structures can also play important, beneficial roles where they are called "functional amyloids" [Otzen and Nielsen, 2008].

Over the last decade, numerous studies have demonstrated that just like globular and unstructured states, the propensity to form amyloids is coded by the amino acid sequence ([Ventura and Villaverde, 2006], [Uversky and Fink, 2004], [Fändrich, 2012], [Shirahama, 1967], [Eanes and Glenner, 1968], [Kirschner et al., 1987], [Serpell et al., 2012]). Based on this data, several computational methods for the prediction of amyloidogenicity have been proposed, using different approachs:

- machine learning on an experimental dataset of amyloidogenic peptides (Waltz [Maurer-Stroh et al., 2010], FISH amyloids [Gasior and Kotulska, 2014])
- methods based on individual amino-acid aggregation score and the composition of amyloidogenic regions (AGGRESCAN [Conchillo-Solé et al., 2007], FoldAmyloid [Garbuzynskiy et al., 2010])
- methods based on individual amino acid aggregation score and propencity to adopt β–structural conformation (Zyggregator [Tartaglia et al., 2008],TANGO [Fernandez-Escamilla et al., 2004])
- pairwise interractions within the β–sheets (PASTA [Trovato et al., 2006], BETASCAN [Bryan et al., 2009])
- estimation of propensity for a protein to be partially unfolded (AmylPred [Hamodrakas et al., 2007], Net-CSSP [Kim et al., 2009])

ArchCandy – a structure-based method for prediction of amyloidogenicity.

Recently, new experimental approaches have shed more light on the details of the 3D structural arrangement of amyloid fibrils. Progress was made by the application of new experimental techniques such as solid state nuclear magnetic resonance, cryoelectron microscopy, scanning transmission electron microscopy mass measurements, and electron paramagnetic resonance

*. Intervenant
†. Corresponding author : francois.richard@crbm.cnrs.fr
‡. Corresponding author : andrey.kajava@crbm.cnrs.fr

spectroscopy, in conjunction with more established approaches such as X-ray fiber diffraction, conventional electron microscopy, and optical spectroscopy ([Margittai and Langen, 2008], [Benzinger et al., 1998], [Goldsbury et al., 2011], [Sharma et al., 2005], [Sachse et al., 2008]). As a result, it has been shown that a majority of structural models of disease-related amyloid fibrils can be reduced to a so called "beta-arcade". This fold represents a columnar structure produced by stacking of beta-strand-loop-beta-strand motifs called "b-arches".

Using this information, we have developed a novel bioinformatics approach, ArchCandy [Ahmed et al., 2015], for the prediction of amyloidogenicity. The benchmark results show the superior performance of our method over the existing programs.

Large scale analysis of amyloidogenic regions in proteins

In this work we used ArchCandy in the large-scale analysis of proteomes to get a global view on the distribution and conservation during evolution of regions with high amyloidogenic potential in a set of evolutionary diverse organisms : 94 references proteoms covering Archaea, Bacteria, Eukaryota as well as viruses.

Direct application of the methods for prediction of amyloidogenicity is not sufficient. Today, it is becoming evident that an accurate estimation of the structural state(s) encoded by a given amino acid sequence requires evaluation of the individual probabilities of the protein to have either soluble 3D structure, an unstructured state, or insoluble structures, as well as the likelihoods of transition between the states of this triad. Therefore, during our analysis we cross-examined several types of data about proteins, such as location of unstructured regions, structured domains, transmembrane regions etc. The matter is that most of the known amyloid-forming regions of proteins are unfolded in their native state. Folded proteins or transmembrane regions may also contain potential amyloidogenic regions, however, as these regions are hidden within the 3D structure, usually, these proteins do not form fibrils. Therefore, in our analysis we exclude predicted amyloidogenic regions candidates that overlap with the putative transmembrane regions or those incompatible with known 3D structures.

In addition to ArchCandy we also analysed proteins of the proteomes by using several other existing computer programs for prediction of amyloidogenicity and these predictions were cross-examined with the ArchCandy results to get the most complete data.

The conclusions of this work will be presented at the conference.

References

[Ahmed et al., 2015] Ahmed, A. B., Znassi, N., Château, M.-T., and Kajava, A. V. (2015). A structure-based approach to predict predisposition to amyloidosis. *Alzheimer's & Dementia*, 11(6):681–690.

[Benzinger et al., 1998] Benzinger, T. L., Gregory, D. M., Burkoth, T. S., Miller-Auer, H., Lynn, D. G., Botto, R. E., and Meredith, S. C. (1998). Propagating structure of Alzheimer's beta-amyloid(10-35) is parallel beta-sheet with residues in exact register. *Proceedings of the National Academy of Sciences of the United States of America*, 95(23):13407–13412.

[Bryan et al., 2009] Bryan, A. W., Menke, M., Cowen, L. J., Lindquist, S. L., and Berger, B. (2009). Betascan: Probable beta-amyloids identified by pairwise probabilistic analysis. *PLoS Computational Biology*, 5(3).

[Conchillo-Solé et al., 2007] Conchillo-Solé, O., de Groot, N. S., Avilés, F. X., Vendrell, J., Daura, X., and Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC bioinformatics*, 8(1):65.

[Eanes and Glenner, 1968] Eanes, E. and Glenner, G. (1968). X-ray diffraction studies on amyloid filaments. *Journal of Histochemistry & Cytochemistry*, 16(11):673–677.

[Fändrich, 2012] Fändrich, M. (2012). Oligomeric intermediates in amyloid formation: Structure determination and mechanisms of toxicity. *Journal of Molecular Biology*, 421(4-5):427–440.

[Fernandez-Escamilla et al., 2004] Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology*, 22(10):1302–6.

[Garbuzynskiy et al., 2010] Garbuzynskiy, S. O., Lobanov, M. Y., and Galzitskaya, O. V. (2010). FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*, 26(3):326–332.

[Gasior and Kotulska, 2014] Gasior, P. and Kotulska, M. (2014). FISH Amyloid - a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. *BMC bioinformatics*, 15(1):54.

[Goldsbury et al., 2011] Goldsbury, C., Baxa, U., Simon, M. N., Steven, A. C., Engel, A., Wall, J. S., Aebi, U., and Müller, S. A. (2011). Amyloid structure and assembly: Insights from scanning transmission electron microscopy. *Journal of Structural Biology*, 173(1):1–13.

[Hamodrakas et al., 2007] Hamodrakas, S. J., Liappa, C., and Iconomidou, V. A. (2007). Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins. *International Journal of Biological Macromolecules*, 41(3):295–300.

[Kim et al., 2009] Kim, C., Choi, J., Lee, S. J., Welsh, W. J., and Yoon, S. (2009). NetCSSP: Web application for predicting chameleon sequences and amyloid fibril formation. *Nucleic Acids Research*, 37(SUPPL. 2):469–473.

[Kirschner et al., 1987] Kirschner, D. A., Inouye, H., Duffy, L. K., Sinclair, A., Lind, M., and Selkoe, D. J. (1987). Synthetic peptide homologous to beta protein from Alzheimer disease forms amyloid-like fibrils in vitro. *Proceedings of the National Academy of Sciences*, 84(19):6953–6957.

[Margittai and Langen, 2008] Margittai, M. and Langen, R. (2008). Fibrils with parallel in-register structure constitute a major class of amyloid fibrils: molecular insights from electron paramagnetic resonance spectroscopy. *Quarterly Reviews Of Biophysics*, 41(3-4):265–297.

[Maurer-Stroh et al., 2010] Maurer-Stroh, S., Debulpaep, M., Kuemmerer, N., de la Paz, M. L., Martins, I. C., Reumers, J., Morris, K. L., Copland, A., Serpell, L., Serrano, L., Schymkowitz, J. W. H., and Rousseau, F. (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods*, 7(3):237–242.

[Otzen and Nielsen, 2008] Otzen, D. and Nielsen, P. H. (2008). We find them here, we find them there: Functional bacterial amyloid. *Cellular and Molecular Life Sciences*, 65(6):910–927.

[Pepys, 2006] Pepys, M. B. (2006). Amyloidosis. *Annual Review of Medicine*, 57(1):223–241.

[Sachse et al., 2008] Sachse, C., Fändrich, M., and Grigorieff, N. (2008). Paired beta-sheet structure of an Abeta(1-40) amyloid fibril revealed by electron microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 105(21):7462–7466.

[Serpell et al., 2012] Serpell, L. C., Fraser, P., and Sunde, M. (2012). X-Ray Fibre Diffraction Studies of Amyloid Fibrils. *Amyloid Proteins : Methods and Protocols*, page 121.

[Sharma et al., 2005] Sharma, D., Shinchuk, L. M., Inouye, H., Wetzel, R., and Kirschner, D. A. (2005). Polyglutamine homopolymers having 8-45 residues form slablike β-crystallite assemblies. *Proteins: Structure, Function and Genetics*, 61(2):398–411.

[Shirahama, 1967] Shirahama, T. (1967). Published June 1, 1967. (15).

[Tartaglia et al., 2008] Tartaglia, G. G., Pawar, A. P., Campioni, S., Dobson, C. M., Chiti, F., and Vendruscolo, M. (2008). Prediction of Aggregation-Prone Regions in Structured Proteins. *Journal of Molecular Biology*, 380(2):425–436.

[Trovato et al., 2006] Trovato, A., Chiti, F., Maritan, A., and Seno, F. (2006). Insight into the Structure of Amyloid Fibrils from the Analysis of Globular Proteins. *PLoS Computational Biology*, 2(12):e170.

[Uversky and Fink, 2004] Uversky, V. N. and Fink, A. L. (2004). Conformational constraints for amyloid fibrillation: The importance of being unfolded. *Biochimica et Biophysica Acta – Proteins and Proteomics*, 1698(2):131–153.

[Ventura and Villaverde, 2006] Ventura, S. and Villaverde, A. (2006). Protein quality in bacterial inclusion bodies. *Trends in Biotechnology*, 24(4):179–185.

Mots clefs : Amyloids, large scale analysis, database

Meta-Repeat Finder, a pipeline to obtain the most complete set of tandem repeats in proteins and its application to large scale analysis across the tree of life

Session séquences protéiques
mardi 28 15h00
Salle des thèses

François Richard [*1,2], Étienne Villain[1,2], Andrey Kajava[1,2]

[1] Centre de Recherche en Biologie cellulaire de Montpellier (CRBM) – CNRS : UMR5237, Université Montpellier II - Sciences et techniques, Université Montpellier I – 1919 route de Mende, F-34293 Montpellier Cedex 5, France
[2] Institut de Biologie Computationnelle (IBC) – Université de Montpellier – 95 rue de la Galéra, F-34095 Montpellier, France

Introduction

Today, the growth of protein sequencing data significantly exceeds the growth of capacities to analyze these data. In line with the dramatic growth of this information and urgent needs in new bioinformatics tools our work deals with the development of new algorithms to better understand the sequence-structure-function relationship. Proteins contain a large portion of periodic sequences representing arrays of repeats that are directly adjacent to each other (Heringa, 1998), so called tandem repeats (TRs). TRs occur in at least 14% of all proteins (Marcotte et al., 1999). Moreover, they are found in every third human protein. Highly divergent, they range from a single amino acid repetition to domains of 100 or more repeated residues. Over the last decade, numerous studies demonstrated the fundamental functional importance of such TRs and their involvement in human diseases, especially cancer (Liggett and Sidransky, 1998; Morin, 1999). Thus, TR regions are abundant in proteomes and are related to major health threats in the modern society. In line with this, understanding of their sequence–structure–function relationship and mechanisms of their evolution promises to lead the identification of targets for new medicines and vaccines.

Results

Speeding up the performance of *Meta-Repeat Finder*

Existing TRs detection methods are optimized for finding TRs, over a specific range of repeat size and degree of their perfection (Kajava, 2012). Therefore, in order to obtain the most complete set of TRs one can imagine running all those tools in parallel to gather all types of repeats. This is the aim of *Meta-Repeat Finder*, a pipeline that we are developing. At present, it implements five different TRs detection methods including T-REKS (Jorda and Kajava, 2009), TRUST (Szklarczyk and Heringa, 2004), HHrepID (Biegert and Söding, 2008), MARCOIL (Delorenzi and Speed, 2002), pfsearch (Schuepbach et al., 2013). One of them, HHrepID, is the most sensitive approach for ab initio identification of long "covert" TRs, relying on Hidden Markov Model (HMM) – HMM comparisons, however, it is highly time consuming, representing a bottleneck of *Meta-Repeat Finder*'s performance in terms of speed. Therefore, to speed up the *Meta-Repeat Finder* we developed the TRDistiller filter (Richard and Kajava, 2014) that rapidly pre-selects proteins with TRs prior to HHrepID search. Discarding up to 20% of the no-TR containing sequences it ends

*. Intervenant

by an enrichment of the dataset in TRs and by doing so leading to a gain of time. After being identified separately by each identification method of the pipeline, TRs undergo validation process by evaluation of the universal score.

Tool for validation of existence of TRs

One of the problems is to distinguish between the sequences that contain highly imperfect TRs and the aperiodic sequences without TRs. The majority of TRs in sequences have repetitive arrangements in their 3D structure. Therefore, the 3D structure of proteins can be used as a benchmarking criterion for TR detection in sequence. According to our benchmark, none of the existing scoring methods are able to clearly distinguish, based on the sequence analysis, between structures with and without 3D TRs (Richard and Kajava, 2015). We developed a scoring tool called Tally (Richard et al., 2016), which is based on a machine learning approach. Tally is able to achieve a better separation between sequences with structural TRs and sequences of aperiodic structures, than existing scoring procedures. It performs at a level of 81% sensitivity, while achieving a high specificity of 74% and an Area Under the Receiver Operating Characteristic Curve of 86%. Tally can be used to select a set of structurally and functionally meaningful TRs from all TRs detected in proteomes.

Large-scale analysis

We selected 94 reference proteomes covering Archaea, Bacteria, Eukaryota as well as viruses and applied our *Meta-Repeat Finder* on their sequences to gather the most complete set of TRs for each proteome. Today, we aim to conduct the census of TR and to address the question of the relationship between their sequence, structure and function as well as the evolutionary mechanisms operating behind the TRs supplementing previous analyses (Andrade et al., 2001; Schaper and Anisimova, 2015; Schaper et al., 2014). Eventually, we want to better understand the relationship between TR and diseases, especially in human using the OMIM database (Amberger et al., 2015). This link has been made partially on specific cases linking the ankyrin repeats and Leucine Rich Repeats to cancer (Cheng et al., 2011; Ha et al., 2013; Jobling et al., 2014; Zhu and Bourguignon, 2000), as well as neurological disorders (Sinibaldi et al., 2004) or collagenopathies with other tandem repeats (Paladin et al., 2015) for instance, but no large scale analyses has been undertaken. Therefore, such a complete study can systematically link TRs to diseases allowing us to see the global picture of this phenomenon and give the opportunity to understand why such proteins play such an important role in diseases.

References

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43:D789–D798.

Andrade, M.A., Perez-Iratxeta, C., and Ponting, C.P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* 134:117–131.

Biegert, A., and Söding, J. (2008). De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinforma. Oxf. Engl.* 24:807–814.

Cheng, Z., Biechele, T., Wei, Z., Morrone, S., Moon, R.T., Wang, L., and Xu, W. (2011). Crystal structures of the extracellular domain of LRP6 and its complex with DKK1. *Nat. Struct. Mol. Biol.* 18:1204–1210.

Delorenzi, M., and Speed, T. (2002). An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* 18:617–625.

Ha, G.-H., Kim, J.-L., and Breuer, E.-K.Y. (2013). Transforming acidic coiled-coil proteins (TACCs) in human cancer. *Cancer Lett.* 336:24–33.

Heringa, J. (1998). Detection of internal repeats: how common are they? *Curr. Opin. Struct. Biol.* 8:338–345.

Jobling, R., D'Souza, R., Baker, N., Lara-Corrales, I., Mendoza-Londono, R., Dupuis, L., Savarirayan, R., Ala-Kokko, L., and Kannu, P. (2014). The collagenopathies: review of clinical phenotypes and molecular correlations. *Curr. Rheumatol. Rep.* 16:394.

Jorda, J., and Kajava, A.V. (2009). T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinforma. Oxf. Engl.* 25:2632–2638.

Kajava, A.V. (2012). Tandem repeats in proteins: From sequence to structure. *J. Struct. Biol.* 179:279–288.

Liggett, W.H., and Sidransky, D. (1998). Role of the p16 tumor suppressor gene in cancer. *J. Clin. Oncol.* 16:1197–1206.

Marcotte, E.M., Pellegrini, M., Yeates, T.O., and Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.* 293:151–160.

Morin, P.J. (1999). beta-catenin signaling and cancer. *BioEssays News Rev. Mol. Cell. Dev. Biol.* 21:1021–1030.

Paladin, L., Tosatto, S.C.E., and Minervini, G. (2015). Structural in silico dissection of the collagen V interactome to identify genotype-phenotype correlations in classic Ehlers-Danlos Syndrome (EDS). *FEBS Lett.* 589:3871–3878.

Richard, F.D., and Kajava, A.V. (2014). TRDistiller: a rapid filter for enrichment of sequence datasets with proteins containing tandem repeats. *J. Struct. Biol.* 186:386–391.

Richard, F.D., and Kajava, A.V. (2015). In search of the boundary between repetitive and non-repetitive protein sequences. *Biochem. Soc. Trans.* 43:807–811.

Richard, F.D., Alves, R., and Kajava, A.V. (2016). Tally: a scoring tool for boundary determination between repetitive and non-repetitive protein sequences. *Bioinformatics* btw118.

Schaper, E., and Anisimova, M. (2015). The evolution and function of protein tandem repeats in plants. *New Phytol.* 206:397–410.

Schaper, E., Gascuel, O., and Anisimova, M. (2014). Deep conservation of human protein tandem repeats within the eukaryotes. *Mol. Biol. Evol.* 31:1132–1148.

Schuepbach, T., Pagni, M., Bridge, A., Bougueleret, L., Xenarios, I., and Cerutti, L. (2013). pfsearchV3: a code acceleration and heuristic to search PROSITE profiles. *Bioinforma. Oxf. Engl.* 29:1215–1217.

Sinibaldi, L., De Luca, A., Bellacchio, E., Conti, E., Pasini, A., Paloscia, C., Spalletta, G., Caltagirone, C., Pizzuti, A., and Dallapiccola, B. (2004). Mutations of the Nogo-66 receptor (RTN4R) gene in schizophrenia. *Hum. Mutat.* 24:534–535.

Szklarczyk, R., and Heringa, J. (2004). Tracking repeats using significance and transitivity. *Bioinforma. Oxf. Engl.* 20(Suppl 1):i311–i317.

Zhu, D., and Bourguignon, L.Y. (2000). Interaction between CD44 and the repeat domain of ankyrin promotes hyaluronic acid-mediated ovarian tumor cell migration. *J. Cell. Physiol.* 183:182–195.

Mots clefs : Protein sequence, Tandem repeats, Large scale analysis, 3D structure

Improving pairwise comparison of protein sequences with domain co-occurrency

Christophe Menichelli [*] [†1,2], Laurent Bréhélin[1,2]

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – CNRS : UMR5506, Université de Montpellier – Bâtiment 5 - 860 rue de St Priest, F-34 095 Montpellier Cedex 5, France

[2] Institut de Biologie Computationnelle (IBC) – Institut national de la recherche agronomique (INRA), Institut de recherche pour le développement [IRD], INRIA, Centre de coopération internationale en recherche agronomique pour le développement [CIRAD], CNRS : UMR5506, Université de Montpellier – 860 rue St Priest, Bâtiment 5, CC05019, F-34 095 Montpellier Cedex 5, France

Session séquences protéiques
mardi 28 15h20
Salle des thèses

Proteins are macromolecules essential for the structuring and functioning of living cells. Domains occupy a key position among the relevant annotations that can be assigned to a protein. Protein domains are sequential and structural motifs that are found in different proteins and in different combinations and, as such, are the functionnal subunits of proteins above raw amino acid level. As a result, domain identification is an essential task in bioinformatics. Two kind of approaches can be used for identifying the domains of a target protein. A powerfull method is the profile analysis approach, also known as sequence-profile comparison. This is an non ab initio method which requires a database of domains. One of the most widely used database is the Pfam database [Finn RD et al., 2016]. In this database, each family of domains is defined from a manually selected and aligned set of protein sequences, which is used to learn a profile hidden Markov model (HMM) of the domain. To identify the domains of a protein, each HMM of the database is used to compute a score that measures the similarity between the sequence and the domain. If this score is above a predefined threshold, the presence of the domain can be asserted in the protein. However this method may miss several domains, when applied to an organism which is phylogenetically distant from the species used to train the HMMs. This may happen for two reasons. First when the models in the database are not adapted to the sequence specificity of the studied organism [Terrapon N et al., 2012]. Second when the domains in the target protein are simply absent from the database. Databases like Pfam were built with sequences that originate mostly from Plant and Unikont super-groups, and very few from the other groups. As a result, the proportion of proteins covered by a Pfam domain in Chromalveolates or Excavates species is almost half that identified in Plants and Unikonts [Ghouila A., 2014]. For example in *Plasmodium falciparum*, which is the organism responsible for the deadliest form of malaria, only 22 % of its residues are covered by a Pfam domain while these statics raise to 44 % for both yeast and human.

When the profile analysis does not offer good enough results, an alternative approach for identifying the domains of a protein is to run an ab initio approach based on sequence-sequence comparison using pairwise comparison tools like FASTA [Pearson WR et al., 1988] or BLAST [Altschul SF et al., 1990]. These tools look for local similarities between a query protein and a database of sequences like Uniprot [The UniProt Consortium, 2015]. They are based on local alignements, and use specific scoring functions for assessing similarities. These scores are then used to estimate a p-value (and e-value) under specific H 0 hypothesis of score distribution. Because sequence-sequence approaches do not include information from other homologous proteins, they are more prone to false positives than sequence-profil approaches. As a result, they are usually used with stringent score thresholds and hence may also miss several homologies. In order to improve sensitivity, different improvement of BLAST were proposed. One can cite for example PSI-BLAST [Altschul SF et al., 1997], which constructs a position-specific score matrix (PSSM)

*. Intervenant

†. Corresponding author : menichelli@lirmm.fr

to perform incremental searches, PHI-BLAST [Zhang Z. et al., 1998], which uses a motif to initiate hits or DELTA-BLAST [Boratyn G. M. et al., 2012], which searches a database of pre-constructed PSSMs before searching a protein-sequence database to yield better homology detection.

Surprisingly, domain co-occurrence has not been used to improve the sensitivity of sequence-sequence approaches so far. Domain co-occurrence is a strong feature of proteins, which relies on the fact that most protein domains tend to appear with a limited number of other domains in the same protein [Vogel C., 2004]. Functionnal studies have shown that domains that co-occur in proteins are more likely to display similar function [Ye Y., 2004] or structural cooperation [Li H., 1996] than domains in separate proteins. A well known example of co-occurrence are domains PAZ and PIWI which are frequently found together: when we look at proteins with the PAZ domain, we can also frequently see the PIWI domain. This information has already been used for improving the sensitivity of sequence-profil approaches [Terrapon N et al., 2009]. However it could also be of great help for the detection of sequence-sequence homology. For example, the attached figure reports homologies found between a protein of *Plasmodium falciparum* and several proteins from Uniprot. Most of these hits have moderate e-values and, taken independently, cannot be considered with high confidence. However, all these cases reveal co-occurrence of the same two or three independent sub-sequences, and thus add evidences to the identified homologies.

The core of our approach is a new scoring function that takes co-occurrence information into account for assessing BLAST hits. This new scoring function allows us to identify interresting hits that would not been considered on the basis of the BLAST results only because of their low e-values. The hits selected this way are then clustered and aligned to define new protein domains. Finally, the alignements are used to train an HMM for each new domain. Applied to *P. falciparum*, our method extends the number of significant hits by 20 %. Using these significant hits, we are able to identify 2,572 new domain occurrences in addition to the 6,443 occurrences identified by the Pfam HMMs thus increasing the residue coverage up to 16.11 % with an estimated false detection rate under 10 %. Among the 2 572 new domains, about half are not similar to any existing Pfam family.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol.* 1990 October 5; 215(3): 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389-3402.

Boratyn, G. M., Schaffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J., & Madden, T. L. (2012). Domain enhanced lookup time accelerated BLAST. *Biology Direct*, 7, 12.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(Database issue), D279-D285.

Ghouila, A., Florent, I., Guerfali, F. Z., Terrapon, N., Laouini, D., Yahia, S. B., … Bréhélin, L. (2014). Identification of Divergent Protein Domains by Combining HMM-HMM Comparisons and Co-Occurrence Detection. *PLoS ONE*, 9(6), e95275.

Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. *Science*. 1996 Aug 2;273(5275):666-9.

Pearson, W. R., Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8), 2444-2448.

Terrapon, N., Gascuel, O., Marchal, & Bréhélin, L. (2009). Detection of new protein domains using co-occurrence: application to *Plasmodium falciparum*. *Bioinformatics* (2009) 25 (23): 3077-3083.

Terrapon, N., Gascuel, O., Marchal, & Bréhélin, L. (2012). Fitting hidden Markov models of protein domains to a target species: application to *Plasmodium falciparum*. *BMC Bioinformatics*, 13, 67.

The UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, 43(Database issue), D204-D212.

Vogel C, Teichmann SA, Pereira-Leal J. The relationship between domain duplication and recombination. *J Mol Biol*. 2005 Feb 11;346(1):355-65. Epub 2004 Dec 23.

Ye Y., & Godzik, A. (2004). Comparative Analysis of Protein Domain Organization. *Genome Research*, 14(3), 343-353.

Zhang, Z., Schaffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V., & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research*, 26(17), 3986-3990. 3

Mots clefs : protein domain, domain co occurrency, BLAST, domain detection

BATfinder : alternative transcript selection in multiple sequence alignments

Héloïse Philippon[1], Alexia Souvane[1], Céline Brochier-Armanet[1], Guy Perrière [*] [†1]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69622 Villeurbanne Cedex, France

Session phylogénie 2
mardi 28 16h30
Amphi Mérieux

The reliability of a molecular phylogeny strongly depends on the quality of the dataset used. For analyses involving sequences in which alternative transcripts are available, it is necessary to select only one transcript per genomic locus. Indeed, lack of alternative transcripts selection will lead to the introduction of biases during the phylogenetic study. Because most of Multiple Sequences Alignment (MSA) algorithms tend to keep the gaps introduced in the early stages of computation, the first bias will occur during the alignment step. Since alternative transcripts are very similar, they are grouped together at the beginning of the process and gaps introduced because of exon skipping events will be kept in the final alignment. So, alternative transcripts will lead to misaligned sites in the flank regions and long gaps insertion in the MSA. The second bias occurs before the tree inference, during the selection of conserved blocks. Programs like Gblocks [1] or BMGE [2] select sites based on their conservation level and the number of gaps presents at a given position. So, alternative transcripts will lead to the overestimation of conservation rates, a same residue being represented more than once in the MSA. A selection of one sequence per genomic locus is thus essential before the inference of a molecular phylogeny.

Two simple approaches to overcome this problem consist in performing a random selection or keeping the longest transcript. The problem is that there is no justification for the former and the latter usually leads to the introduction of many gaps in the alignment, which can be problematic when considering sites homology. There are also two tools dedicated to automated transcript selection: PALO [3] and Guidance [4]. The first one uses a sequence length criteria and the second one alignment scores. Nevertheless, PALO is specifically designed to be used on protein sequences from the Ensembl database while Guidance is a rather general tool devoted to MSA quality analysis.

In that context, we developed BATfinder (Best Alternative Transcript finder) in order to provide a tool specifically devoted to the selection of alternative transcripts. BATfinder uses the same scoring function as the one implemented in Guidance but is faster and allows introducing special options to penalize gaps and/or short transcripts. Indeed, it appears that the choice of an appropriate alternative transcript requires to take into account three criteria: i) the isoform selected must have the highest possible similarity with sequences from closely related species; ii) it must minimize the number of gaps introduced in the alignment; and iii) it must be long enough. The options implemented in BATfinder allows to tune the balance between those three criteria.

BATfinder minimal requirement is a dataset of unaligned homologous protein sequences in Fasta format. Optionally, the user can provide a file in which the information on transcripts locus tag is given. In this case, BATfinder will also create a file in Fasta format containing the filtered dataset (*i.e.* in which only the transcript having the best score for a given gene is kept).

BATfinder is based on the Sum-of-Pairs (SP) score introduced by Carrillo et al. [5]. The program is divided into three main steps. The first one consists in the alignment of the input protein dataset to generate a reference alignment. The second step is the generation of a set of

*. Intervenant

†. Corresponding author : guy.perriere@univ-lyon1.fr

perturbed alignments built through a bootstrap procedure applied on the reference alignment. Finally, the third step is the computation of the SP score itself using those perturbed alignments. Due to the scoring methodology, the reference alignment and the perturbed alignments must be computed with the same MSA program. For each sequence, BATfinder returns a score comprised between 0 and 1 representing how well this sequence is aligned relatively to the others in the reference alignment. The alternative transcripts having the higher score being the one selected because they are the least disruptive.

An alternative faster option (-DS) is also available for large datasets. Instead of computing perturbed alignments, it simply sum up all pairwise distances of a transcript to the other sequences of the input dataset. The alternative transcript having the smaller sum of distances is selected as the most similar to the other homologs. To take into account the difference of length between alternative transcripts, this option uses a modified p-distance (observed divergence) for which a gap is considered as a supplementary character state. In that way, an alternative transcript that lack a region present in the others sequences, will have greater distances and not be selected.

To test the performances of BATfinder, we built datasets containing homologs of 45 proteins involved in the human autophagy pathway. We used the procedure and databases described in [6] for gathering the different sets of homologs. Among the 45 sets of homologs collected, four were very large and we were unable to align them with MAFFT [7] on our 16 Gbytes Virtual Machine (VM) used for testing, so we decided to discard them. The remaining 41 datasets contained from 42 to 1,484 sequences and the corresponding reference MSAs built with MAFFT ranged from 280 to 11,175 sites. Alternative transcripts represented from 2.62 % to 33.54 % of those datasets content.

For each dataset, we aligned the filtered output generated by BATfinder using MAFFT. Then, the corresponding phylogenetic trees were inferred by SeaView using *p*-distance and the BioNJ algorithm. We compared those trees to the one inferred on the alignments generated when the longest transcripts were used. For that purpose, we used the sum of all branch length of the trees as criteria, our hypothesis being that better alignments will results in trees having a smaller lengths, especially when using the *p*-distance as the measure of evolutionary distance.

On average, BATfinder filtering allows to obtain alignments resulting in significantly shorter trees than the ones obtained with the longest transcript selection. Indeed, Wilcoxon signed rank test for means comparison gives significant results when using default parameters ($P < 10^{-4}$) and the -gap option ($P = 1.19 \times 10^{-3}$). When looking closely at the results, we found that a shorter tree is obtained with BATfinder for 38 datasets among 41. The parameters allowing obtaining the shortest tree with BATfinder were: i) the default parameters for 18 datasets; ii) the -gap option for 13 datasets; iii) the combination of the -gap and -short options for five datasets; and iv) the -short option for two datasets. Optimal parameters setting thus strongly depends on dataset characteristics and we recommend to try the different possibilities before making a choice.

As they use the same scoring scheme, we wanted to compare the performances of BATfinder and Guidance in terms of speed only. For both programs we used MAFFT for computing the MSAs and JTT as the amino acid substitution model (Guidance default options). Also, the number of bootstrap replicates was set to 20. Computation were performed on a eight CPUs Linux CentOS VM cadenced at 2.6 GHz and having 16 Gbytes of RAM. As both programs are multithreaded, we used an increasing number of CPU to see the effect of parallelization level on their relative performances. When only one or two CPUs were used, Guidance outperforms BATfinder. On the other hand, BATfinder is usually faster than Guidance when the number of CPUs is larger or equal to three and the difference increases with this number. This is due to the fact that BATfinder has been optimized for parallel computing, especially the part devoted to evolutionary distances computation. As an illustration, BATfinder is slower than Guidance when using one or two CPUs for the dataset containing the alternative transcripts G13 homologs but, with the increasing number of CPUs, it becomes up to two times faster. For the dataset containing the SH3B4 homologs, BATfinder is from 3.6 (one CPU) to 13.3 (eight CPUs) times faster than Guidance. With four

CPUs, this represents a gain in computation time of about eight hours.

Implemented in C/C++ and easy to install, BATfinder is a freely available software than runs on Linux and MacOSX operating systems. It provides a much better alternative than the selection of the longest transcript and gives better results in term of alignment quality in the framework of phylogenetic reconstruction. Optimized for parallel computing it is also faster than its only direct equivalent when multi-threading is enabled.

References

[1] J. Castresana. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, 17(4):540–552, 2000.

[2] A. Criscuolo and S. Gribaldo. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC. Evol. Biol.*, 10:210, 2010.

[3] J. L. Villanueva-Cãnas, S. Laurie and M. M. Albà. Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol. Evol.*, 5(2):457–467, 2013.

[4] O. Penn, E. Privman, G. Landan, D. Graur and T. Pupko. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.*, 27:1759–1767, 2010.

[5] H. Carrillo and D. Lipman. The multiple sequence alignment problem in biology. *SIAM J. Appl. Math*. 48:1073–1082, 1988.

[6] H. Philippon, C. Brochier-Armanet and G. Perrière. Evolutionary history of phosphatidylinositol-3-kinases ancestral origin in eukaryotes and complex duplication patterns. *BMC Evol. Biol.* 15, 226, 2015.

[7] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, 30(14):3059–3066, 2002.

Mots clefs : Alternative transcripts, Bootstrap, Sum of Pairs, Phylogeny

Processus stochastiques avec sauts sur arbres : détection de changements adaptatifs

Paul Bastide * †[1,2], Mahendra Mariadassou[2], Stéphane Robin[1]

[1] UMR MIA-Paris – AgroParisTech, Institut National de la Recherche Agronomique - INRA, Université Paris-Saclay – 16 rue Claude Bernard, F-75 231 Paris Cedex 05, France
[2] Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut National de la Recherche Agronomique - INRA, Université Paris-Saclay – INRA Unité MaIAGE Bât. 233 Domaine de Vilvert, F-78 352 Jouy-en-Josas Cedex, France

Session phylogénie 2
mardi 28 16h50
Amphi Mérieux

Résumé court

En écologie comparative et évolutive, les traits quantitatifs d'un jeu d'espèces peuvent être vus comme le résultat d'un processus stochastique courant le long d'un arbre phylogénétique. Cette modélisation permet de prendre en compte les corrélations produites par une histoire évolutive partagée. Le processus stochastique est choisi afin de de capturer les mécanismes qui gouvernent l'évolution d'un trait. Les écologues préfèrent ainsi le processus d'Ornstein-Uhlenbeck (OU) au Mouvement Brownien (BM), plus simple mais moins réaliste. Le processus OU modélise la sélection naturelle s'opérant sur un trait par un mécanisme de rappel vers une valeur centrale, interprétée comme optimale dans un environnement donné. On s'intéresse ici à des changements de niche évolutive qui auraient entraîné un changement abrupt dans la valeur de cet optimum, et dont il s'agit de retrouver la position sur l'arbre. À partir des mesures d'un trait pour un jeu d'espèces liées par un arbre phylogénétique connu, on se propose de construire, d'étudier, et d'implémenter efficacement un modèle à données incomplètes permettant d'inférer simultanément la position des sauts et la valeur des paramètres. Les sauts sur l'arbre induisent naturellement une classification des espèces actuelles en groupes cohérents avec la phylogénie et définis par une même valeur optimale du trait. Au vu des données, seule cette classification est identifiable, ce qui pose problème pour la localisation exacte des sauts sur l'arbre. On se propose alors de dénombrer, d'une part, les allocations non-identifiables équivalentes, et, d'autre part, les solutions distinctes identifiables. Cette dernière quantité nous sert alors à calibrer une pénalité de sélection de modèle, pour laquelle on est capable de montrer une inégalité de type oracle dans le cas univarié.

Contexte

Notre objectif est ici d'étudier les traces laissées par des changements adaptatifs brutaux sur les valeurs actuelles des caractères d'un ensemble d'espèces liées par un arbre phylogénétique. Ces sauts ancestraux sont provoqués par des changements de niches écologiques, induites par exemple par des changements climatiques. Dans toute la suite, on suppose que l'arbre phylogénétique est donné, fixe, et calibré en temps, donc ultramétrique.

Modèle

On modélise l'évolution de la valeur des traits des différentes espèces considérées par un processus stochastique courant sur l'arbre phylogénétique. La valeur du processus sur une branche ancestrale de l'arbre représente la valeur prise par le trait de cette espèce au cours du temps. Lorsque l'on arrive à un événement de spéciation, c'est-à-dire lorsque la branche se divise en

*. Intervenant
†. Corresponding author : paul.bastide@agroparistech.fr

deux branches distinctes, le processus est également divisé en deux processus indépendants de même loi et partant du même point, modélisant chacun le trait d'une des deux espèces filles nouvellement formées. Le processus le plus simple est le Mouvement Brownien (BM), introduit dans ce cadre par Felsenstein [4]. Il ne représente cependant qu'un bruit pur, inapte à capturer des phénomènes de sélection. Pour prendre en compte ce phénomène, on a recours à un processus d'Ornstein-Uhlenbeck (OU), comme proposé initialement par Hansen [6]. Suivant ce modèle, l'évolution d'un trait le long d'une lignée est définie par une équation différentielle stochastique comportant, en plus d'un terme stochastique Brownien, un terme déterministe de rappel vers une valeur centrale, interprétée comme étant l'optimum primaire du trait, défini de manière mécanique par la niche écologique dans laquelle évolue l'espèce. La valeur du trait est attiré par cet optimum avec une vitesse contrôlée par un paramètre de rappel, vu comme une force de sélection. Pour juger de l'intensité de la sélection, on peut comparer le temps de demie-vie phylogénétique, défini comme le temps nécessaire pour que la moyenne du processus soit à mi-chemin entre sa valeur initiale et sa valeur optimale [6], à la hauteur totale de l'arbre. Si cette demi-vie est très courte par rapport au temps d'évolution total, cela signifie que les espèces trouvent rapidement un équilibre autour de leur optimum après une perturbation.

Saut adaptatif

Dans ce modèle, un changement adaptatif est alors représenté par un saut dans la valeur de l'optimum primaire pour une espèce donnée, et donc pour tous ses descendants, qui héritent de la nouvelle valeur de l'optimum. L'objectif est donc de retrouver les branches ancestrales de l'arbre sur lesquelles un tel changement a eu lieu.

État de l'art

Ce travail s'inscrit dans le champ de les méthodes comparatives phylogénétiques, qui sont aujourd'hui en plein développement [11], grâce notamment à la disponibilité récente de plus en plus d'arbres phylogénétiques de bonne qualité. La question de la détection de sauts sur l'arbre a reçu beaucoup d'attention ces dernières années [3,7,8,10,12], et a été traitée dans plusieurs cadres statistiques. L'originalité de la méthode décrite ici est de proposer une procédure d'inférence rigoureuse basée sur le maximum de vraisemblance, avec une description fine des modèles identifiables, et un critère de sélection de modèle solide pour lequel on peut exhiber une inégalité oracle dans le cas univarié. Cette méthode, dans le cadre univarié, est décrite en détail dans [2].

Problème d'identifiabilité

La position des sauts sur l'arbre produit de manière naturelle un certain nombre de groupes d'espèces, chacun défini par une histoire évolutive propre et la distribution du trait associée. Dans le cas d'un OU sur un arbre ultramétrique, on montre que seuls ces groupes d'espèces aux feuilles sont identifiables. Plusieurs allocations différentes des sauts sur l'arbre peuvent ainsi induire la même distribution aux feuilles, et ne sont donc pas distinguables. Il est donc possible que deux scénarios évolutifs différents ne puissent pas être départagés sur la base des observations disponibles.

Parcimonie

Le premier problème qui se pose est un problème de sur-paramétrisation. Par exemple, sans modifier la répartition des groupes aux feuilles, on peut ainsi ajouter un saut n'importe où dans l'arbre, puis l'annuler immédiatement par deux sauts contraires sur ses branches filles. Un premier remède consiste donc à ne considérer que les solutions qui sont minimales en terme de nombre de sauts. Une telle solution est dite parcimonieuse, et le groupement aux feuilles qu'elle produit ne saurait être produit par une solution concurrente comportant moins de sauts. Une modification

des algorithmes classiques de Fitch et Sankoff (décrits dans [5], chapitre 2) permet de compter et d'énumérer l'ensemble de ces configurations. La position des sauts sur l'arbre pouvant avoir une valeur pour l'interprétation historique de l'évolution des traits d'un ensemble d'espèces, il est important de fournir à l'utilisateur l'ensembles des configurations équivalentes concurrentes pour expliquer une distribution donnée du trait parmi les espèces actuelles.

Nombre de modèles identifiables

En supposant que chaque saut donne lieu à un nouvel optimum (hypothèse de non-homoplasie), l'ensemble des solutions identifiables à K sauts est en bijection avec l'ensemble des classifications en $K+1$ groupes dites « compatible avec l'arbre », c'est-à-dire obtenues par un processus de sauts. Un algorithme récursif nous permet de compter ces classifications. On constate ainsi que le nombre de modèles à K sauts distincts dépend en général de la topologie de l'arbre, sauf si ce dernier est binaire. Dans ce cas, ce nombre est donné par le nombre de sauts parmi le nombres de branches moins le nombre de sauts. La connaissance du nombre de solutions identifiables à nombre de sauts fixés nous permet de connaître la complexité de l'espace des modèles, et ainsi de dériver une procédure de sélection du nombre de sauts à positionner sur l'arbre.

Sélection de modèle

La procédure d'inférence statistique que nous proposons est basée sur un algorithme EM (Expectation-Maximization) qui tire avantage de la structure particulière des données en arbre pour trouver la solution maximisant la vraisemblance parmi tous les modèles comportant un nombre de sauts donné. On obtient ainsi une solution possible du problème pour chaque valeur du nombre de sauts. Cependant, comme le nombre de paramètre croît avec le nombre de sauts autorisés, ne garder que la solution avec la plus grande vraisemblance conduirait à choisir systématiquement la solution avec le plus de sauts possibles. Pour éviter ce travers, on a recours à une sélection basée sur une vraisemblance pénalisée. Une pénalité, liée à la complexité de l'espace des modèles telle que calculée précédemment, est ajouté à la vraisemblance. Le modèle réalisant le maximum de la vraisemblance pénalisée est alors celui réalisant le meilleur compromis entre pouvoir explicatif, et nombre de paramètres. La pénalité est dérivée à partir de la méthode décrite dans [1], et assure que le modèle sélectionné vérifie une inégalité oracle non asymptotique.

Gigantisme insulaire chez les Chéloniens

Les chéloniens sont une sous-classe de reptiles dont les seuls représentants actuels sont les tortues. Ils sont présents dans divers habitats partout dans le monde. Jaffe et al. [9] ont recensé les tailles de carapaces pour 226 espèces de tortues, marines, terrestres, insulaires ou d'eau douce, dont l'arbre phylogénétique est connu. Appliquée à ce jeu de donné, notre méthode sélectionne une solution comportant 5 sauts. On remarque que les espèces marines sont séparées dans un groupe à part, ainsi que les espèces insulaires, à quelques exceptions près, compatibles avec l'arbre. La valeur optimale pour ce dernier groupe est de 66 cm, contre 38 cm à l'origine, ce qui étaye bien l'hypothèse d'un gigantisme insulaire. Le temps de demi-vie trouvé est de 4.6 millions d'années, soit 2.3 % de la hauteur de l'arbre, ce qui indique une forte sélection.

Extension au multivarié

La procédure décrite ci-dessus peut s'étendre au cas où plusieurs traits, au lieu d'un seul, sont mesurés aux feuilles de l'arbres, pour les espèces actuelles. On impose d'abord à tous les traits de sauter en même temps sur l'arbre, ce qui nous ramène exactement dans le cadre développé pour le cas univarié pour ce qui est de l'identifiabilité des modèles considérés. Pour l'inférence proprement dite, la méthode précédente peut être étendue facilement sous réserve de faire l'une des deux

hypothèses suivantes. La première, explorée dans [8,10], est de considérer que tous les traits sont indépendants, avec chacun leur propre variance, et leur propre force de sélection. La seconde, que nous avons explorée plus particulièrement, est de relâcher l'hypothèse d'indépendance en permettant une matrice de variance covariance complète entre les différents traits, au prix d'une hypothèse supplémentaire sur la force de sélection, supposée identique pour tous les traits.

References

[1] Baraud, Y., Giraud, C., and Huet, S. (2009) Gaussian Model Selection with an Unknown Variance. *The Annals of Statistics*, 37(2), 630–672.

[2] Bastide, P., Mariadassou, M. and Robin, S. (2015) Detection of adaptive shifts on phylogenies using shifted stochastic processes on a tree. *arXiv*:1508.00225 (soumis, en révision)

[3] Butler, M.A. and King, A.A. (2004) Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist*, 164(6):683-695.

[4] Felsenstein, J. (1985) Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1-15.

[5] Felsenstein, J. (2004) *Inferring Phylogenies*. Sinauer Associates, Suderland, USA.

[6] Hansen, T.F. (1997) Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51(5):1341-1351.

[7] Ho, L. and Ané, C. (2014) Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods in Ecology and Evolution* 5(11):1133-1146.

[8] Ingram, T. and Mahler, D.L. (2013) SURFACE: detecting convergent evolution from comparative data by fitting Ornstein-Uhlenbeck models with stepwise Akaike Information Criterion. *Methods in Ecology and Evolution*, 4(5):416-425.

[9] Jaffe, A.L., Slater, G.J. and Alfaro, M.E. (2011) The evolution of island gigantism and body size variation in tortoises and turtles. *Biology letters*.

[10] Khabbazian, M., Kriebel, R., Rohe, K. and Ané, C. (2016) Fast and accurate detection of evolutionary shifts in Ornstein–Uhlenbeck models. *Methods in Ecology and Evolution*.

[11] Pennell, M.W. and Harmon, L.J. (2013). An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Ann. N. Y. Acad. Sci.*, 1289:90-105.

[12] Uyeda, J.C. and Harmon, L.J. (2014). A Novel Bayesian Method for Inferring and Interpreting the Dynamics of Adaptive Landscapes from Phylogenetic Comparative Data. *Syst. Biol.*

Mots clefs : Processus stochastiques, Ornstein Uhlenbeck, Segmentation, Sauts adaptatifs, Phylogénie, Sélection de modèle

Dating with transfers

Adrián Davín [*] [†1], Gergely Szöllősi[2], Éric Tannier[1,3], Bastien Boussau [‡ 1], Vincent Daubin [§ 1]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Eötvös University – Hongrie
[3] INRIA Rhône-Alpes (INRIA Grenoble Rhône-Alpes) – INRIA – ZIRST 655 avenue de l'Europe, Montbonnot, F-38 334 Saint Ismier Cedex, France

Session phylogénie 2
mardi 28 17h10
Amphi Mérieux

To reconstruct the timing of the diversification of life on Earth, biologists combine fossil evidence with inferences drawn from the comparison of genome sequences. These inferences are based on the molecular clock or on a softer version of it, the relaxed molecular clock. This approach consists of estimating the divergence between sequences and then, assuming that mutations occur clock-wise, trying to determine the age of the ancestral sequence. This method can be refined by using diverse models that relax the hypothesis of constant pace of evolution and consider that sequences can evolve at different speeds. Some models assume that these rates of evolution are independent along the different branches of the tree relating the different sequences of DNA; some others consider that the rates are correlated among related branches, so the rate of a given branch is inherited to some extent from the parental one. Which is of these methods perform best is still heavily debated. In spite of the sophistication of the different models, calculating these rates is not a trivial problem and the best estimates of divergence time have usually very wide confidence intervals. To overcome this problem scientists can use fossils, that can be independently dated using methods such as stratigraphy or radiometry. Fossils are useful because they provide external information that can be used to constrain the positions of the nodes in a species tree, improving the accuracy of the estimates of the molecular clock. Combining relaxed molecular clock estimates and fossil is in active field of research in phylogenetics [1].

However, fossils are extremely scarce in the geological record. For about 80 % of the history of life, all organisms were unicellular, which means that finding fossils becomes an almost impossible task. Bones and hard shells are easy to be preserved but they just became frequent after the Cambrian explosion, when all the major animal clades appear at sudden. Before that Earth was dominated by bacteria and to a minor extent, small eukaryotes. These organisms are extremely small organisms with no hard parts that can fossilize easily. On top of that, for the few existent fossils we have there is very little certainty about the clades to which they belong, since morphological features cannot be used to place them in a phylogenetic tree. This means that if we are interested in studying what happened in the distant past, we have very little help coming from fossils and we must rely almost exclusively in the information conveyed by the DNA. As we previously stated, this is a hard problem since the estimates of the molecular clock can vary widely. We need accurate calculations if we want to know for example when did Eukaryotes diversify or when did cyanobacteria appear on Earth.

To overcome these problems, we propose a new method of dating, based on the DNA sequence complementary to the molecular clock. Lateral gene transfer (LGT) is a common and almost universal phenomenon in nature, where different species (sometimes even species belonging to different domains) exchange genes. This can be detected using differences between species trees and gene trees. We do this using ALE, a method to reconcile species trees and gene trees that allows

*. Intervenant
†. Corresponding author : adrian.arellano-davin@univ-lyon1.fr
‡. Corresponding author : bastien.boussau@univ-lyon1.fr
§. Corresponding author : vincent.daubin@univ-lyon1.fr

detecting gene duplication, transfers and loss with high accuracy [2]. ALE takes distribution of gene trees to consider the uncertainty in the tree topology and estimates event rates by taking into account these gene trees. By doing this, it leads to better estimates of lateral gene transfers than by using methods that rely only on the comparison between gene trees and species trees, which have been shown in simulations to consistently estimate an incorrect number of transfers.

These gene transfers events contain information that can be used to order the divergence of different clades, since an existing clade can only donate a gene to other contemporary clades. Put in other words, if we detect a transfer between A and B necessarily means that the ancestors of A are older than any descendant of B. This same type of analysis can be performed over many thousand families to detect a large number of transfers that are then converted to a large number of node order constraints. This complements molecular dating analyses for a time when we don't have any fossils to use and every possible source of information must be used [3].

We analyzed several data sets to investigate whether the dating information carried by transfers agree with the information carried by the relaxed molecular clock. For each data set, we built species trees using concatenates of universal genes alignments and bayesian inference. We then computed gene tree distributions and inferred transfer events using the software ALE. We find that in all cases, dates based on relaxed molecular clocks are more consistent with the relative constraints coming from the transfers we detect than random trees. Further, among relaxed molecular clock estimates, we find that a particular model of rate evolution, where branchwise rates are independently drawn from a Gamma distribution, agrees consistently better with the transfer-based constraints than other models of rate evolution.

Our results show that transfers carry a signal for dating species trees that is compatible with and therefore can advantageously complement existing methods based on relaxed molecular clock. Further, they suggest an approach for choosing between different models of the rate of molecular evolution

References

[1] Lepage, T., Bryant, D., Philippe, H., & Lartillot, N. (2007). A general comparison of relaxed molecular clock models. *Molecular biology and evolution*, 24(12):2669-2680.

[2] Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, É., & Daubin, V. (2013). Efficient exploration of the space of reconciled gene trees. *Systematic biology*, syt054.

[3] Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, É., & Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences*, 109(43):17513-17518.

Mots clefs : Dating, molecular clock, phylogenetics, LGT, ALE

Indexer un ensemble de séquences ADN annotées

Tatiana Rocher [*1], Mikael Salson [†1], Mathieu Giraud [‡1]

[1] Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL) – INRIA – Université Lille 1, Bâtiment M3 extension Avenue Carl Gauss, F-59 655 Villeneuve d'Ascq, France

Session séquences nucléiques
mardi 28 16h30
Salle des thèses

Les structures d'indexation sont largement utilisées en bioinformatique, par exemple pour rechercher de courtes séquences dans un génome (BWA, Bowtie). Mais comment prendre en compte des annotations lors de cette recherche ? Nous souhaitons indexer un ensemble de séquences ADN annotées, telles que les séquences recombinées codant pour les récepteurs lymphoïdes.

Cet article présente une méthode d'indexation compressée de séquences ADN annotées permettant une recherche rapide d'informations grâce à la transformée de Burrows-Wheeler ainsi qu'un Wavelet Tree.

Introduction : recherches de séquences annotées

L'immunité adaptative est le mécanisme par lequel notre corps se défend de manière spécifique aux infections. Les lymphocytes B et T jouent un rôle clé dans cette immunité. À plusieurs endroits de leur ADN, ces cellules effectuent un mécanisme de recombinaison V(D)J, mécanisme donnant des milliards de possibilités différentes à partir d'un répertoire de quelques gènes. Un réarrangement V(D)J est par exemple V402 4/ACTG/0 J112. Cela signifie que la séquence ADN est composée :

— du gène V402 dont les 4 dernières lettres ont été tronquées,
— de la séquence ACTG,
— et du gène J112.

La diversité des réarrangements V(D)J implique la diversité des récepteurs des lymphocytes T tout comme des anticorps produits par les lymphocytes B.

Aujourd'hui, grâce au séquençage à haut-débit couplé à une analyse bioinformatique, il est possible d'avoir une vue précise d'une population de lymphocytes à un instant donné. Plusieurs logiciels, comme IMGT HighV-QUEST, IgBlast, Vidjil ou MiXCR peuvent quantifier les populations de lympocytes suivant leurs recombinaisons V(D)J, en immunologie fondamentale tout comme en hématologie des leucémies. En effet, le diagnostic et le suivi des leucémies aigües demandent d'analyser lymphocytes et lymphoblastes suivant leur recombinaisons V(D)J.

Pour mieux analyser la population d'un patient ou comparer des populations de lymphocytes de plusieurs patients, nous avons besoin d'une vue globale de l'ensemble des réarrangements VDJ du patient ainsi que d'une fonction de recherche des séquences très rapide. Est-il possible de réaliser un index stockant des séquences d'ADN (telles que ATGCGAT…CGATCGA, de taille 96) munies d'annotations (telles que V402 aux positions 1-64 et J112 aux positions 68-96) ?

Parmi les structures de données déjà existantes, une structure de type tableau des suffixes [9] ou arbre des suffixes [8], est trop volumineuse sur les données utilisées. La transformée de Burrow-Wheeler [3] et la compression LZ-77 [7] sont de très bons outils de compression de fichiers, mais l'entropie du texte augmente lorsque nous y ajoutons les annotations et sa compression devient moins bonne. Dans le cas des séquences de recombinaison VDJ, les séquences de gènes sont trop différentes pour pouvoir utiliser un FM-index for similar string [6].

*. Intervenant
†. Corresponding author : mikael.salson@univ-lille1.fr
‡. Corresponding author : mathieu.giraud@univ-lille1.fr

Nous proposons ici un nouvel index, inspiré d'un FM-index [5], permettant de stocker de telles séquences avec leurs annotations et de répondre en temps linéaire à des requêtes d'association annotation/séquence.

Méthodes

La structure

Notre structure utilise une transformée de Burrows-Wheeler (BWT) pour sauvegarder les séquences, un Wavelet Tree (WT) pour leur associer leurs annotations, et un vecteur de bits pour lier les deux.

La transformée de Burrows-Wheeler (BWT) [3] est un algorithme permettant de réorganiser les lettres d'un texte. Le texte transformé est la concaténation de la dernière lettre de chacune des rotations du texte, triées dans l'ordre lexicographique. Cette transformée sans perte d'information permet une lecture du texte aussi rapide que sur le texte original, et le texte transformé peut mieux se compresser.

Le Wavelet Tree (WT) [4] est un arbre. Les feuilles stockent une information. Les nœuds internes sont composés d'un vecteur de bits. Les bits 0 indiquent que la feuille correspondant à l'information de cette position se trouve dans le sous-arbre gauche, les bits 1, dans le sous-arbre droit. Cet arbre permet de rassembler dans une feuille toutes les occurences d'une information et de faire des recherches facilement.

Nous appelons annotation l'information que nous souhaitons ajouter comme commentaire sur une partie du texte. L'annotation doit être associée à une portion des lettres du texte (par exemple : annotation 1 allant des lettres 12 à 54). De plus, ces annotations peuvent être hiérarchisées. Par exemple, pour les recombinaisons VDJ, une annotation peut être une famille de gènes (V), un gène (V2), ou un allèle (V201).

Soit T, le texte de taille n incluant l'ensemble des séquences que nous souhaitons stocker. Nous transformons T avec la transformée de Burrow-Wheeler : BWT(T). Puis nous ajoutons un vecteur d'annotations A correspondant aux annotations des lettres de BWT(T) : $A[i] = a_{\{}\text{BWT}(T)[i]\}$, avec i étant une position dans A et a_j étant l'annotation de la lettre j. (Le vecteur A n'est utilisé que lors de la construction). Une même annotation est souvent sur un même facteur, donc après application de la tranformée, plusieurs ensembles d'annotations sont consecutifs. Nous associons au texte transformé un vecteur de bit B de taille n avec les conditions suivantes : $B[i] = 0$ si $A[i] = A[i-1]$, $B = 1$ sinon, et $B[0] = 1$. Chaque annotation correspondant à un bit 1 est mise dans la racine du WT. Nous ajoutons aux feuilles de notre WT un entier indiquant le nombre d'apparitions de l'annotation dans le texte.

Optimisation du WT

Nous mettons deux contraintes sur le WT : optimiser la structure de l'arbre pour avoir des requêtes plus rapides, et hiérarchiser les annotations. Afin d'optimiser le temps de réponse à une requête, nous utilisons la forme d'un arbre de Huffman [2]. Les feuilles des annotations les plus fréquentes sont placées plus haut dans l'arbre.

La hiérarchie des annotations des recombinaisons VDJ se représente par un arbre k-aire. Les annotations ayant la même précision, par exemple V, D et J, sont à la même profondeur de l'arbre. Le sous-arbre descendant du nœud V ne contient que des nœuds de la catégorie V ayant une meilleure précision : V1, V2*01…

Nous transformons l'arbre k-aire (muni des fréquences d'apparition des nœuds) en WT, arbre binaire, en lui appliquant au maximum une structure de Huffman. Pour cela, nous parcourons les différentes profondeurs de l'arbre, de la plus profonde vers la moins profonde, pour organiser les fils du nœud courant en utilisant une structure d'un arbre de Huffman.

Tous les vecteurs de bits (le vecteur lien et les vecteurs du WT) sont associés aux fonctions rank et select. La fonction rank_b(V, i) renvoie le nombre de fois où le bit b apparait dans le préfixe [1, i] du vecteur de bits V. La fonction select_b(V, j) renvoie la position i du jième bit b dans le vecteur de bits V. Toutes deux s'effectuent en temps constant dans le vecteur [10].

Taille de la structure

Soient n, la taille de l'ensemble des séquences, l, le nombre d'annotations présentes dans l'ensemble des séquences analysées, et z, le nombre d'annotations différentes de l'annotation précédente dans la BWT.

La BWT produit un texte dont la compression varie en fonction de l'entropie du texte T. La taille du texte compressé est donc n * H_k(T) avec H_k(T) étant l'entropie d'ordre k du texte T. L'arbre a une taille totale de z * log(l). La taille totale de la structure est alors : O(n*H_k(T)+ n + z * log(l)) Dans le pire des cas, l'ensemble des annotations est utilisé et, après transformation du texte, deux annotations adjacentes ne sont jamais identiques, donc z=n. Dans ce cas là, la racine du WT est de taille n.

En pratique, le répertoire immunologique d'un individu présente souvent une ou des recombinaisons VDJ présentes en plusieurs exemplaires : entre 1 % et 5 % chez un individu sain, et entre 5 % et 80 % chez une personne atteinte de leucémie. La taille théorique est donc largement surestimée. Par exemple, pour un texte constitué de 10^6 lettres, la structure aura une taille maximum de 13×10^6 bits (2 bits par lettres pour la BWT, 1 bit par lettre pour le vecteur lien et un WT maximum : 10 fois la taille du vecteur lien), mais nous pouvons espérer une taille de 10^6 bits (une compression de 50 % lors de la BWT, des annotations économisant 9 lettres sur 10 et certaines annotations très présentes donnant un WT très déséquilibré).

Requêtes possibles

Cette structure permet de répondre efficacement à plusieurs requêtes liant séquences et annotations.

Soit les abréviations suivantes :
— r : taille d'une séquence,
— l : nombre total de feuilles/annotations,
— b_f : nombre de bits sur la feuille correspondant à l'annotation f (au plus z).

1) Le nombre de séquences possédant une annotation : ce nombre est connu par un simple accès à l'entier stocké dans la feuille possédant cette annotation. Cette recherche se fait en temps $O(\log(l))$. Cette requête permet d'avoir une vue d'ensemble des proportions de présence des annotations dans l'ensemble des séquences, comme la proportion de séquences présentant le gène V4.

2) L'ensemble des annotations associées à une séquence S appartenantà T : pour chaque lettre de S, nous cherchons l'annotation correspondante dans le WT. La complexité totale est : $O(r \times log(l))$. Nous pouvons éviter certains parcours de l'arbre en recherchant l'annotation d'une lettre toutes les x lettres, et ainsi réduire la complexité. De plus, nous pouvons arrêter la recherche de l'annotation d'une séquence dans un nœud interne de l'arbre lorsque nous aurons obtenu le niveau de précision voulu de l'annotation.

3) La liste des séquences possédant une annotation f : nous récupérons tous les bits de la feuille associée à l'annotation f. Puis nous remontons l'arbre pour chacun de ces bits pour trouver les séquences associées à ce bit. Enfin, nous parcourons la séquence dans la BWT pour trouver son identifiant. La complexité totale est : $O(b_f * log(l) + (r * y))$, avec y étant l'ensemble des lettres possédant l'annotation f. Cette fonction a une complexité assez grande principalement à cause de la taille de b_f et à la redondance des séquences trouvées à lire. Cette requête permet par exemple de récupérer toutes les séquences ayant le gène V4, pour les aligner et les étudier.

Perspectives

Nous avons proposé une structure permettant de stocker des séquences ADN ainsi que des annotations associées à certaines positions. Appliquée aux recombinaisons VDJ, cette structure permet d'avoir une vision d'ensemble du répertoire immunitaire d'une personne, en connaissant l'ensemble des séquences lymphoïdes ainsi que les gènes qui les composent. Nous allons à présent procéder à l'implémentation de la structure de données, à l'évaluation de celle-ci sur des données réelles, et optimiser certains algorithmes de la structure pour les rendre spécifiques aux données VDJ, en particulier pour pouvoir rechercher toutes les séquences qui ont une annotation spécifique telle que V102/D3/J402. Nous souhaitons ensuite donner une estimation de la qualité des données à l'utilisateur. Nous voulons pouvoir lui indiquer les différences entre les gènes présents sur les séquences analysées et ceux de référence proposés par IMGT/GENE-DB. Couplée au logiciel Vidjil [1], cette structure de donnée sera utilisée pour faire des comparaisons de répertoire immunologique intra- et inter-patients.

Références

[1] M. Giraud, M. Salson, et al.. Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC Genomics*, vol 15, 2014.

[2] D.A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the I.R.E.*, 1952.

[3] M. Burrows, D. J. Wheeler. A block-sorting lossless data compression algorithm. *Digital SRC Research*, 1994.

[4] R. Grossi, A. Gupta et J.S. Vitter. High-order entropy-compressed text indexes. *Symposium on Discrete Algorithms (SODA 2003)*, 2003.

[5] P. Ferragina et G. Manzini. Opportunistic Data Structures with Applications. *Symposium on Foundations of Computer Science (FOCS 2000)*, 2000.

[6] J.C. Na, H. Kim, H. Park, T. Lecroq, M. Léonard, L. Mouchard, K. Park. FM-index of alignment: A compressed index for similar strings. *Theoretical Computer Science*, 2015.

[7] J. Ziv et A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, volume 23, numéro 3, pages 337-343, 1977.

[8] P. Weiner. Linear pattern matching algorithms. *IEEE Symposium on Switching and Automata Theory*, 1973.

[9] U. Manber et G. Myers. Suffix arrays: a new method for on-line string searches. *Symposium on Discrete Algorithms (SODA 1990)*, 1990.

[10] G. Jacobson. Space-efficient static trees and graphs. *Symposium on Foundations of Computer Science (FOCS 1989)*, 1989, pages 549–554.

[11] IMGT. Base de données IMGT des gènes VDJ.`http://www.imgt.org/genedb`, Accessed: 2016-04-22.

Mots clefs : indexation, transformée de Burrows, Wheeler, Wavelet Tree, immunologie, hématologie, recombinaison VDJ, répertoire immunitaire

Impact de la recherche d'amorces mutées sur les résultats d'analyses métagénomiques

Aymeric Antoine-Lorquin [*] [†1], Frédéric Mahé[2], Micah Dunthorn[3], Catherine Belleannée [‡1]

[1] DYLISS (INRIA - IRISA) – INRIA, Université de Rennes 1, CNRS : UMR6074 – Campus de Beaulieu, F-35 042 Rennes Cedex, France
[2] CIRAD – Centre de coopération internationale en recherche agronomique pour le développement : UPR39 – France
[3] Technical University of Kaiserslautern (TU Kaiserslautern) – PO Box 3 049 67.653 Kaiserslautern, Allemagne

Session séquences nucléiques
mardi 28 16h50
Salle des thèses

Introduction

En métagénomique ciblée, une problématique récurrente concerne la quantité de lectures réellement exploitables en sortie d'un séquenceur à haut débit. L'une des étapes influençant cette quantité est la détection dans chaque séquence des amorces utilisées pour amplifier le gène ciblé. Huse et al. (2010) ont émis l'hypothèse que ne retenir que les séquences disposant d'amorces parfaites (non-mutées par une erreur de séquençage) permettrait d'améliorer la qualité globale d'un échantillon. Cette approche, basée sur l'idée que les séquences des amorces doivent dominer numériquement à l'issue de l'étape de PCR, présente l'avantage de ne pas nécessiter l'utilisation d'outils complexes, puisqu'il est possible de rechercher les amorces parfaites à l'aide de simples expressions régulières, naturellement prises en charge par de nombreux langages de programmation (python, perl, ruby, etc.). Néanmoins, cette stratégie élimine probablement aussi des séquences correctes. C'est pourquoi, maintenant qu'il existe des outils capables d'éliminer *a posteriori* les séquences les moins fiables (tel que SWARM de Mahé et al. (2015b)), nous avons voulu savoir si le fait de rechercher des amorces potentiellement mutées permettait d'augmenter le nombre de séquences exploitables par échantillon et si cette augmentation impactait les résultats obtenus au terme de l'analyse métagénomique.

Matériel et méthodes

Jeux de données. Nous avons travaillé sur 9 échantillons biologiques, dans le cadre d'une étude caractérisant la biodiversité des sols tropicaux chez les eucaryotes unicellulaires (Mahé et al. (2015a)). Les échantillons ont été séquencés à la fois en Roche/454 et en Illumina MiSeq, afin de pouvoir constater l'impact de la technologie sur les résultats (i.e. obtient-on les mêmes séquences avec les deux technologies ? Les résultats sont-ils les mêmes en Roche/454 et en Illumina Miseq ?). Le séquençage a ciblé la région V4 de la sous-unité 18S de l'ARN ribosomique (Stoeck et al. (2010)). En effet, cette région possède une portion hypervariable spécifique de chaque espèce et encadrée par deux segments de séquences extrêmement conservés utilisables en tant qu'amorces universelles (ci-après amorces *Forward* et *Reverse*). La détection et l'élimination des amorces dans les échantillons séquencés permet donc d'isoler la partie spécifique, appelée amplicon.

Pour l'ensemble des 9 échantillons, le séquençage Roche/454 a produit 310 375 séquences, contre 5 223 138 séquences sous Illumina MiSeq.

*. Intervenant
†. Corresponding author : aymeric.antoine-lorquin@irisa.fr
‡. Corresponding author : catherine.belleannee@irisa.fr

Recherche des amorces. La recherche des amorces parfaites s'est faite à l'aide d'expressions régulières recherchées en Python (CCAGCA[CG]C[CT]GCGGTAATTCC pour V4F et T[CT][AG]ATCAAGAACGAAAGT pour V4R) ; les séquences obtenues forment l'ensemble des *amplicons Regex* (nombre total d'amplicons Regex Roche/454 : 39 917, Illumina Miseq : 274 801). La recherche des amorces mutées parmi les séquences non regex s'est faite à l'aide de l'outil de *pattern matching* grammatical Logol (Belleannée et al. (2014)) ; les séquences obtenues forment l'ensemble des *amplicons Logol* (nombre total d'amplicons Logol Roche/454 : 2 520, Illumina Miseq : 20 558).

Calcul de la proximité de deux ensembles de séquences. Le test de dissimilarité de Bray-Curtis permet de visualiser sur un graphe la similarité de deux ensembles de séquences. Il est basé sur les comparaisons 2 à 2 des profils de séquences de deux ensembles de taille identique. Les séquences Regex étant bien plus nombreuses que les séquences Logol, la valeur finale de dissimilarité a été obtenue en faisant la moyenne de 10 000 calculs de dissimilarité de sous-échantillons aléatoires de 1 000 séquences Regex ou Logol, et ce indépendamment pour chaque échantillon biologique.

Regroupement des séquences par similarité. Une clustérisation a été faite avec l'outil SWARM (Mahé et al. (2015b)) sur la totalité des amplicons Regex et Logol et ce séparément pour chaque technologie. Chaque cluster constitué par SWARM contient des séquences proches les unes des autres à quelques substitions près. La règle de validation d'un cluster d'amplicons a été la suivante : pour être conservé, un cluster doit regrouper un minimum de 3 séquences ou au moins 2 séquences provenant de 2 échantillons différents. Chaque cluster valide forme un OTU (*Operationnal taxonomic unit*, unité taxonomique opérationnelle) souvent considéré comme représentant une espèce biologique.

Résultats

Modèle d'amorces mutées. Le modèle de mutation défini par l'expert est le suivant : les amorces Forward et Reverse peuvent posséder jusqu'à 2 substitutions OU jusqu'à 1 insertion-délétion. Par ailleurs, l'amorce Reverse peut être partiellement tronquée sur ses 2 nucléotides terminaux sans que cela compte comme une délétion (i.e. on autorise l'amorce Reverse a être légèrement incomplète, avant de considérer la présence de mutations).

Il n'est pas simple de définir de tels modèles uniquement avec les expressions régulières. En effet, les expressions régulières nécessitent de définir explicitement chaque possibilité recherchée. Par exemple, autoriser deux substitutions sans apriori de position sur un mot de taille n crorrespond à n x (n-1)/2 variants ; soit 190 mots différents pour une séquence de taille 20, donc un nombre d'expression régulière explosif en fonction de la taille du mot. C'est pour cette raison que nous avons utilisé un outil de *pattern matching* permettant d'exprimer des expressions régulières approchées, Logol, qui permet de rechercher facilement des variants de séquences, notamment en autorisant l'ajout de propriétés aux modèles, tel que le nombre d'erreurs autorisées par rapport à une référence. Ainsi, Logol peut couvrir les 190 modèles d'expressions régulières de l'exemple précédent avec un unique modèle d'expression.

Détection des séquences avec amorces mutées. La recherche des amorces parfaites permet de récupérer respectivement 90,2 % et 82,7 % des séquences totales. La recherche d'amorces mutées permet de capturer respectivement 8,3 % et 7,1 % de séquences additionnelles (pour un total respectif de 98,5 % et 89,8 %). La recherche des amorces mutées permet donc d'ajouter une quantité non-négligeable de séquences (+25 619 séquences en Roche/454 et +368 260 séquences en Illumina MiSeq).

Validation des séquences avec amorces mutées

Deux situations peuvent se présenter pour unamplicon Logol : soit il est déjà connu (typiquement, il est identique à unamplicon Regex) et donc tout autant biologiquement envisageable ; soit il est nouveau et la question se pose de savoir s'il ne s'agit pas d'un amplicon trop lourdement muté

ou chimérique, ne reflétant donc pas la réalité biologique. Pour répondre à cette interrogation, nous avons regardé si l'ensemble des amplicons Logol formait une population comparable à la population Regex, puis nous avons observé le devenir individuel de chaque amplicon.

Comparaison globale des amplicons Logol et Regex. L'utilisation du test de dissimilarité de Bray-Curtis a permis de vérifier que les amplicons Logol sont relativement similaires aux amplicons Regex. Les résultats montrent d'une part que les amplicons Logol sont très proches des amplicons Regex et d'autre part, que les amplicons Logol et Regex obtenus via la même technologie de séquençage sont plus proches entre eux qu'avec leurs homologues de la technologie de séquençage alternative pour un échantillon donné (par exemple, les amplicons Regex/Illumina sont plus proches des amplicons Logol/Illumina que des amplcions Regex/454). Les amplicons Logol sont donc comparables aux amplicons Regex, malgré leurs amorces mutées.

Caractérisation individuelle des amplicons Logol. Dans le cadre de notre analyse, chaque amplicons Logol correspond à l'une des 4 situations suivantes après la clustérisation par SWARM :

- Cas 1) L'amplicon Logol appartient à un cluster non-valide, i.e. est isolé. Il n'a alors pas d'impact sur les résultats finaux car il en sera éliminé. Les amplicons chimériques font par exemple partie de cette catégorie
- Cas 2) L'amplicon Logol s'ajoute à un cluster Regex valide. Il augmente ainsi l'abondance totale du cluster mais ne modifie pas le nombre d'OTU identifiés. Cela valide tout de même l'amplicon.
- Cas 3) L'amplicon Logol s'ajoute à un cluster Regex non-valide et le rend valide. Il permet l'identification d'un nouvel OTU.
- Cas 4) L'amplicon Logol appartient à un cluster valide purement Logol. Il permet l'identification d'un nouvel OTU.

Les deux derniers cas sont particulièrement intéressants, puisqu'ils modifient concrètement le résultat de l'analyse, en ajoutant de nouveaux OTU.

En Roche/454, 9 % des séquences Logol appartiennent à un cluster non-valide (contre 5 % pour les séquences Regex) et sont donc rejetées. La clustérisation aboutit à 4 432 OTU, dont 2 059 contiennent des séquences Logol (46,5 %). Pour 1 632 OTU, il s'agit d'une augmentation de l'abondance totale d'OTU déjà détectés avec les séquences Regex. Les 427 autres OTU sont de nouveaux OTU (dont 205 uniquement constitués de séquences Logol).

En Illumina, 24 % des séquences Logol appartiennent à un cluster non-valide (contre 20 % pour les séquences Regex) et sont donc rejetées. La clusterisation aboutit à 28 377 OTU, dont 12 693 contiennent des séquences Logol (44,7%). Pour 10 835 OTU, il s'agit d'une augmentation de l'abondance totale d'OTU déjà détectés avec les séquences Regex. Les 1 858 autres OTU sont de nouveaux OTU (dont 937 uniquement constitués de séquences Logol).

La récupération des séquences disposant d'amorces mutées a donc permis l'identification de 10 à 7 % de nouveaux OTU dans les échantillons (respectivement en 454/Roche et Illumina).

Conclusion

La recherche des amorces mutées permet d'améliorer de façon non-négligeable la sensibilité d'une analyse métagénomique en augmentant le rappel parmi les séquences d'un échantillon.

Bien sûr, cette recherche nécessite l'utilisation de moyens adaptés. Il existe un certain nombre d'outils disponibles pour mettre en œuvre la recherche des amorces mutées, tels que CutAdapt (Martin, 2011), simple et très rapide mais assez peu flexible, ou Logol, beaucoup plus lent, qui permet un contrôle complet sur les spécificités du modèle. En elle-même, cette recherche est simple à inclure dans les pipelines métagénomiques existants.

Un post-filtrage par clustérisation permet d'éliminer, parmi les nouveaux candidats, les séquences isolées et de ne conserver que les séquences similaires aux séquences à amorces parfaites. Ces séquences supplémentaires permettent d'accroître la sensibilité des analyses métagénomiques,

en permettant la détection de nouveaux OTU (+7 à +10 %, dans notre étude, en fonction de la technologie de séquençage), que ce soit en augmentant l'abondance de séquences détectées en nombre insuffisant pour être validées ou que ce soit en détectant des séquences totalement nouvelles qui n'étaient pas visibles auparavant.

Références

Belleannée, C., Sallou, O., and Nicolas, J. (2014). "Logol: Expressive Pattern Matching in Sequences. Application to Ribosomal Frameshift Modeling". *Pattern Recognition in Bioinformatics*, number 8626 in Lecture Notes in Computer Science, pp 34–47. Springer International Publishing. `https://hal.inria.fr/hal-01059506v1`

Bray, J. Roger, and J. T. Curtis. "An Ordination of the Upland Forest Communities of Southern Wisconsin". *Ecological Monographs* 27.4 (1957): 326–349. Web.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). "Ironing out the wrinkles in the rare biosphere through improved OTU clustering". *Environmental Microbiology*, 12(7):1889–1898.

Mahé, F., Mayor, J., Bunge, J., Chi, J., Siemensmeyer, T., Stoeck, T., Wahl, B., Paprotka, T., Filker, S., and Dunthorn, M. (2015a). "Comparing High-throughput Platforms for Sequencing the V4 Region of SSU-rDNA in Environmental Microbial Eukaryotic Diversity Surveys". *Journal of Eukaryotic Microbiology*, 62(3):338–345.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015b). "Swarm v2: highly-scalable and high-resolution amplicon clustering". *PeerJ*, 3:e1420.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". *EMBnet.journal*, 17(1):pp. 10–12.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H.-W., and Richards, T. A. (2010). "Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water". *Molecular Ecology*, 19 Suppl 1:21–31.

Mots clefs : pattern matching, Illumina, Roche/454, 18S ribosomal RNA

Sequencing a large plant Y chromosome using the MinION

Cécile Fruchard *[1], Nelly Burlet[1], Roman Hobza[2], Gabriel Marais[1]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Institute of Experimental Botany – Laboratory of Molecular Cytogenetics and Cytometry ; Centre of the Region Hana for Biotechnical and Agricultural Research, Sokolovska 6, CZ-77200, Olomouc, Czech Republic / République tchèque

Session séquences nucléiques
mardi 28 17h10
Salle des thèses

Sequencing Y chromosomes remains one of the greatest challenges of current genomics. In sharp contrast with the > 300 fully sequenced eukaryotic genomes, only a handful of Y chromosomes have been sequenced to date. The non-recombining Y chromosome tends to accumulate repeats (TEs and amplicons) which renders the assembly using small-read sequencing technologies virtually impossible. We aim at sequencing the Y chromosome of *Silene latifolia*, a well-studied dioecious plant, which represents a real challenge as the Y, in this species, is 550 Mb long and probably comprises a very large fraction of repeats. We are using a hybrid approach, combining both Illumina pair-end sequencing and Oxford Nanopore MinION sequencing to sequence and assemble the *S. latifolia* Y chromosome. As part of the MinION Access Programme, we are currently testing this device using bacterial pores for single-molecule electronic sequencing. We will present preliminary results on using the MinION sequencing device to help with the de novo assembly of complex chromosomes.

Mots clefs : sex chromosomes, nanopore sequencing, *Silene latifolia*

*. Intervenant

Keynote

Session génomique des populations
mercredi 29 09h00
Amphi Mérieux

Integrative population genomics of mosquito-arbovirus interactions

Louis Lambrecht [*1],

[1] Institut Pasteur, Department of Genomes and Genetics, Paris, France

Mosquito-borne transmission of human pathogens such as dengue and Zika viruses depends on a complex biological interaction between the virus and the insect vector. Both mosquitoes and arboviruses in nature consist of genetically diverse populations. In this talk, I will illustrate how the integrated genomic analysis of virus and mosquito populations can provide insights into the ecological and evolutionary processes underlying mosquito-borne transmission of arboviruses.

*. Intervenant Invité

tess3r : un package R pour l'estimation de la structure génétique des populations spatialisées

Kevin Caye [*1], Olivier François [†2], Michel Olivier[3]

[1] Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG) – CNRS : UMR5525 – Domaine de la Merci, F-38 706 La Tronche, France
[2] TIMC-IMAG – Université Joseph Fourier - Grenoble I, CNRS : UMR5525 – Grenoble, France
[3] Grenoble Images Parole Signal Automatique (GIPSA-lab) – Université Stendhal - Grenoble III, Université Pierre Mendès-France - Grenoble II, Université Joseph Fourier - Grenoble I, CNRS : UMR5216, Institut Polytechnique de Grenoble - Grenoble Institute of Technology, Université Pierre-Mendès-France - Grenoble II – Gipsa-lab - 961 rue de la Houille Blanche - BP 46, F-38 402 Grenoble Cedex, France

Session génomique des populations
mercredi 29 10h00
Amphi Mérieux

Une étape importante lors de l'étude de données de grande dimension est la recherche d'une structure de faible dimension mettant en évidence les principales caractéristiques des données. Cette étape est très importante en génétique des populations. En particulier elle est utile pour détecter les signatures laissées par l'histoire démographique de la population étudiée. De nombreuses méthodes ont été développées pour estimer la structure génétique des populations dans des échantillons où les individus sont issus du métissage de plusieurs sous-populations. La plupart des jeux de données issus d'espèces naturelles présentent une forte autocorrélation spatiale. Quand elle est disponible il est donc intéressant d'utiliser l'information spatiale pour estimer la structure de population. Cela permet de représenter la répartition spatiale des sous-populations de manière continue. De plus, on sait que l'adaptation des individus à leur milieu induit des différences de fréquence allélique entre les différentes sous-populations. Il est possible d'utiliser l'estimation de la structure de population pour détecter des gènes potentiellement sous sélection environnementale. Enfin il est important de noter que la taille des jeux données récoltés par les généticiens des populations à largement explosée ces dernières années. Il est donc nécessaire que les outils d'analyse s'adaptent à cette tendance.

Le logiciel tess3r est implémenté sous la forme d'un package du logiciel libre de traitement des données et d'analyse statistique R. Le package R tess3r permet d'estimer les coefficients de métissage individuels et les fréquences alléliques de chaque sous-population à partir de la matrice de génotype et des coordonnées spatiales des individus. La méthode repose sur un problème de factorisation de matrice régularisé par un graphe afin de garantir la continuité spatiale de l'estimation des coefficients de métissage individuels. Le problème d'optimisation est résolu à l'aide d'un algorithme alternant une phase de résolution d'un problème des moindres carrées et une phase de projection sur le polyèdre des contraintes. Dans le cadre de notre problème cet algorithme permet un bon compromis entre la vitesse de convergence de l'algorithme et son temps d'exécution. En pratique, tess3r permet d'estimer la structure de population en un temps de calcul de l'ordre de l'heure sur des matrices allant jusqu'à mille individus et un million de gènes. De plus, le package propose des fonctions pour projeter sur une carte les coefficients de métissage ainsi calculés. Enfin, à partir de l'estimation des fréquences alléliques de chaque sous-population et des coefficients de métissage individuels, le package tess3r calcule une statistique de test de type Fst pour chaque gène. Cette statistique peut être utilisée pour effectuer un balayage pangénomique dans le but de détecter des gènes potentiellement responsables d'une adaptation à l'environnement.

En résumé, l'intérêt du package tess3r est de permettre une estimation efficace de la structure des populations spatialisées et d'une statistique en s'intégrant dans l'environnement de travail

*. Intervenant
†. Corresponding author : olivier.francois@imag.fr

R. L'utilisateur peut alors profiter des autres packages de R pour la visualisation des résultats ou encore la gestion des formats de données.

Mots clefs : Génétique des populations, Analyse de structure, Balayage pangénomique

Prediction and characterization of ciliary proteins by comparative genomics

Yannis Nevers [*] [†1], Megana Prasad[2], Laetitia Poidevin[1], Kirsley Chennen[1], Alexis Allot[1], Arnaud Kress[1], Raymond Ripp[1], Hélène Dollfus[2], Olivier Poch[1], Odile Lecompte[1]

[1] Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube) – ENGEES, Institut National des Sciences Appliquées [INSA] - Strasbourg, université de Strasbourg, CNRS : UMR7357 – 300 bd Sébastien Brant - BP 10413, F-67 412 ILLKIRCH Cedex, France
[2] Laboratoire de Génétique Médicale, Faculté de Médecine/Université de Strasbourg – Inserm U3949 – STRASBOURG, France

Session génomique des populations
mercredi 29 10h50
Amphi Mérieux

Context

Cilium is an organelle that protrudes from eukaryotic cells and is both a common cellular movement effector and a center of reception and integration of extracellular signals. Across the eukaryotic domain, cilia share the same structural basis: a membrane-covered extension of the microtubule cytoskeleton, the axoneme, that elongates from a microtubule organizing center, the basal body (Gerdes et al., 2009). However, while those general concepts hold in most ciliated organisms, cilia are subject to an important variability in terms of number, length or molecular composition between eukaryotic species, and even between different developmental stages or tissues within the same organism. In Vertebrates, cilia are historically divided into motile and primary cilia on the basis of structural and functional features. Motile cilia act as effectors of movement and are responsible for sperm motility and fluid flow along multiciliated epithelium surfaces while primary cilia, present in most cells, have essential roles in developmental pathways, cell cycle regulation… With such diversified range of functions, defects in cilia are linked to a wide panel of human genetic disorders, the ciliopathies (Badano et al., 2006). These diseases show an important phenotypic diversity: from organ-specific pathologies (such as kidney in polycystic kidney diseases) to pleiotropic diseases with complex phenotypes (as Bardet Biedl Syndrome). Despite significant progresses in the last fifteen years, there is still much to do to uncover the totality of mechanisms and genes behind the complexity of ciliopathies and ciliary processes.

In that context, high-throughput studies have been undertaken to identify new ciliary genes, in particular comparative genomic approaches exploiting the singular evolutionary history of cilia. The organelle was likely present in the last eukaryotic common ancestor and can be found in all major eukaryotic lineages but has also been subject to multiple independent losses during evolution. Thus, identification of genes present in ciliated species and absent in non-ciliated species by phylogenetic profiling can constitute an efficient way to predict genes functionally linked to cilia (Li et al., 2004). Moreover, cilia subfunctions have been lost in some lineages (e.g. nematodes are only able to construct immotile cilia), which could open the way to the characterization of cilia proteins at the functional level but also demands to distinguish biological heterogeneity from technical noise.

To tackle that challenge, it is necessary to optimize both the construction and the analysis of phylogenetic profiles used in comparative genomic approaches. The construction step consists in identifying orthology relationships between genes of eukaryotic species to generate

*. Intervenant
†. Corresponding author : yannis.nevers@etu.unistra.fr

presence/absence profiles (Pellegrini et al., 1999). Thus, the profile accuracy relies on the selection of well-annotated genomes as well as in choosing a robust orthology prediction method adapted to distant eukaryotic genomes. The analysis step necessitates to identify genes with a "ciliary" phylogenetic distribution. Studies dedicated to cilia usually rely on knowledge-guided methods (Avidor-Reiss et al., 2004; Li et al., 2004, Hodges et al., 2011) but more general predicting approaches have recently been applied to ciliary proteins: an agglomerative grouping of proteins based on profile similarity (Dey et al., 2015), and CLIME a machine learning method based on evolutionary models (Li et al., 2014). However, all these approaches have been applied on different data sets and their relative strength and weaknesses are yet to be determined.

In that context, we capitalized on previous works and on our orthology prediction program OrthoInspector (Linard et al., 2015) to develop an integrative approach of prediction and exploration of ciliary genes. First, we constructed phylogenetic profiles of human genes in 100 carefully selected eukaryotic genomes using OrthoInspector and analyzed them using three independents methods. Definition of positive and negative reference sets allowed the objective assessment of each method. On this basis, we combined the different strategies to define a set of 276 ciliary proteins. Then, in-depth analysis of ciliary protein profiles allowed us to classify them into functional modules based on their evolutionary histories.

Results and discussion

Definition of reference sets

To allow an objective comparison of different methods of ciliary proteins prediction, we generated a set of known ciliary genes (positive reference set) and set of genes with no link to cilia (negative reference set). The positive set included the Ciliary Gold Standard (van Dam et al., 2013), an expert curated set of ciliary genes and 75 additional genes annotated with Gene Ontology terms related to cilium, for a total of 377 cilia related genes.
The negative set was more challenging to obtain considering the large panel of processes related to cilia. To avoid inclusion of genes with a still undeciphered link to cilium, we selected functional pathways from Reactome (Croft et al., 2014) that includes neither genes of our positive set nor genes interacting with two or more genes of this set according to the STRING database (Franceschini et al., 2013). Genes from these pathways constituted our negative set of 1,754 genes.

Comparison of prediction methods

Phylogenetic profiles of the 20,193 human proteins in 64 ciliated and 36 non-ciliated species representative of all major eukaryotic lineages were generated on the basis of orthology relationship predicted by OrthoInspector (Linard et al., 2015).The presence/absence profiles were analyzed independently using three methods.

We first developed an empirical score: briefly, positives values are assigned to proteins present in ciliated species of each of our six major eukaryotic lineages and negatives values assigned to proteins present in non-ciliated species. Using that scoring method, we retrieved 370 proteins including 122 proteins of our positive set (true positives) but also 3 proteins of our negative reference set (false positives).
Secondly, we performed a hierarchical clustering of phylogenetic profiles resulting in 14 clusters that contain from 327 to 2,766 proteins. Gene ontology enrichment analysis revealed that one of the 14 clusters was greatly enriched in terms linked to cilia (for example "ciliary part" with an enrichment of $2.07 \cdot ^{-110}$). Those 327 proteins contained 120 true positives and two false positives.
Finally, we used the CLIME algorithm (Li et al., 2014) to predict ciliary proteins based on an evolutionary model derived from a training set, here the Ciliary Gold Standard, and an user provided phylogenetic tree. 178 proteins were predicted as part of informative extended evolutionary modules, including 90 true positives and no false positive.

It is worth noting that only a portion of the positive set is recovered, whatever the considered method. This is expected since ciliary genes involved in multiple processes as well as recent ciliary genes do not exhibit a typical evolutionary ciliary signature and cannot be detected by comparative genomics on a presence/absence basis. Overall, 442 genes were predicted by one or more approaches, including 131 true positives, and 5 false positives. The overlap was considerable, with 276 genes predicted by at least two methods. Interestingly, 116 of these 276 genes are genes of the positive reference set and none of them belong to the negative reference set. As such, these 276 genes constitute a robust prediction of ciliary genes, with higher sensitivity and specificity than previous comparative genomic studies.

Functional characterization

We analyzed the phylogenetic profiles and functional annotations of the 276 predicted ciliary genes to delineate evolutionary modules linked to functional processes. Considering the presence/absence patterns among Metazoa, three classes of evolutionary histories emerged within selected genes.

The first evolutionary module is composed of 91 genes conserved in most ciliated species, including all metazoan lineages. 70 of those genes have a documented ciliary role and most of them are essential to cilia assembly and compartmentalization. Notably, this module regroups most components of important ciliary complexes: IFT-B (15 genes out of 17), IFT-A (6 out of 6) and BBSome (8 out of 10), whose disruption causes pleiotropic ciliopathies. The 21 remaining genes have no documented ciliary function but represent, by association, outstanding candidates for novel causative genes in ciliopathies and essential ciliary genes.

The second evolutionary module is composed of 87 genes lost in nematodes, but conserved in others metazoans. Nematodes develop only immotile cilia, and accordingly, known genes of this set are involved in movement of motile cilia and include notably 23 of the 32 known responsible genes of ciliopathies linked to motility defect, i.e. primary ciliary dyskinesia. Consequently, the 20 genes of that set without known ciliary function are potentially linked to cilium motility and may be involved in ciliary dyskinesia.

The last evolutionary module gathers the 98 remaining genes, including 51 genes with a documented ciliary role. Overall, their profiles show a loss in most Ecdysozoa (nematodes and arthropods), with some exceptions among insects. Functional characterization of that set is not trivial but analysis of their STRING interaction networks link them to centrosome, in agreement with prior observations suggesting a simplification of centriole in Ecdysozoa (Woodland and Fry, 2008).

Conclusion

We established the phylogenetic distribution of human genes in 100 eukaryotic species and analyzed them using three independent methods to predict ciliary genes. Accuracy of each method was objectively assessed using a positive and negative reference gene sets, available for subsequent omics studies of cilia. Using a combined methodology of phylogenetic profiling, we predicted a set of 276 ciliary genes exhibiting both a highly significant enrichment in cilia related genes and no false positives. 87 predicted genes have currently no documented ciliary function and are thus promising ciliary and ciliopathies genes candidates.

The 276 predicted genes were further studied and categorized into three modules according to particularities of their evolutionary histories among ciliated species. These evolutionary differences correlate with gene implication in different cilium processes, allowing us to propose a functional role for new predicted ciliary genes. To validate our predictions and investigate the function of ciliary candidates, 41 uncharacterized proteins from the three evolutionary modules were selected for subcellular localization assays in human cells. Experiments of immunocytochemistry in cultured RPE cells are currently underway.

References

Avidor-Reiss, T., Maer, A.M., Koundakjian, E., Polyanovsky, A., Keil, T., Subramaniam, S., and Zuker, C.S. (2004). Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. *Cell* 117:527–539.

Badano, J.L., Mitsuma, N., Beales, P.L., and Katsanis, N. (2006). The ciliopathies: an emerging class of human genetic disorders. *Annu. Rev. Genomics Hum. Genet.* 7:125–148.

Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* :42:D472–D477.

Van Dam, T.J., Wheway, G., Slaats, G.G., SYSCILIA Study Group, Huynen, M.A., and Giles, R.H. (2013). The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. *Cilia* 2:7.

Dey, G., Jaimovich, A., Collins, S.R., Seki, A., and Meyer, T. (2015). Systematic Discovery of Human Gene Function and Principles of Modular Organization through Phylogenetic Profiling. *Cell Rep.*

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41:D808–D815.

Gerdes, J.M., Davis, E.E., and Katsanis, N. (2009). The Vertebrate Primary Cilium in Development, Homeostasis, and Disease. *Cell* 137:32–45.

Hodges, M.E., Wickstead, B., Gull, K., and Langdale, J.A. (2011). Conservation of ciliary proteins in plants with no cilia. *BMC Plant Biol.* 11:185.

Li, J.B., Gerdes, J.M., Haycraft, C.J., Fan, Y., Teslovich, T.M., May-Simera, H., Li, H., Blacque, O.E., Li, L., Leitch, C.C., et al. (2004). Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. *Cell* 117:541–552.

Li, Y., Calvo, S.E., Gutman, R., Liu, J.S., and Mootha, V.K. (2014). Expansion of biological pathways based on evolutionary inference. *Cell* 158:213–225.

Linard, B., Allot, A., Schneider, R., Morel, C., Ripp, R., Bigler, M., Thompson, J.D., Poch, O., and Lecompte, O. (2015). OrthoInspector 2.0: Software and database updates. *Bioinforma. Oxf. Engl.* 31:447–448.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96:4285–4288.

Woodland, H.R., and Fry, A.M. (2008). Pix proteins and the evolution of centrioles. *PloS One* 3:e3778.

Mots clefs : comparative genomics, ciliopathy, cilium, evolution

In silico experimental evolution provides independent and challenging benchmarks for comparative genomics

Priscila Biller[1], Éric Tannier[2,3], Guillaume Beslon[3,4], Carole Knibbe *[3,5]

[1] University of Campinas [Campinas] (UNICAMP) – São Paulo, Brésil
[2] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[3] BEAGLE (Insa Lyon / INRIA Grenoble Rhône-Alpes / UCBL) – INRIA, Institut National des Sciences Appliquées [INSA] - Lyon, Université Claude Bernard - Lyon I (UCBL) – Antenne INRIA Lyon la Doua Bâtiment CEI-1, 66 boulevard Niels Bohr, F-69 603 Villeurbanne, France
[4] Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) – Institut National des Sciences Appliquées [INSA], CNRS : UMR5205 – France
[5] Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) – Université Claude Bernard - Lyon I (UCBL), CNRS : UMR5205 – France

Session génomique des populations
mercredi 29 11h10
Amphi Mérieux

The following extended abstract is a highlight of [7].

A common concern in all evolutionary studies is the validity of the methods and results. Results relate to events that were supposed to occur in a deep past (up to 4 billion years) and they have no other trace today than the present molecules used by comparative methods.

As we cannot travel back in time to verify the results, there are several ways to assess the validity of molecular evolution studies: theoretical considerations about the models and methods (realism, consistency, computational complexity, robustness, model testing, ability to generate a statistical support or a variety of the solutions) [23], coherence with fossil records [25], or ancient DNA [11], or empirical tests when the solution is known, on experimental evolution [17] or simulations. Each method has its caveats. Models for inference have to adopt a compromise between realism, consistency and complexity. Ancient DNA is rarely available, usually not in an assembled shape. Fossils are also rare and provide a biased sampling of ancient diversity. Experimental evolution is expensive, time- consuming and limited in the number of generations it can provide.

Simulation is the most popular validation tool. Genome evolution can be simulated in silico for a much higher number of generations than in experimental evolution, much faster and at a lower cost. All the history can be recorded in details, and compared with the inference results. A problem with simulations, however, is that they necessarily oversimplify genome evolution processes. Moreover, very often, even if they are designed to be used by another team for inference [4, 15, 14, 10, 22], they encode the same simplifications as the inference methods. For example, only fixed mutations are generated because only these are visible by inference methods; selection is tuned to fit what is visible by the inference methods; genes are often evolutionary units in simulations because they are the units taken for inference. Everything is designed thinking of the possibilities of the inference methods.

This mode of ad-hoc simulation has been widely applied to test estimators of rearrangement distances, and in particular inversion distances [9, 12, 5, 21, 6]. The problem consists in comparing two genomes and estimating the number of inversions (a rearrangement that reverses the reading direction of a genomic segment) that have occurred in the evolutionary lineages separating them. To construct a solution, conserved genes or synteny blocks are detected in the two genomes, and a number of inversions explaining the differences in gene orders is estimated. A lot of work has

*. Intervenant

consisted in finding shortest scenarios [13]. Statistical estimations need a model. The standard and most used model depicts genomes as permutations of genes and assumes that an inversion reverses a segment of the permutation, taken uniformly at random over all segments. When simulators are designed to validate the estimators, they also use permutations as models of gene orders, and inversions on segments of this permutations, chosen uniformly at random. Estimators show good performances on such simulations, but transforming a genome into a permutation of genes is such a simplification from both parts that it means nothing about any ability to estimate a rearrangement distance in biological data [8].

We propose to use simulations that were not designed for validation purposes. It is the case, in artificial life, of in silico experimental evolution [18], and in particular of the Aevol platform [19, 3]. Aevol contains, among many other features, all what is needed to test rearrangement inference methods. The genomes have gene sequences and non coding sequences organized in a chromosome, and evolve with inversions, in addition to substitutions, indels, duplications, losses, translocations. Rearrangements are chosen with a uniform random model on the genome, which should fit the goals of the statistical estimators, but is different from a uniform random model on permutations [8].

We tested 10 different estimators of inversion distance found in the literature, one shortest path estimator and 9 statistical estimators on 18 different datasets generated by Aevol. The difference with ad-hoc simulations is striking. Whereas good results were largely reported for ad-hoc simulations, most estimators completely fail to give a close estimate in a vast majority of conditions. As soon as the true number of events exceeds about n/3 (where n is the number of genes), most estimators significantly underestimate the true value. This highly contrasts with the claimed performances of these estimators. For example the shortest path estimator is supposed to have great chance of giving the right value up to n/2 [16], while all statistical estimators have been tested on simulations and reported to give the right value far above n [9, 20, 12, 5, 21, 2, 6, 8].

We argue, based on the differences in performances of some estimators, that our datasets are not artefactually difficult (nor purposely made difficult), and that the poor results encountered here are susceptible to reflect real results on biological data. Indeed part of the failure of the estimators can be explained by this ignorance of intergene sizes, because the only one handling intergene sizes performs significantly better. We investigated this further in [8].

Part of the discrepancy between the true value and the estimated value still remains unexplained. The complexity of the real scenarios probably blurs the signal that estimators are able to capture. But again, this complexity is not a specificity of Aevol, and is probably encountered in biological data. So by this simple experiment we can worry that none of the existing estimators of rearrangement distance would be able to produce a plausible value on real genomes.

We tested only the estimation of the number of inversions. But only with the runs we have already computed, a lot more can be done: estimation of the proportion of translocations (transposition of a block of DNA at an other locus) as in [1], or estimating both inversions and duplications as in [24]. For the moment the sequences are made of 0s and 1s, which is not a problem to study gene order, but can be disturbing for sequence analyses. This way of coding sequences is on another hand a good sign that Aevol was not developed for benchmarking purposes. In a close future, nucleotidic and proteic sequences with the biological alphabet will be added to extend the benchmarking possibilities of the model.

Also we worked with only one lineage, and compare only two genomes here (final versus ancestral), because Aevol currently evolves only one population at a time. A useful addition will be speciation processes, in order to be able to compare several genomes.

As a final note, we would like to point out the singular kind of interdisciplinarity experimented in this study. Obviously communities from comparative genomics and artificial life have to work together in order to make such results possible. But, on the opposite, these results are only possible because both communities first worked in relative isolation. If they had defined their

working plans together, spoke to each other too often or influenced each other's way of thinking evolutionary biology, the work would have lost some value. Indeed, what makes the difficulty here for comparative genomicists is that they have to infer histories on data for which they have no stranglehold on the processes, just as for biological data, but on which they also have the correct answer, just not as for biological data.

References

[1] N Alexeev, R Aidagulov, and MA Alekseyev. A computational method for the rate estimation of evolutionary transpositions. *Bioinformatics and Biomedical Engineering*, pages 471–480, 2015.

[2] N Alexeev and Max A. Alekseyev. Estimation of the true evolutionary distance under the fragile breakage model. *Arxiv*, 2015.

[3] Berenice Batut, David P. Parsons, Stephan Fischer, Guillaume Beslon, and Carole Knibbe. In silico experimental evolution: a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14 (S15):S11, 2013.

[4] R G Beiko and R L Charlebois. A simulation test bed for hypotheses of genome evolution. *Bioinformatics*, 23(7):825–831, April 2007.

[5] Nathanal Berestycki and Rick Durrett. A phase transition in the random transpo- sition random walk. *Probability Theory and Related Fields*, 136:203–233, 2006.

[6] Priscila Biller, Laurent Guéguen, and Éric Tannier. Moments of genome evolution by double cut-and-join. *BMC Bioinformatics*, 16, 2015.

[7] Priscila Biller, Carole Knibbe, Guillaume Beslon, and Éric Tannier. Comparative genomics on artificial life. In *Proceedings of Computability in Europe (CiE) 2016, LNCS*. Springer, 2016.

[8] Priscila Biller, Carole Knibbe, Laurent Guéguen, and Éric Tannier. Breaking good: accounting for the diversity of fragile regions for estimating rearrangement distances. *Genome Biology and Evolution*, 2016, in press.

[9] Alberto Caprara and Giuseppe Lancia. Experimental and statistical analysis of sorting by reversals. In David Sankoff and Joseph H. Nadeau, editors, *Comparative Genomics*, pages 171–183. Springer, 2000.

[10] Daniel A. Dalquen, Maria Anisimova, Gaston H. Gonnet, and Christophe Dessimoz. ALF – a simulation framework for genome evolution. *Mol Biol Evol*, 29(4):1115–1123, Apr 2012.

[11] Wandrille Duchemin, Vincent Daubin, and Éric Tannier. Reconstruction of an ancestral *Yersinia pestis* genome and comparison with an ancient sequence. *BMC Genomics*, 16 Suppl 10:S9, 2015.

[12] Niklas Eriksen and Axel Hultman. Estimating the expected reversal distance after a fixed number of reversals. *Advances in Applied Mathematics*, 2004.

[13] G. Fertin, A. Labarre, I. Rusu, É. Tannier, and S. Vialette. *Combinatorics of Genome Rearrangements*. MIT press, London, 2009.

[14] William Fletcher and Ziheng Yang. Indelible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*, 26(8):1879–1888, Aug 2009.

[15] B G Hall. Simulating DNA Coding Sequence Evolution with EvolveAGene 3. *Molecular Biology and Evolution*, 25(4):688–695, February 2008.

[16] Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Foundations of Computer Science, 1995*. Proceedings., 36th Annual Symposium, 1995.

[17] D. M. Hillis, J. J. Bull, M. E. White, M. R. Badgett and I. J. Molineux. Experimental phylogenetics: generation of a known phylogeny. *Science*, 255(5044):589–592, Jan 1992.

[18] Thomas Hindré, Carole Knibbe, Guillaume Beslon, and Dominique Schneider. New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nature Reviews Microbiology*, 10:352–365, May 2012.

[19] Carole Knibbe, Antoine Coulon, Olivier Mazet, Jean-Michel Fayard, and Guillaume Beslon. A long-term evolutionary pressure on the amount of noncoding DNA. *Molecular Biology and Evolution*, 24(10):2344–2353, Oct 2007.

[20] B. Larget, D. L. Simon, and J.B. Kadane. On a bayesian approach to phyloge- netic inference from animal mitochondrial genome arrangements (with discussion). *Journal of the Royal Statistical Society, B*, 64:681–693, 2002.

[21] Yu Lin and Bernard M E. Moret. Estimating true evolutionary distances under the DCJ model. *Bioinformatics*, 24(13):i114–i122, Jul 2008.

[22] Diego Mallo, Leonardo De Oliveira Martins, and David Posada. Simphy: Phyloge- nomic simulation of gene, locus, and species trees. *Syst Biol*, Nov 2015.

[23] M. Steel and D. Penny. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol Biol Evol*, 17(6):839–850, Jun 2000.

[24] KM Swenson, M Marron, JV Earnest-DeYoung, and BME Moret. Approximating the true evolutionary distance between two genomes. *Journal of Experimental Algorithmics*, 12, 2008.

[25] Gergely J. Szollosi, Bastien Boussau, Sophie S. Abby, Éric Tannier, and Vincent Daubin. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci U S A*, 109(43):17513–17518, Oct 2012.

Mots clefs : simulation, genome evolution, inversion distance, intrachromosomal rearrangements, benchmark, individual based modeling, comparative genomics

Comment la reconstruction de génomes ancestraux peut aider à l'assemblage de génomes actuels

Yoann Anselmetti [*1,2], Vincent Berry[3,4], Cedric Chauve[5], Annie Chateau[3,4], Éric Tannier[2,6], Sèverine Bérard[1,3,4]

Session génomique des populations mercredi 29 11h30 Amphi Mérieux

[1] Institut des Sciences de l'Évolution - Montpellier (ISEM) – CNRS : UMR5554, Institut de recherche pour le développement [IRD] : UMR226, Université Montpellier II - Sciences et techniques – Place E. Bataillon CC 064, F-34 095 Montpellier Cedex 05, France
[2] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[3] Institut de Biologie Computationnelle (IBC) – CNRS : UMR5506, Université Montpellier II - Sciences et techniques – 95 rue de la Galéra, F-34 095 Montpellier, France
[4] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université Montpellier II - Sciences et techniques, CNRS : UMR5506 – CC 477, 161 rue Ada, F-34 095 Montpellier Cedex 5, France
[5] SFU Discrete Mathematics Group (SFU-DMG) – Dept. Mathematics, SFU 8888 University Drive Burnaby, BC, V5A 1S6, Canada
[6] INRIA Rhône-Alpes (INRIA Grenoble Rhône-Alpes) – INRIA – ZIRST 655 avenue de l'Europe, Montbonnot, F-38 334 Saint Ismier Cedex, France

Introduction

L'avènement des *NGS* et l'accès à un nombre croissant de génomes ces dernières années dans les bases de données permettent d'entreprendre des études sur l'histoire évolutive complexe de la structure des génomes. Pour cela, la plupart des méthodes d'analyse nécessitent des génomes complètement assemblés. Cependant, les génomes disponibles dans les bases de données sont souvent incomplètement assemblés et la plupart d'entre eux restent à l'état de *permanent draft genomes* [1]. Pour illustration, dans Ensembl (version 84) les génomes sont en moyenne constitués de 2 882 fragments, appelés contigs (écart-type=2977) et 58 des 69 espèces présentes dans la base de données ont leur génome composé de plus de 100 *contigs*. L'assemblage de génomes est un problème difficile qui consiste à ordonner et orienter des fragments d'ADN issus du séquençage afin de reconstruire les chromosomes composant le génome.

Il existe des méthodes de génomique comparative adaptées à la problématique d'assemblage afin d'améliorer l'ordre et l'orientation de marqueurs génomiques dans les génomes actuels à l'aide d'un ou plusieurs génomes de référence [2-6]. Parmi les méthodes dites multireference-guided assembly, certaines comme treecat [4], RACA [5] et celle d'Aganezov et al. [6] utilisent également les liens de parentés avec la structure des génomes de référence pour pondérer les apports de ces derniers dans l'assemblage du génome fragmenté (phylogeny-guided assembly). La méthode présentée dans cet article, ARt-DeCo (*Assembly Recovery through DeCo*) [7], s'inscrit dans cette catégorie de méthodes. Elle permet de réduire la fragmentation des génomes actuels composant un jeu de données, en prédisant des adjacences entre gènes situés aux extrémités de leurs contigs à l'aide de la composition des autres génomes du jeu de données. ARt-DeCo s'appuie sur la phylogénie des marqueurs génomiques utilisés mais n'a aucune contrainte quant à l'unicité et l'universalité de ceux-ci.

*. Intervenant

Dans un premier temps, nous présenterons succinctement le principe d'ARt-DeCo. Ensuite nous exposerons des simulations permettant d'évaluer la capacité d'ARt-DeCo à retrouver les adjacences de gènes. Enfin, nous détaillerons des applications d'ARt-DeCo sur des données réelles.

Principes d'ARt-DeCo

Pour simplifier cet exposé, considérons que les marqueurs génomiques utilisés sont des gènes. On définit deux gènes comme adjacents s'ils sont situés sur un même fragment génomique sans aucun autre gène entre eux. ARt-DeCo est basé sur l'algorithme DeCo (Detection of Coevolution) [8] qui permet la reconstruction de l'histoire évolutive des adjacences de gènes et donne ainsi accès à une estimation de la structure des génomes ancestraux.

La méthode prend en entrée un ensemble d'arbres phylogénétiques de gènes, les adjacences de gènes observées dans les génomes actuels ainsi que l'arbre phylogénétique des espèces. Elle applique un principe de parcimonie basé sur les coûts de cassure et de création d'adjacences pour calculer une histoire évolutive des adjacences de moindre coût. Une telle histoire de coût minimum est calculée par une méthode de programmation dynamique basée sur l'exploration des arbres phylogénétiques de gènes.

La méthode DeCo ne tient compte que des adjacences présentes dans les assemblages des bases de données. Or, on sait qu'une adjacence peut être réelle mais non observée situé à la jonction de deux contigs non ordonnés et orientés. ARt-DeCo a été conçu pour permettre d'inférer, les adjacences de génomes actuels non présentes dans les bases de données, en plus d'inférer les adjacences ancestrales (comme DeCo). Cette inférence prend en compte une probabilité pour deux gènes d'être adjacents en fonction du degré de fragmentation du génome auquel ils appartiennent. ARt-DeCo s'autorise à inférer l'existence de ces adjacences si elles contribuent à une histoire de coût minimum.

Expérience et résultats biologiques

Nous présentons ici les résultats obtenus avec ARt-DeCo. Les deux premières expériences ont permis de valider l'approche générale. Ensuite, nous analysons de façon qualitative une adjacence prédite par ARt-DeCo, puis nous analysons les résultats préliminaires d'une variante de la méthode sur le jeu d'anophèles permettant l'apport de données de séquençage dans la reconstruction d'histoires évolutives.

Simulations de fragmentation

Pour évaluer la capacité de la méthode à prédire des adjacences de gènes présentes dans les génomes mais non répertoriées dans les bases de données (adjacences qualifiées ici de « réelles »), nous avons aléatoirement occulté certaines adjacences et testé la capacité d'ARt-DeCo à les retrouver. Ce premier jeu de données est composé de 18 génomes d'anophèles récemment séquencés et assemblés [9], et de 11 534 arbres de gènes incluant 172 585 gènes disponibles sur la base de données VectorBase.

Les 18 espèces ont été fragmentées aléatoirement avec divers pourcentages d'adjacences occultées (0,1 %, 0,5 %, 1 %, 5 %, 10 %, 25 %, 50 % et 75 %) et chaque expérience répliquée 30 fois.

À partir des nouvelles adjacences proposées par ARt-DeCo, nous avons calculé le rappel et la précision de la méthode. Le rappel indique la proportion d'adjacences occultées qui ont été retrouvées. La précision correspond à la proportion d'adjacences correctes (i.e., occultées) parmi les adjacences prédites.

Les résultats (cf. Figure 1) montrent que la précision la plus faible constatée est de l'ordre de 80 %, obtenue pour le plus faible pourcentage d'adjacences occultées (0,1 %) et augmente graduellement jusqu'à atteindre un plateau avec une valeur fluctuant autour de 92 % à partir de 1

% d'adjacences occultées. Le rappel est de 69,75 % pour le plus faible pourcentage d'adjacences occultées (0,1%), et il décroît au fur et à mesure que plus d'adjacences sont occultées, jusqu'à atteindre 12,18 % pour le cas extrême où 75 % des adjacences sont occultées.

On observe donc que pour des génomes faiblement fragmentés (de 0,1 à 0,5 %), ARt-DeCo retrouve de l'ordre de 69 % des adjacences occultées mais infère également une proportion non négligeable d'adjacences non présentes dans les assemblages initiaux (jusqu'à _~20%). Pour des génomes plus fragmentés, ARt-DeCo retrouve une plus faible proportion d'adjacences mais avec une précision supérieure à 90 %.

Pour analyser plus finement l'effet d'ARt-DeCo, nous avons effectué la même expérience mais en simulant des cassures chez une seule espèce à la fois. Pour cela, trois espèces placées à des positions différentes dans la phylogénie des anophèles ont été choisies :

— *Anopheles gambiae* localisée en profondeur dans l'arbre,
— *Anopheles minimus* située en profondeur dans l'arbre mais avec peu d'espèces proches,
— *Anopheles albimanus* en position d'outgroup par rapport aux autres espèces de l'arbre.

Pour les trois espèces (cf. Figure 2), la précision et le rappel décroissent légèrement lorsque la fragmentation artificielle augmente :

— Pour *A. gambiae* :
  — Précision : Max : 100 % | Min : 87,37 %
  — Rappel : Max : 70 % | Min : 68 %
— Pour *A. minimus* :
  — Précision : Max : 99,43 % | Min : 95,54 %
  — Rappel : Max : 54 % | Min : 53,06 %
— Pour *A. albimanus* :
  — Précision : Max : 99,49 % | Min : 93,68 %
  — Rappel : Max : 38,89 % | Min : 37,37 %

On observe que pour les trois espèces, on obtient une bonne précision fluctuant entre 100 et 87,37 %. Le rappel plus élevé chez *A. gambiae* que chez les autres espèces peut s'expliquer par le voisinage d'espèces proches dans la phylogénie. Comme le faible rappel de *A. albimanus* peut être expliqué par sa position d'*outgroup* dans l'arbre des espèces.

En conclusion, on observe que la performance d'ARt-DeCo à lier des contigs n'est pas la même suivant la proportion de génomes fragmentés dans le jeu de données et le degré de fragmentation de ces génomes. Pour le cas, où l'ensemble des génomes est fragmenté (cf. Figure 1), les résultats montrent que ARt-DeCo obtient un meilleur compromis rappel/précision pour des génomes moyennement fragmentés et retrouve 333 vraies adjacences sur 340 adjacences prédites lorsque 493 sont occultées. Tandis que dans le cas où une seule espèce est fragmentée (cf. Figure 2), ARt-DeCo est plus performant pour de faibles fragmentations, chez *A. gambiae* pour 9 adjacences occultées, ARt-DeCo en prédit 7 toutes valides.

Passage à l'échelle

Pour déterminer la capacité de l'algorithme ARt-DeCo à travailler sur de grands jeux de données, nous l'avons appliqué aux 69 espèces eucaryotes de la base de données Ensembl (version 79). Ce jeu est composé de 20 279 arbres de gènes contenant 1 222 543 gènes codant pour des protéines et 1 023 492 adjacences chez les génomes actuels. Une grande proportion des génomes actuels sont fortement fragmentés, dont le génome du wallaby (*Macropus eugenii*) composé de 12 704 *contigs*. L'algorithme prédit 36 445 nouvelles adjacences sur l'ensemble des espèces du jeu de données en $\approx$ 18h sur un ordinateur de bureau.

L'analyse des résultats montre que plus les génomes sont fragmentés plus l'assemblage est amélioré, c'est-à-dire que le ratio du nombre d'adjacences prédites sur nombre d'adjacences manquantes est plus élevé dans ces génomes là (cf. [7]).

Analyse détaillée d'une adjacence prédite

L'algorithme ARt-DeCo peut également être combiné avec l'algorithme DeClone [10] qui permet d'explorer l'ensemble des solutions parcimonieuses co-optimales d'ARt-DeCo et ainsi fournir un score à une adjacence. Ce score correspond à la proportion de scénarios dans lesquels l'adjacence a été prédite par ARt-DeCo et mesure ainsi un support de confiance compris entre 0 et 1.

Sur le jeu de données que nous considérons maintenant, qui contient 39 placentaires, ARt-DeCo prédit 22 675 nouvelles adjacences dont 95 % ont un support > 0,9.

Parmi elles, une nouvelle adjacence proposée chez le panda (*Ailuropoda melanoleuca*) a été analysée en détail. L'analyse chez les espèces proches du voisinage plus large des gènes impliqués dans cette adjacence a montré de fortes similarités avec la situation dans le génome du panda (cf. Figure 3). L'analyse du voisinage des gènes concernés conforte donc la forte confiance dans l'existence de l'adjacence prédite. Cette confiance est renforcée par l'inférence par ARt-DeCo d'autres adjacences homologues à l'adjacence du panda chez trois autres espèces (*Ochotona princeps*, *Tupaia belangeri* & *Dipodomys ordii*) toutes avec des supports > 0,99.

Une analyse systématique reste à mener pour évaluer l'ensemble des prédictions.

Intégration des données de séquençage

Récemment, l'algorithme ARt-DeCo a été amélioré pour permettre l'intégration des liens de *scaffolding* dans son calcul. Ces liens de scaffolding prédits par des logiciels comme BESST [11], apportent des informations de structure non présentes dans les bases de données. Combinés aux adjacences connues, ils permettent une reconstruction d'histoires évolutives d'adjacences plus complètes augmentant ainsi le pouvoir prédictif d'ARt-DeCo.

Des données de séquençage *paired-end* et *mate-pair* sont disponibles pour le jeu de données des 18 anophèles. Les 34 542 liens de *scaffolding* calculés par BESST sur ce jeu de données ont permis à ARt-DeCo d'inférer 5 894 nouvelles adjacences représentant ≈25 % des prédictions.

Conclusion

ARt-DeCo est une méthode de prédiction d'adjacences de gènes basée sur la phylogénie qui s'attache à réduire la fragmentation des assemblages de génomes actuels. Intuitivement, ces génomes peuvent se corriger mutuellement si les zones génomiques fragmentées sont différentes d'un génome à l'autre. ARt-DeCo s'appuie sur un ensemble de marqueurs pour lequel les histoires évolutives sont connues. Il est capable de gérer de larges jeux de données, même si ceux-ci contiennent des gènes dupliqués. ARt-DeCo est un pas en avant dans la réduction de la fragmentation des génomes séquencés. Les simulations et l'analyse de cas réels montrent que les adjacences proposées sont fiables.

Références

[1] GOLD database : `https://gold.jgi.doe.gov/statistics/`

[2] Lu et al., (2014). CAR : contig assembly of prokaryotic draft genomes using rearrangements. *BMC Bioinformatics*, 15:381–390.

[3] Bosi et al., (2015). MEDUSA : a multi-draft based scaffolder. *Bioinformatics*, 31(15):2443–51.

[4] Husemann & Stoye (2010). Phylogenetic comparative assembly. *Algorithms for Molecular Biology*, 5(1):3–14.

[5] Kim et al., (2013). Reference-assisted chromosome assembly. *Proceedings of the National Academy of Sciences (PNAS)*, 110(5):1785–90.

[6] Aganezov et al., (2015). Scaffold assembly based on genome rearrangement analysis. *Computational Biology and Chemistry*, 57(August):46–53.

[7] Anselmetti et al. (2015) Ancestral gene synteny reconstruction improves extant species scaffolding, *BMC genomics*. 16(Suppl 10):S11.

[8] Bérard et al., (2012) Evolution of gene neighborhoods within reconciled phylogenies, *Bioinformatics*. 28:i382–i388.

[9] Neafsey et al., (2015). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science*, 347(6217):1258522–1 – 1258522–8.

[10] Chauve et al.(2014). Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach. In *Lecture Notes in Computer Science (Advances in Bioinformatics and Computational Biology)* (Vol. 8826 LNBI, pp. 49–56).

[11] Sahlin et al., (2014). BESST - Efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics*, 15(1):281

Mots clefs : reconstruction de génomes ancestraux, assemblage de génomes, adjacences de gènes, évolution, parcimonie, programmation dynamique

Keynote

Session évolution moléculaire
mercredi 29 13h30
Amphi Mérieux

Towards population-level microbiome monitoring : the Flemish Gut Flora Project

Jeroen Raes [*1],

[1] Bioinformatics and (eco-)systems biology lab, Department of Microbioloby and Immunology, Rega Institute Leuven, Belgium

Alterations in the gut microbiota have been linked to various pathologies, ranging from inflammatory bowel disease and diabetes to cancer. Although large numbers of clinical studies aiming at microbiome-based disease markers are currently being performed, our basic knowledge about the normal variability of the human intestinal microbiota and the factors that determine this still remain limited. Here, I will present a large-scale study of the gut microbiome variation in a geographically confined region (Flanders, Belgium). A cohort of >5000 individuals from the normal population is sampled for microbiome analysis and extensive metadata covering demographic, health- and lifestyle-related parameters is collected. Based on this cohort, a large-scale cross-sectional study of microbiome variability in relation to health as well as parameters associated to microbiome composition is being performed. In this presentation, I will discuss our experiences in large-scale microbiome monitoring, show how the development of dedicated computational approaches can assist in microbiome analysis and interpretation, and first results coming out of this effort.

*. Intervenant Invité

FTAG Finder : un outil simple pour déterminer les familles de gènes et les gènes dupliqués en tandem sous Galaxy

Bérengère Bouillon [*] [†1], Franck Samson[1], Étienne Birmelé[2], Loïc Ponger[3], Carène Rizzon[1]

[1] Laboratoire de Mathématiques et Modélisation d'Évry (LaMME) – Université d'Évry-Val d'Essonne, CNRS : UMR8071, ENSIIE, Institut national de la recherche agronomique (INRA) – I.B.G.B.I, 23 boulevard de France, F-91 037 Évry, France
[2] Laboratoire de Mathématiques Appliquées (MAP5 Paris Descartes) – CNRS : UMR8145 – Sorbonne Paris Cité, 45 rue des Saints-Pères, F-75 270 Paris Cedex 06, France
[3] Laboratoire Structure et Instabilité des Génomes – Muséum National d'Histoire Naturelle (MNHN), CNRS : UMR7196, Inserm : U1154, Sorbonne Universités – Muséum National d'Histoire Naturelle - CP26, 43 rue Cuvier, F-75 231 Paris Cedex 05, France

Session évolution moléculaire
mercredi 29 14h40
Amphi Mérieux

La duplication est un mécanisme qui joue un rôle important dans l'acquisition de nouveaux gènes et de nouvelles fonctions [1]. Une variété de mécanismes est impliquée dans l'apparition de gènes dupliqués, tels les crossing-over inégaux, la duplication par rétrotransposition ou bien encore la polyploïdie [2]. Parmi les différents types de gènes dupliqués, les gènes dupliqués en tandem ou TAG (Tandemly Arrayed Genes) sont peu étudiés pour le moment et les processus impliqués dans leur maintien sont mal compris. Ce sont des groupes de gènes paralogues adjacents sur le chromosome qui représentent une part importante des génomes [3,4,5]. Ils peuvent représenter selon les estimations jusqu'à 15 % des gènes chez l'Arabette [4] et 17.1 % chez le rat [5]. On sait aussi qu'ils présentent des caractéristiques spécifiques comparativement aux autres gènes dupliqués, par exemple ils sont surreprésentés en fonctions de stress principalement chez les plantes [3,4]. La dynamique des TAG au sein des génomes restant encore mal comprise, il en résulte qu'il n'y a pas de consensus établi quant à leur définition, et la plupart des études tendent à en prendre plusieurs en compte. Ainsi, dans le but de proposer un outil automatique permettant de déterminer les TAG de manière simple et en fonction des différentes définitions, nous avons développé le pipeline FTAG Finder : « Family and TAG Finder ». Ce pipeline consiste en trois étapes majeures (Cf. Figure 1) : a) déterminer les paires de gènes homologues b) construire les familles de gènes, c) puis à partir d'elles, déterminer les TAG.

Détermination des paires de gènes homologues : Nous avons choisi une définition évolutive pour déterminer les familles de gènes, à savoir que les membres d'une même famille représentent un groupe de gènes homologues. Trois outils permettent de déterminer les homologies :

i) Le premier outil du pipeline consiste à réaliser un BLASTp du protéome de l'espèce contre lui-même.

ii) Puis pour chaque paire de gènes, une fusion des hits de mêmes paires protéiques est réalisée afin d'obtenir la longueur totale et la similarité globale des régions alignées. Le meilleur hit entre deux protéines est sélectionné et fusionné avec le meilleur hit suivant de la même paire protéique si le chevauchement entre ces deux hits est inférieur à une valeur *n* définie par l'utilisateur (définie à 12 acides aminés par défaut). Ce processus est répété jusqu'à ce que toutes les paires soient rassemblées. Pour chaque paire de gènes donnée, sont alors calculées la valeur globale de couverture des hits pris en compte et la valeur moyenne de similarité. Le meilleur score parmi les hits conservés est également affiché en sortie.

*. Intervenant
†. Corresponding author : berengere.bouillon@hotmail.fr

iii) Ces résultats sont ensuite filtrés selon différents critères de similitude de séquence, couverture d'alignement, score d'alignement, et type d'isoforme (optionnel). L'utilisateur a le choix de définir lui-même les valeurs de ces critères ou de conserver les valeurs par défaut. Cette étape du pipeline permet ainsi à partir d'un fichier de séquences protéiques au format FASTA, de déterminer les paires de gènes homologues et de délivrer le résultat sous forme d'un fichier texte contenant une paire de gènes homologues par ligne (colonnes : gène 1 / gène 2 / score d'alignement du meilleur hit).

Construction des familles de gènes : Les relations d'homologie entre paires de gènes homologues sont ensuite codées sous la forme d'un graphe : les gènes représentent les nœuds, et les relations d'homologies les arêtes. Une famille de dupliqués va ainsi correspondre à un ensemble de nœuds densément connectés, appelé aussi communauté, c'est-à-dire que la plupart des relations d'homologies ont été retenues à l'intérieur de la famille. Parmi les nombreux algorithmes permettant de déterminer de tels ensembles, nous appliquons celui du Walktrap [6]. Celui-ci repose sur un système de marche aléatoire qui part du principe qu'une marche aléatoire courte tend à rester dans la même communauté. Il construit ainsi une hiérarchie de partitions des gènes puis choisi celle de modularité maximale, la modularité [7] étant un indice favorisant à la fois une forte densité *dans* chaque communauté et un faible nombre d'arêtes *entre* communautés.

Nous avons également introduit un nouveau paramètre qui donne la possibilité à l'utilisateur d'imposer une densité minimale pour les familles retenues, via la fraction minimale du nombre de gènes de la famille avec lesquels chaque gène doit être homologue. L'algorithme du Walktrap est alors appliqué récursivement jusqu'à ce que cette densité soit obtenue dans chaque famille de gènes retenue. Le fichier de sortie contient la liste des gènes appartenant à une famille avec le numéro de cette dernière (colonnes : gène / numéro de famille).

Détermination des TAG : L'étape finale du pipeline consiste à déterminer les TAG à partir du jeu de données de familles obtenu et du/des chromosomes à étudier. Pour un chromosome donné, l'outil recherche pour chacun des gènes quels sont ceux qui appartiennent à la même famille et qui lui sont proches, c'est-à-dire séparés au maximum d'un nombre x de « spacers » ou gènes intercalants. Les spacers sont les gènes n'appartenant pas à la famille considérée, donc non homologues à celle-ci. On peut ainsi proposer plusieurs définitions de TAG selon le nombre de gènes intercalants autorisés entre les membres de la même famille de gènes. Par exemple, la définition TAG0 n'autorise aucun spacer, alors que la définitionTAG5 autorise au maximum 5 spacers entre les différents membres de la même famille pour déterminer les TAG.

Le nombre maximal de gènes intercalants autorisés est défini par l'utilisateur qui a aussi la possibilité de calculer plusieurs définitions de TAG en même temps (Cf. Figure 2). Le fichier de sortie contient alors la liste de l'ensemble des gènes appartenant à des TAG en fonction des différentes définitions choisies par l'utilisateur pour chaque chromosome. Les numéro de la famille à laquelle appartiennent les gènes, ainsi que le brin sur lequel se situe le gène sont indiqués (colonnes : gène / brin / famille / définition TAG 1 / … / définition TAG n).

Application à l'Arabette : La version TAIR10 du protéome d'*Arabidopsis thaliana* contenant 35 386 protéines a été utilisée pour construire 3 jeux de données de familles de gènes. Des seuils de densité minimale de 30 %, 50 % et 70 % on été utilisés, et ce, à partir des homologues sélectionnés à 30 % de similitude et 70 % de couverture d'alignement. Les résultats obtenus entre les 3 densités étant homogènes, nous ne présenterons que les résultats pour une densité de 50 % dans ce résumé (Cf. Figure 3). Nous avons également comparé ces résultats à une précédente étude faite sur l'arabette [4] ou pour un total de 25 972 gènes ils obtenaient entre 40.4 et 67.0 % de dupliqués en fonction de la stringence de la définition, et 7.8 à 11.7 % de TAG parmi ces dupliqués en n'autorisant aucun « spacer ». Notre étude, qui a été réalisée sur une version plus récente du génome contenant 27 416 gènes (TAIR10), montre des résultats similaires, à savoir 64.6 % de dupliqués et 12.3 % de TAG.

Nous avons aussi calculé le temps total d'exécution du pipeline, qui est d'environ 2h22 (dont

2h20 pour le BLAST lancé avec 6 processus, 1 processus pour les autres outils), le tout sur une machine virtuelle Unix de 4 Go de RAM et 4 CPU.

Conclusion

Le workflow que nous proposons s'inscrit dans l'objectif de déterminer les familles de gènes et les TAG de manière simple pour l'intégralité du génome d'une espèce donnée. Nous avons également choisi de l'intégrer dans Galaxy sous la forme d'un workflow. Galaxy est à l'heure actuelle une plate-forme bioinformatique largement diffusée qui propose de nombreux outils principalement pour l'analyse de données NGS, mais aussi de plus en plus pour d'autres aspects de la bioinformatique. Cette plate-forme possède une interface commune à tous les outils et simple d'utilisation, permettant à quiconque ne possédant pas nécessairement des compétences informatiques de l'utiliser en toute simplicité. Elle facilite également le travail des développeurs qui n'ont pas besoin de créer d'interface pour leurs outils. Intégrer le pipeline FTAG Finder dans Galaxy permet ainsi de le rendre accessible à tous librement et d'automatiser le processus pour faciliter la gestion de grandes masses de données. Il laisse néanmoins la possibilité à l'utilisateur de se servir des outils séparément lorsqu'il dispose déjà d'une partie des données et ne souhaite pas tout recommencer, comme par exemple déterminer uniquement les TAG à partir d'une sélection de familles de gènes personnelle. Ainsi, chaque étape clef du pipeline mène à la production d'un fichier texte - gènes homologues, familles de gènes, TAG - qui peuvent être utilisés séparément. De plus, les paramètres proposés dans les différents outils de FTAG Finder ont été pensés pour être facilement abordables même par un utilisateur peu averti, et une documentation claire a été mise en place pour chaque outil dans l'interface Galaxy.

FTAG Finder est fonctionnel chez l'arabette avec la syntaxe des identifiants TAIR. Toutes les étapes du pipeline sont paramétrables par l'utilisateur et il n'existe pas d'outils équivalent sur Galaxy notamment pour la détermination des TAG. Une amélioration du pipeline est en cours et permettra de gérer des identifiants de protéines et de gènes possédant une syntaxe complexe.

Références

[1] Wolfe and Li, Molecular evolution meets the genomics revolution, *Nat. Genet. 33(Suppl.): 255–265* ; 2003

[2] Ohno, Evolution by gene duplication, 1970 ; *Springer*

[3] Hanada and *al.*, Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli, *Plant Physiology* ; 2008

[4] Rizzon and *al.*, Striking similarities in the genomic distribution of Tandemly Arrayed Genes in Arabidopsis and Rice, *PLOS Computational Biology* ; 2006

[5] Shoja and Zhang, A roadmap of Tandemly Arrayed Genes in the genomes of Human, Mouse and Rat, *Mol. Biol. Evol.* ; 2006

[6] Pons and Latapy, Computing communities in large networks using random walks, IS-CIS'05, *Lecture Note in Computer Science* ; 2005

[7] Girvan and Newman, Community structure in social and biological networks, *PNAS* ; 2002

Mots clefs : Galaxy, Gènes dupliqués en tandem, Duplication, Génomique comparative, Workflow

Evolution of gene regulation in 20 mammals

Camille Berthelot [*1,2], Diego Villar[3], Duncan T. Odom[3], Paul Flicek[1]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) – Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK/Royaume-Uni
[2] Institut de Biologie de l'École Normale Supérieure (IBENS) – CNRS : UMR8197, Inserm : U1024 – 46 rue d'Ulm, F-75 005 Paris, France
[3] University of Cambridge, Cancer Research UK - Cambridge Institute (CRUK-CI) – Royaume-Uni

Session évolution moléculaire
mercredi 29 15h00
Amphi Mérieux

Despite long-standing efforts from the genomics community, our understanding of gene expression regulation in mammalian genomes remains incomplete. This is in a large part because unlike coding sequences, non-coding regulatory elements cannot be fully uncovered using sequence characteristics and conservation alone, so that traditional sequence-based approaches have proved inadequate to describe the regulatory network of mammalian genomes. Previous work on the evolutionary dynamics of transcription factor binding and histone modifications has shown that mammalian regulatory sequences seem to be fast-evolving and exhibit high turn-over rates. However, the underlying properties that rule these dynamics and their consequences on gene expression are poorly understood.

We report here the results of a large-scale functional genomics study in liver tissue from 20 species spanning the breadth of the mammalian phylogenetic tree from monotremes to primates, including many species that have not been previously studied with genome-wide methods such as cetaceans. We used ChIP-seq to produce a high-resolution description of the genomic landscape marked by two types of histone modifications: H3K4me3 (associated with active promoters) and H3K27ac (associated with active promoters and enhancers), as well as RNA-seq to document gene expression levels. The project explores the conservation, birth, loss and turn-over of promoters and enhancers and their consequences on gene expression in liver, a representative, largely homogeneous tissue shared by all mammals.

We observe that while the total number of active regulatory regions in liver remains largely unchanged over mammalian evolution, the genomic locations of these sites and especially of active enhancers have evolved markedly. Using statistical models, we describe the rates and patterns of evolutionary divergence of mammalian promoters and enhancers as well as the emergence of new regulatory elements. We then investigate the genomic properties and functional annotations of mammalian regulatory regions exhibiting either high conservation or high plasticity in an effort to connect regulatory evolution with functional significance. Lastly, we explore how changes in local regulatory context affect gene expression, and attempt to reconcile the apparent evolutionary stability of gene expression in mammalian liver with the high plasticity of the regulatory regions that control their expression.

Mots clefs : functional genomics, epigenomics, comparative genomics, transcriptional regulation, promoters, enhancers

*. Intervenant

Evolution of internal eliminated sequences in *Paramecium*

Diamantis Sellis [*1], Frédéric Guérin[2], Olivier Arnaiz[3], Walker Pett[1], Nicole Boggetto[4], Sascha Krenek[5], Thomas Berendonk[5], Arnaud Couloux[6], Jean-Marc Aury[6], Karine Labadie[7], Sophie Malinsky[8], Simran Bhullar[8], Éric Meyer[8], Linda Sperling[3], Sandra Duharcourt[4], Laurent Duret[1]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Institut Jacques Monod (IJM) – Université Paris VII - Paris Diderot, CNRS : UMR7592 – Université Paris Diderot, Bât. Buffon, 15 rue Hélène Brion, F-75 205 Paris Cedex 13, France
[3] Institut de Biologie Intégrative de la Cellule (I2BC) – CNRS : UMR9198, CEA, Université Paris Sud - Paris XI – Institut de Biologie Intégrative de la Cellule - CNRS Avenue de la Terrasse Bât. 26, F-91 198 Gif-sur-Yvette Cedex, France
[4] ImagoSeine, Institut Jacques Monod – ImagSeine – UMR CNRS 7592, Université Paris-Diderot, 15 rue Hélène Brion, F-75 205 Paris Cedex 13, France
[5] TU Dresden, Institute of Hydrobiology, Dresden, Germany/Allemagne
[6] Genoscope - Centre national de séquençage – CEA, Genoscope – 2 rue Gaston Crémieux CP5706, F-91 057 Évry Cedex, France
[7] Commissariat à l'Énergie Atomique (CEA) – CEA – France
[8] École Normale Supérieure de Paris (ENS Paris) – 45 rue d'Ulm, F-75230 Paris Cedex 05, France

Session évolution moléculaire
mercredi 29 15h20
Amphi Mérieux

The genome of *Paramecium* undergoes a remarkable reorganization during the development of the macronucleus. Tens of thousands of mostly short sequences are interspersed in the micronuclear genome, often interrupting protein reading frames. During the macronuclear development, these internal eliminated sequences (IESs) are precisely excised and thus the macronuclear genes are fully functional. The role, if any, of IESs, their evolutionary history and dynamics as well as the molecular mechanism(s) involved in their excision are still far from resolved. As a first step towards exploring these questions we here performed a large scale systematic reconstruction of the evolutionary history of IESs in *Paramecium*. We focused on 7 species from the aurelia species complex and used *P. caudatum* as an outgroup. We sequenced the micronuclear and, when not available, the macronuclear genomes of each species and annotated genes and IESs. We specifically focused on conserved genes that can be aligned across species in order to find homologous IESs. We inferred the phylogenies of each gene family and reconciled the resulting gene trees with the species tree. Using a Bayesian approach we inferred the ancestral states of presence and absence for each IES. The result is an unprecedented detailed description of the evolutionary history of tens of thousands of IESs. Preliminary results validate previous models of IES evolution. We find that there was a wave of insertion of IESs after the split of the aurelia species complex from *P. caudatum* and subsequently the rate of gain and loss of IESs was significantly reduced. Our detailed description of IES evolutionary history also enables us to compare the age of acquisition of different IESs with various genomic properties. For example we find that recently inserted IESs are on average longer. We believe that our results will fuel further studies to test models and gain new insights on the molecular mechanisms of genomic rearrangements in the developing macronucleus.

Mots clefs : Evolution, genomics, ciliates, Paramecium, internal eliminated sequences, IES

*. Intervenant

IntegronFinder reveals unexpected patterns of integron evolution

Jean Cury [*] [†1], Thomas Jové[2], Bertrand Néron[3], Marie Touchon[1], Eduardo Rocha[1]

[1] Génomique évolutive des microbes – Institut Pasteur de Paris, CNRS : UMR3525 – 28 rue du Dr Roux, F-75 724 Paris Cedex 15, France
[2] Université de Limoges – Inserm : UMRS1092 – F-87 000 Limoges, France
[3] Centre d'informatique pour la biologie – Institut Pasteur de Paris – 25 rue du Dr Roux, F-75 015 Paris, France

Session évolution moléculaire
mercredi 29 15h40
Amphi Mérieux

Integrons recombine gene arrays and favor the spread of antibiotic resistance. They are composed of an integrase and a cluster of gene cassettes (typically an Open Reading Frame flanked by 2 recombination sites, named *attC* sites). The integrase integrates novel cassettes and shuffles the cluster of existing gene cassettes. A subset of the cluster is actually expressed while the rest of it is costless for the organism. Integrons may have very high adaptive value, e.g., they drive the acquisition of antibiotic resistance in many bacteria. However, their broader adaptive roles remain mysterious, partly due to lack of computational tools.

We made program – IntegronFinder – to identify integrons with high accuracy and sensitivity. It uses a combination of methods to identify the two key features of the integron, namely, the integron-integrase, intI, and the recombination site, *attC*, as well as some well-known promoters. The identification of the *attC* sites was the crux of the integron detection methods in bacterial genomes because their sequence is very poorly conserved making classical DNA annotation methods, such as blast query or HMM profile, ineffective. Since the secondary structure of *attC* sites is conserved (and essential for function) we built a covariance model for it. The resulting program was written in Python and can be used as a standalone application (available at `https://github.com/gem-pasteur/Integron_Finder`), or as a web-server (available at `http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::integron_finder`).

We extensively benchmarked the program using curated datasets and simulations. We show that the program is not sensitive to the G+C percent of the replicon, is able to find the vast majority of known *attC* sites (88 % of sensitivity), and has a very low false positive rate (ranging from 0.03 FP/Mb to 0.72 FP/Mb, *i.e.* less than one site per genome on average). The high sensitivity and low false positive rate of the program makes its use on draft genome and metagenomes suitable. This could highly increase the number of cassette described so far.

The availability of a covariance model to identify *attC* sites shed new light on integron evolution. We searched for integrons, integron-integrases lacking *attC* sites (In0), and clusters of *attC* sites lacking a neighboring integron-integrase (CALIN) in bacterial genomes. We found that all these elements are much more frequent in genomes of intermediate size (4-6 Mb). This might be the result of the combined effect of the frequency of transfer (increasing with genome size) and selection of compact genetic elements (decreasing with genome size). Second, Integrons were not found in some key phyla, such as α-Proteobacteria, whereas they were abundant in sister-phyla such as γ- and β-proteobacteria and we could find them in much more distant clades. This might reflect selection against cell lineages that acquire integrons. Third, the similarity between *attC* sites is proportional to the number of cassettes in the integron. Suggesting a model where the creation of novel cassettes is proportional to the number of cassettes in the integron. Finally, we found a

*. Intervenant
†. Corresponding author : `jean.cury@pasteur.fr`

completely unexpected high frequency of CALIN in genomes lacking integron-integrases or their remains. They constitute a large novel pool of cassettes lacking homologs in the databases. We propose that they represent an evolutionary step between the acquisition of genes within integrons and their stabilization in the new genome upon integron loss.

Mots clefs : Integrons, Computational biology, Genomics, Bacteria, Evolutionary biology

Encoding genomic variation using De Bruijn graphs in bacterial GWAS

Magali Jaillard [*1,2], Maud Tournoud[2], Leandro Ishi[1], Vincent Lacroix[1], Jean-Baptiste Veyrieras[2], Laurent Jacob[1]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 VILLEURBANNE Cedex, France
[2] Bioinformatics Research Department – BIOMÉRIEUX – Chemin de l'Orme, F-69 280 MARCY L'ÉTOILE, France

Session études d'association
mercredi 29 14h40
Salle des thèses

Motivation

Antimicrobial resistance has become a major worldwide public health concern, as illustrated by the increase of hospital acquired infections on which empirical treatments fail because of multi-resistant strains [11]. This worrisome situation calls for a better comprehension of the genetic bases of existing and novel resistance mechanisms. Genome-wide association studies (GWAS) aim at linking phenotypes to genetic determinants, and seem appropriate for this purpose. Over the last four years, thanks to the access to rich panels of bacterial genomes and phenotypic data, bacterial GWAS have shown encouraging results for new genetic markers discovery [1, 4, 5, 7, 15].

Most existing bacterial association studies use approaches developed for human GWAS for encoding genome variation and look at SNPs that are called using a reference genome [1, 4, 7]. However a suitable reference is not always available, in particular for species with a high genome plasticity which have a large accessory genome. The accessory genome is the part of the genome not found in all strains (usually of the same species), and is made of genetic material acquired independently from inheritance, by horizontal gene transfer. For highly plastic species, this accessory genome can represent more than a quarter of the complete genome, leading to manifold genomes which vary by their size and content [8].

To account for the variation in genes content, some studies also use the presence or absence of genes represented in the studied panel as candidates [5]. However, genetic determinants linked to regulation patterns may be located in non-coding regions, and thus will be missed by approaches relying on this representation. Thus, the way to encode genomic variation in bacterial GWAS remains an open question [13].

To get around this issue, some studies have considered sets of k-mers (i.e. overlapping sequences of length k), as candidate determinants [5, 15]. Contrarily to SNP-based approaches, k-mers describe the genomes diversity without requiring an alignment to a reference genome. They are however highly redundant as each single locus is represented by several overlapping k-mers. Moreover, the number of distinct k-mers contained in a set of genomes increases with the value of k, and reaches easily tens of millions features leading to very high memory, time loads and complexity in feature interpretation.

We propose here a representation of the genomic variation between bacterial genomes which accounts for both coding and non-coding haplotypes of the core and accessory genome, while remaining computationally tractable. Our representation is based on De Bruijn graphs [3] (DBG), which are widely used for *de novo* genome assembly [17] and variant calling [10]. We show that using the nodes of the compressed DBG rather than fixed length k-mers as candidate determinants reduces drastically the computational burden of data preprocessing and yields more interpretable

*. Intervenant

results. We also conjecture that it may help better estimate the population structure which is crucial to avoid spurious association in GWAS.

Material and Methods

DBGs are directed graphs representing overlaps between a set of strings. More specifically in the context of genomic sequences, the DBG nodes are all unique *k*-mers extracted from the sequences and an edge is drawn between nodes if the (*k*-1)-length suffix of a node equals the (*k*-1)-prefix of another node. If a *k*-mer overlaps two other *k*-mers but these two *k*-mers do not overlap each other, we obtain a fork pattern in the graph. When both branches of the fork join again into one shared *k*-mer, this draw a bubble pattern representing a polymorphic region such as a SNP. DBGs are thus well suited to describing variants [10].

Interestingly, these graph can be compressed, by first using a unique node to store a *k*-mer sequence and its reverse complement, and then merging linear paths, i.e. sequences of nodes not linked to more than two other nodes. This compression is done without loss of information, because it only affects redundant information [14]. Thus, the nodes of the compressed DBG can be thought of as haplotypes of variable length in different regions of the genomes: coding and non-coding regions as well as core and accessory genome. In the remaining, we denote by "node" a node of the compressed graph.

We worked on a panel of 282 strains of *Pseudomonas aeruginosa* species, sequenced on Illumina HiSeq 2500 and assembled using a modified version of the IDBA_UD assembler [12]. Antibiotic resistance phenotypes were obtained by broth dilution assays complemented with Vitek testing (bioMérieux, Marcy-l'Étoile, France), for several drugs commonly used in *P. aeruginosa* infections, including amikacin [16]. Clinical and Laboratory Standards Institute (CLSI) guidelines were applied on the resistance data to determine susceptibility or non-susceptibility.

We assumed that most of the sequencing errors had been removed by the assembly step. We built a single DBG from the 282 genomes, with a *k*-mer length of 41 pb, using kissplice software, version 2.3.1 [14]. All resulting *k*-mers and DBG nodes were then mapped to the original genome assemblies using bowtie2 [9] in order to build a presence/absence matrix of *k*-mers and nodes. Duplicated features were removed from the matrix and features with a minor allele frequency (MAF) below 2 % were filtered.

Years of experience of human GWAS have taught us that spurious associations can be detected if no precaution is taken to correct for population structure [2, 18]. Since population structure can be very strong within bacterial strains [5, 6], we estimated this structure from the genotypes and included it in the univariate linear models for amikacin resistance. From these models, we computed a p-value per candidate (either k-mers or DBG nodes) and compared the p-value distribution of amikacin known determinants to the p-value distribution of other candidates.

Results

k is the key parameter for the graph resolution level. A small value (below 20 bp) will generate words of low complexity which will be highly repeated in the genomes, creating thus numerous loops in the graph. This latter will then hardly be compressed: for *k*=15, we counted twice more *k*-mers than nodes (34 M versus 15 M). When *k* increases, k-mers are less repeated within each genome, leading to better levels of graph compression. However, the number of *k*-mers required to describe the complete polymorphism of all the genomes at a given position will increase with the number of polymorphic positions contained in the *k*-mer and thus will increase with *k*. Thus for *k*=41, we obtained 62.5 M *k*-mers and 2.2 M nodes.

This important difference in the number of features obtained for nodes and *k*-mers makes a real difference for the data preprocessing step: while creating the node presence/absence matrix takes few hours, it takes several days for the *k*-mer matrix creation, and requires a lot of memory

(> 250 Go). The last step, removing duplicated features is particularly demanding even though the final de-duplicated matrices are identical for nodes and *k*-mers (nodes are compressed sets of *k*-mers, without loss of information). The percent of duplicated features illustrates how *k*-mers are highly redundant. Indeed, while duplicates represent 28 % of the MAF-filtered nodes, they represent 96 % of the MAF-filtered *k*-mers.

k-mer and node features showed a comparable enrichment of low p-values for amikacin known determinants in our linear model. However, each *k*-mer feature typically corresponds to a large number of distinct *k*-mers, that have been filtered out in the previous step. Some of these redundant *k*-mers represent the same haplotype while others represent distinct haplotypes in perfect linkage disequilibrium. For example if two SNPs are in perfect disequilibrium, i.e., each allele of the first SNP is always associated with the same allele of the second SNP, then the *k*-mers covering either SNP have the exact same pattern of presence/absence across genomes.

A single p-value is computed for the *k*-mer retained to represent this pattern in the linear model, and if this p-value is below the chosen significance level, all the corresponding *k*-mers need to be considered for interpretation or validation, e.g. by mapping them on genome annotation. On the other hand, these two SNPs could correspond to four DBG nodes. The same significantly associated presence/absence pattern will therefore correspond to a much lower number of more interpretable features: longer haplotypes which by construction are at the right resolution for the set of considered genomes. In practice for *k*=41, the node length varied from *k* to 104,553 bp, with a median value of 57 pb.

In addition to the preprocessing time and memory load, and interpretation aspects, using nodes rather than *k*-mers may yield better estimates of the population structure. Indeed, this estimate takes into account the duplication within features and, while duplicated nodes only represent biological duplicates, i.e. regions in perfect linkage disequilibrium, duplicates within *k*-mers also account for neighbor sequence overlaps. Validating our conjecture that DBG nodes provide better population structure estimates than *k*-mers and lead in turn to more power for detecting genetic determinants will require simulation of synthetic genomes from a given phylogeny and will be the subject of future work.

These encouraging results suggest that compressed De Bruijn Graphs nodes are well suited to describing genetic determinants of bacterial resistance in particular for bacterial species with high plasticity.

References

[1] M. T. Alam, *et al.* Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. *Genome biology and evolution*, 6(5):1174–1185, 2014.

[2] D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[3] d. N. Bruijn. A combinatorial problem. *Proceedings of the KoninklijkeNederlandse Akademie van Wetenschappen.* Series A, 49(7):758, 1946.

[4] C. Chewapreecha, *et al.* Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet*, 10(8):e1004547, 2014.

[5] S. G. Earle, *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology*, page 16041, 2016.

[6] D. Falush and R. Bowden. Genome-wide association mapping in bacteria? *Trends in microbiology*, 14(8):353–355, 2006.

[7]M. R. Farhat, *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant mycobacterium tuberculosis. *Nature genetics*, 45(10):1183–1189, 2013.

[8] V. L. Kung, E. A. Ozer, and A. R. Hauser. The accessory genome of pseudomonas aeruginosa. *Microbiology and Molecular Biology Reviews*, 74(4):621–641, 2010.

[9] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. methods*, 9(4):357–359, 2012.

[10] Y. Le Bras, *et al.* Colibread on galaxy: a tools suite dedicated to biological information extraction from raw ngs reads. *GigaScience*, 5(1):1, 2016.

[11] S. T. Micek, *et al.* An international multicenter retrospective study of pseudomonas aeruginosa nosocomial pneumonia: impact of multidrug resistance. *Crit Care*, 19(219.10):1186, 2015.

[12] Y. Peng, *et al.* Idba-ud: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, 2012

[13] T. D. Read and R. C. Massey. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med*, 6(11):109, 2014.

[14] G. A. Sacomoto, *et al.* Kissplice: de-novo calling alternative splicing events from rna-seq data. *BMC bioinformatics*, 13(Suppl 6):S5, 2012.

[15] S. K. Sheppard, *et al.* Genome-wide association study identifies vitamin b5 biosynthesis as a host specificity factor in campylobacter. *Proceedings of the National Academy of Sciences*, 110(29):11923–11927, 2013.

[16] A. van Belkum, *et al.* Phylogenetic distribution of crispr-cas systems in antibiotic-resistant pseudomonas aeruginosa. *mBio*, 6(6):e01796–15, 2015.

[17] W. Zhang, *et al.* A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PloS one, 6(3):e17915, 2011.

[18] X. Zhou and M. Stephens. Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nature methods*, 11(4):407, 2014.

Mots clefs : GWAS, De Bruijn graph, bacteria, genomic variation

needlestack : an highly scalable and reproducible pipeline for the detection of ctDNA variants using multi-sample deep next generation sequencing data

Session études d'association
mercredi 29 15h00
Salle des thèses

Tiffany Delhomme [*1], Patrice Avogbe[1], Graham Byrnes[2], James Mckay[1], Matthieu Foll [†1]

[1] Genetic Cancer Susceptibility Group, Section of Genetics, International Agency for Research on Cancer – IARC (WHO) – France
[2] Biostatistics Group, Section of Genetics, International Agency for Research on Cancer – IARC (WHO) – France

Every cancer arises as a result of the acquisition of a series of DNA sequence abnormalities, including somatic mutations, *i.e.* mutations occurring in a non germ cell. The comprehensive characterization of somatic mutations by screening cancer genomes can help to understand cancer appearance and progression but also to identify accurately predictive biomarkers. In 2015, the International Cancer Genome Consortium launched a large benchmarking operation with the objective of identifying and resolving issues of somatic mutation calling [1]. The conclusion of this study was that detecting somatic mutations in cancer genomes remains an unsolved problem. The problem is exacerbated when trying to identify mutations in circulating tumor DNA (ctDNA), due to a low proportion of tumor DNA among the total amount of circulating free DNA (cfDNA). Indeed, Next Generation Sequencing (NGS) error level can reach this low proportion, and somatic variant calling from ctDNA is like finding a needle in a needlestack.

Here we introduce needlestack, an ultra sensitive variant caller based on the idea that analysing several samples together can help estimate the distribution of sequencing errors to accurately identify variants present in very low proportion. Contrary to most existing algorithms, needlestack can deal with both single nucleotide substitutions (SNVs) and short insertions or deletions. At each position and for each candidate variant, we model sequencing errors using a robust negative binomial (NB) regression, with a linear link and a zero intercept, and detect variants as being outliers from this error model. More specifically, let $i=1 \dots N$ be the index of the sample taken from a sequenced panel of size N, j the analysed position and k the potential alteration, with *ins* and *del* covering respectively every insertion and deletion observed in the data at position j. Let denote the total number of sequenced reads at position j for the sample i, the reads count supporting alteration k and the error rate at position j. At each position and for each candidate variant, we model, for each sample i identified by its coverage, the sequencing error distribution using a negative binomial (NB) regression:

with the overdispersion parameter and.

Genetic variants are detected as being outliers from this error model. To avoid these outliers biasing the regression we adapted a recently published robust estimator for the negative binomial regression [2] based on a robust estimation of the overdispersion parameter. We calculate for each sample a p-value for being a variant (*i.e.* outlier from the regression) that we further transform into q-values using the Benjamini and Hochberg method [4] to account for multiple testing and control the false discovery rate.

*. Intervenant
†. Corresponding author : follm@iarc.fr

We compared the performance of needlestack with both MuTect [5], one of the most widely used somatic variant caller, and Shearwater [6], a variant caller based on a similar idea developed at the Wellcome Trust Sanger Institute. We benchmarked our method using BamSurgeon [7] to introduce SNVs at varied variant allele fractions (VAFs) in Ion Torrent Proton deep sequencing data of *TP53* exons from 125 cfDNA samples (1,501 base pairs with a median coverage of 11,344 X). In a first scenario, we added 1,000 SNVs in total at random positions in the gene, and in a second "hotspot" scenario, we introduced SNVs at 40 random positions, each mutation being added to a set of 25 samples randomly chosen each time. In both cases, we repeated the process ten times. Each sample has been sequenced twice for validation, with each *in-silico* mutation being introduced in the two technical duplicates. Variants calls were compared using Receiver Operator Curves, with and without taking into account technical duplicates for additional error correction. We also studied the influence of the sequencing error rate as estimated by needlestack and of the variant allelic fractions (VAF, denoting the fraction of reads carrying the mutation) on the sensitivity of the three methods. In this case, the statistic cutoff for each method was chosen in order to achieve the same specificity among the three methods (1 false positive per sample over the whole *TP53* region sequence). We then measured ratios of sensitivity () and ratios of detectable VAFs () between needlestack and both Shearwater and MuTect.

We show that sequencing each sample as a technical duplicate is an essential step in these methods to remove errors that are likely to be introduced by laboratory procedures prior to the sequencing (PCR errors). needlestack outperforms existing variant callers with sensitivities 1.6 and 2.3 times higher than Shearwater and MuTect respectively. needlestack is able to detect variants with allelic fractions more than 10 times smaller than Shearwater and 20 times smaller than MuTect. The difference with the two other methods is even stronger in the case of the hotspot scenario. Sequencing errors generated by the Ion Torrent Proton sequencer vary by two orders of magnitude over the 1,501 positions we sequenced. Unlike other methods, needlestack can detect VAF as low as 0.01 % when the sequencing error rate is low ($< 10^{-4}$).

needlestack is deployed using the Nextflow [3] pipeline environment, allowing high scalability, portability and reproducibility by providing a Docker container image and versioned source code on GitHub. Parallelization is achieved by splitting the genomic positions to analyse in chunks that are then run in parallel. In term of performance, needlestack can analyse a set of 20 whole exome sequenced samples in 15 hours when launched on 200 computing cores.

References

[1] Alioto TS, Buchhalter I, Derdak S, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun*. 2015;6:10001.

[2] Aeberhard WH, Cantoni E, Heritier S. Robust inference in the negative binomial regression model with an application to falls data. *Biometrics*. 2014;70(4):920-31.

[3] Paolo Di Tommaso et al. A novel tool for highly scalable computational pipelines. `https://figshare.com/articles/A_novel_tool_for_highly_scalable_computational_pipelines/1254958`.

[4] Yoav Benjamini and Yosef Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*. 1995;57(1):289-300.

[5] Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-9.

[6] Gerstung M, Papaemmanuil E, Campbell PJ. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics*. 2014;30(9):1198-204.

[7] Ewing AD, Houlahan KE, Hu Y, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods*. 2015;12(7):623-30.

Mots clefs : bioinformatics, cancer genomics, variant calling, somatic mutations, computational pipeline, circulating free DNA

De novo identification, differential analysis and functional annotation of SNPs from RNA-seq data in non-model species

Hélène Lopez-Maestre [*] [†1], Lilia Brinza[2], Camille Marchet[3], Janice Kielbassa[4], Sylvère Bastien[1], Mathilde Boutigny[1], David Monin[1], Adil El Filali[1], Claudia Cararetto[5], Cristina Vieira[1], Franck Picard[1], Natacha Kremer[1], Fabrice Vavre[1], Marie-France Sagot[6], Vincent Lacroix [‡1]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] IRT – BIOMÉRIEUX, SANOFI Recherche – France
[3] GENSCALE (INRIA - IRISA) – École normale supérieure (ENS) - Cachan, Université de Rennes 1, CNRS : UMR6074, INRIA – Campus de Beaulieu, F-35 042 Rennes Cedex, France
[4] Centre Léon Bérard – CRLCC Léon Bérard – 28, rue Laennec, F-69 373 Lyon Cedex 08, France
[5] Department of Biology (UNESP) – São Paulo State University, São José do Rio Preto, Brésil
[6] ERABLE/BAOBAB (LBBE Lyon / INRIA Grenoble Rhône-Alpes) – CNRS : UMR5558, INRIA, Université Claude Bernard - Lyon I, Laboratoire de Biométrie et Biologie Évolutive – Laboratoire de Biométrie et Biologie Évolutive. UMR CNRS 5558 Campus de La Doua - Université Claude Bernard - Lyon 1 Bâtiment Grégoire Mendel - 16 rue Raphaël Dubois 69 100 Villeurbanne – INRIA Grenoble - Rhône-Alpes Inovallée, 655 avenue de l'Europe, Montbonnot, F-38 334 Saint Ismier Cedex, France

Session études d'association
mercredi 29 15h20
Salle des thèses

Introduction

SNPs (Single Nucleotide Polymorphisms) are genetic markers whose precise identification is a prerequisite for association studies. Methods to identify them are currently well developed for model species (McKenna et al. 2010, Li et al. 2009), but rely on the availability of a (good) reference genome, and therefore cannot be applied to non-model species. They are also mostly tailored for whole genome (re-)sequencing experiments, whereas in many cases, transcriptome sequencing can be used as an alternative which already enables to identify SNPs. In RNA-seq data, only SNPs from transcribed regions can be targeted, but they arguably correspond to those with a more direct functional impact. Using RNA-seq technology largely reduces the cost of the experiment, and the obtained data concurrently mirror gene expression, the most basic molecular phenotype. RNA-seq experiments may also provide very high depth at specific loci and therefore allow to discover infrequent alleles in highly expressed genes.

In this work, we present a method for the de novo identification, differential analysis and annotation of variants from (only) RNA-seq data, that allows to work with non-model species.

Methods

KisSplice (Sacomoto et al. 2012) is a software initially designed to find alternative splicing events (AS) from RNA-seq data, but which also outputs indels and SNPs. The key concept is that a SNP corresponds to a recognisable pattern (Peterlongo et al. 2010, Iqbal et al. 2012 and Uricaru et al. 2015), called a bubble, in a de Bruijn graph built from the reads. The nodes of a de Bruijn graph are words of length k, called k-mers. There is an edge between two nodes if the suffix

*. Intervenant
†. Corresponding author : helene.lopez@univ-lyon1.fr
‡. Corresponding author : vincent.lacroix@univ-lyon1.fr

of length k-1 of the first k-mer is identical to the prefix of length k-1 of the second k-mer. The de Bruijn graph that is built from two alleles of a locus will therefore correspond to a pair of paths in the graph, which form the bubble. Unlike AS events and indels, bubbles generated by SNPs have two paths of equal length. Thus, KisSplice consists in essentially 3 steps: (i) building the de Bruijn graph from the RNA-seq reads; (ii) enumerating all bubbles in this graph; and (iii) mapping the reads to each path of each bubble to quantify the frequency of each variant.

The whole pipeline we present here consist in several steps. 1) First we identify and quantify the SNPs with KisSplice, 2) as KisSplice provides only a local context around the SNPs, a reference can be built with Trinity. 3) SNPs can be positioned on whole transcripts (some SNPs that do not map on the transcripts of Trinity, are harder to study but can still be of interest). 4) We propose a statistical method, called KissDE, to find condition-specific SNPs (even if they are not positioned) out of all SNPs found. 5) Finally, we can also predict the amino acid change for the positioned SNPs, and intersect these results with condition-specific SNPs using our package Kissplice2reftranscriptome.

It takes as input RNA-seq reads from at least 2 conditions (e.g., the modalities of the phenotype) with at least 2 replicates each, and outputs variants associated with the condition and their annotation (variant in non-coding region or in coding region, are the two variant synonymous or non synonymous). The method does not require any reference genome, nor a database of SNPs. It can therefore be applied to any species for a very reasonable cost. Individuals can be pooled prior to sequencing, if not enough material is available for sequencing from one individual.

Results

We first evaluated our method using RNA-seq data from the human Geuvadis project (Lappalainen et al. 2013). The great advantage of this dataset is that SNPs are well annotated, since the selected individuals were initially included in the 1000 genomes project, and a functional annotation of SNPs is available (dbSNP). This enables to clarify what is the precision and recall of our method and the impact of several parameters to the performance of SNP calling, but also how it compares to widely used mapping methods using a reference genome or an assembled transcriptome (as in Van Belleghem et al. 2012, Pante et al. 2012 and GATK Best practices for RNAseq).

We could evaluate the prediction of the functional impact for 81 % of SNPs (present in dbSNP and having functional impact reported), and our predictions were correct for 96 % of these cases.

When comparing KisSplice to other methods, we could assess that both the precision and recall are lower than methods using a reference genome but larger than the one using the assembled transcriptome.

However, the recall of KisSplice can be improved by lowering the minimum allele frequency authorized from 5% to 2%, allowing KisSplice to reach a better recall than methods using a reference genome in highly expressed regions. In the same way, the lack of recall of methods using an assembled transcriptome can be improve by keeping only longest isoform of each gene.

We then applied our method in the context of non-model species.

First we focused on *Asobara tabida*, an hymenoptera that exhibits contrasted phenotypes of dependance to its symbiont. Using RNA-seq data from two extreme modalities of the phenotype, we were able to establish a catalog of SNPs, stratify them by functional impact, and assess which SNPs had a significant change of allele frequency across modalities. We further selected cases for experimental validation, and were able to confirm RT-PCR and Sanger sequencing that the SNPs were indeed condition specific through.

Then we applied our method on two recently diverged *Drosophila* species, *D. arizonae* and *D. mojavensis*. In this case our method identifies differences of one nucleotide which are not SNPs

but divergences. On this system also, we were able to validate experimentally that the loci we identify were truly divergent.

Conclusion

Using human RNA-seq data, we show that, although our method does not require a reference genome, its accuracy is close to methods that use one. We also show that our method can find SNPs in the context of alternative splicing, a key feature which enables it to outperform methods that call SNPs from reads mapped to an assembled transcriptome.

The method is distributed as a pipeline (`http://kissplice.prabi.fr/TWAS/`) and can be used for any species to annotate SNPs and predict their impact on proteins. We further enable to test for the association of the identified SNPs with a phenotype of interest.

References

GATK Best Practices: Calling variants in RNAseq, `https://www.broadinstitute.org/gatk/guide/article?id=3891` posted on 2014-03-06, last updated on 2014-10-31.

Iqbal, Zamin, Caccamo, Mario, Turner, Isaac, et al. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, 2012, vol. 44, no 2, p. 226-232.

Lappalainen, Tuuli, Sammeth, Michael, Friedländer, Marc R., et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 2013, vol. 501, no 7468, p. 506-511.

Li, Heng, Handsaker, Bob, Wysoker, Alec, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009, vol. 25, no 16, p. 2078-2079.

McKenna, Aaron, Hanna, Matthew, Banks, Eric, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 2010, vol. 20, no 9, p. 1297-1303.

Pante, Eric, Rohfritsch, Audrey, Becquet, Vanessa, et al. SNP detection from de novo transcriptome sequencing in the bivalve Macoma balthica: marker development for evolutionary studies. *PLoS One*, 2012, vol. 7, no 12, p. e52302.

Peterlongo, Pierre, Schnel, Nicolas, Pisanti, Nadia, et al. Identifying SNPs without a reference genome by comparing raw reads. In : *String Processing and Information Retrieval*. Springer Berlin Heidelberg, 2010. p. 147-158.

Sacomoto, Gustavo AT, Kielbassa, Janice, Chikhi, Rayan, et al. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC bioinformatics*, 2012, vol. 13, no Suppl 6, p. S5.

van Belleghem, Steven M., Roelofs, Dick, van Houdt, Jeroen, et al. *De novo* transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle *Pogonus chalceus* (Coleoptera, Carabidae). *PloS one*, 2012, vol. 7, no 8, p. e42605.

Mots clefs : SNP, RNA, seq, method, reference, free, quantification, annotation, pooled, data

Prediction of ionizing radiation resistance in bacteria using a multiple instance learning model

Manel Zoghlami [*] [†1,2], Sabeur Aridhi [‡1,3], Haithem Sghaier[4], Mondher Maddouri[2], Engelbert Mephu Nguifo[1]

[1] Laboratoire d'Informatique, de Modélisation et d'optimisation des Systèmes (LIMOS) – Institut Français de Mécanique Avancée, Université Blaise Pascal - Clermont-Ferrand II, Université d'Auvergne - Clermont-Ferrand I, CNRS : UMR6158 – Bât. ISIMA Campus des Cézeaux BP 10025, F-63 173 Aubière Cedex, France
[2] Laboratoire en Informatique en Programmation Algorithmique et Heuristique (LIPAH) – Tunisie
[3] Aalto University, School of Science (Aalto) – Finlande
[4] National Center for Nuclear Sciences and Technology (CNSTN) – Tunisie

Session études d'association
mercredi 29 15h40
Salle des thèses

This work was recently published in [Aridhi et al. (2016)]. The original paper could be found in the following link: http://online.liebertpub.com/doi/10.1089/cmb.2015.0134.

Abstract

Ionizing-radiation-resistant bacteria (IRRB) are important in biotechnol-ogy. In this context, in silico methods of phenotypic prediction and genotype-phenotype relationship discovery are limited. In this work, we analyzed basal DNA repair proteins of most known proteome sequences of IRRB and ionizing-radiation-sensitive bacteria (IRSB) in order to learn a classifier that correctly predicts this bacterial phenotype. We formulated the problem of predicting bacterial ionizing radiation resistance (IRR) as a multiple-instance learning (MIL) problem and we proposed a novel approach for this purpose. We provide a MIL-based prediction system that classify a bacterium to either IRRB or IRSB. The experimental results of the proposed system are satisfactory with 91.5% of successful predictions. Datasets and an implementation of the approach are freely available at http://www.isima.fr/mephu/IRR/.

Introduction

To date, genomic databases indicate the presence of thousands of genome projects. However, limited computational works are available for the prediction of bacterial IRR and consequently the rapid determination of useful microorganisms for several applications (bioremediation of radioactive wastes…). As mentioned in a previous paper [Sghaier et al., (2008)], we consider IRRB as non-spore-forming bacteria that can protect their cytosolic proteins from oxidation and tolerate many DNA double-strand breaks (DSBs) after exposure to high, acute ionizing radiation (IR). Partly, it seems that the shared ability of IRRB to survive the damaging effects of IR is the result of positively selected basal deoxyribonucleic acid (DNA) repair pathways and high intracellular manganese concentration.

In this work, we study basal DNA repair proteins of IRRB and IRSB to develop a bioinformatics approach for the phenotype prediction of IRR. Thus, we consider that each studied bacterium is represented by a set of DNA repair proteins. Due to this fact, we formalize the problem of predicting IRR in bacteria as a MIL problem in which bacteria represent bags and repair proteins

*. Intervenant
†. Corresponding author : manel.zoghlami@gmail.com
‡. Corresponding author : sabeur.aridhi@gmail.com

of each bacterium represent instances. Many MIL algorithms have been developed to solve several problems such as Diverse Density, Citation-kNN and MI-SVM.

Our proposed prediction approach uses a local alignment technique to measure the similarity between protein sequences of the studied bacteria. To the best of our knowledge, this is the first work which proposes an in silico approach for phenotypic prediction of bacterial IRR.

MIL-ALIGN algorithm

The proposed algorithm focuses on discriminating bags by the use of local alignment technique to measure the similarity between each protein sequence in the query bag and corresponding protein sequence in the different bags of the learning database. Informally, the algorithm works as follows:

— For each protein sequence in the query bag, MIL-ALIGN computes the corresponding alignment scores.

— Group alignment scores of all protein sequences of query bacterium into a matrix S. Line i of S corresponds to a score vector of protein pi against all associated proteins.

- Apply an aggregation method to S in order to compute the final prediction result. A query bacterium is predicted as IRRB (respectively IRSB) if the aggregation result of similarity scores of its proteins against associated proteins in the learning database is IRRB (respectively IRSB).

We implemented two aggregation methods to be used with MIL-ALIGN: the Sum of Maximum Scores method and the Weighted Average of Maximum Scores method.

— Sum of Maximum Scores (SMS). For each protein in the query bacterium, we scan the corresponding line of S which contains the obtained scores against all other bacteria of the training database. The SMS method selects the maximum score among the alignments scores against IRRB (which we call maxR) and the maximum score among the scores of alignments against IRSB (which we call maxS). It then compares these scores. If maxR is greater than maxS, it adds maxR to the total score of IRRB (which we call totalR(S)). Otherwise, it adds maxS to the total score of IRSB (which we call totalS(S)). When all selected proteins were processed, the SMS method compares total scores of IRRB and IRSB. If totalR(S) is greater than totalS(S), the prediction output is IRRB. Otherwise, the prediction output is IRSB.

— Weighted Average of Maximum Scores (WAMS). With the WAMS method, each protein pi has a given weight wi. For each protein in the query bacterium, we scan the corresponding line of S which contains the obtained scores against all other bacteria of the training database. The WAMS method selects the maximum score among the scores of alignments against IRRB (which we call maxR(S)) and the maximum score among the scores of alignments against IRSB (which we call maxS(S)). It then compares these scores. If the maxR(S) is greater than maxS(S), it adds maxR(S) multiplied by the weight of the protein to the total score of IRRB and it increments the number of IRRB having a max score. Otherwise, it adds maxS(S) multiplied by the weight of the protein to the total score of IRSB and it increments the number of IRSB having a max score. When all the selected proteins were processed, we compare the average of total scores of IRRB (which we called avgR(S)) and the average of total scores of IRSB (which we called avgS(S)). If avgR(S) is greater than avgS(S), the prediction output is IRRB. Otherwise, the prediction output is IRSB.

Experiments

Data set

Information on complete and ongoing IRRB genome sequencing projects was obtained from the GOLD database. We initiated our analyses by retrieving orthologous proteins implicated

in basal DNA repair in IRRB and IRSB with sequenced genomes. Proteins of the bacterium *Deinococcus radiodurans* were downloaded from the UniProt web site. PrfectBLAST tool was used to identify orthologous proteins. Proteomes of other bacteria were downloaded from the NCBI FTP web site. For our experiments, we constructed a database containing 28 bags (14 IRRB and 14 IRSB). Table 1 presents the used IRRB and IRSB. Each bacterium contains 25 to 31 instances which correspond to proteins implicated in basal DNA repair in IRRB (see Table 2). Datasets and an implementation of the approach are freely available at `http://www.isima.fr/mephu/IRR/`.

Results

In order to simulate traditional setting of machine learning in the context of prediction of IRR in bacteria, we conducted a set of experiments with MIL-ALIGN by selecting just one protein for each bacterium in the learning set. Each experiment consists of aggregating alignment scores between a protein sequence of a query bacterium and the corresponding protein sequences of each bacterium in the learning database. We present in Table 3 learning results with the traditional setting of machine learning. The LOO-based evaluation technique was used to generate the presented results. As shown in Table 3, we conducted 31 experiments (with 31 proteins). Results show that the use of our algorithm with just one instance for each bag in the learning database allow good accuracy values.

In order to study the importance of considering the problem of predicting bacterial IRR as a multiple instance learning problem, we present in Table 5 the experimental results of MIL-ALIGN using a set of proteins to represent the studied bacteria. For each set of proteins and for each aggregation method, we present the accuracy, the sensitivity and the specificity of MIL-ALIGN. We notice that the WAMS aggregation method was used with equally weighted proteins. We used the LOO-based evaluation technique to generate the presented results.

We notice that the use of the whole set of proteins to represent the studied bacteria allows good accuracy accompanied by high values of sensitivity and specificity. This can be explained by the pertinent choice of basal DNA repair proteins to predict the phenotype of IRR. The high values of specificity presented by MIL-ALIGN indicate the ability of this algorithm to identify negative bags (IRSB). Using all proteins, we have 92.8 % of accuracy and specificity. We do not exceed these values in all the cases of mono-instance learning presented in Table 3. As shown in Table 5, the SMS aggregation method allows better results than the WAMS aggregation method using the whole set of proteins to represent the studied bacteria. Using the other subsets of proteins (DNA polymerase, replication complex and other DNA-associated proteins) to represent the bacteria, SMS and WAMS present the same results.

In order to study the correctly classified bacteria with the MIL, we computed for each bacterium in the learning database the percentage of experiments that succeed to classify the bacterium (see Table 4). As shown in Table 4, more than 89 % of tested bacteria show successful predictions of 100 %. This means that we succeed to correctly predict the IRR phenotype of those bacteria. On the other hand, the results illustrated in Table 4 may help to understand some characteristics of the studied bacteria. In particular, the IRRB *M. radiotolerans* and the IRSB *B. abortus* present a high rate of failed predictions. It means that in most cases, *M. radiotolerans* is predicted as IRSB and *B. abortus* is predicted as IRRB; the former is an intracellular parasite [Halling et al. (2005)] and the latter is an endosymbiont of most plant species [Fedorov et al. (2013)]. A probable explanation for these two failed predictions is the increased rate of sequence evolution in endosymbiotic bacteria [Wool and Bromham (2003)]. As our training set is composed mainly of mem-bers of the phylum *Deinococcus–Thermus*; expectedly, the *Deinococcus* bacteria (B2-B7) present a very low rate of failed predictions.

Conclusion

In this paper, we addressed the issue of prediction of bacterial IRR phenotype. We have considered that this problem is a multiple-instance learning problem in which bacteria represent bags and repair proteins of each bacterium represent instances. We have formulated the studied problem and described our proposed algorithm MIL-ALIGN for phenotype prediction in the case of IRRB. By running experiments on a real dataset, we have shown that experimental results of MIL-ALIGN are satisfactory with 91.5 % of successful predictions.

In the future work, we will study the performance of the proposed approach to improve its efficiency, particularly for endosymbiont bacteria. Also, we will study the use of a priori knowledge to improve the efficiency of our algorithm. This a priori knowledge can be used to assign weights to proteins during the learning step of our approach. A notable interest will be dedicated to the study of other proteins that can be involved into the high resistance of IRRB to the IR and desiccation, two positively correlated phenotypes.

References

H. Sghaier, K. Ghedira, A. Benkahla, et al. Basal DNA repair machinery is subject to positive selection in ionizing-radiation-resistant bacteria. *BMC Genomics*, 9:297, 2008.

N.D. Fedorov, G.A. Ekimova, N.V. Doronina, and Y.A. Trotsenko. 1-aminocyclopropane-1-carboxylate (acc) deaminases from *Methylobacterium radiotolerans* and *Methylobacterium nodulans* with higher specificity for acc., *FEMS Microbiol Lett.*, 343(1):70-76, 2013.

S. Aridhi, H. Sghaier, M. Zoghlami, M. Maddouri, and E. Mephu Nguifo. Prediction of ionizing radiation resistance in bacteria using a multiple instance learning model. *Journal of Computational Biology*, 23(1):10-20, 2016.

S.M. Halling, B.D. Peterson-Burch, B.J. Bricker, R.L. Zuerner, Z. Qing, L. Li, V. Kapur, D.P. Alt, and S.C. Olsen. Completion of the genome sequence of *Brucella abortus* and comparison to the highly similar genomes of *Brucella melitensis* and *Brucella suis*. *J Bacteriol*, 187(8):2715-2726, 2005.

M. Woolfit, and L. Bromham. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol. Biol. Evol.*, 20:1545–1555, 2003.

Mots clefs : Bacterial ionizing radiation resistance, multiple instance learning, prediction, classification

Decoding regulatory landscapes in cancer

Stein Aerts [*1],

[1] Department of Human Genetics – KU Leuven Leuven, Belgique

Session biologie des systèmes et réseaux d'interaction
jeudi 30 09h00
Amphi Mérieux

I will discuss several applications of epigenomic profiling in cancer and how these data can be combined with regulatory sequence analysis to decipher gene regulatory networks controlling cellular states in cancer. The first application uses a cancer model in Drosophila where we evaluate different methods for open chromatin profiling and infer functional networks driven by AP-1 and STAT. In the second application we compare different phenotypic states in human melanoma, and show how decoding the regulatory landscape in each state provides novel insight into the gene networks that underlie clinically relevant events in melanoma, such as phenotype switching, invasion and resistance to therapy. In a final case study we explore massively parallel enhancer reporter assays and deep learning methods to understand what distinguishes functional enhancers from other bound genomic regions that have no regulatory activity, using human TP53 as a model transcription factor.

*. Intervenant Invité

Studying microRNAs with a system biology approach : inferring networks, visualizing genome wide data and predicting microRNA functions

Session biologie des systèmes et réseaux d'interaction
jeudi 30 10h00
Amphi Mérieux

Laurent Guyon [*] [†1], Ricky Bhajun [‡1], Amandine Pitaval[1], Éric Sulpice[1], Stéphanie Combe[1], Patricia Obeid[1], Vincent Haguet[1], Christian Lajaunie[2], Xavier Gidrol [§1]

[1] Laboratoire de Biologie à Grande Échelle, Biomics (BGE - UMR 1038) – Université Grenoble Alpes, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble, Inserm – CEA -Grenoble/BIG Laboratoire Biologie à Grande Échelle – 17 rue des Martyrs, F-38 054 Grenoble Cedex 9, France
[2] Centre de Bioinformatique (CBIO) – MINES ParisTech - École nationale supérieure des Mines de Paris, Institut Curie – 35 rue Saint-Honoré, F-77 300 Fontainebleau, France

MicroRNAs (miRs) are roughly twenty-two nucleotide long non-coding RNAs present in all superior eukaryotes and there are presently 2,588 mature miRs identified in human (miRBase v21, [1]). miRs regulate gene expression by degrading messenger RNA (mRNA) or preventing translation of targeted mRNAs. In animals, each microRNA can target many mRNAs due to partial complementarity. Rules have been determined to predict which messenger RNAs can be targeted by a given microRNA. In particular, the "seed sequence", defined as the nucleotides 2-7 from the 5' end of the mature miR, plays an important role and should show near perfect complementarity. Many prediction algorithms have been proposed, such as TargetScan or Diana-microT, to predict miR-mRNA pairs. Each microRNA is predicted to target up to a few thousands genes, and also more than half of protein coding genes are predicted to be regulated by many microRNAs.

We used miR-mRNA pair predictions to infer a microRNA network in which each node is a microRNA, and a link exists between two nodes if they share enough targets. A multi-threshold approach was conducted to define the most informative network. The first microRNA network was built using Diana-microT v3, which comprises 555 microRNAs in human. Using the meet-min metric to define the percentage of shared targets between two microRNAs, and using a threshold of 50 %, we defined a network with a density of 0.02 % and with 4 isolated nodes. The threshold 50 % corresponds to the maximum of the "betweenness centrality" and the minimum of the "clustering coefficient". We further analyzed the topology of the network by isolating the hubs. There are 11 microRNA hubs that are connected in the network to at least 50 other nodes. These hubs form two subnetworks, which we denoted as "assorted clubs". For each club, we performed Gene Ontology (GO) enrichment for the shared predicted targets. The first club, comprising of 8 microRNAs, was predicted to regulate essentially gene expression. The second club, comprising of hsa-miR-612, hsa-miR-661 and hsa-miR-940, was predicted to regulate principally protein coding genes involved in small GTPase signaling. We found that the whole network is clearly organized in three distinct but connected subnetworks: the first one with 315 microRNAs all connected to the assorted club 1, the second one with 129 microRNAs all connected to the assorted club 2, and the third one with 89 microRNAs connected to both clubs. Only 11 miRs are not linked to either club.

*. Intervenant
†. Corresponding author : laurent.guyon@cea.fr
‡. Corresponding author : ricky.bhajun@live.fr
§. Corresponding author : xavier.gidrol@cea.fr

Interestingly, clear GO enrichment appeared for the two first subnetworks, and the GO terms are similar to the corresponding assorted club. As a result, we denoted these two subnetworks as "modules" [2].

The particular structure of the network with two clear parts implicated in two different main biological functions, led us to propose a model of gene expression network organized with a Lavallière-tie architecture, in which we integrate the microRNA network. The Lavallière-tie is a tie knot organized with four wings around a central knot corresponding to the transcriptional/translation machinery. Three wings are composed of molecular objects: genes, proteins and non-coding RNAs including the microRNA network. The fourth wing manages the information flow through signal transduction and epigenetics mechanisms. The two modules of the microRNA network would regulate the transcriptional/translation machinery knot and the signaling wing respectively [3].

The inference of the microRNA networks is based on the prediction of the mRNA targets. As all the dedicated algorithms are known to have both false positive and false negative predictions, we tested the robustness by comparing networks inferred after TargetScan and Diana-MicroT prediction of microRNAs targets. These two algorithms were chosen as they use two different databases to predict the 3' UTR sequences of coding genes: UCSC database for Targetscan and Ensembl for Diana-microT [4,5]. Indeed, the proportion of common targets of the two prediction algorithms forms a bimodal distribution with one peak around 55 % and the other around 80 %. Despite these discrepancies among predictions and the fact that TargetScan v6.2 deals with three times more microRNAs than Diana-microT v3, we showed that graph properties of both networks are very similar. Besides, the network is also organized around the two modules dealing with transcription and signal transduction. The same Gene Ontologies are also enriched for both assorted clubs.

To further investigate our approach, experimental validations were performed for the associated club 2 dealing with signal transduction. We measured phenotypic changes after over-expression of hsa-miR-612, hsa-miR-661 and hsa-miR-940. As expected, all three microRNAs affect cytoskeleton organization, but differently. Ectopic expression of miR-661 led to dense and contracted actin stress fibers, whereas the two other ones led to relaxed and fewer stress fibers as compared to controls. Wound healing experiments showed faster closing time after over-expression of miR-661, and the opposite phenotype for the other miRNAs. Transwell experiments showed similar behavior with more cells able to cross a membrane with 5 micron pores after over-expression of miR-661, and fewer cells after over-expression of miR-640 and miR-612. Finally, we have shown that miR-940 is downregulated in breast cancer, as supported by three different experiments performed by colleagues and reported in the gene expression omnibus (GEO) database.

The network was visualized using Cytoscape [6] and was displayed using the force-directed layout, in which nodes are considered as physical objects with a repulsion force between them, and with a stronger attraction force between connected nodes. A cluster is defined as a subnetwork of nodes largely interconnected to each other and less connected to nodes outside of the cluster. Here, microRNAs in a cluster share many targets and may act in synergy. As a result, it is interesting to visualize clusters of co-expressed microRNAs in a given condition.

Additionally, we have defined another complementary network, in which genomic distance links two microRNAs if they are close enough on the same chromosome. This network helps to quickly visualize clusters of microRNAs on the genome. We also defined 2 networks using the seed sequence as the basis of the metric. The first one links microRNAs having the exact same "seed", that is the exact same nucleotide sequence in positions 2 to 7. The second is defined after an adjacency matrix built using a more complex distance. Here, the distance between two microRNAs with the exact same seed is very small, and a shift or mismatch increases the distance. We are now building a web interface to share the visualization of various microRNA data, including differential expression, on the different networks. The interest of visualizing microRNA data on the different

networks will be demonstrated, using both expression data and high content screening scores.

References

[1] A. Kozomara and S. Griffiths-jones, "miRBase: annotating high confidence microRNAs using deep sequencing data", *Nucleic Acids Res.*, no. November, pp. 68–73, 2013.

[2] R. Bhajun, L. Guyon, A. Pitaval, E. Sulpice, S. Combe, P. Obeid, V. Haguet, I. Ghorbel, C. Lajaunie, and X. Gidrol, "A statistically inferred microRNA network identifies breast cancer target miR-940 as an actin cytoskeleton regulator", *Sci. Rep.*, vol. 5, p. 8336, Jan. 2015.

[3] R. Bhajun, L. Guyon, and X. Gidrol, "MicroRNA degeneracy and pluripotentiality within a Lavalliere tie architecture confers robustness to gene expression networks", *Cell. Mol. Life Sci.*, no. 2, 2016.

[4] W. Ritchie, S. Flamant, and J. E. J. Rasko, "Predicting microRNA targets and functions: traps for the unwary", *Nat. Methods*, vol. 6, no. 6, pp. 397–8, Jun. 2009.

[5] M. Maragkakis, P. Alexiou, G. L. Papadopoulos, M. Reczko, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, V. A. Simossis, P. Sethupathy, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou, "Accurate microRNA target prediction correlates with protein repression levels", *BMC Bioinformatics*, vol. 10, pp. 1–10, 2009.

[6] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks" *Genome Res.*, pp. 2498–2504, 2003.

Mots clefs : microRNA, network inference, network analysis, data visualization

Handling the heterogeneity of genomic and metabolic networks data within flexible workflows with the PADMet toolbox

Session biologie des systèmes et réseaux d'interaction
jeudi 30 10h50
Amphi Mérieux

Marie Chevallier [*] [†1], Meziane Aite[1], Jeanne Got[1], Guillaume Collet[1], Nicolas Loira[2], Maria Paz Cortes[2], Clémence Frioux[1], Julie Laniau[1], Camille Trottier[1], Alejandro Maass[2], Anne Siegel [‡1]

[1] DYLISS (INRIA - IRISA) – INRIA, Université de Rennes 1, CNRS : UMR6074 – Campus de Beaulieu, F-35 042 Rennes Cedex, France
[2] Centre de Modélisation Mathématique / Centro de Modelamiento Matemático (CMM) – Chili

A main challenge of the era of fast and massive genome sequencing is to transform sequences into biological knowledge. The reconstruction of metabolic networks that include all biochemical reactions of a cell is a way to understand physiology interactions from genomic data. In 2010, Thiele and Palsson described a general protocol enabling the reconstruction of high-quality metabolic networks. Since then several approaches have been implemented for this purpose, such as Model Seed (*Henry et al., 2010*), the Cobra ToolBox (*Becker et al., 2007*) and the Raven Toolbox (*Agren et al., 2013*). These methods rely mainly on drafting a first metabolic network from genome annotations and orthology information followed by a gap-filling step. More precisely, in the case of exotic species the lack of good annotations and poor biological information result in incomplete networks. Reference databases of metabolic reactions guide the filling process in order to check whether adding reactions to a network allows compounds of interest to be produced from a given growth media. As a final objective, as soon as the network is considered to be complete enough, functional studies are undergone, often relying on the constraint-based paradigm derived from the Flux Balance Analysis (FBA) framework (*Orth et al., 2010*).

The high diversity of input files and tools required to run any metabolic networks reconstruction protocol represents an important drawback. Genomic data is often required, provided in different formats: either annotated genomes, and/or protein sequences, possibly associated with trained Hidden Markov Models. In addition, most approaches require reference metabolic networks of a template organism. Dictionaries mapping the reference metabolic databases to the gene identifiers corresponding to the studied organism may be required. As a main issue, it appears very difficult to ensure that input files agree among them. Such a heterogeneity produces loss of information during the use of the protocols and generates uncertainty in the final metabolic model. Here we introduce the PADMet-toolbox which allows conciliating genomic and metabolic network information. The toolbox centralizes all this information in a new graph-based format: PADMet (PortAble Database for Metabolism) and provides methods to import, update and export information. For the sake of illustration, the toolbox was used to create a workflow, named AuReMe, aiming to produce high-quality genome-scale metabolic networks and eventually input files to feed most platforms involved in metabolic network analyses. We applied this approach to two exotic organisms and our results evidenced the need of combining approaches and reconciling information to obtain a functional metabolic network to produce biomass.

The main concept underlying the PADMet-toolbox is to provide solutions that ensure the consistency, the internal standardization and the reconciliation of the information used within

*. Intervenant
†. Corresponding author : marie.chevallier@inria.fr
‡. Corresponding author : anne.siegel@irisa.fr

any workflow that combines several tools involving metabolic networks. In other words, all the information is stored in a single light and human-readable database that can be easily updated. The latter is then used as a single cornerstone which spreads the information all along the workflow. To that goal, the genomic and metabolic information about an organism is depicted using an oriented graph wherein nodes are linked by relations. A PADMet file consists in three parts. Firstly, the "Nodes" part stores information about each node *e.g.* reactions, metabolites, pathways, genes… Then, the "Relations" part depicts all relations between nodes *e.g.* consumption and production connections between metabolites and reactions. Finally, the "Policy" part introduces constraints satisfied by both nodes and relations. For instance, the "consumption" connection necessarily involves 'metabolite' nodes and 'reaction' nodes. This format can be viewed as an extension of both the internal format used by the Biocyc database (*Caspi et al., 2014*), and the SBML format (*Hucka et al. 2003)*, with a very precise definition of fields avoiding the possible confusion generated by SBML fields. It is strongly inspired by RDF with additional flexibility and an easier readability, which is an attractive feature for biologists.

The PADMet-toolbox is wrapped as a Python library that can manipulate and operate the PADMet format. It includes several methods to represent, reconstruct from multiple data sources, analyze and compare metabolic networks. One main procedure consists in creating a reference PADMet from one or several external database such as Metacyc. The latter can be updated, corrected, or enriched with additional nodes and external relationships. The metabolic network to be studied is then represented as a subset of the PADMet-reference database by importing manually created lists of reactions or a SBML file whose identifiers are automatically mapped on the reference PADMet. Based on the latter, several networks can be merged into a combined network while preserving the consistency and the traceability of the data. Moreover, the PADMet-toolbox enables the analysis of a metabolic network, *e.g.* reports about the contents in terms of reactions, metabolites, pathways. The toolbox can also perform the exploration and visualization of data: the whole network as a wiki and subnetworks (*i.e.* pathways) as pictures. Finally, PADMet can be exported to SBML to feed any tool using such format.

In order to illustrate, we used the PADMet-toolbox to implement the AuReMe workflow: AUtomatic REconstruction of MEtabolic networks based on genomic and metabolomic information. The workflow performs high quality metabolic network reconstruction based on sequence annotation and orthology as well as metabolomic data. It includes three main components that can be run in parallel or subsequently. The first part of the reconstruction process consists in the extraction of information from genome annotations. It is run by the Pathway-Tools applications (*Karp et al., 2010*) and output files are converted in PADMet format. The second independent part of the workflow consists in the creation of a metabolic network based on orthology between the studied species and a taxonomically-close template species with a curated metabolic network. The chosen tool to perform this orthologue-based network reconstruction was the Pantograph software (*Loira et al., 2012*), itself based on consensus scores from the InParanoid (*Remm et al., 2001*) and OrthoMCL tools (*Li et al., 2003*). The result of Pantograph is a SBML file that is also converted in PADMet format. Both networks are then merged in one unique and unified network to benefit from the internally standardized annotations of the reference database. The last independent part of the workflow consists in gap-filling a metabolic network (either a network resulting of any of the two previous steps, the merged one or a newly imported one). The gap-filling step aims to complete the network with computed predicted reactions to ensure that a set of metabolic compounds can be produced from the growth media of the studied organism. It is implemented in the Meneco tool which uses a topological criteria to assess producibility. As above, the toolbox is able to handle inputs and outputs of Meneco and also manual curation performed by the user, who can remove reactions or add particular ones rather than the whole Meneco output. Therefore, the workflow is flexible enough to be iterated both automatically and manually through expert curation until the result of the reconstruction is considered relevant. The PADMet-toolbox enables to check the properties of the network at any time of the reconstruction and facilitates the transition to other metabolic network analysis platforms such as the Matlab

Cobra and Raven toolboxes. Both the toolbox and the workflow are available as a Docker image to facilitate their distribution among the scientific community.

The AuReMe workflow was applied to two case-studies. We voluntarily selected species which are distantly related to common model organisms *i.e.* species whose genome or transcriptome annotations cannot deserve a very detailed and specific attention as it was the case for the model species. Therefore, despite the efforts provided for their annotation, genomes of these species contain many genes of unknown function which may generate uncertainties in the network reconstruction process. We considered the cases of an extremophile bacteria (*Acidithiobacillus ferrooxidans* strain Wenelen) and of an eukaryota macro-alga (*Ectocarpus siliculosus*). For each case, the reference PADMet database was fed with the Metacyc 18.5 database content. PADMet tools enabled to unify and compare the results from the various AuReMe reconstruction steps: Pantograph orthology-based network, Pathway-Tools annotation-based network, merging of both, with or without a Meneco gap-filling step. An artificial biomass reaction, based on all the metabolites known to be produced by the organism, was created using PADMet tools to quantitatively simulate the functionality of the organism in FBA, which is a way of assessing the quality of reconstructed metabolic networks.

Along the AuReMe workflow analyses, we noticed that annotations and orthology information was insufficient to make a functional metabolic network, as all the targets are not producible with any individual approach nor the merged network. This leaded to non-production of the biomass and needed to be offset by the Meneco gap-filling step. However, the complementarity of both the annotation-based reconstruction and the orthology based one is an interesting feature. For *A. ferrooxidans*, the resulted network from the merging of Pathway-Tools and Pantograph approaches consisted of 1778 reactions. 44% of them were specific to the first network whereas 23% were specific to the second, illustrating this complementarity and supporting the fact that all reconstruction steps are thus important.

The added-value of the combination of tools in the workflow was also noticeable by comparing the number of unproducible metabolic targets remaining at each step. Merging annotation-based and orthology-based networks increased the number of metabolites topologically producible from the media. Indeed, in the case of *E. siliculosus*, the Pantograph network was only able to produce 2 metabolic compounds out of the 50 experimentally evidenced for this species. The Pathway-tools network produced only 5 metabolic targets. The Pantograph and Pathway-tools merged network was able to produce 27 metabolic compounds. After the Meneco gap-filling step, all compounds were topologically producible. Comparing the reactions added by the gap-filling step to the orthology-based, the annotation-based and the merged networks is an additional way to assess the added-value of the workflow and of the combination of both approaches.The non-merged networks need a greater number of added reactions to produce all the metabolic targets than the merged one. Furthermore, Meneco added a relatively small number of reactions compared to the size of the initial network (less than 5%) but they all were necessary to make the network functional towards the metabolic targets. According to our results, the *E. siliculosus* orthology-based network needed 108 reactions to reach the producibility of metabolic targets. To reach the same producibility, the annotation-based network needed less reactions (66). Considering the resulted network obtained by the merging of both networks, only 42 reactions were necessary to enable the topological producibility of all compounds.

After AuReMe reconstruction, both the bacterial and algal networks were functional, *i.e.* able to produce biomass according to the FBA formalism. Interestingly, performing gap-filling of *A. ferrooxidans* after each reconstruction step was powerful enough to produce biomass. However, gap-filling the merged networks rather than the individual ones is more relevant as less putative reactions need to be added to make the target compounds producible. Moreover, the analysis of the flux distribution in each network indicated that a larger subnetwork (more routes) can be used to produce biomass in the merged gap-filled network rather than in the two others independently.

Together, our analyses suggest that the combination of tools in the workflow through a local

flexible database is a way to reconstruct high-quality metabolic networks. Indeed it lowers the number of reactions that are supported neither by genetic annotation nor by orthology evidence. The uncertainty related to the biological relevance of the reconstruction is thus reduced and less expert-performed manual curation is needed. The AuReMe workflow based on the PADmet toolbox allows handling the heterogeneity of metabolic networks data and preserving the consistency of information in a light and efficient way. It can interact with the various existing *in silico* platforms that aim to reconstruct and/or analyze metabolic networks. In the landscape of numerous metabolic network reconstruction and analyses methods, the PADMet toolbox and AuReMe workflow are flexible solutions to conciliate data and facilitate its processing by biologists and bioinformaticians.

Mots clefs : data homogenisation, genome, scale metabolic networks, reconstruction workflow, exotic species

Comparing transcriptomes to probe into the evolution of developmental program reveals an extensive developmental system drift

Session biologie des systèmes et réseaux d'interaction
jeudi 30 11h10
Amphi Mérieux

Coraline Petit [*1], Carine Rey[1,2], Anne Lambert[3], Manon Peltier[3], Sophie Pantalacci [†1], Marie Sémon [‡1]

[1] Laboratoire de Biologie Moléculaire de la Cellule (LBMC) – CNRS : UMR5239, Institut national de la recherche agronomique (INRA) : UR5239, Université Claude Bernard - Lyon I (UCBL), École Normale Supérieure (ENS) - Lyon – ENS de Lyon, 46 allée d'Italie, F-69 364 Lyon Cedex 07, France
[2] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[3] Institut de Génomique Fonctionnelle de Lyon (IGFL) – CNRS : UMR5242, Institut national de la recherche agronomique (INRA) : UA1288, Université Claude Bernard - Lyon I (UCBL), École Normale Supérieure (ENS) - Lyon – École Normale Supérieure de Lyon, 46 allée d'Italie, F-69 364 Lyon Cedex 07, France

Background

Understanding how differences in phenotype arise from modification of developmental programs is a key issue in evolutionary developmental biology. Several examples show that developmental programs can also evolve in the absence of phenotypic change, a phenomenon known as developmental systems drift (True & Haag, 2001; Kiontke *et al.*, 2007; Felix MA, 2012; Wotton *et al.*, 2015). In this context, whether the degree of change in developmental systems correlates with the amount of change in the final morphology is not a trivial question. But how to quantify differences in developmental paths to address this question? Comparative transcriptomics of developing embryos of different species has become a tool of interest to compare their development. Large-scale patterns of inter-specific divergence of development have been extracted through such transcriptomic comparisons, and periods of particular conservation, the so-called "hourglass", have been identified in early development (Domazet-loso and Tautz, 2010; Kalinka *et al.*, 2010; Quint e*t al.*, 2012; Levin *et al.*, 2016). We use this approach to examine how the evolution of development, as probed by the transcriptome, is coupled to the morphological evolution in tooth. The tooth is a long-standing model in evolutionary and evolutionary developmental biology. Molars have complex shapes, which can change extensively over short evolutionary time, or instead stay morphologically stable over long periods, as shown by their extensive fossil record. Their morphological evolution is influenced by diet-mediated selection, and also constrained by the need for corresponding upper and lower teeth to occlude in order to correctly perform their function. Furthermore, lower and upper molars develop into different, complementary shapes using largely the same sets of genes. This induces a strong developmental constraint on their evolution, as only mutations producing occluding shapes should be retained by selection. These strong selective and developmental constraints could, in theory, result in two opposite patterns for the evolution of development. In the first model, "developmental conservation", the strong constraints would limit the capacity for change of developmental paths. Thus developmental path would be well conserved, except in the case of teeth with markedly different morphologies. In this case, we would expect to observe the conservation of gene expression in the tooth germs across species, and a departure from this conservation for teeth with markedly different final morphologies. The

*. Intervenant
†. Corresponding author : sophie.pantalacci@ens-lyon.fr
‡. Corresponding author : marie.semon@ens-lyon.fr

second hypothesis, "developmental system drift", proposes that the development has diverged between the two lineages, despite the similarity of the resultant phenotype. The divergence in the underlying molecular basis of development may be rapid, if in each lineage, different mutations have accumulated, that triggered the selection of different compensatory mutations to maintain the occlusion. In this case, gene expression differences during development should be numerous, and uncorrelated to differences in tooth morphologies.

Our models are the first upper and lower molars of mouse and hamster. Molar crown shape is characterized by the number of cusps present, their size and connections. Cusps, which are hills at the surface of the tooth crown, are patterned sequentially during development. Since mouse and hamster diverged from their common ancestor about 25 MY ago, crown shape has changed in both lineages. However, lower molars in both lineages, and the upper molar of hamster, have kept the ancestral number of cusps (6). Only the upper molar in mouse has experienced a drastic change in morphology, with the addition of 2 supplementary main cusps. Do developmental differences follow morphological differences (the "developmental conservation" hypothesis above)? In this case, upper and lower molars should differ more in mouse than in hamster, while upper molars should be more different between species than lower molars. Alternatively, is developmental drift so high that no link can be made between the degree of developmental difference and the type of molars being compared (the "developmental systems drift" hypothesis above)?

Results and Discussion

Firstly,, we documented the sequence of cusp addition in upper and lower molars of both species. The results are not immediately interpretable in favor of one or the other of the previous hypotheses. As expected, the developmental sequence differs between the two upper molars in mouse and hamster, which have a very different morphology. We also observe specificities of upper versus lower molars, which indicates a certain degree of conservation of a lower versus an upper developmental program. This favors the "developmental conservation" model. But minor changes are also observed between the lower molars, which have the same number of cusps. The later is expected in the "developmental system drift" model.

Next, we compared the transcriptomes of upper and lower molars of mouse and hamster, at 8 different developmental stages spanning the complete sequence of cusp acquisition. These data show common trends in the development of all molars. Multivariate analysis and clustering of temporal profiles shows that there is a common temporal dynamic for all molars (11 % of total variation). Yet, to some extent, upper and lower molars have their specificities since early morphogenesis, in term of both cellular composition and the expression of specific genes. The upper-lower difference is bigger in the mouse than in the hamster (shown by multivariate analysis and number of genes differentially expressed) which is in line with the excess of differences in morphologies. In both species, the difference between upper and lower molars peaks early, at end of the patterning of the second cusp and the third one starts to be patterned. This excess of upper-lower differences has the same duration and the same intensity in both species. However, this peak happens slightly earlier in mouse than in hamster. These observations favor the "developmental conservation" model in expression data. Moreover, we noted that the peak of upper-lower differences is anticipated in mouse. Based on our knowledge of molar formation, we propose that this heterochronic change could be responsible for the patterning of the two supplementary main cusps in mouse upper molar.

However, this is not the main information carried in the transcriptome data. The main axis of variation (54 % of total variance in a multivariate analysis) in the transcriptomic data is accounted for by species effect, such that, in the transcriptomes, the mouse lower molars are more similar to the mouse upper molars, than they are to their hamster counterpart. Such a strong "species effect" is a hallmark of rapid evolution of expression levels. We do not think that this is an experimental bias, because the estimation of expression level was performed with great care, at both the design

and analysis levels. We limited batch effects by preparing the libraries and running the sequencing for both species jointly.

We also designed a dedicated pipeline to compare expression data across different species (one of which, without available genome): orthologous genes were identified by a phylogenetic approach and the expression level was estimated on the parts of the genes that were found in both species.

Having discarded the possibility that this effect is due to experimental bias, an obvious explanation for this finding would be neutral evolution in gene expression (Yanai *et al.*, 2004; Khaitovich *et al.*, 2004). In this case, the levels of expression of most genes, except for key players would be under very weak selection, and free to evolve. But we think this unlikely for two reasons. First, this neutral model has been contradicted by recent studies comparing organ transcriptomes in distant mammalian species, that have shown that the expression levels of most genes are under the constraints of purifying selection (Brawand *et al.*, 2011). Second, the "species effect", with much more difference between than within species, is visible not only for expression levels, but also for gene expression profiles during development. Using clustering analyses, we found that the expression profiles are more similar within species (50 % similarity between upper and lower temporal clusters) than between organs of the same type (25 % similarity between molars of the same type in different species). This finding supports the idea that not merely the expression levels at a given time, but the temporal profiles of gene expression over developmental time are also changing rapidly between species. This massive change in developmental expression profiles suggests that developmental processes are different between the two species, which is in accordance with the "developmental system drift" model.

In conclusion, we have analysed the development path of upper and lower first molar in two rodent species, using both molecular markers, to follow cusp acquisition, and transcriptomic time series, to probe developmental states at the molecular level. The extent of difference in lower versus upper molar development within one species does correlate with the extent of difference present in final morphology. However, the specificity of lower versus. upper molar development, although present, is drowned out by the rapid evolution of development, which is species specific in term of expression levels and profiles. This pattern is best explained by extensive "developmental system drift".

Mots clefs : comparative transcriptomics, evolution of developmental program, developmental systems drift

Single-cell analysis reveals a link between cell-to-cell variability and irreversible commitment during differentiation

Angélique Richard *[1], Loïs Boullu[2,3,4], Ulysse Herbach[1,3,5], Arnaud Bonnaffoux[1,5,6], Valérie Morin[7], Élodie Vallin[1], Anissa Guillemin[1], Nan Papili Gao[8,9], Rudiyanto Gunawan[9,8], Jérémie Cosette[10], Ophélie Arnaud[11], Jean-Jacques Kupiec[12], Thibault Espinasse[5], Sandrine Gonin-Giraud[1], Olivier Gandrillon[1,5]

[1] Laboratoire de Biologie Moléculaire de la Cellule (LBMC) – CNRS : UMR5239, Institut national de la recherche agronomique (INRA) : UR5239, Université Claude Bernard - Lyon I (UCBL), École Normale Supérieure (ENS) - Lyon – ENS de Lyon, 46 allée d'Italie, F-69 364 Lyon Cedex 07, France
[2] DRACULA Grenoble Rhône-Alpes/Institut Camille Jordan – Université Claude Bernard - Lyon I, CNRS : UMR5534, CNRS : UMR5208, INRIA, Institut Camille Jordan – France
[3] Institut Camille Jordan (ICJ) – Institut National des Sciences Appliquées [INSA], École Centrale de Lyon, Université Claude Bernard - Lyon I (UCBL), CNRS : UMR5208, Université Jean Monnet - Saint-Étienne – Bât. Jean Braconnier n°101, 43 boulevard du 11 novembre 1918, F-69 622 Villeurbanne Cedex, France
[4] Département de Mathématiques et de statistiques de l'Université de Montréal – Canada
[5] DRACULA (INRIA Grenoble Rhône-Alpes / Institut Camille jordan) – Université Claude Bernard - Lyon I (UCBL), CNRS : UMR5534, CNRS : UMR5208, INRIA, Institut Camille Jordan – France
[6] The Cosmo compagny – Start-up – France
[7] Centre de génétique et de physiologie moléculaire et cellulaire (CGphiMC) – CNRS : UMR5534, Université Claude Bernard - Lyon I (UCBL) – Université Claude Bernard Lyon 1, Bâtiment Gregor Mendel, 16 rue Raphaël Dubois, F-69 622 Villeurbanne Cedex, France
[8] Swiss Institute of Bioinformatics – Suisse
[9] Institute for Chemical and Bioengineering, ETH Zurich – Suisse
[10] Genethon – Institut National de la Santé et de la Recherche Médicale - INSERM – France
[11] RIKEN - Center for Life Science Technologies (Division of Genomic Technologies) - CLST (DGT) – Japon
[12] Centre Cavaillès – Inserm, École Normale Supérieure de Paris - ENS Paris – France

Session biologie des systèmes et réseaux d'interaction
jeudi 30 11h30
Amphi Mérieux

Models of cell differentiation have been proposed in which cells switch from one differentiation state to another through stochastic dynamics characterized by a peak in gene expression variability at the point of fate commitment (Huang, 2011). The present summarized work aims to test experimentally these assumptions by analysing the differentiation process quantitatively and qualitatively.

Gene expression was measured by high-throughput RT-qPCR at the population and single-cell level using a microfluidic-based device and the *BioMark*™HD System (Fluidigm), that allow analysing the expression of 96 genes in 96 samples at the same time. Gene expression data were generated from an original cell differentiation model constituted of avian erythrocytic progenitors (Gandrillon *et al.*, 1999). To analysed gene expression during the differentiation process, these cells were collected or isolated at several differentiation time-points. These datasets were then explored using a multivariate statistical analysis including several dimensionality reduction algorithms.

Cell-to-cell heterogeneity, hidden behind the averaging effect in populations was revealed at the single-cell level. An entropy value was calculated per time-point to measure such heterogeneity and assess its evolution during the differentiation process. It resulted in a significant increase of the entropy value during the first hours of the differentiation process, peaking at 8-24h, and decreas-

*. Intervenant

ing at 33h until 72h. Interestingly, such increase in cell-cell variability preceded an irreversible commitment to differentiation and an increase in cell size variability.

Using different statistical approaches and tools for analysing these important population and single-cell-based datasets, we obtained strong evidences to support a cell differentiation model in which cells follow timed stochastic dynamics through a phenotype switch towards a stable gene expression pattern.

References

[1] Huang, S. (2011). Systems biology of stem cells : three useful perspectives to help overcome the paradigm of linear pathways. *Philosophical transactions Series A, Mathematical, physical, and engineering sciences* 366:2247-59.

[2] Gandrillon, O., Schmidt, U., Beug, H., and Samarut, J. (1999). TGF-beta cooperates with TGF-alpha to induce the self-renewal of normal erythrocytic progenitors: evidence for an autocrine mechanism. *Embo J* 18:2764-2781.

Mots clefs : Single, cell, Stochasticity, Differentiation

Exploring the dark matter of proteomes using fold signatures.

Session bioinformatique structurale
jeudi 30 13h30
Amphi Mérieux

Isabelle Callebaut[1],

[1] IMPMC, UMR 7590, CNRS, Université Pierre & Marie Curie, Paris, France

A significant part of protein sequences does not match any known domain, as stored in dedicated databases. Different evolutionary hypotheses have been proposed to explain this "dark matter of the protein universe" (Levitt 2009). Lack of domain annotation may result from the difficulty of identifying domain signatures, due to a too high divergence of sequences or from the fact that sequences are novel, restricted to a taxon or a species, and therefore are difficult to annotate due to the lack of abundant, sequenced and close sister species. These two groups of "dark protein sequences" are also called "orphan" sequences. They include globular domains, but also disordered regions, which play key roles in regulatory and signaling processes. Over the last years, many efforts have been made for characterizing this dark matter of proteomes, especially by developing tools for the detection of remote relationships and for the prediction of disordered regions.

Here, I will describe tools we have developed for delineating, in a comprehensive and automated way, foldable regions, where order can be predicted, and identifying fold signatures, which are much more conserved than sequences. These tools are based on physicochemical and structural properties of protein folds and only use the information of a single amino acid sequence, without the prior knowledge of homologs. They were used for quantifying the amount of orphan domains and orphan proteins in whole proteomes and analyze their specific properties, as well as for investigating disordered regions, especially when combined with other disorder predictors. Hence, different flavors of disorder can be distinguished, from fully disordered states to segments able to undergo disorder to order transitions. The developed tools also give insightful information for identifying novel families of domains, starting from orphan sequences, and for characterizing their structures, functions and evolution. These approaches, combined to the consideration of experimental data, will be illustrated with a few examples of proteins associated with diseases and involved in genome dynamics and ion transport.

Mots clefs : Structural bioinformatics, Sequence alignment, Protein structure, Protein structure prediction, Fold recognition

Homology-modeling of complex structural RNAs

Wei Wang [*1], Matthieu Barba[2], Philippe Rinaudo[1], Alain Denise [†1,2], Yann Ponty [‡3]

[1] Laboratoire de Recherche en Informatique (LRI) – CNRS : UMR8623, Université Paris Sud – LRI - Bâtiments 650-660, Université Paris-Sud, F-91 405 ORSAY Cedex, France
[2] Institut de Biologie Intégrative de la Cellule (I2BC) – CNRS : UMR9198, Université Paris Sud Orsay, CEA – Bâtiment 400, Université Paris-Sud, F-91 405 ORSAY Cedex, France
[3] Laboratoire d'informatique de l'école polytechnique [Palaiseau] (LIX) – CNRS : UMR7161, Polytechnique - X – Route de Saclay, F-91 128 PALAISEAU Cedex, France

Session bioinformatique structurale
jeudi 30 15h00
Amphi Mérieux

Abstract

Aligning macromolecules such as proteins, DNAs and RNAs in order to reveal, or conversely exploit, their functional homology is a classic challenge in bioinformatics, with far-reaching applications in structure modelling and genome annotations. In the specific context of complex RNAs, featuring pseudoknots, multiple interactions and non-canonical base pairs, multiple algorithmic solutions and tools have been proposed for the structure/sequence alignment problem. However, such tools are seldom used in practice, due in part to their extreme computational demands, and because of their inability to support general types of structures. Recently, a general parameterized algorithm based on tree decomposition of the query structure has been designed by Rinaudo et al. We present an implementation of the algorithm within a tool named LiCoRNA. We compare it against state-of-the-art algorithms. We show that it both gracefully specializes into a practical algorithm for simple classes pseudoknot, and offers a general solution for complex pseudoknots, which are explicitly out-of-reach of competing softwares.

Introduction

Since Thomas R. Cech discovered that RNA is able to catalyze chemical reaction (Cech 1985), increasing exciting experimental results demonstrated the versatility of RNA and its importance in many cellular processes. Non-coding (nc) RNA has increasingly been shown to be a major player in all cell processes, notably in gene regulation (Zimmerman and Dahlberg 1996) (Sleutels, Zwart, and Barlow 2002)(Willard and Salz 1997)(Eddy 2001). Like proteins, ncRNAs molecules fold into complex three-dimensional structures which are essential to their function. Therefore, one cannot fully understand the biological process without a structurally-aware annotation of ncRNAs.

Briefly, modeling RNA structure relies on two complementary approaches, homology modeling and ab initio modeling. Here we focus on homology modeling, also known as sequence-structure alignment. In recent years, there has been an increasing amount of literatures on RNA sequence-structure alignment for secondary structure prediction including pseudoknots. Matsui et al. (Matsui, Sato, and Sakakibara 2004) proposed pair stochastic tree adjoining grammars (PSTAG) for aligning and predicting RNA structure including Lyngso & Pederson (L&P) pseudoknots (Lyngsø and Pedersen 2000). Han et al. (Han, Dost, and Bafna 2008) contributed an algorithm for Jabbari & Condon (J&C) pseudoknots (Jabbari et al. 2007). However, the resulting algorithm was complex, and practical tools such as PAL (Han, Dost, and Bafna 2008)

*. Intervenant
†. Corresponding author : alain.denise@u-psud.fr
‡. Corresponding author : yann.ponty@lix.polytechnique.fr

only support L&P pseudoknots. Another proposed method is based on profile context-sensitive hidden Markov models (profile-csHMMs) by Yoon et al. (Yoon and Vaidyanathan 2008). Profile-csHMMs have been proven more expressive, and were shown to handle J&C pseudoknot. Here, we developed a fully general method for the sequence-structure comparison, which is able to take as input any type of pseudoknotted structures. In the following we briefly present the algorithm, its implementation and some preliminary results in comparison with other programs.

Material & Methods

Model and algorithmic foundations

Since details of the LiCoRNA sequence-structure alignment algorithm has been published in (Rinaudo et al. 2012), we only briefly describe the algorithm. A query RNA sequence/structure (A) is represented as a general arc-annotated sequence, in which vertices represent nucleotides, and edges represent canonical interactions and backbone adjacencies. Our goal is then to align a query RNA A to a target RNA sequence (B), in a way that minimizes an overall cost function, depending on sequence similarity, base-pair similarity, and structure conservation. More specifically, our objective cost function includes terms for base and interaction substitutions, calculated with RIBOSUM85-60 as described by Klein and Eddy (Klein and Eddy 2003). Gap penalties are computed using two affine cost functions for loops and helices. Since helices are generally more conserved than loop regions, an optimal alignment is less likely to feature gaps in stacked regions. Accordingly, the opening gap penalty within stacked regions is set to twice that of loop regions (200 and 100, respectively). The elongation gap penalty are set to 50 and 20 respectively.

Our alignment algorithm critically relies on the concept of tree decomposition, which we now remind.

Definition 1 (Tree decomposition of an arc-annotated sequence). Given an arc-annotated sequence A=(S, P), a tree decomposition of A is a pair (X,T) where X={X1,…, XN} is a family of subsets of positions {i, i ∈ [1,n]}, n = length(S), and T is a tree whose nodes are the subsets Xr (called bags), satisfying the following properties:

Each position belongs to a bag: $\exists r \in [1,N]$ Xr=[1,n].

Both ends of an interaction are present in a bag: $\forall (i,j) \in P, \exists r \in [1, N], \{i, j\} \in Xr$.

Any two consecutive positions are both present in a bag: $\forall i \in [1, n-1], \exists r \in [1, N], \{i, i+1\} \in Xr$.

For every Xr and Xs, $r, s \in [1, N]$, $Xr \cap Xs \subset Xt$ for all Xt on the shortest path between Xr and Xs.

Figure 1 (in the supplementary file) shows a tree decomposition for a pseudoknot-free and L&P pseudoknot arc-annotated sequence. Once a tree decomposition has been built for the query RNA, a dynamic programming algorithm is used to find the optimal alignment between the query and a given target sequence.

Implementation aspects

Finding the optimal tree width and tree decomposition for a general graph is a NP-hard problem (Bodlaender and al. 1986). Fortunately, efficient heuristic algorithms algorithms have been proposed for computing upper/lower bounds on the treewidth in a constructive fashion (Bodlaender and Koster, 2010 and 2011). We used LibTW, a Java library implementing various tree decomposition algorithms (`http://www.treewidth.com/`). In particular, we used the GREEDYDEGREE and GREEDYFILLIN heuristics (van Dijk, van den Heuvel, and Slob 2006) because of their dominant behavior observed upon previous empirical studies.

To accelerate the dynamic programming algorithm without losing accuracy, we implemented the M constraints as illustrated by Uzilov et al (Uzilov, Keegan, and Mathews 2006). The M parameter reduces the computational cost through restricting the scanned region instead of the whole length in the target sequence for a particular base. To be more precise, suppose L1 and L2 are respectively the total length of the query structure A and target sequence B, and m and n are the nucleotide indices in A and B respectively. The following inequality must be satisfied for allowing m and n to be aligned together:

The default value for M is the difference between the two sequence lengths; and if the difference is less than 6, we set M to be 6.

Benchmark and applications

First, we compared our generic program LiCoRNA with the three available state-of-the-art programs which handle pseudoknots: PSTAG (Matsui, Sato, and Sakakibara 2004), PCSHMM (Yoon and Vaidyanathan 2008) and PAL (Han, Dost, and Bafna 2008). The dataset used in our experiments combines data from the RFAM database (Nawrocki, Burge, and Bateman 2014) and the PseudoBase database (Taufer et al. 2009). RFAM is a collection of RNA families in which all the sequences are aligned and all families are annotated with secondary structures using covariance model (CM) method. However, CM cannot effectively model pseudoknots and therefore, there is no reliable pseudoknot annotation for each family. On the other hand, PseudoBase provides reliable pseudoknot annotations for single sequences. Infernal (Nawrocki and Eddy 2013) was used here to find the most similar RFAM families with default E-value cutoff of 0.0001 for each pseudoknot sequence in PseudoBase database. Pseudoknot annotations were added into the corresponding RFAM families. Overall, 14 families with different kinds of pseudoknot were obtained. For each family in our dataset, we chose in turn each of its members, along with its pseudoknotted consensus, as the query sequence to predict the secondary structure of the other members.

Table 1 (in the supplementary file) reports the comparison of the three state-of-the-art implementations and our software for each RFAM pseudoknotted family in our benchmark set. Average fractional identity (AFI) of pairwise alignment and Sensitivity/specificity analysis have been performed, assuming correctness of the RFAM alignment. The fractional identity represents the alignment identity between the test and reference alignment, that is the number of identities divided by the length of the alignment. This parameter is calculated by the tool CompalignP that is distributed with BRAliBase 2.1 (Wilm, Mainz, and Steger 2006). Good alignment performance is demonstrated by being close to 1. The predicted structure is evaluated by Specificity = TP / (TP + FP) and Sensitivity = TP / (TP + FN), where TP (true positive) represents the number of correctly predicted base pairs, FP (false positive) represents the number of predicted base pairs which are not in the annotated structure, and FN (false negative) represents the number of base pairs in the annotated structure that are not predicted. The fact that the parameter Specificity and Sensitivity are close to 1 indicates good performances.

Table 1 shows that LiCoRNA results are generally equivalent or better than results of its competitors. Notably, the AFI is almost always better for LiCoRNA than for any of the other programs. It must be noted that some of the structures predicted by PAL were corrected, since they contained non Watson-Crick and non Wobble base pairs, even though the reference structures contained only Watson-Crick and Wobble basepairs. Moreover, LiCoRNA is the only program which gives an alignment for the last three structures whose pseudoknots belong to the Rivas & Eddy (R&E) class (Rivas and Eddy 1999), the most complex class of pseudoknots.

Use case : Realignment of PK RFAM families

LiCoRNA can also be used to curate RFAM pseudoknotted alignments. Indeed, RFAM alignments are retrieved and aligned using covariance models, a grammar formalism that discards

crossing interactions. Some pseudoknots are recovered post-facto by an iterative modeling strategy, but some are very likely missed due to the lack of structural awareness of the initial alignment. However, once an experimental structure is known, it can be mapped onto the alignment, and each sequence be realigned in the light of the new structural constraint.

For instance, we took the query sequence X63538.1/1434-1495 (A) and target sequence DQ445912.1/27322-27383 (B) (Figure 3A) in the Corona_pk3 RFAM family (RF00165). Sequence A has a validated L&P pseudoknot in PseudoBase with PKB255. Our realignment reveals some discrepancy with the initial RFAM alignment, but the overall predicted structure for sequence B is conserved. However, in Figure 3B (in the supplementary file), there is quite large discrepancy between RFAM and our realignment. The query and target sequence are AAWR02040610.1/2027-2086 (PseudoBase number PKB258) and ACTA01044722.1/650-709 in RF_site5 family, respectively. We hope that a systematic realignment will allow to reveal or refute an evolutionary pressure towards the preservation of a functional pseudoknot.

Conclusion

In this work, we introduced the LiCoRNA software (aLignment of Complex RNAs), a program that gives a sequence-structure alignment for two RNAs in the presence of arbitrary pseudoknots. The program is based on a tree decomposition of query sequences and a general parameterized algorithm to compute the optimal alignment. Notably, and contrarily to other state-of-the-art programs, LiCoRNA supports any type of pseudoknotted structure as the query. Besides use-cases mentioned in this abstract, interesting applications would include scanning for homologs of a single structured RNA sequence within whole genomes, possibly from unassembled NGS data. Besides, by incorporating the scoring function (Stombaugh et al. 2009) using all canonical and non-canonical base pairs, we could extend our algorithm to 3D structure alignment.

Mots clefs : RNA, structure, sequence alignment, pseudoknots

ThreaDNA : a simple and efficient estimation of DNA mechanical contribution to protein sequence preferences at the genomic scale

Session bioinformatique structurale
jeudi 30 15h20
Amphi Mérieux

Sam Meyer *[1]

[1] INSA Lyon – Institut National des Sciences Appliquées (INSA) - Lyon – France

Many DNA-binding proteins induce substantial mechanical distortions into the double helix : these include architectural proteins such as histones in eukaryotes or nucleoid-associated proteins in prokaryotes, as well as numerous transcriptional regulators. Because DNA's mechanical properties depend on the sequence, the energetic cost of these deformations can represent a significant contribution to the sequence selectivity of DNA binding proteins. In contrast to direct sequence readout mechanisms, where amino-acids directly contact the basepair residues and which is predominant for proteins recognizing highly specific sequences, this indirect readout mode is crucial for proteins that bind DNA in a less specific manner, but still present well-defined sequence preferences required in their biological functions.

Usual bioinformatics descriptions based on nucleotide occurrences (e.g. sequence logos), which are well-adapted for the former recognition mode, can be irrelevant for the latter, where e.g. dinucleotides rather than nucleotides are determinant in the process (or even beyond) [1]. Protein binding affinities can also be affected by modifications of the physical state of the double-helix such as supercoiling, a mechanical process generated in all transcription and replication events [2]. Thus, estimating sequence-dependent deformation energies may prove important in our understanding of the physical and functional organization of the genome.

These energies can be successfully computed from all-atomic molecular dynamics simulations [3], but this requires a considerable computational effort, and specific expertise is required to properly model a protein of interest. For many proteins stiffer than DNA, coarser descriptions were proposed, where the protein is treated implicitly, and deformation energies are estimated with a coarse-grained, nanoscale elastic model of DNA [4]. Each base or base-pair is treated as a rigid body, and sequence-dependent elastic constants between these bodies were parametrized either from experimental approaches [5] or DNA molecular dynamics simulations [1]. Thus, knowing the DNA conformation within the bound complex, it is possible to compute the associated energy with little computational requirement for a given DNA sequence, allowing analyses of entire genomes [6]. This approach was developed and applied mostly for predictions of nucleosome binding, with the complexed conformation being treated with different levels of complexity [6,7,8,9,10]. The software proposed here relies on an improvement of the "threading" method [7,8]: while this conformation was treated as sequence-independent, and taken from a single high-resolution experimental structure, we propose to use a combination of such structures to sample (a limited part of) the conformational landscape. This method allows to keep the extreme simplicity and computational efficiency of the "threading" algorithms, while effectively reproducing many features of more involved algorithms [10]. In particular, it is predictive of nucleosomal high-affinity sequences as well as genome-wide nucleosome positions [11], with a computing time of less than a minute for the yeast genome on a desktop computer.

But importantly, while many studies have dealt with the nucleosome, the approximations used in the algorithm make it possible to apply the program on any DNA-binding protein of interest, for which a high-resolution structure of the complex is available. In the case where sequence-

*. Intervenant

specific contacts are present, the computed energy then represents the mechanical contribution to the binding free energy. These approximations are strictly valid when the base-pair structure of the DNA is not broken, but previous works on HIV integrase [12] suggests that the computed profiles might still be qualitatively valuable for extreme deformations resulting in breaks in the double-helix.

The program was written with a user-friendly graphical user interface running natively on most Linux and MacOS distributions, and is also available as a tool in the Galaxy project; in most cases, computing the deformation energy profile of a new protein along a whole genome requires only a few steps for a non-specialist on an usual computer.

We have analyzed different bacterial transcriptional regulators for which a high-resolution structure is available. The deformation energy associated to the experimentally determined binding sites (as collected in the RegulonDB database) were always significantly lower than average, demonstrating that DNA mechanical energy represents a detectable contribution in their binding mechanism. The extent of this contribution then depends on the particular protein. For the global regulator CRP, sequences of most experimental binding sites are among the most favorable of the whole genome with respect to the strong deformation induced by the protein: the mechanical contribution might therefore the dominant feature of the binding profile. For the nucleoid-associated protein Fis, the same is true to a lesser extent: direct contacts at the (—7,+7) basepairs as well as indirect readout of the central non-contacted region both contribute significantly to the profile.

Finally, the program includes the response of the profile to torsional stress, an essential mechanical feature that gives rise to supercoiling. Although the computed effects cannot be immediately be compared to the levels of supercoiling measured *in vivo*, for CRP, they suggest that this stress shifts the mechanical energies by values comparable to those associated to sequence variations. Thus, inhomogeneous superhelical profiles along the genome would redistribute the protein equally strongly as sequence-dependent inhomogeneities.

References

[1] Pasi M, et al. (2014), *Nucleic Acids Res.* 42:12272-12283

[2] Kouzine F, Gupta A, Baranello L, Wojtowicz D, Ben-Aissa K, et al. (2013) *Nat Struct Mol Biol* 20:396–403.

[3] Paillard G, Lavery R, (2004) *Structure* 12:113-122

[4] Becker N B, Wolff L, Everaers R, (2006) *Nucl Ac Res* 34:5638-49

[5] Olson W K, Gorin A A, Lu X J, Hock L M and Zhurkin V B (1998) *Proc. Natl Acad. Sci. USA* 95:11163–8

[6] Morozov A V, Fortney K, Gaykalova D A, Studitsky V M, Widom J and Siggia E D (2009) *Nucleic Acids Res* 37:4707–4722

[7] Xu F and Olson W K (2010) *J. Biomol. Struct. Dyn.* 27:725–39

[8] Deniz M, Flores O, Battistini F, P erez A, Soler-Lopez M, Orozco M, (2011) *BMC Genomics* 12:489

[9] Fathizadeh A, Besya A B, Ejtehadi M R and Schiessel H (2013) *Eur. Phys. J. E* 36:1–10

[10] Meyer S, Everaers R (2015) *J Phys Cond Mat* 27.

[11] Kaplan N et al. (2009) *Nature* 458:362-366

[12] Naughtin M et al. (2015) *PLoS ONE* 10(6):e0129427

Mots clefs : biophysics, genomics, regulation, computational biology

Towards structural models for the Ebola UTR regions using experimental SHAPE probing data

Afaf Saaidi [*1,2], Delphine Allouche[3], Bruno Sargueil [†3], Yann Ponty [‡1,2]

Session bioinformatique structurale
jeudi 30 15h40
Amphi Mérieux

[1] Laboratoire d'informatique de l'école polytechnique (LIX) – CNRS : UMR7161, Polytechnique - X – Route de Saclay, F-91 128 Palaiseau Cedex, France
[2] AMIB (INRIA Saclay - Île de France) – Université Paris XI - Paris Sud, CNRS : UMR8623, Polytechnique - X, INRIA – Bâtiment Alan Turing, Campus de l'École Polytechnique, 1 rue Honoré d'Estienne d'Orves, F-91 120 Palaiseau, France
[3] Laboratoire de cristallographie et RMN biologiques (LCRB) – CNRS : UMR8015, Université Paris V - Paris Descartes – Faculté de Pharmacie, 4 avenue de l'Observatoire, F-75 270 Paris Cedex 06, France

Abstract

Next-Generation Sequencing (NGS) technologies have opened new perspectives to refine the process of predicting the secondary structure(s) of structured non-coding RNAs. Herein, we describe an integrated modeling strategy, based on the SHAPE chemistry, to infer structural insight fromdeep sequencing data. Our approach is based on a pseudo-energy minimization, incorporating additional information from evolutionary data (compensatory mutations) and SHAPE experiments (reactivity scores) within an iterative procedure. Preliminary results reveal conserved and stable structures within UTRs of the Ebola Genome, that are both thermodynamically-stable and highly supported by SHAPE accessibility analysis. Keywords. RNA, Ebola, secondary structure prediction, SHAPE chemistry,High-throughput sequencing, compensatory mutations.

[An extended version of this abstract, including a figure, is available at https://hal.inria.fr/hal-01332469]

Context

The Ebola virus causes an acute, serious illness which is often fatal if untreated. A promising research direction would be to selectively interfere with the expression of Ebola genes. This requires a mechanical understanding, at a molecular level, of the processes underlying gene regulation. Many such processes revolve around the presence of specific secondary structure elements [7] in untranslated regions [2]. RNA structure prediction can either be performed computationally from thermodynamics principles in the presence of only the sequence [10], or via a comparative approach when an aligned set of homologous sequences is available [1]. Experimentally, such data can be supplemented by chemical probing data, leading to more accurate predictions in the context of RNA structure determination [3]. Our approach is based on a pseudo-energyminimization, incorporating additional information from evolutionary data (compensatory mutations) and SHAPE experiments (reactivity scores) within an iterative procedure. RNA folding can be approached in a less deterministic setting, by considering the outcome of the folding process as a dynamic ensemble of structures. In this vision, a sequence actively fluctuates between many alternative conformations through time, leading to a probability distribution over the set of structures. The challenge is to take advantage from those probabilities to determine the most probable structures that verify the set of experimental constraints and thus help in revealing the Ebola virus replication mechanism.

*. Intervenant
†. Corresponding author : bruno.sargueil@parisdescartes.fr
‡. Corresponding author : yann.ponty@lix.polytechnique.fr

SHAPE probing principles. SHAPE (Selective 2'Hydroxyl Acylation analyzed by Primer Extension) chemistry is a prominent experimental method, when used in combination with structural modeling methods, it can lead to finding reliable structures [9]. This experimental technique is based on the addition of reagents that interactwith the phosphate-sugar backbone on particularly flexible positions, inducing the formation of an adduct. Upon exposure to the reverse-transcriptase (RT), the adduct either causes the mutation of the nucleotide (SHAPE-MAP [9]), or causes the early detachment of the RT (SHAPE-CE [3] , SHAPE-Seq [5]). A quantification of these effects allows to assign a Reactivity score to each position to refer to its ability to interact with the reagent. Unpaired nucleotides are predominantly reactive compared to the paired bases, the latter being constrained by their Hydrogen bonds. Hence, the reactivity of a nucleotide can be used to informab initio structure modeling, or to validate homology-based predictions.

SHAPE MAP reactivities In the context of SHAPE-Map, reactivities are computed by comparing three mutation rates observed in different experimental conditions: presence (Shape)/absence (Control) of SHAPE reagent, and in the absence of structure (Denatured). The reactivity of a position n corresponds to the ratio of mutation rate from SHAPE experiment, after deducting the spontaneous mutations (Control), and the Denatured mutation rate.

Material and methods

Dataset

The Ebola virus genome is 19 kb long, with seven open reading frames in the order 3' NP-VP35-VP40-GP-VP30-VP24-L 5', encoding for structural proteins, envolope glycoprotein, nucleoprotein, non structural proteins and viral polymerase. For each frame, we study consecutively the 5' and 3' non coding regions, this results into 14 sequences ranging from 3'NP to 5'L. The Shape experiments are performed using 1M7 as reagent (weeks et al [8]). We also evaluated the differential SHAPE experiment that uses an additional reagent NMIA, in order to increase the accuracy of RNA structural models. The experiments are followed by a sequencing and mapping process [8]. Each position in the RNA sequence is characterized by its occurrence from the throughput sequencing and thus by a mutation rate. ShapeMapper program(weeks et al [8]), after being adapted to Single End reads input, is subsequently called to calculate reactivity scores.

Modeling tools

We use SHAPE reactivities within an iterative modeling strategy based on a combination of predictive methods: 1. Free-Energy minimization. RNAfold predicts the most thermodynamically stable structure compatible with an RNA sequence [4]. It performs an exact optimization, using dynamic programming, of the free-energy within the Turner energy model (Zuker-Stiegler algorithm [10]). 2. Partition function analysis. RNAfold also allows analyzing the folding landscape at the thermodynamic equilibrium. An efficient dynamic-programming scheme (McCaskill algorithm [6]) produces the base-pairing probability matrix, at the Boltzmann equilibrium. Those probabilities are used to provide a support for predicting base-pairs. Our current method is based on finding the more reliable pairwises according to their probabilities, then use them as structural constraints to recalculate the energy values. 3. Comparative analysis. RNAalifold computes the consensus structure for a set of – previously aligned – homologous sequences. It optimizes a credibility score (Hofacker algorithm [1]), which primarily depends on compensatorymutations.

Main workflow

The present workflow relies on a conservative integrated approach where sets of structures are retrieved from different methods in order to detect similar substructures. Methods 1, 2 are called to build a first set of structures. The method 3 is used to reveal the most credible base pairs according

to the internal scoring scheme of RNAalifold. These base-pairs are then considered as constraints within a new run of RNAfold, thus a second ensemble of structures is built. Substructures that are found in the both datasets are kept and the pairing probabilities that are higher than a certain threshold are introduced as structural constraints for the next run. This process is performed iteratively until no additional bases is predicted. Contrasting with typical methods, which include SHAPE data in the iterative modeling strategy, we chose to keep such a data for a final validation step. In this last step, the compatibility between predicted single/double stranded regions and SHAPE induced accessibilities is asessed. The resulting substructures are eminently supported by the SHAPE reactivities.

[A figure can be found in an extended PDF version at https://hal.inria.fr/hal-01332469]

Preliminary results and discussion

Our preliminary suggest the existence of conserved and stable hairpin loops in the UTR regions of the Ebola genome are structured into hairpin-loop substructures stemming from the exterior region. Moreover, we found little evidence for complex tree-like structures that are the landmark of structural ncRNAs. One such example is the VP24 ncRNA. The method proposed features the most conserved stem-loop substructures. The first stem-loop of the 5'-UTR is highly reminiscent of typical structures found in viral genomes. By analogy, we hypothesize that the role of this substructure is to protect the RNA from being degraded by nucleases. The SHAPE data is generally consistent with our predictions, and indeed labels as accessible most single-stranded regions in predicted hairpins. Conversely, helices are labeled as inaccessible. We plan to further improve our modeling strategy by including Boltzmann sampling and structure-based clustering instead of our current voting mechanisms. We also wish to include additional information, such as SHAPE experimental data using different reagents, possibly including probing on different homologous sequences to establish a SHAPE-informed structural consensus.

Acknowledgment

This work is supported by the FRM Fondation de la RechercheMedicale. The authors would like to thank Steven Busan (Department of Chemistry, University of North Carolina), Alice Heliou (AMIB, INRIA saclay), Vladimir Reinharz (McGill University, Montreal) and Jules Desforges (Faculté de biologie et médecine, Lausanne) for their valuable feedback and support.

Bibliography

[1] S H Bernhart, I L Hofacker, S Will, A R Gruber, and P F Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics, 9:474, 2008.

[2] S M Crary, J S Towner, J EHonig, T R Shoemaker, and S T Nichol. Analysis of the role of predicted RNA secondary structures in Ebola virus replication. Virology, 306(2):210–218, 2003.

[3] Fethullah Karabiber, Jennifer LMcGinnis, Oleg V Favorov, and Kevin M Weeks. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. RNA (New York, N.Y.), 19(1):63–73, 2013.

[4] Ronny Lorenz, Stephan H Bernhart, Christian zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. Algorithms for Molecular Biology, 6(1):26, 2011.

[5] Julius B. Lucks. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. Nucleic acids research, 42(21), 2014.

[6] J S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers, 29(6-7):1105–1119, 1990.

[7] Masfique Mehedi, Thomas Hoenen, Shelly Robertson, Stacy Ricklefs, Michael A. Dolan, Travis Taylor, Darryl Falzarano, Hideki Ebihara, Stephen F. Porcella, and Heinz Feldmann. Ebola Virus RNA Editing Depends on the Primary Editing Site Sequence and an Upstream Secondary Structure. PLoS Pathogens, 9(10), 2013.

[8] Matthew J Smola, Greggory M Rice, Steven Busan, Nathan A Siegfried, and Kevin M Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPEMaP) for direct, versatile and accurate RNA structure analysis. Nature protocols, 10(11):1643–1669, 2015.

[9] Kevin A Wilkinson, Edward J Merino, and Kevin M Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nature protocols, 1(3):1610–6, 2006.

[10] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Research, 9(1):133–148, 1981.

Mots clefs : RNA, Ebola, secondary structure prediction, SHAPE chemistry, High, throughput sequencing, compensatory mutations

Inferring gene regulatory networks from single-cell data : a mechanistic approach

Ulysse Herbach [*] [†1,2,3], Arnaud Bonnaffoux[2,3], Thibault Espinasse[1], Olivier Gandrillon[2,3]

[1] Institut Camille Jordan (ICJ) – Institut National des Sciences Appliquées [INSA], École Centrale de Lyon, Université Claude Bernard - Lyon I (UCBL), CNRS : UMR5208, Université Jean Monnet - Saint-Étienne – Bât. Jean Braconnier n°101, 43 boulevard du 11 novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Laboratoire de Biologie et Modélisation de la Cellule (LBMC) – CNRS : UMR5239, Institut national de la recherche agronomique (INRA) : UR5239, Université Claude Bernard - Lyon I (UCBL), École Normale Supérieure (ENS) - Lyon – ENS de Lyon, 46 allée d'Italie, F-69 364 Lyon Cedex 07, France
[3] Inria team Dracula (Inria Grenoble Rhône-Alpes / Institut Camille Jordan) – CNRS : UMR5534, CNRS : UMR5208, INRIA, Institut Camille Jordan, Université Claude Bernard - Lyon I – Institut Camille Jordan Université Claude Bernard Lyon 1 43 boulevard du 11 novembre 1918, F-69 622 Villeurbanne Cedex France, France

Session données cellules uniques
jeudi 30 14h40
Salle place de l'école

The behaviour of a cell population largely depends on decisions made at the individual cell level, resulting from a complex interaction network between genes within each cell. Recent technologies make it possible to get some traces (mRNA levels) of the activity of this network in individual cells [1], where biological stochasticity is preserved from averaging [2]. Though it is still a challenging task, taking full advantage of this type of data could take gene network inference (the problem of reconstructing the network from the generated traces) to the next level by allowing the use of a mechanistic approach to gene interactions.

Here we focus on the construction of a class of stochastic dynamical models for a set of interacting genes that should be relevant both from a mathematical and from a biological perspective. The starting point is a simple yet rich model of single gene expression, the well known "two-state model", for which one can compute analytically – and infer – the stationary distribution [3]. We can describe the production of mRNA and proteins with a two-state-like model, the paramaters of which will now depend on other genes. This provides a general network model where each link between two genes is directed and has an explicit biochemical interpretation in terms of chemical reactions. From such a viewpoint, inferring a network is equivalent to fitting the model to data: this confers higher interpretability than one would usually have with more abstract statistical approaches. Above all, it allows to go back and forth between biological experiments and simulations.

However, the inference problem has a high-dimensional nature and the computational cost of a naive "brute force" fitting (i.e. performing a great number of simulations with different parameters values) is prohibitive. Hence it is fundamental to get simple analytical expressions for some statistical objects describing the model (e.g. its stationary distribution), at least in an approximate way. As a first step towards such results, we replace the underlying model for each gene with a piecewise deterministic Markov process (PDMP) in which promoter states stay discrete but where the mRNA and protein quantities are continuous [4]. The resulting network model (also a PDMP) is an appealing trade-off between the "perfect" molecular description and the long-established deterministic "rate equations" for gene regulation, which indeed can be seen as an approximation of this model.

We then use a variational approach – closely related to the self-consistent proteomic field theory introduced in [5] – to explicitly compute an approximation of the network stationary

*. Intervenant
†. Corresponding author : ulysse.herbach@ens-lyon.fr

distribution. This approximation, theoretically valid when the interactions between genes are slower than the dynamics of the promoters, turns out to be very robust in practice and thus can be used to derive a promising statistical model for the data, where stochasticity is not just noise but also contains information.

The present work should eventually provide an efficient way to infer gene networks in the order of one hundred genes from single cell transcriptomics data, with oriented, quantified gene-gene interactions, endowed with a molecular interpretation.

References

[1] T. Kouno, M. de Hoon, J. Mar, Y. Tomaru, M. Kawano, P. Carninci, H. Suzuki, Y. Hayashizaki, and J. Shin, *Temporal dynamics and transcriptional control using single-cell gene expression analysis*, Genome Biology, 14 (2013), R118.

[2] V. Shahrezaei and P. S. Swain, *The stochastic nature of biochemical networks*, Curr. Opin. Biotechnol., 19 (2008), pp. 369–374.

[3] J. Peccoud and B. Ycart, *Markovian Modelling of Gene Product Synthesis*, Theoretical Population Biology, 48 (1995), pp. 222–234.

[4] Y. T. Lin and T. Galla, *Bursting noise in gene expression dynamics: linking microscopic and mesoscopic models*. J. R. Soc. Interface, 13 (2016), 20150772.

[5] A. M. Walczak, M. Sasai and P. G. Wolynes, *Self-consistent proteomic field theory of stochastic gene switches*. Biophysical Journal 88 (2005), 828–850.

Mots clefs : stochastic gene expression, biochemical networks, Markov processes

Factorization of count matrices with application to single cell gene expression profile analysis

Ghislain Durif*[1], Franck Picard[1], Sophie Lambert-Lacroix[2]

Session données cellules uniques
jeudi 30 15h00
Salle place de l'école

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG) – CNRS : UMR5525, Université Joseph Fourier - Grenoble I – Domaine de la Merci, F-38 706 La Tronche, France

For nearly 20 years, sequencing technologies have been on the rise, producing more and more data often characterized by their high dimensionality, meaning when the number of covariates like genes is far larger than the number of observations. Analyzing such data is a statistical challenge and requires the use of dimension reduction approaches. Compression methods show particular abilities concerning data interpretation, for instance to expose latent structure such as complex multi-correlation between genes. Especially, projection-based methods generally solve a problem of matrix factorization. For instance, the principal component analysis (PCA) corresponds to a singular value decomposition (SVD) of the data matrix. The number of genes being too high, the interest is to project the data into a lower dimensional space, and construct new and less numerous features that would summarize the information within the data. Such approaches are widely used and suitable for data visualization or clustering. However the nature of genomic data produced by Next Generation Sequencing (NGS) like gene expression profiles is very specific with count matrices. This calls for the development of specific compression methods, that do not supposed the data to be (relatively) Gaussian. The idea is to use model-based approaches, choosing distributions that match the specificity of the data.

Thanks to the last advances in molecular biology, biologists are now able to sequence the genetic material of a single cell, giving stunning insights on cellular diversity and cellular evolution within a tissue. Such data are not just counts, like standard RNA-seq data for instance, but show very particular patterns. Indeed, in single cell sequencing data, a null value in read counts may refer to an absence of read or to a failure in the experiment due to the short amount of genetic material available in a single cell. Data characterized by these dropout events are called zero-inflated, corresponding to an artificial amplification of the zeros. This phenomenon is less likely to occur when sequencing genetic material from a population of cells because a dropout event in a particular cell is likely to be counter-balanced by reads from other cells. The particular zero-inflated counts are a pitfall for standard methods dealing with NGS data. The huge amount of zeros create artificial correlations between genes, and it becomes necessary to infer the proportion of true zeros, to avoid false interpretation.

We focus on compression methods that are suitable for count data in high dimensions. Following this guideline, we developed a matrix factorization approach, suitable for count data and that can be adapted to zero-inflated data. We propose a Gamma-Poisson factor model. In particular, the data matrix X is supposed to depend on latent factors or components (that describe the latent structure). The entries of X are supposed to follow a Poisson distribution, that is appropriate to model counts. Following the principle of generalized PCA, the matrix of Poisson intensities L is factorized into a product of two parameter matrices U and V. Each one respectively quantifies the contributions of the observations and variables to the latent factors. To account for the covariance structure within the data, and especially the possible correlation between covariates (e.g. genes), we use a Bayesian approach. We introduce gamma priors onto the entries of the parameter matrices

*. Intervenant

U and V, so that the Poisson rates depend on linear combination of products of Gamma distribution. This constitutes a more complete and flexible model than for instance Non-Negative Matrix Factorization, based on a Poisson latent factor model without priors, assuming independence between covariates. An additional interest is that the Gamma-Poisson distribution also model over-dispersion, which often characterizes NGS data. Such latent factor model can be extended to zero-inflated case, by using Bernoulli latent variables (taking 0-1 values) indicating whether each zero comes from the Poisson (true zero) or from a dropout event. The distribution modeling the count becomes a mixture between a Dirac in zero and the Poisson distribution.

In this context, the parameter inference is tricky. The likelihood optimization and the EM algorithm are intractable. There does not exist closed-form expressions for the latent variable posteriors. An alternative would be to infer parameters of the model by simulating these posteriors using Markov Chain Monte Carlo (Gibbs sampling or Metropolis-Hastings algorithm), however such approaches are expensive regarding computation time. Instead we use variational inference, a method to approximate the (intractable) parameter posteriors by (tractable) factorizable distributions. This approximation in term of Kullback–Leibler divergence is equivalent to optimizing a lower bound on the marginal likelihood or model evidence. Thanks to the conjugacy in the exponential family, the optimization to infer posterior parameters becomes tractable, through fixed-point algorithms or gradient descent. Such approach appears to be scalable and very efficient computationally, a particular interest when analyzing high dimensional data. In the case of zero-inflated data, the variational inference approach allows to estimate the proportion of dropouts as any other parameter.

Eventually, we asses the performance of our method on simulated and experimental data sets. In particular, we give an illustration of its interest for data visualization. We use expression profiles of single T cells during and after an immune response to a yellow fever vaccine shot, considering thousands of genes and hundreds of cells.

Mots clefs : Compression, Counts, Data visualization, Factor model, Matrix Factorization, NGS, Single cell, Zeroinflated data

Drugs modulating stochastic gene expression affect the erythroid differentiation process

Anissa Guillemin [*1], Sam Meyer[2], Ronan Duchesne[3], Angélique Richard[1,4], Roy Dar[5,6], Fabien Crauste[7,8], Sandrine Giraud[1,4], Olivier Gandrillon[1,9]

Session données cellules uniques
jeudi 30 15h20
Salle place de l'école

[1] Laboratoire de Biologie Moléculaire de la Cellule (LBMC) – Centre National de la Recherche Scientifique - CNRS, Université de Lyon, École Normale Supérieure (ENS) - Lyon, Institut National de la Recherche Agronomique (INRA) – France
[2] INSA Lyon – Institut National des Sciences Appliquées (INSA) - Lyon – France
[3] M2 Complex Systems – École Normale Supérieure (ENS) - Lyon – ENS Lyon 46 allée d'Italie, F-69 007 Lyon, France
[4] UCBL – Université de Lyon – France
[5] Carl R. Woese Institute for Genomic Biology – États-Unis
[6] Center for Biophysics and Quantitative Biology – États-Unis
[7] Institut Camille Jordan (CNRS UMR5298) – CNRS : UMR5208 – 43 boulevard du 11 novembre 1918, F-69 622 Villeurbanne, France
[8] Dracula – INRIA – 56 boulevard Niels Bohr, F-69 603 Villeurbanne, France
[9] CNRS – Centre national de la recherche scientifique - CNRS (France), Université de Lyon, Institut National de Recherche en Informatique et en Automatique (INRIA), INSA - Institut National des Sciences Appliquées – France

It has now been firmly established that isogenic cells, whether prokaryotic or eukaryotic, display an heterogeneous behavior within an homogeneous environment. This stochastic dimension is a probabilistic phenomenon which challenges the deterministic view of the genetic program (Gandrillon *et al.* 2012). Such cell-to-cell differences were shown to be involved in embryonic developmental process (Rue and Martinez Arias 2015), in the heterogeneity of the active immune cells (Duffy *et al.* 2012, Gerlach *et al.* 2013), in the resistance to antibiotics (Balaban 2011) and cancer treatment (Kreso *et al.* 2013) or in decision-making by HIV (Dar *et al.* 2014). The main source of this variability arises in eukaryotic cells from the transcriptional process (Elowitz *et al.* 2002, Ozbudak *et al.* 2002) through stochastic gene expression (SGE).

The ultimate goal of our group is to understand how a metazoan cell takes the decision to differentiate, in relation with SGE. Its role during the differentiation process was first suggested without firm support from experimental evidence (Kupiec 1997, Hoffmann *et al.* 2008). In such a view, the differentiation process can be described by the theory of dynamical systems (Huang 2010), where cells are pictured as particles moving around in a state space. Each cell state is then represented by a vector of gene expression values. When defined gene expression values change for a cell, the cell seen as a particle "moves" in the formal space. In such a view, the self-renewal state would be an attractor state, as a valley, from which cells would have to "escape" in order to enter in a differentiation process (Rebhahn *et al.* 2014). They could do so by increasing SGE, thereby moving along trajectories toward a differentiated state, seen as another attractor state. In this view, SGE could positively participate in taking decision for each cell and could be a determining factor in the commitment and progression in differentiation process. We obtained experimental evidence that is fully compatible with this view (Richard *et al.*, in preparation), but the conclusive demonstration for a causative role for SGE is still lacking.

Recently, studies were conducted on the role of SGE in HIV infection. In human lymphocytes, the influence of a large variety of drugs was measured on a reporter gene expression under the control of HIV LTR promoter (Dar *et al.* 2014). This allowed the authors to identify drugs that were modulating the amount of LTR-based noise in reporter expression.

*. Intervenant

We decided to assess to which extent these drugs were specific for the tat system or whether they could be used for modulating noise in gene expression in a different system. For this we used the 6C2 reporter cell line that are AEV-transformed chicken erythroid progenitor cells. These cells harbor in their genome one copy of the mCherry reporter gene under the control of CMV promoter (Viñuelas *et al.* 2012, Viñuelas *et al.* 2013). The effect of the ten most efficient drugs in the LTR system was assessed on 5 different 6C2 clones, distinguished by different chromosomal locations of the inserted reporter gene. We obtained clear evidence that two of those drugs, Artemisinin and Indomethacin, significantly decreased the variability (measured by the normalized variance of the fluorescence distribution) of the reporter gene expression in these chicken cells. This shows that drugs affecting SGE in one system can also be useful in a very different setting. In order to get a better understanding of the molecular action of the drugs, we fitted a two-state model of gene expression to this data (Viñuelas *et al.* 2013, Arnaud *et al.* 2015). This led to the conclusion that the main effect of the two drugs consisted in increasing burst size, thereby both increasing the mean value as well as decreasing the normalized variance of the fluorescence distribution.

We then asked whether or not such a drug-induced modulation in the amplitude of SGE could affect a differentiation process. To that end, we turned to a cellular system related to the 6C2 cells, but that has the ability to differentiate, the T2EC (Gandrillon *et al.* 1999). Those are primary chicken erythroid progenitor cells that can either be maintained in self-renewal or induced to differentiate, by changing the external medium. We observed that both drugs were able to significantly reduce the differentiation process in T2EC. In order to relate this effect to the modulation of SGE, we performed single cell transcriptomics on T2EC treated with 1μM of Artemisinin for 24h, or left untreated. Using entropy as a measure for the intensity of cell-to-cell variability, we could show a significant homogenization of gene expression levels under Artemisinin treatment. This provides a first evidence that in a physiologically relevant cellular system, the modulation of SGE can result in impairment of the differentiation process.

The question then arises as to what might be the cellular basis for such an effect. Indeed one could envision that a reduction of the differentiation rate would have a similar effect at the cellular level than an increase in the rate of death of mature cells. In order to disentangle those different possibilities, we started to draft a dynamical model of erythroid cell differentiation, based upon ODEs describing the dynamics of cellular compartments and their entry and exit rates. Parameters of this model are being estimated in the control case, before being applied to the case of cells treated with the drugs in the future.

References

Arnaud, O., S. Meyer, E. Vallin, G. Beslon and O. Gandrillon (2015). "Temperature-induced variation in gene expression burst size in metazoan cells." BMC Mol Biol 16: 20.

Balaban, N. Q. (2011). "Persistence: mechanisms for triggering and enhancing phenotypic variability." Curr Opin Genet Dev 21(6): 768-775.

Dar, R. D., N. N. Hosmane, M. R. Arkin, R. F. Siliciano and L. S. Weinberger (2014). "Screening for noise in gene expression identifies drug synergies." Science 344(6190): 1392-1396.

Duffy, K. R., C. J. Wellard, J. F. Markham, J. H. Zhou, R. Holmberg, E. D. Hawkins, J. Hasbold, M. R. Dowling and P. D. Hodgkin (2012). "Activation-induced B cell fates are selected by intracellular stochastic competition." Science 335(6066): 338-341.

Elowitz, M. B., A. J. Levine, E. D. Siggia and P. S. Swain (2002). "Stochastic gene expression in a single cell." Science 297(5584): 1183-1186.

Gandrillon, O., D. Kolesnik-Antoine, J. J. Kupiec and G. Beslon (2012). "Chance at the heart of the cell." Progress in Biophysics & Molecular Biology 110: 1-4.

Gandrillon, O., U. Schmidt, H. Beug and J. Samarut (1999). "TGF-beta cooperates with TGF-alpha to induce the self-renewal of normal erythrocytic progenitors: evidence for an autocrine mechanism." Embo J 18(10): 2764-2781.

Gerlach, C., J. C. Rohr, L. Perie, N. van Rooij, J. W. van Heijst, A. Velds, J. Urbanus, S. H. Naik, H. Jacobs, J. B. Beltman, R. J. de Boer and T. N. Schumacher (2013). "Heterogeneous differentiation patterns of individual CD8+ T cells." Science 340(6132): 635-639.

Hoffmann, M., H. H. Chang, S. Huang, D. E. Ingber, M. Loeffler and J. Galle (2008). "Noise-driven stem cell and progenitor population dynamics." PLoS ONE 3(8): e2922.

Huang, S. (2010). "Cell Lineage Determination in State Space: A Systems View Brings Flexibility to Dogmatic Canonical Rules." PLOS Biol 8(5): e1000380.

Kreso, A., C. A. O'Brien, P. van Galen, O. I. Gan, F. Notta, A. M. Brown, K. Ng, J. Ma, E. Wienholds, C. Dunant, A. Pollett, S. Gallinger, J. McPherson, C. G. Mullighan, D. Shibata and J. E. Dick (2013). "Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer." Science 339(6119): 543-548.

Kupiec, J. J. (1997). "A Darwinian theory for the origin of cellular differentiation." Mol Gen Genet 255(2): 201-208.

Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman and A. van Oudenaarden (2002). "Regulation of noise in the expression of a single gene." Nat Genet 31(1): 69-73.

Rebhahn, J. A., N. Deng, G. Sharma, A. M. Livingstone, S. Huang and T. R. Mosmann (2014). "An animated landscape representation of CD4+ T-cell differentiation, variability, and plasticity: insights into the behavior of populations versus cells." Eur J Immunol 44(8): 2216-2229.

Rue, P. and A. Martinez Arias (2015). "Cell dynamics and gene expression control in tissue homeostasis and development." Mol Syst Biol 11: 792.

Viñuelas, J., G. Kaneko, A. Coulon, G. Beslon and O. Gandrillon (2012). "Toward experimental manipulation of stochasticity in gene expression." Progress in Biophysics & Molecular Biology((in the press)).

Viñuelas, J., G. Kaneko, A. Coulon, E. Vallin, V. Morin, C. Mejia-Pous, J.-J. Kupiec, G. Beslon and O. Gandrillon (2013). "Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts." BMC Biology 11: 15.

Mots clefs : stochastic gene expression, differentiation, erythroid, dynamical system, drugs, model, ODE

Quality control of the transcription by Nonsense-Mediated-mRNA Decay (NMD) revealed by TSS-RNAseq analysis

Christophe Malabat [*1,2], Frank Feuerbach[2], Laurence Ma[3], Cosmin Saveanu[2], Alain Jacquier[2]

[1] Centre de Bioinformatique, biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur de Paris, Centre National de la Recherche Scientifique - CNRS – Hub de bioinformatique & biostatistique Centre de Bioinformatique, biostatistique et Biologie Intégrative Institut Pasteur 25-28 rue du Docteur Roux, F-75 724 Paris Cedex 15, France

[2] Unité de Génétique des Interactions Macromoléculaires (GIM) – Institut Pasteur de Paris, Centre National de la Recherche Scientifique - CNRS – Unité de Génétique des Interactions Macromoléculaires Institut Pasteur 25-28 rue du Docteur Roux, F-75 724 Paris Cedex 15, France

[3] Plate-Forme Génomique – Institut Pasteur de Paris – Plate-Forme Génomique, Institut Pasteur, Paris, France

Session données cellules uniques
jeudi 30 15h40
Salle place de l'école

Nonsense-mediated mRNA decay (NMD) is a translation-dependent RNA quality-control pathway targeting transcripts such as messenger RNAs harbouring premature stop-codons or short upstream open reading frame (uORFs). Our transcription start sites (TSSs) analysis of *Saccharomyces cerevisiae* cells deficient for RNA degradation pathways revealed that about half of the pervasive transcripts are degraded by NMD, which provides a fail-safe mechanism to remove spurious transcripts that escaped degradation in the nucleus. Moreover, we found that the low specificity of RNA polymerase II TSSs selection generates, for 47 % of the expressed genes, NMD-sensitive transcript isoforms carrying uORFs or starting downstream of the ATG START codon. Despite the low abundance of this last category of isoforms, their presence seems to constrain genomic sequences, as suggested by the significant bias against in-frame ATGs specifically found at the beginning of the corresponding genes and reflected by a depletion of methionines in the N-terminus of the encoded proteins.

Mots clefs : RNA quality control, S. cerevisiae, chromosomes, evolutionary biology, genes, genomics, non, coding RNAs, non, sense mediated mRNA decay, nuclear RNA degradation, transcription initiation, RNA, seq, TSS

*. Intervenant

Vidjil, une plateforme pour l'analyse interactive de répertoire immunologique

Marc Duez[1], Mathieu Giraud [* †2,3], Ryan Herbert[2,3], Tatiana Rocher[*2,3], Mikael Salson [* ‡2,3], Florian Thonier[4]

[1] School of Social and Community Medicine, University of Bristol – Royaume-Uni
[2] Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL) – INRIA, Université des Sciences et Technologies de Lille - Lille I, CNRS : UMR9189 – Université Lille 1, Bâtiment M3 extension Avenue Carl Gauss, F-59 655 Villeneuve d'Ascq, France
[3] BONSAI (INRIA Lille - Nord Europe) – CNRS : UMR9189, Université Lille I - Sciences et technologies, INRIA – France
[4] Hôpital Necker - Enfants malades – Université Paris V - Paris Descartes, Assistance publique - Hôpitaux de Paris (AP-HP) – 149, rue de Sèvres, F-75 743 Paris Cedex 15, France

Session bioinformatique pour la santé
jeudi 30 14h40
Salle des thèses

Introduction

La diversité immunologique des lymphocytes B et T provient principalement des recombinaisons V(D)J dans la région appelée CDR3. Ces recombinaisons sont des marqueurs utiles de pathologies : dans la leucémie, ils sont utilisés pour quantifier la maladie résiduelle durant le suivi des patients (Cavé 1998). Le séquençage à haut débit permet de réaliser un séquençage profond de répertoire immunitaire, c'est-à-dire précisément de ces zones CDR3. Connaître le répertoire immunitaire, c'est en particulier identifier et quantifier les *clones* qui proviennent d'une réponse immunitaire ou d'une pathologie.

Les méthodes bioinformatiques dédiées au séquençage de répertoire immunologique (Rep-Seq) doivent prendre en compte la spécificité du mécanisme de recombinaison V(D)J. Depuis plus de 30 ans, plusieurs outils ont été développés par IMGT (`http://www.imgt.org/`). Récemment de nouveaux outils sont apparus et sont capables de traiter plusieurs millions de séquences (Ye 2013, Thomas 2013, Bolotin 2013, Yan 2014, Bolotin 2015, Kuchenbecker 2015). D'autres programmes permettent une analyse plus en profondeur des résutats ainsi qu'une visualisation de la population de lymphocytes (Darzentas 2016, Nazarov 2015, Schaller 2015).

Avec la diffusion des séquenceurs à haut-débit, il y a une demande pour des logiciels disposant d'une interface facile à utiliser sans expertise bioinformatique, adapté à une utilisation quotidienne en routine clinique ou en recherche. Actuellement, le logiciel le plus proche de cette définition est IgGalaxy (Moorhouse 2014) bien qu'il se repose sur des logiciels qui ne sont pas spécifiquement dédiés à l'analyse de millions de séquences. De plus l'utilisateur doit téléverser lui-même les résultats qu'il a obtenu par ailleurs avec IMGT/V-QUEST, ce qui rend IgGalaxy moins pratique.

Nous présentons la plateforme Vidjil (`http://www.vidjil.org/`), utilisable en routine hospitalière ou en laboratoire de recherche. La plateforme est constituée d'algorithmes efficaces et d'une application web flexible reposant sur la visualisation et l'annotation d'un ou plusieurs échantillons couplée à une base de données qui stocke des métadonnées sur les patients et les échantillons. La plateforme permet aussi l'analyse avec des logiciels complémentaires (pour le moment : IMGT/V-QUEST, IgBlast, Blast et MiXCR). À notre connaissance, il s'agit de la première plateforme open-source autonome pour le Rep-Seq, partant des fichiers de séquence bruts

*. Intervenant
†. Corresponding author : `mathieu.giraud@univ-lille1.fr`
‡. Corresponding author : `mikael.salson@univ-lille1.fr`

jusqu'à l'analyse, l'annotation et le stockage des informations. Ouverte au public fin 2014, la plateforme est aujorud'hui utilisée régulièrement par une vingtaine de laboratoires d'hématologie ou d'immunologie en France et dans le monde.

Conception et mise en œuvre

Algorithme à haut-débit

La plateforme Vidjil n'est pas spécifique à un algorithme particulier mais pourrait lancer n'importe quel programme pour le Rep-Seq pour autant qu'il produise un fichier JSON compatible (voir `http://rbx.vidjil.org/browser/doc/format-analysis.html`). Par exemple, elle utilise aussi le logiciel MiXCR. Cependant la plateforme a été initialement conçue pour l'algorithme de Vidjil, présenté à JOBIM 2013 puis publié dans (Giraud, Salson 2014).

L'algorithme de Vidjil, codé en C++, traite des données de séquençage à haut-débit. À l'aide d'une méthode fondée sur les graines, il détecte des recombinaisons V(D)J dans les séquences et les regroupent en clones. L'idée principale est que le regroupement est fait très rapidement et que la désignation V(D)J, plus longue, est réalisée *après* le regroupement sur chacun des clones. Cela rend l'analyse extrêmement rapide parce que, dans la première phase, aucun alignement n'est réalisé. L'ensemble des locus humains codant pour des immunoglobulines et des récepteurs des lymphocytes T sont reconnus, y compris quelques réarrangements incomplets ou irréguliers (Dh/Jh, Intron et KDE en IgK, Vd/Ja, Dd2/Dd3…). Vidjil analyse aussi des répertoires d'autres espèces, comme la souris et le rat.

Application web

La partie client de l'application web est développée en Javascript / d3.js. En entrée, l'application web prend des données analysées par l'algorithme de Vidjil (ou par d'autres pipelines) sous forme d'un fichier Json. Ce fichier contient diverses informations sur les clones principaux, avec en particulier leur abondance et leur désignation V(D)J.

L'application web est composée de différentes vues : une liste de clones, une représentation des clones en grille ou en histogramme, une liste de séquences, et, lorsqu'il y a plusieurs points de suivi, un graphe au cours du temps (voir la figure jointe). Sur la grille, chaque clone est représenté par une bulle. Les axes de la grille, représentant par défaut les gènes V et J, sont configurables pour réaliser différentes statistiques sur la population. Un clic sur un clone n'importe où dans l'application sélectionne le clone dans toutes les vues, en particulier en affichant sa séquence et l'alignant éventuellement contre d'autres séquences.

D'autres logiciels font aussi de la visualisation de résultats d'analyse Rep-Seq, tels que VdjViz ou ARReST/Interrogate. L'originalité de l'application web de Vidjil est d'avoir facilement accès à la fois à des programmes d'analyse efficaces et d'explorer en détail certains clones pour se rapprocher le plus possible de la pratique clinique hématologique.

L'ensemble de l'application web est ainsi pensée pour être en interaction avec l'utilisateur qui peut annoter, étiqueter ou corriger certains clones et les transmettre à d'autres programmes d'analyse : IMGT/V-QUEST (Brochet 2008), IgBlast (Ye 2013) et Blast (Karlin 1990). L'utilisateur peut aussi éditer manuellement la désignation V(D)J pré-calculée.

Une opération particulière est la *fusion* de clones similaires décidée par l'utilisateur qui souhaite regrouper des séquences avec quelques différences provenant d'imprécisions technologiques (PCR, séquençage) ou d'hypermutations somatiques. Pour aider cette décision, l'application web propose un outil d'alignement multiple et une représentation 2D de l'ensemble des clones reposant sur une distance d'alignement (algorithme tSNE, (van der Maaten 2008).) L'application web permet enfin de normaliser des séries de données et de générer des rapports pour les dossiers patients.

Serveur avec base de données patient et expériences

L'application web a aussi une partie serveur (Python, web2py). Après authentification, les utilisateurs créent des fiches correspondant à des patients ou des expériences, téléversent des jeux de reads, et ensuite lancent directement Vidjil ou d'autres programmes d'analyse. Via le serveur, il est possible soit de visualiser les échantillons d'un même patient, soit d'afficher des échantillons de patients différents ou provenant de différents programmes ou paramétrages. Les fichiers de résultats sont privés mais peuvent être partagés avec d'autres utilisateurs ou rendus publics, éventuellement après anonymisation des données personnelles. Les utilisateurs peuvent sauvegarder leurs annotations.

Lorsque plusieurs utilisateurs lancent un programme d'analyse en même temps, le serveur les met dans une file d'attente jusqu'à ce qu'un processeur soit disponible. Vidjil est particulièrement économe en puissance de calcul : le serveur public de test (app.vidjil.org) est une machine peu puissante avec deux Intel(R) Core(TM) i5-2400 CPU et 16Go de RAM. Les temps d'analyse sont compatibles avec un travail quotidien de recherche ou de clinique. La version 2016.03 de l'algorithme traite 1 Gbp en moins de 5 minutes pour des données simples. Les recombinaisons V(D)J de différents types ou incomplètes demandent plusieurs itérations et le traitement peut être jusqu'à 10 fois plus lent. En 2015, 95% des jobs envoyés sur le serveur ont été traités en moins de 10 minutes. La principale limite sur le serveur est l'espace disque pour stocker les fichiers de séquences. Enfin, le serveur contient des outils de maintenance et d'administration (monitoring, de sauvegarde et de notifications).

Développement et intégration continue

Vidjil est un logiciel stable, développé en license libre GPLv3 avec un répertoire git public (plus de 4 000 commits par 9 développeurs les 24 derniers mois).

Dans une démarche d'intégration continue (Jenkins) et de releases régulières, nous avons ajouté systématiquement des tests à Vidjil : plus de 1 400 tests, unitaires et fonctionnels, visent les trois composants, algorithme, application web et serveur. En particulier, pour les tests fonctionnels de l'algorithme, nous avons rassemblé une collection d'une centaine de séquences manuellement annotées pour tester la dénomination V(D)J de la seconde partie de l'algorithme, mais aussi la détermination du type de recombinaison au cours de la première partie. Les séquences sont de difficultés différentes. Ces désignations ont été vérifiées à la main, éventuellement avec d'autres outils bioinformatiques. Elles ont été fournies par Yann Ferret et Aurélie Caillault (CHRU Lille), Gary Wright (GOSH, NHS, Londres) et des développeurs de Vidjil.

Quelques utilisations de Vidjil

Vidjil peut être utilisé localement ou sur notre serveur public de test (app.vidjil.org). Depuis son ouverture en octobre 2014, 50 laboratoires de 11 pays ont soumis des données, pour un total de plus de 6 milliards de séquences dans plus de 3 000 échantillons, avec une moyenne de 1 400 000 reads par échantillon. Une vingtaine de ces laboratoires envoient régulièrement des séquences.

Les données proviennent prinicpalement de séquenceurs Illumina Mi-Seq et Ion Torrent, et sont entrées soit brutes, soit après pré-traitement par des logiciels comme PEAR (Zhang 2014) ou pRESTO (Van der Helden 2014). Les jeux de données contiennent de moins de 0.1% (capture avec de nombreuses sondes, RNA-Seq) à plus de 95% (primers spécifiques) de recombinaisons V(D)J.

1) Plusieurs laboratoires travaillent sur les leucémies aigües lymphoblastiques (LAL), en fort lien avec la clinique. À Lille, depuis début 2015, Vidjil est utilisé en situation de routine pour le diagnostic des patients de LAL (Ferret, Caillault 2016). 125 patients ont été suivis au cours de

l'année 2015. Les laboratoires de Rennes, Montpellier, Bruxelles, Bristol, Prague (Kotrova 2015) et Bergame testent aussi Vidjil dans ce contexte.

2) À Paris, au laboratoire d'hématologie de l'hôpital Necker, plusieurs projets sont menés : étude sur la normalité des CDR3, étude de populations de lymphocytes T chez des patients sains en fonction de leur localisation dans le thymus.

3) Plusieurs utilisateurs mènent des projets de RNA-Seq et se servent de Vidjil pour analyser ou filtrer les recombinaisons immunologiques au milieu d'autres données, tels que l'hôpital de Lyon (S. Huet), l'Institut Gustave Roussy (Villejuif), ou l'université McGill (Montréal, Canada).

4) Vidjil est aussi utilisé pour des données d'autres organismes, comme à Göttingen (Allemagne) dans des études sur le répertoire de la souris et du rat (Linker 2015, Fischer 2016).

Nous avons d'autres utilisateurs réguliers dont nous ne connaissons pas l'objet des recherches. Certains laboratoires se servent directement de Vidjil en ligne de commande. Enfin, un workshop en mars 2016 a réuni 35 utilisateurs et développeurs de Vidjil (`http://www.vidjil.org/workshop-2016.html`).

Perspectives

Le séquençage à haut-débit permet des analyses efficaces et plus complètes de répertoire immunitaire. Vidjil a été conçu pour aider les cliniciens et les chercheurs à analyser leurs données de manière autonome. Aujourd'hui, c'est la seule plateforme d'analyse de Rep-Seq avec une application web autonome, du fichier de séquence jusqu'à l'analyse sauvegardée en lien avec une base de données de patients ou d'expériences.

La plateforme Vidjil est en évolution constante et nous sommes en lien régulier avec nos utilisateurs. Débuté à Lille, le développement de Vidjil est aujourd'hui assuré en partie par les hôpitaux de Bristol et de Paris-Necker.

Nous continuons à améliorer l'algorithme, pour mieux analyser des recombinaisons particulières. Cependant, la stratégie de développement de la plateforme web est de donner accès à plusieurs logiciels Rep-Seq. Nous proposons déjà des liens vers IMGT/V-QUEST, IgBlast et Blast, et permettons de lancer soit l'algorithme Vidjil soit MiXCR. D'autres logiciels seront prochainement intégrés pour encore mieux répondre aux besoins des immunologues et hématologues. Enfin, nous travaillons aussi sur l'installation et l'administration du serveur : deux hôpitaux français sont en train d'installer leur propre instance du serveur.

Nous remercions tous les utilisateurs de Vidjil pour leurs commentaires et leurs retours qui contribuent grandement à l'amélioration de la plateforme.

Pour tout contact : `contact@vidjil.org`

Mots clefs : immunologie, hématologie, répertoire immunologique, lymphocytes, immunoglobulines, séquençage haut, débit, RepSeq

Identification des régions régulatrices de l'intégrine beta-8 grâce à l'ATAC-Seq

Marie-Laure Endale Ahanda [*] [†1], Mathilde Boucard-Jourdin[1], Sébastien This[1], Helena Païdassi [‡1]

[1] Centre International de Recherche en Infectiologie (CIRI) – École Normale Supérieure (ENS) - Lyon, Université Claude Bernard - Lyon I (UCBL), CNRS : UMR5308, Inserm : U1111 – 21 avenue Tony Garnier, F-69 365 Lyon Cedex 07, France

Session bioinformatique pour la santé
jeudi 30 15h00
Salle des thèses

Les maladies inflammatoires chroniques de l'intestin (MICI), telles que la maladie de Crohn et la recto-colite hémorragique, représentent plus de 200 000 malades en France. Leur prévalence est élevée dans les pays dits industrialisés et semble s'y stabiliser alors qu'elle s'accroît fortement dans les états en voie d'industrialisation. Ces pathologies invalidantes se caractérisent par une inflammation chronique du tube digestif entraînant une forte altération de la qualité de vie des malades. Cette inflammation est associée à une dérégulation de la réponse immunitaire dirigée contre la flore intestinale commensale induisant la rupture de l'équilibre entre réponses pro et anti-inflammatoires.

Dans un tissu sain, cet équilibre est maintenu, entre autres, par la régulation de la réponse immunitaire par des cellules T dites régulatrices (Tregs). Ces dernières sont induites en périphérie par les cellules dendritiques sous la dépendance de la cytokine immunorégulatrice TGF-beta (Transforming Growth Factor-beta). Or, cette cytokine est produite sous une forme inactive. Cependant, les cellules dendritiques sont capables d'activer le TGF-beta par l'intermédiaire de l'intégrine alpha-v-beta-8, ce qui joue un rôle clé pour le maintien de l'homéostasie immunitaire intestinale. Pour preuve, chez les souris, l'absence de cette intégrine à la surface des cellules dendritiques entraîne le développement de colites sévères spontanées associé à une perte des Tregs intestinales.

Or l'expression de beta-8 est très variable d'une sous-population de cellules dendritiques à l'autre. Dans l'intestin, on peut distinguer à l'aide des marqueurs CD103 et CD11b, trois sous-populations de cellules dendritiques, les CD103+ CD11b- dont il a été montré qu'elles exprimaient préférentiellement beta-8, les CD103+ CD11b+ et les CD103- CD11b+ qui toutes deux n'expriment que de faibles niveaux de l'intégrine à l'état basal. Dans la rate, les cellules dendritiques exprimant le marqueur CD8-alpha ont le même lignage que la sous-population intestinale CD103+ CD11b- et ne sont capables d'exprimer beta-8 que lorsqu'elles sont stimulées avec des facteurs mimant l'environnement muqueux de l'intestin tels que des agonistes des Toll-like receptors (TLRs) ou l'acide rétinoïque. En revanche, dans les cellules dendritiques spléniques CD11b+ qui partagent la même origine que les CD103+ CD11b+ intestinales, l'expression de beta-8 est infime et n'est pas induite par ces facteurs. Ces observations mettent en évidence une régulation fine de l'expression de cette intégrine sous l'influence du lignage et de facteurs environnementaux. Néanmoins, les mécanismes moléculaires de cette régulation restent à découvrir.

Nous avons donc choisi d'établir le profil épigénétique des différentes sous-populations de cellules dendritiques intestinales et spléniques, dans l'optique d'identifier des éléments régulateurs associés à une forte expression de l'intégrine et/ou définissant la potentialité d'exprimer l'intégrine à haut niveau.

*. Intervenant
†. Corresponding author : marie-laure.endale-ahanda@inserm.fr
‡. Corresponding author : helena.paidassi@inserm.fr

Nous avons donc réalisé et séquencé des librairies d'ATAC-Seq (Assay for Transposase-Accessible Chromatin-Sequencing) pour chaque sous-population de cellules dendritiques. Cette technologie, permet d'identifier les régions régulatrices du génome fixées par des facteurs de transcription et ainsi accessibles à la transposase. Huit librairies spléniques et douze librairies issues des cellules dendritiques intestinales ont été générées (quatre réplicats pour chaque sous-population). Après un contrôle qualité, le trimming des adaptateurs et des reads en fonction de leur qualité et la suppression des séquences redondantes, les séquences ont été alignées sur le génome murin (version mm9) et le peak calling a été effectué avec le logiciel MACS2.

Dans un premier temps nous avons observé que les pics d'ATAC-Seq sont corrélées avec les marques régulatrices activées (enhancers actifs). En effet, en utilisant des données publiques établissant les profils de mono-méthylation de la lysine 4 et de l'acétylation de la lysine 27 de l'histone H3 (H3K4me1 et H3K27ac) obtenus par Chromatin Immunoprecipitation Sequencing (ChIP-Seq) de cellules dendritiques dérivées de moelle osseuse, nous avons défini le profil des enhancers de différentes sous-population de cellules dendritiques. Un r^2 de 0,5 est observé entre les profils d'ATAC-Seq et les marques H3K4me1 caractéristiques des enhancers, tandis qu'un r^2 d'environ 0,65 est observé entre les profils d'ATAC-Seq et les marques H3K27ac (régions régulatrice actives). Enfin, 90 % des enhancers actifs sont co-localisés avec au moins un pic d'ATAC-Seq. La technologie de l'ATAC-Seq semble donc adaptée pour identifier les régions régulatrices actives dans notre système.

Nous avons ensuite effectué différentes analyses différentielles à l'aide du package edgeR. La comparaison des cellules dendritiques CD8-alpha+ versus CD11b+ de la rate a permis de mettre en évidence 17 120 pics préférentiellement identifiés dans les CD8-alpha+ tandis que 16 986 sont identifiés dans les CD11b+. Les motifs correspondant aux sites de liaisons des facteurs de transcription PU.1 et IRF sont préférentiellement identifiés chez les CD8-alpha+. La sur-représentation d'un demi-site Retinoic Acid Response Element (RARE) est également retrouvée dans 44,53 % des séquences. En revanche, la séquence consensus du site de liaison du facteur de transcription RUNX est sur-représenté dans les séquences spécifiques des CD11b+. Le fait que ces facteurs de transcription aient été impliqués précédemment dans l'établissement ou le maintien du lignage de ces cellules dendritiques, valide la stratégie que nous avons choisie pour identifier les facteurs déterminants de la régulation de l'intégrine beta-8.

Concernant, le gène codant pour l'intégrine beta-8 (Itgb8), nous avons identifié à -74 kb de son promoteur une marque spécifique des CD8-alpha+. De plus, cette région pourrait contenir un demi-site RARE pouvant lier les récepteurs à l'acide rétinoïque (RAR). Cette région est également accessible préférentiellement dans la sous-population intestinale CD103+ CD11b- par rapport aux cellules CD103+ CD11b+, suggérant ainsi qu'elle dépend du lignage de ces cellules.

En se concentrant sur la région autour d'Itgb8, nous avons identifié quatre marques supplémentaires, préférentiellement identifiées dans la sous-population CD103+ CD11b-. L'une de ces marques contient un site RARE et une autre un site de fixation pour IRF8. La région du site d'initiation de la transcription est accessible préférentiellement dans les CD103+ CD11b- et les CD103- CD11b+ et deux sites de liaisons de NF-κB y ont été identifiés.

L'analyse différentielle entre d'une part les sous-populations intestinales CD103+ CD11b- et CD103+ CD11b+, d'autre part les sous-populations spléniques CD8-alpha+ et CD11b+, révèle plus de 15 000 pics spécifiques du microenvironnement intestinal quand moins de 11000 pics sont identifiés préférentiellement dans la rate. La recherche de motifs sur-représentés, restreinte aux pics dont la densité de reads fait partie des 25 % plus élevés, identifie le site de liaison de la sous-unité p65 du facteur de transcription NF-κB (nuclear factor-kappa B) comme étant sur-représenté dans les cellules intestinales. Ceci se vérifie particulièrement dans les régions distales (enhancers) dans lesquelles plus de 57 % des séquences ont un site NF-κB-p65 alors qu'au niveau des régions promotrices, l'enrichissement est plus faible ($\approx$ 50 %) mais reste significatif. Les régions distales montrent également un enrichissement pour les sites de liaison des facteurs de transcription

BATF3 (Basic Leucine Zipper Transcription Factor, ATF-Like 3) et IRF (Interferon Regulatory Factor), mais également pour un motif partiel de type RARE. Les régions promotrices montrent un enrichissement pour les motifs associés aux interferons, à PU.1, BATF3 et STAT.

Ainsi, nous avons construit des librairies d'ATAC-Seq pour établir le profil épigénétique de deux sous-populations de cellules dendritiques de rate et de trois sous-population de cellules dendritiques intestinales. Ces données nous ont permis d'identifier de potentielles régions régulatrices dans ces cellules et nous a notamment permis d'identifier des régions qui pourraient être impliquées dans la régulation de l'intégrine beta-8. En effet, les conditions que nous avons choisies nous permettent de distinguer les régions dépendant du lignage cellulaire et celles établies en réponse à des stimuli. Nous avons notamment identifié une marque à -74 kb du promoteur d'Itgb8 accessible préférentiellement dans les sous-population CD8-alpha et CD103+ CD11b-. On peut toutefois noter que cette région est également accessible à un niveau moindre dans les CD103- CD11b+. Nous avons également identifié des régions accessibles préférentiellement dans les cellules dendritiques intestinales par rapport aux cellules spléniques, suggérant leur induction par des stimuli environnementaux. Nous avons prédit que les régions régulatrices dans le locus d'Itgb8 sont capables de fixer le récepteur à l'acide rétinoïque et cette observation est en accord avec l'induction de l'expression de beta-8 par ce métabolite du rétinol. La présence de sites capables de lier la sous-unité p65 de la protéine NF-κB et accessibles à la transposase préférentiellement dans les cellules intestinales, réaffirme l'importance de la signalisation des TLRs dans la régulation de beta-8. En effet, il a été montré que le facteur de transcription NF-κB est un des médiateurs de la voie de signalisation des TLRs, notamment dans les cellules dendritiques.

Finalement, nous avons montré qu'au niveau du locus d'Itgb8, on trouve des régions régulatrices qui semblent être activées en fonction du lignage cellulaire tandis que d'autres semblent l'être en réponse à des stimuli. Il semblerait que l'acide rétinoïque et le facteur de transcription NF-κB puissent directement se fixer sur des enhancers à proximité d'Itgb8 et contribueraient ainsi à réguler son expression. Ces données ouvrent de nouvelles pistes d'investigation pour identifier des moyens d'agir sur la régulation de cette intégrine pour le rétablissement de l'homéostasie immunitaire chez des patients atteints de MICI.

En conclusion, nous avons montré que l'ATAC-Seq est une méthode puissante pour identifier les régions régulatrices actives dans des populations de cellules présentes en faible proportion dans les tissus. Cette technique d'ATAC-seq ayant l'avantage d'avoir une meilleure résolution que la méthode de ChIP-Seq pour une quantité de matériel près de vingt fois moindre, son utilisation est promise à un bel avenir pour établir et étudier le profil épigénétique de cellules rares ou difficiles à isoler.

Mots clefs : ATAC, Seq, Epigénétique, Immunité intestinale, Cellules dendritiques

Focused handprint of asthma using data from the U-biopred project

Romain Tching Chi Yen [*1], Bertrand De Meulder[1], Diane Lefaudeux[1], Jeanette Bigler[2], Craig Wheelock[3], Ratko Djukanovic[4], Frédéric Baribaud[5], Charles Auffray [†1], U-Biopred Study Group [6]

[1] European Institute for Systems Biology and Medicine (EISBM) – Université Claude Bernard-Lyon I - UCBL – Lyon, France
[2] Amgen – Seattle, États-Unis
[3] Bioanalytical Chemistry Research Laboratory in Inflammatory Metabolomics – Karoinska Institutet, Stockholm, Suède
[4] National Institute for Health Research Respiratory Biomedical Research Unit – University of Southampton, Southampton, Royaume-Uni
[5] Systems pharmacology and biomarkers – Janssen R&D, Springhouse, États-Unis
[6] Pulmonology – Academisch Medisch Centrum, Amsterdam, Pays-Bas

Session bioinformatique pour la santé
jeudi 30 15h20
Salle des thèses

Severe asthma and the U-biopred project

Severe asthma

Asthma is one of the most common, yet well-controlled, chronic diseases in the world, with an estimated 235 million people (according to WHO statistics in 2013 [1]) from all ages and conditions suffering from it worldwide. However, about 5 to 10 % of the asthmatic population have what is known as difficult-to-treat severe asthma [2]. For these patients, the existing treatments are either inefficient, or very high doses are needed to control symptoms and at the cost of deleterious side effects. Furthermore, frequent exacerbations which often lead to hospitalization and the high level of treatment represent a large part of the total asthma-related healthcare expenses.

Moreover, is has been recently highlighted that what has been diagnosed as severe asthma is most likely a collection of phenotypes with different pathobiological mechanisms under the umbrella name of "severe asthma". Independent studies [3,4] have indeed managed to distinguish different subphenotypes among patients with asthma. Those results open the path for further investigation and could explain why some patients are unresponsive to existing treatments, and lead to the development of new personalised treatments.

The U-biopred project

The Unbiased BIOmarkers for the PREDiction of respiratory disease outcomes project (U-biopred [5]) is a 6-year (2009-2015) European-wide research project using information and samples from adults and children to learn more about different types of asthma and ensure better diagnosis and treatment for each patient. It brought together scientists from 41 partners from universities, research institutes, pharmaceutical industries and small companies. Patients with asthma and patient representative organisations were involved, providing the patients' perspective in the development of the project. The main objectives of the program were to:

- improve understanding of severe asthma phenotypes;

- determine how it differs from patient to patient;

*. Intervenant
†. Corresponding author : cauffray@eisbm.org

- uncover new information and generate new hypotheses that could lead to the creation of effective new treatments for patients that do not respond to the currently available ones.

Objectives

One of the objectives of the U-biopred project was to identify sub-phenotypes of severe asthma through the analysis of several high-throughput data from different omics platforms and sample types [6]. Although some results have already been produced [7], they have yet to be confirmed and thoroughly exploited, which is the aim of our work.

By identifying such "handprints", that is to say biomarker signatures derived from the combination of clinical data and high-dimensional biomarker data collected within multiple technical platforms (the signature derived from a single platform representing a "fingerprint"), we aim at reaching a deeper understanding of the disease. That may not only reveal new targets from treatment-resistant patients, but also allow to find out minimal (and potentially non-invasive) measurements needed to determine the sub-phenotype of a given patient. This would bring about an easier diagnosis method for the clinicians, and greatly contribute to the quickly developing field of personalised medicine.

Data

From the wealth of high quality data produced within the U-biopred project, four omics platforms were chosen as our focus, as the number of patients for which those omics data were available was the largest. These platforms cover different aspects of the biology:

- gene expression (Affymetrix® HT HGU-133+PMonly - 54,175 probesets measured);

- serum protein abundance (focused: SomaLogic® SOMAscan assay - 1,129 targets; nonfocused: serum UPLC MS/MS - 147 identified proteins in more than 40% of patients);

- lipid abundance (focused: urine eicosanoids panel - 11 eicosanoids).

Under the U-biopred project, measures were taken for hundreds of patients, among which there were 227 adult patients who had the four types of omics measured in blood-related samples. The "handprint" analysis was consequently conducted on those 227 patients.

Added to this, over 700 clinical variables (age, allergies, well-being, medicine consumption, among others) were obtained for each patient, thus providing a large amount of information on the patients' conditions on all scales [8].

Methods

In order to integrate, compare and combine heterogeneous data from different platforms, the workflow of the analysis is constituted of several important steps.

Data retrieval

The clinical and omics data produced in the project has been warehoused in a tranSMART database instance. Quality-controlled and batch effect corrected data matrices from all platforms were recovered either directly from the database interface or from the data exchange platform put in place for large file transfers.

Feature reduction

In this study, and particularly for the gene expression data, the large number of measurements makes analysis heavily resource-consuming even though some measurements may be irrelevant to

the analysis. As a consequence, a first step for the analysis consisted in reducing the size of the gene expression matrix as well as of the focused serum protein abundance matrix.

First, in the gene expression matrix, the internal control and non-annotated probesets were removed. The mean and standard deviation distributions for the remaining probesets were also computed and all probesets for which both mean and standard deviation were below the first quartile of their respective distribution were also removed. For computational capabilities issues, the probesets that had the smallest mean were also removed in order to limit the total number of analysed probesets to 30,000.

The second step to reduce the analysed features was the use of correlation networks, which are increasingly being used in bioinformatics applications. For example, Weighted Gene Co-expression Network Analysis (WGCNA, [9]) is a systems biology method for describing the genes correlation pattern across microarray samples and other functional genomics experiments. WGCNA can be used for finding clusters (modules) of highly correlated genes; for summarizing such clusters using the module eigengene or an intramodular hub gene; for relating modules to one another and to external sample traits (using eigengene network methodology) for further investigation; and for calculating module membership measures. Correlation networks facilitate network-based gene screening methods that can be used to identify candidate biomarkers or therapeutic targets. These methods have been successfully applied in various biological contexts, *e.g.* cancer, mouse genetics, yeast genetics, and analysis of brain imaging data.

We used this method implemented in the WGCNA R software package in order to identify in the gene expression data as well as in the focused serum protein abundance data which probesets (respectively proteins) were most related to the clinical traits of interest and which were not. As a consequence we could focus on the ones that held most of the information and reduce the size of the analysed data down to about 15,000 probesets for the Affymetrix® data and around 600 proteins for the SomaLogic® data. This reduced the noise from the data and improved both the speed of the subsequent computational processes and the interpretation made from the results of the whole analysis.

Omics data fusion

Recent technologies have made it cost-effective to collect diverse types of genome-wide data. Computational methods are needed to combine these data to create a comprehensive view of a given disease or biological process. Similarity Network Fusion (SNF, [10]) adresses this problem by constructing networks of samples (*e.g.*, patients) for each available data type and then efficiently fusing these into a single network that represents the full spectrum of underlying data. For example, to create a comprehensive view of a disease given a cohort of patients, SNF computes patient similarity networks obtained from each of their data types separately and then fuses them, taking advantage of the complementarity in the data. This is exactly what is needed for the integration of the data from the different omics platforms and has proven rather efficient in the mentioned preliminary studies.

Clustering

The final step in the discovery of sub-phenotypes for severe asthma is to identify common patterns from the combined network of the patients to form clusters. It is also necessary to assess the stability of such clusters to guarantee as much reliability for the results as possible.

In conjunction with resampling techniques, consensus clustering [11] provides a method to represent the consensus across multiple runs of a clustering algorithm and to assess the stability of the resulting clusters. The method can also be used to represent the consensus over multiple runs of a clustering algorithm with random restart (such as K-means, model-based Bayesian clustering, SOM, etc.), so as to account for its sensitivity to the initial conditions. Finally, it provides a visualisation tool to inspect cluster number, membership, and boundaries.

Identification of biomarkers

To determine biomarkers for each severe asthma sub-phenotype, a statistical comparison between clusters has been run in order to identify different molecular patterns. Then an enrichment analysis with g:profiler [12] was used to detect biological processes that were different between clusters of patients. Mapping differential features from the results on the asthma disease map developed in our laboratory, first in the U-BIOPRED project and now in the eTRIKS project, to compare against the state of the art knowledge allowed for deeper understanding of the sub-phenotypes. Finally, machine learning will be used to identify necessary and sufficient molecular signatures to characterise the clusters.

Comparison with existing handprints

During the U-biopred project, an unbiased (*i.e.* without feature reduction) handprint analysis on the same datasets was conducted [7] and results of this focused handprint analysis will be compared to it. Similarities between the two approaches will help improving confidence in the existence of sub-phenotypes of asthma in the U-biopred dataset.

Conclusion

We present a systems biology data analysis conducted on a selected dataset from the U-biopred project, using state-of-the-art methods and algorithms. The results of this analysis, which are currently in the process of being validated and interpreted, will help to identify the sub-phenotypes of severe asthma and allow development of new treatments for these severe patients.

References

[1] Asthma Page (`http://www.who.int/mediacentre/factsheets/fs307/en/`). World Health Organisation, April 2016.

[2] R. S. Irwin, F. J. Curley, and C. J. French. Difficult-to-control asthma. Contributing factors and outcome of a systematic management protocol. *Chest*, 103(6):1662–1669, June 2003.

[3] P. Haldar, I. D. Pavord, D. E. Shaw et al. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med*, 178(3):218–24, August 2008.

[4] W. C. Moore, D. A. Meyers, S. E. Wenzel et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med*, 181:315–323, February 2010.

[5] Home Page (`http://www.europeanlung.org/en/projects-and-research/projects/u-biopred/home`). *UBIOPRED*, April 2016.

[6] C. E. Wheelock, V. M. Goss, D. Balgoma et al. Application of 'omics technologies to biomarker discovery in inflammatory lung diseases. *Eur Respir J*, 42:802–825, September 2013.

[7] B. De Meulder, D. Lefaudeux, J. Bigler et al. The first U-BIOPRED blood handprint of severe asthma. *European Respiratory Journal*, 46(suppl 59):PA4889, September 2015.

[8] D. E. Shaw, A. R. Sousa, S. J. Fowler et al. Clinical and inflammatory characteristics of the European U-BIOPRED adult severe asthma cohort. *Eur Respir J*, September 2015.

[9] P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, December 2008.

[10] B. Wang, A. M. Mezlini, F. Demir et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*, 11:333–7, March 2014.

[11] S. Monti, P. Tamayo, J. Mesirov. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52:91–118, July 2003.

[12] J. Reimand, T. Arak, and J. Vilo. g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res*, 39:W307–15, July 2011.

Mots clefs : severe asthma, systems medicine, omics, data integration, UBIOPRED

Démonstrations

Searching algorithm for type IV effector proteins (S4TE) 2.0 : tool for Type IV effectors prediction

Christophe Noroy [*] [†1], Adela Chávez[1], Nathalie Vachiery[1], Thierry Lefrançois[1], Damien Meyer [‡1]

Session démos 1
mardi 28 16h30
Salle place de l'école

[1] Contrôle des maladies animales exotiques et émergentes [Montpellier] (CMAEE) – Institut national de la recherche agronomique (INRA), Centre de coopération internationale en recherche agronomique pour le développement (CIRAD) : UMR15 – Campus international de Baillarguet, T A A-15 / G, F-34 398 Montpellier Cedex 5, France

Bacterial pathogens have evolved numerous strategies to corrupt, hijack or mimic cellular processes in order to survive and proliferate. Among those strategies, Type IV effectors (T4Es) are proteins secreted by pathogenic bacteria to manipulate host cell processes during infection. They are delivered into eukaryotic cells in an ATP-dependent manner by a specialized multiprotein complex, the type IV secretion system. T4Es contain a wide spectrum of features such as eukaryotic-like domains, localization signals or a C-terminal translocation signal. A combination of these 14 features enables prediction of T4Es in a given bacterial genome. In this research, we implemented a workflow, called Searching Algorithm for Type IV Effector proteins 2.0 (S4TE 2.0), to provide a comprehensive computational tool for accurate prediction and comparison of T4Es with a web-based graphical user interface. Applications range from characterizing effector features and identifying potential T4Es to analyzing effectors localization among the genome, according to G+C composition and local gene density. Following upload of Genbank files, bacterial genomes can be analyzed with default or user parameters. Furthermore, each feature can be searched independently making S4TE2.0 a useful tool to analyze a genome. Finally, S4TE 2.0 allows the comparison of putative T4Es repertoires among up to four bacterial strains. The software identifies T4Es orthologs between compared strains and returns the Venn diagram and lists of genes for each intersection. Last, we added interactive new features to offer the best visualization experience of the localization of identified candidate T4Es, including hyperlinks to NCBI and Pfam databases. S4TE 2.0 has been conceived to evolve rapidly with the publication of new experimentally validated T4Es, which will reinforce the predictive power of the algorithm. Our computational methodology is general and can be applied to the identification of a wide spectrum of bacterial effectors that lack sequence conservation but have similar amino acid characteristics. This approach will provide highly valuable information about bacterial host-specificity, virulence factors and to identify host targets for the development of new anti-bacterial molecules.

Mots clefs : S4TE2.0, prediction, effectors, T4SS, genome, workflow

*. Intervenant
†. Corresponding author : noroy.christophe@cirad.fr
‡. Corresponding author : meyer.damien@cirad.fr

DockNmine, a web portal to compare virtual and experimental interaction data

Jean Lethiec[1], Caroline Roze[1], Alan Amossé[1], Stéphane Téletchéa [*1]

[1] Unité de Fonctionnalité et Ingénierie des Protéines (UFIP) – Université de Nantes, CNRS : UMR6286 – 2 rue de la Houssinière, Bâtiment 25, F-44 322 Nantes Cedex 3, France

Session démos 1
mardi 28 16h50
Salle place de l'école

Determining and predicting detailed protein-protein or protein-ligand interactions is still a challenging task relying a lot on the expertise of the scientist in charge of the study. Better achievements are attainable when the expert is able to embrace all the literature related to the system under inspection, but this literature mining is time consuming and lacks of formal retrievable data. A small amount of existing binding data are for example available in general purpose databases like in the PubChem Projet [1], but many databases have to be crawled to gather only a fraction of the experimentally determined binding constants and activation / inhibition activities. One of the objectives of dockNmine platform is to automate some data retrieval in existing databases and to offer the possibility to its user to manually add the missing 70-90 % data to incorporate properly the proper information into a central repository. This addition can also included *in-house* data not available to the public, with a proper control access to the *work-in-progress*.

Although this first curation step is tedious, it is of extreme importance to enrich and benchmark predictions amenable via virtual screening methods. Once this initial set of data is properly entered in the database, it becomes more easily to derive from any existing molecules subfamily of chemical entities, to classify them and to assess their putative action on a given target. This incorporation allows to overcome the two major difficulties rising when one is willing to compare docking predictions with experiments, i.e. the lack of exact transferability of experimental data between experiment types and laboratories and the lack of standardization in molecule names.

Another feature of dockNmine is to ease the parsing of docking results via automated routines, linking automatically experimental data to the previous annotations coming from experiments. This co-occurrence of virtual and experimental data can be visualized in dockNmine in any browser without external plugins, and an interactive mapping of ligand-receptor interactions is proposed to the user. Classical comparisons are available for ligands (ROC, linear, scatter plot…) and many can be implemented if needed.

The recent developments of dockNmine have allowed to predict the ligand category using supervised and unsupervised methodologies. On datasets of 65 to 90 known ligand-protein affinities, each predictor is able to properly classify cross-validation ligands into "good", "medium" and "bad" ligands with a success rate of 70 % on average. These methodologies are implement using scikit-learn [2] and RDKit [3] and can be applied to any ligands entered into the dockNmine portal if sufficient ligands are available. An extension to docking predictions are envisioned.

The dockNmine platform will be presented at JOBIM 2016, and a companion web site will be set up soon to facilitate its evaluation.

Ce développement est soutenu dans le cadre du projet PIRAMID par le financement régional des Pays de la Loire.

*. Intervenant

References

[1] Wang, Y., Xiao, J., Suzek, T. O., et al. (2012) *Nucleic Acids Res.*, 40:D400–D412, PubChem's BioAssay Database.

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011) *J. Mach. Learn. Res.*, Scikit-learn: Machine Learning in Python.

[3] RDKit `http://rdkit.org/`.

Mots clefs : Data mining, Docking, Virtual Screening, Protein, Machine Learning, Python, Django

RiboDB : a dedicated database of prokaryotic ribosomal proteins

Frédéric Jauffrit [*1,2], Simon Penel[1], Jean-Pierre Flandrois [†1], Carine Rey[1,3], Manolo Gouy[1], Jean-Philippe Charrier[2], Stéphane Delmotte[1], Céline Brochier-Armanet [‡1]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] bioMérieux, Département de Recherche Technologique – BIOMÉRIEUX – 376 chemin de l'orme, F-69 280 Marcy l'Étoile, France
[3] Laboratoire de Biologie Moléculaire de la Cellule (LBMC) – CNRS : UMR5239, Institut national de la recherche agronomique (INRA) : UR5239, Université Claude Bernard - Lyon I (UCBL), École Normale Supérieure (ENS) - Lyon – ENS de Lyon, 46 allée d'Italie, F-69 364 Lyon Cedex 07, France

Session démos 1
mardi 28 17h10
Salle place de l'école

Since the end of the 70's, phylogenies of prokaryotes have been mainly relying on the analysis of the RNA component of the small subunit of the ribosome or a small set of housekeeping genes. The resulting phylogenies have provided interesting but partial information on the evolutionary history of these organisms because the corresponding genes do not contain enough phylogenetic signal to resolve all nodes of the Bacteria and Archaea domains. Thus, many relationships, and especially the most ancient and the most recent ones, remained elusive.

The recent burst of complete genome sequencing projects have made a lot of protein markers available as an alternative to SSU rRNA. Among protein markers, ribosomal proteins (r-proteins) are increasingly used as an alternative to ribosomal rRNA to study prokaryotic systematics.

However, their routine use is difficult because r-proteins are often not or wrongly annotated in complete genome sequences, and there is currently no dedicated exhaustive database of r-proteins.

Here we present RiboDB, a weekly updated, comprehensive database allowing the fast and easy retrieval of r-protein sequences.

The RiboDB database is built from the reannotation of publicly available complete prokaryotic genome sequences using a dedicated annotation engine. The RiboDB engine combines sequence similarity and profile based approaches for better sensitivity.

RiboDB is based on the ACNUC database system. This system allows users to retrieve RiboDB sequences from a multitude of interfaces, ranging from clients for all major platforms to APIs for many languages including C, python and R.

A dedicated RiboDB website is also available at `http://ribodb.univ-lyon1.fr/`. This website allows users to select one or more taxa of interest and a set of r-proteins of interest and returns the sequences in a format optimized for phylogenetic studies.

Mots clefs : sequence database, ribosomal proteins, annotation, phylogeny

*. Intervenant
†. Corresponding author : `jean-pierre.flandrois@univ-lyon1.fr`
‡. Corresponding author : `celine.brochier-armanet@univ-lyon1.fr`

Méta-analyse de données transcriptomiques avec metaMA et metaRNAseq sous Galaxy

Samuel Blanck [*1], Guillemette Marot[1,2]

Session démos 2
mercredi 29 14h40
Salle place de l'école

[1] Centre d'Études et de Recherche en Informatique Médicale (CERIM) – CHRU Lille, Université de Lille – E.A. 2694, Faculté de Médecine - Pôle Recherche, 1 place de Verdun, F-59 045 Lille Cedex, France
[2] MODAL (INRIA Lille - Nord Europe) – INRIA – France

Les technologies de puces à ADN et de séquençage à haut débit telles que le RNAseq sont très utilisées pour les analyses de données transcriptomiques. Cependant, pour des questions de coût, peu de réplicats biologiques sont inclus dans les études, ce qui affecte la capacité de détection de vrais transcrits différentiellement exprimés. Les meta-analyses offrent donc la possibilité d'augmenter la puissance statistique et d'accroitre la pertinence des résultats. Les packages R metaMA et metaRNAseq permettent de réaliser des méta-analyses sur des données provenant de différentes études d'analyse différentielle d'expression de gènes. Toutefois l'utlisation de ces outils nécessitant une bonne maitrise du langage R, nous avons intégré les fonctionnalités offertes par ces packages R à la plate-forme web d'analyse de données Galaxy, afin d'en faciliter la manipulation par les biologistes. En plus des meta-analyses, cette suite d'outils permet d'accéder aux données de la base GEO, de réaliser des contrôles qualités et des analyses individuelles d'expression différentielle de gènes basées sur les packages R limma et Deseq2. De plus, l'interface conviviale de Galaxy permet de mener ces différents traitements de façon aisée et les résultats annotés sont facilement visualisables et exportables dans des formats communs.

Mots clefs : Meta analyse, transcriptomique, Galaxy

*. Intervenant

SHAMAN : a shiny application for quantitative metagenomic analysis

Amine Ghozlane [*1], Stevenn Volant [*†1]

[1] Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur de Paris – France

Session démos 2
mercredi 29 15h00
Salle place de l'école

Quantitative metagenomic is broadly employed to identify genera or species associated with several diseases. These different studies are all based on the targeted sequencing of 16S/18S/ITS rDNA or on random sequencing of whole-community DNA. Quantitative data are obtained by mapping the reads against operational taxonomic units (OTU) or a gene catalog. The data generated can then be analysed quantitatively using R packages (metagenomeseq, edgeR) or with web-interfaces (Shiny-phyloseq, Phinch) that do not integrate the statistical modeling. The lack of easy access methods providing both the statistical modeling and the visualisation constitutes a critical issue to address this type of analysis. Here we present SHAMAN, a Shiny-based application integrating the metagenomic data (a count matrix for each sample and a table assigning a taxonomical annotation to each feature), the experimental design, the statistical model and a dynamic-interface dedicated to the differential analysis.

SHAMAN process is divided into three steps : normalisation, modelisation and visualisation. The count matrix is normalised at the OTU/gene level using the DESeq2 normalisation method and then, based on the experimental design, a generalised linear model is applied to detect differences in abundance at the considered taxonomic level.

SHAMAN provides diagnostic plots to check the quality of the modelisation and visualisation that highlight the differences in abundance that have been identified by the statistical analysis. Diversity plots are also proposed to illustrate the results from a more global view.

SHAMAN is freely accessible through a web interface at
`http://shaman.c3bi.pasteur.fr/`

Mots clefs : Quantitative metagenomic, web interface, statistical analysis

*. Intervenant
†. Corresponding author : `stevenn.volant@pasteur.fr`

Heat-seq : a web-application to contextualize a high-throughput sequencing experiment in light of public data

Session démos 2
mercredi 29 15h20
Salle place de l'école

Guillaume Devailly [*] [†1], Anna Mantsoki[1], Anagha Joshi[1]

[1] The Rolsin Institute, University of Edinburgh (UOEDIN) – The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian EH25 9RG Scotland, UK/Royaume-Uni

With the establishment of better protocols and decreasing costs, high-throughput sequencing experiments such as RNA-seq or ChIP-seq are now accessible even to small experimental laboratories. However, comparing one or few experiments generated by an individual lab to the vast amount of relevant data available in public domain might be hindered due to lack of bioinformatics expertise. Though several user friendly tools allow such comparison gene or promoter level, a genome-wide picture is missing. We developed Heat*seq, a free, open-source web-tool that allows comparison at genome-wide scale of any experiments provided by the user to public datasets (RNA-seq, ChIP-seq and CAGE experiments from Bgee, Blueprint epigenome, CODEX, ENCODE, FANTOM5, modEncode and Roadmap epigenomics) in human, mouse and drosophila. Correlation coefficients amongst experiments is displayed as an interactive correlation heatmaps. Users can thus identify clusters of experiments in public domain similar to their experiment in minutes through a user-friendly interface. This fast interactive web-application uses the R/shiny framework allowing the generation of high-quality figures and tables that can be easily downloaded in multiple formats suitable for publication.

Heat-seq is freely available at
`http://www.heatstarseq.roslin.ed.ac.uk/`

Mots clefs : séquencage haut débit, RNA, seq, ChIP, seq, CAGE, encode, modEncode, roadmap, blueprint, fantom5, codex

*. Intervenant

†. Corresponding author : `guillaume.devailly@rolsin.ed.ac.uk`

Évolution moléculaire à la carte avec bio++

Laurent Guéguen [*1], Julien Yann Dutheil[2]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – Université Claude Bernard - Lyon I, CNRS : UMR5558 – UCB Lyon 1 - Bâtiment Grégor Mendel, 43 boulevard du 11 novembre 1918, F-69 622 VILLEURBANNE Cedex, France
[2] Institut des Sciences de l'Évolution de Montpellier (ISEM) – Université Montpellier II - Sciences et techniques : UMS5554 – France

Session démos 2
mercredi 29 15h40
Salle place de l'école

Bio++ est une suite de librairies en C++, dédiées à l'analyse de séquences biologiques, plus particulièrement dans un contexte évolutif.

Basée sur ces librairies, une suite de programmes, appelée bppsuite, permet à un utilisateur d'effectuer différentes analyses à partir de séquences alignées. Par exemple, il peut optimiser une modélisation au maximum de vraisemblance, inférer des séquences ancestrales ou encore simuler des processus évolutifs.

De manière à permettre à un utilisateur de déclarer facilement quel type d'analyse et de modélisation il désire effectuer, nous avons conçu une syntaxe claire de déclaration de modèles, de processus évolutifs, de types de vraisemblances. De très nombreux modèles d'évolution moléculaire, issus de la littérature, sont disponibles (modèles de nucléotides, de codons ou d'acides aminés). Dans la version publiée en 2013 [1], il était possible de proposer à loisir différents modèles dans un arbre.

Depuis, de nombreux développements ont été effectués dans Bio++. Ainsi, il est possible d'intégrer simultanément différentes topologies et données, et de chercher à optimiser la vraisemblance sur l'ensemble. Par exemple, l'utilisateur peut chercher à modéliser au mieux l'évolution dans un alignement en lui proposant plusieurs topologies, et donc révéler des changements topologiques le long de cette donnée. Dans ce contexte, un lissage par HMM peut être intégré dans le modèle, de manière à révéler des segments avec des topologies différentes. Aussi, autre exemple, il est possible de partitionner a priori les données, et de proposer différents modèles sur ces données (par exemple avec le même arbre, mais pas forcément), pour évaluer au mieux un paramètre en intégrant des données hétérogènes. Dans un souci d'une meilleure modélisation, ces modèles et arbres peuvent partager leurs paramètres.

L'objectif de cette présentation est d'illustrer, via plusieurs exemples, les diverses fonctionnalités accessibles dans bppsuite, fonctionnalités qui pourront permettre aux biologistes de tester des hypothèses d'évolution moléculaire bien plus riches et plus adaptées à leurs problématiques spécifiques, que les programmes « clefs en main » actuellement disponibles.

Bio++ et bppsuite sont accessibles via le site :
`http://biopp.univ-montp2.fr/wiki/index.php/Main_Page`

Dans ce site, l'utilisateur la documentation est accessible, des exemples, ainsi qu'un forum d'aide.

Références

[1] Gueguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette N, Bigot T, Fournier D, Pouyet F, Cahais V, Bernard A, Scornavacca C, Nabholz B, Haudry A, Dachary L, Galtier N,

*. Intervenant

Belkhir K, Dutheil J Y (2013). Bio++: Efficient Extensible Libraries and Tools for Computational Molecular Evolution. *Molecular Biology and Evolution*, vol. 30, pp.1745-1750.

Mots clefs : modeling, molecular evolution, software

Posters

Automatic detection of abnormal plasma cells

Elina Alaterre [*] [†1], Sébastien Raimbault[1], Jean-Michel Garcia[1], Guilhem Requirand[2], Jérôme Moreaux[3]

Poster 1

[1] HORIBA Medical ABX – Horiba – 390 rue du Caducée, F-34 090 Montpellier, France
[2] CHU Montpellier – CHRU Montpellier – France
[3] Institut de génétique humaine (IGH) – CNRS : UPR1142 – 141 rue de la Cardonille, F-34 396 Montpellier Cedex 5, France

Le myélome multiple (MM) ou maladie de Kahler est une hémopathie maligne caractérisée par l'accumulation de plasmocytes tumoraux (lymphocytes B) dans la moelle osseuse, et l'accumulation d'une immunoglobuline monoclonale complète ou d'une chaîne légère monoclonale. Cette pathologie touche 2 000 nouveaux patients par an en France, 19 000 en Europe, 19 000 aux Etats-Unis. L'espérance de survie médiane des patients d'âge inférieur ou égal à 65 ans est de 6-7 ans. L'objectif de ces travaux est de détecter automatiquement les populations plasmocytaires normales et tumorales impliqués dans cette hémopathie à partir de fichiers résultant d'une analyse de moelle osseuse en cytométrie en flux.

Nous avons conçu un programme analysant les fichiers FCS générés par les logiciels d'acquisition des cytomètres en flux. Ces fichiers contiennent toutes les données pour chaque paramètre calculé ou mesuré pour l'ensemble des événements (cellules) qui passent dans la fenêtre de détection du cytomètre. Dans un premier temps, l'ensemble de ces données sont récupérées et compensées grâce à la matrice de compensation. Les données transformées peuvent alors permettre de tracer des graphiques multidimensionnelles sur des échelles linéaires, logarithmiques ou bi-exponentielles pour identifier des populations d'intérêt. La discrimination de ces populations est réalisée grâce à des fenêtres fixes, lorsque les paramètres cellulaires sont stables entre les échantillons (taille ou morphologie des cellules), ou des fenêtres automatiques, quand les paramètres cellulaires sont soumis à des variations biologiques inter-patients (expression de marqueurs). Les fenêtrages automatiques sont réalisés grâce à des transformations rotationnelles, des calculs de moyennes de fluorescence ou encore des recherches de vallées entre histogrammes.

Les résultats de la détection automatique des plasmocytes ont été comparés à ceux obtenus manuellement suite à l'analyse de moelle osseuse de patients atteints de myélome multiple dans deux laboratoires distincts, le laboratoire « Suivi des Thérapies Innovantes » (CHU Montpellier) (n = 66) et HORIBA Medical (Montpellier) (n = 39) respectivement sur deux automates différents, le Cyan (Beckman Coulter) et le LSR Fortessa (BD Biosciences).

La détection automatique des plasmocytes normaux et tumoraux grâce à notre programme permet d'éviter l'analyse subjective des résultats souvent observée en cytométrie, de diminuer le temps et du coût du test ainsi que d'augmenter la reproductibilité.

Mots clefs : myélome multiple, plasmocytes, fichiers FCS, analyse multidimensionnelle, cytométrie en flux

*. Intervenant
†. Corresponding author : elina.alaterre@horiba.com

Analyse bioinformatique du rôle des G-quadruplexes dans la régulation de la transcription

Poster 2

Marianne Bordères [*1], David Martin[2], Cyril Esnault[2], Jean-Christophe Andrau[2]

[1] Parcours Bioinformatique, Connaissances, Données du Master Sciences & Numérique pour la Santé – Université de Montpellier – 2 place Eugène Bataillon, F-34 090 Montpellier, France
[2] Institut de Génétique Moléculaire de Montpellier (IGMM) – CNRS : UMR5535 – 1919, route de Mende, F-34 293 Montpellier, France

Les îlots CpG sont des régions du génome fréquemment associées aux promoteurs [1], très enrichies en CpG et le plus souvent non méthylées. La méthylation de ces îlots contribue à la stabilisation d'un état fermé de la chromatine ce qui entraîne une répression de l'expression des gènes. À l'inverse, une hypométhylation des C dans les îlots CpG aurait tendance à contribuer à une déplétion nucléosomale qui favorise le recrutement de la machinerie transcriptionnelle. De plus, dans ces îlots, on trouve des séquences riches en GC prédites comme formant des structures secondaires de l'ADN appelées G-quadruplexes (PG4s) [2]. Celles-ci apparaissent fréquemment en amont et en aval du site de départ de la transcription. Contrairement aux boîtes TATA qui sont présentes dans seulement 3 % des gènes chez les mammifères, les PG4s le sont dans plus de 60 % des promoteurs [3]. Les G4s sont des structures d'ADN simple brin très stables formant quatre feuillets, chacun constitués de guanines. Une corrélation entre les régions peu denses de la chromatine dans les îlots CpG et la présence de PG4s prédits a été mise en évidence au laboratoire. L'hypothèse est donc que les G4s jouent un rôle majeur dans l'expression des gènes, en favorisant l'accessibilité des promoteurs aux facteurs de transcription.

Néanmoins, la séquence consensus actuelle des G4s reste assez floue et contient de nombreux espaceurs dont la quantité peut varier. Un de nos objectifs a été d'affiner cette séquence. Pour cela, des données de MNase-Seq, une technique permettant de cartographier l'ADN associé aux nucléosomes, ont été exploitées. Après alignement des séquences sur un génome de référence, nous avons déterminé l'ADN associé aux nucléosomes et la fraction qui en est déplétée. A partir des régions déplétées, une recherche de motif a été réalisée pour affiner la séquence consensus des G-Quadruplexes avec des outils d'alignement local avec 'gap'. Ces analyses ont été effectuées sur des données du laboratoire et sur des données publiées par d'autres groupes dans d'autres organismes, ajoutant une dimension évolutive à la problématique biologique. Nous souhaitons maintenant poursuivre ce projet avec l'analyse de données non publiées du laboratoire issues d'une nouvelle technique permettant d'isoler et de séquencer directement les G4s *in vivo.*

Remerciements

Je remercie le département informatique de la Faculté des Sciences de l'Université de Montpellier (`http://deptinfofds.univ-montp2.fr/`) ainsi que le Labex Numev (`http://www.lirmm.fr/numev/`) pour avoir accepté de financer ma participation à JOBIM.

Références

[1] Fenouil, R. *et al.* CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res.* 22 :2399–2408 (2012).

*. Intervenant

[2] Schwarzbauer, K., Bodenhofer, U. & Hochreiter, S. Genome-wide chromatin remodeling identified at GC-rich long nucleosome-free regions. *PloS One* 7 :e47924 (2012).

[3] Bedrat, A., Lacroix, L. & Mergny, J.-L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.* gkw006 (2016). doi :10.1093/nar/gkw006.

Mots clefs : Bioinformatique, MNase, seq, G, Quadruplexes, Recherche de motif, Promoteur, Ilots CpG

Implementation d'une methode d'imputing dans l'analyse de l'association des genes candidats de maladies complexes

Poster 3

Yannick Cogne [*1], Corinne Lautier[1], Sara Haydar[2], Nicoleta Baculescu[3], Florin Grigorescu [†4]

[1] Université Montpellier (UM) – Université Montpellier II - Sciences et Techniques du Languedoc – France
[2] Université Montpellier (UM) – Université Montpellier I – France
[3] "Carol Davila" University of Medicine and Pharmacy (UMF) – Roumanie
[4] Institut National de la Santé et de la Recherche Médicale – INSERM – France

Les études de GWAS représentent une extraordinaire opportunité d'identification des gènes responsables des maladies complexes comme l'obésité et le diabète de type 2 (DT2). Un des facteurs limitants dans cette entreprise est la densité de génotypage des marqueurs SNP (*single nucleotide polymorphism*) à un locus donné ou au niveau du génome entier. De plus, malgré des succès incontestables dans ce domaine (par exemple la découverte d'environ 12 gènes dans le DT2), les GWAS demeurent sous-optimales, car l'association ne peut expliquer l'héritabilité dans la population. Une des stratégies possibles pour améliorer le diagnostic consiste à augmenter la densité des SNP pour un locus donné ou d'effectuer des séquençages à très haute densité afin d'identifier des nouveaux SNP rares, mais à effet majeur. Comme le génotypage par son coût et l'effort humain reste un facteur limitant, notamment pour de très grandes populations (> 100 000 individus), les procédures bioinformatiques d'imputation se sont imposées avec des résultats remarquables. Ces méthodes varient par l'algorithme et la stratégie utilisés comme par exemple, l'utilisation du prétraitement des fichiers d'entrée (pré-phasing) ou l'utilisation des filtres après l'imputation (R^2).

L'objectif de cette étude a été d'effectuer l'association génétique cas/témoins avant et après imputation afin d'observer comment cette procédure bioinformatique peut compléter ou améliorer l'association et la cartographie des haplotypes à un locus donné. Pour cette raison, nous avons exploré d'abord auparavant la *diversité haplotypique* et la répartition des blocs de *déséquilibre de liaison* (DL) à un locus candidat. Un objectif secondaire a été celui d'observer si la procédure d'imputation altère les différences qui existent entre des populations d'origine ethnique différentes. En effet, une des divergences majeures qui se profile entre deux populations est la répartition des blocs de DL, en fonction des sites de recombinaisons. Nous avons donc choisi deux populations génétiquement distantes comme celles du Sud de la France et celles d'Europe Centrale.

Choix des logiciels

Plusieurs programmes sont disponibles pour l'imputation, tels que F-Impute, Impute2, MACH et BEAGLE qui diffèrent par l'algorithme et le temps de calcul. Par rapport aux autres stratégies, celle de BEAGLE a attiré notre attention car ce programme utilise une stratégie graphique de comparaison des haplotypes de la population à étudier avec celle de la population de référence (données de 1000 Génomes). Cette stratégie est donc la mieux adaptée pour la comparaison de populations d'ethnies différentes. D'autres programmes (Impute2 ou MACH) utilisent un temps de calcul plus long et requièrent plus de mémoire en fonction du nombre d'individu étudiés. Enfin

*. Intervenant
†. Corresponding author : florin.grigorescu@inserm.fr

une autre méthode (F-Impute) est basée sur les relations interindividuelles (larges blocs de DL) et est adaptée pour des études familiales. De plus, la comparaison des protocoles (« *benchmark* ») de ces outils a permis de constater que la précision des programmes dépend de plusieurs facteurs. Sur des échantillons réduits et avec un faible nombre de marqueurs SNP initiaux, BEAGLE se montre moins précis par rapport à Impute2. Cependant, la tendance est inversée lorsqu'un plus grand nombre de marqueurs et d'individus sont analysés. Pour cette raison nous avons opté pour le logiciel BEAGLE, qui en outre est simple d'utilisation. En effet, ce logiciel (BEAGLE v4.1) utilise des fichiers standard VCF comme entrée et sortie, ce qui facilite son utilisation par une équipe de biologistes.

Populations et gène étudié

Pour cette étude, parmi les SNP génotypés par GWAS, notre intérêt s'est focalisé vers les SNP des gènes candidats communs des maladies neurodégénératives et de la résistance à l'insuline. Ainsi, nous avons sélectionnés 60 gènes, impliqués principalement dans la cascade de signalisation de l'insuline et de l'IGF-1 (*insulin-like growth factor-1*). Afin, d'évaluer l'impact de l'imputation sur la discrimination des populations d'origine ethniques différentes, nous avons choisi une population Française constituée de 329 individus (243 contrôles, 89 cas) et une population Roumaine de 188 individus (86 contrôles, 102 cas) atteintes de syndrome métabolique (SMet) avec insulino-résistance. Parmi tous les gènes analysés, le gène Grb10 présente une forte variabilité du nombre de SNP significativement plus fréquent dans l'une ou l'autre des populations avant et après imputation, et a donc été sélectionné pour répondre aux questions que l'on se pose sur le rôle de l'imputation dans les études l'associations des haplotypes.

Résultats.

Dans une première étape, nous avons analysé la structure des blocs de DL dans les deux populations au locus du gène Grb10 (*growth factor receptor bound protein* 10) sur Chr7, en utilisant le logiciel HAPLOVIEW qui par la méthode statistique de Gabriel et al. 2002, permet de définir les blocs de DL. Ainsi, dans la population française 15 blocs de DL ont été identifiés alors que dans la population roumaine le gène Grb10 possède 14 blocs de DL. Dans cette structure, avant l'imputation, il y a 5 SNP statistiquement plus prévalent chez les cas ($P < 0,05$, χ^2) et 17 SNP dans la population roumaine. Il est a remarquer que l'imputation à partir des SNP de référence du projet 1000 Génomes, 8 et 12 SNP imputés et différents des précédents sont significativement plus prévalent. Ces résultats impliquent une analyse plus détaillée de la prévalence des SNP en fonction des blocs de DL, voir des nouveaux blocs en changeant la structure des blocs.

Dans une deuxième étape, si on choisit comme modèle la population française avant imputation, les 5 SNP associés (plus fréquents chez les cas) sont répartis sur 2 blocs, 1 sur B1, 1 sur B2 alors que les 3 autres SNP sont extérieurs aux blocs. Le bloc B1 est constitué de 7 SNP, permettant la reconstitution de 7 haplotypes (H) différents pour ce bloc (PHASE contenu dans HAPLOVIEW). L'haplotype H2 est protecteur (plus fréquent chez les contrôles $P < 0,0185$), alors que l'haplotype H4 est pathogène (plus fréquent chez les cas $P < 0,0387$). Ces résultats sont concordant avec l'analyse individuelle de la prévalence des allèles du SNP d'intérêt dans la population. Pour le bloc B2 qui est constitué de 8 SNP à partir des quels on a reconstitué 7 haplotypes différents, une autre situation a été observée. En effet, l'haplotype H4 de ce bloc est pathogène ($P < 0,0151$) mais le signal est divergent par rapport au SNP seul. Autrement dit, dans le bloc B2 l'haplotypage apporte une information supplémentaire.

Après imputation, on ajoute aux SNP précédemment décrits 8 nouveaux SNP associés. Ainsi, ces 13 SNP se répartissent sur 3 nouveaux blocs, 8 sur le bloc B1 (les 8 SNP associés imputés), 1 sur le bloc B2 (B1 avant imputation) et 1 sur le bloc B4 (B2 avant imputation). Après imputation, le bloc B1 est constitué de 25 SNP qui forment 7 haplotypes différents. H2 est protecteur ($P < 0,0081$) et

H4 pathogène (P < 0,0465), ce qui est concordant avec l'association des SNP seuls. Le bloc B2 a été enrichi de 2 SNP distribués sur 7 haplotypes, mais aucun de ces haplotypes n'est significatif. On perd donc après imputation la significativité dans ce bloc pour un SNP associé. Le bloc B4 est constitué de 33 SNP et forme 9 haplotypes différents. On observe que H6 est pathogène (P < 0,04) et est le seul haplotype de ce bloc porteur de l'allèle pathogène du SNP associé individuellement. Ce résultat permet de montrer que grâce à l'imputation, l'information d'association pour un SNP pathogène est retrouvée.

Cette méthode d'analyse a été appliquée à la population Roumaine, pour laquelle des résultats similaires à ceux identifiés chez les Français, ont été observés. Par soucis de clarté, nous décrirons ici uniquement le bloc le plus pertinent pour notre interprétation biologique de l'effet de l'imputation. Ainsi, parmi les 17 SNP associés et répartis sur 5 des 14 blocs défini pour la population Roumaine, le bloc B2 est constitué de 16 SNP dont 7 sont pathogènes et forment 12 haplotypes. Les haplotypes H3 et H11 sont pathogènes (P < 0,034 et P < 0,038 respectivement). Après imputation le bloc B3 est porteur de 49 SNP dont 17 sont pathogènes (le bloc B2 est entièrement retrouvé dans le bloc B3) et forment 13 haplotypes. Ce bloc B3 imputé, intègre donc l'information du bloc B2 avant imputation. On observe qu'après imputation l'haplotype H3 est pathogène (P < 0,0229). Il est intéressant de noter que le paramètre d'association est meilleur après imputation (avant imputation P < 0,038, et après imputation P < 0,0229).

Conclusion.

Nos résultats montrent que l'association pathogène d'un SNP considéré individuellement est conservé sur l'haplotype porteur de l'allèle pathogène de ce SNP. Les blocs de DL définis permettent une cartographie de ces SNP qui va être différentes après imputation. Ainsi, après imputation, on va observer un regroupement des SNP pathogènes dans des blocs plus grands. Ces blocs vont permettre de mettre en évidence des haplotypes plus longs qui auront une meilleure valeur d'association (pathogène ou protecteur). D'autre part, l'imputation ne masque pas l'information génétique spécifique d'ethnie. En effet, l'imputation de SNP à un même locus reste différente d'une population à l'autre, comme nous vous l'avons montré en comparant la population Française et Roumaine. L'utilisation de méthode d'imputation de SNP va donc nous permettre dans l'analyse de GWAS, de densifier l'information à un locus d'intérêt et nous permettre de mieux caractériser les haplotypes pathogènes. De plus, pour les pathologies complexes, cette spécification génétique permettra de mieux sélectionner et/ou de stratifier les populations ayant avec un phénotype observable similaire.

Mots clefs : Imputation, GWAS, Génétique

Qualitative assessment of single-cell RNA-seq data for the robust detection of subpopulations of cells and their characteristic gene signatures

Poster 4

Ewen Corre [*1], Antonio Rausell[1]

[1] Clinical Bioinformatics Laboratory – Institut Imagine – 24 boulevard du Montparnasse, F-75 015 Paris, France

The human immune response is highly heterogeneous across individuals and can be partly explained by genetic factors. From a less explored perspective, heterogeneity is also found across the individual cells from the same subject, despite sharing a common genetic background. More strikingly, some autoimmune disorders caused by haploinsufficient genetic variants present incomplete penetrance both at cellular and clinical level, *i.e.*: only a fraction of the carriers develop the disease despite the fact that both patients and healthy carriers present only a fraction of their cells with a defective phenotype. High-dimensional single-cell approaches may help uncovering intra-individual cell-to-cell differences that could explain the onset and progression of immune diseases presenting incomplete penetrance. Single-cell RNA-seq performs the unbiased transcriptional profiling of large numbers of cells. However, a number of biases and biological and technical noise make its computational analysis challenging. In this work, we present a new statistical pipeline for the unsupervised analysis of single-cell RNA-seq data aiming at the identification of heterogeneous subpopulations of cells within a sample and the automatic assessment of the gene signatures characterizing each group. The pipeline relies on a Multiple Correspondence Analysis (MCA)-based approach (Rausell *et al* 2010) who proved valuable to cluster pancreatic tumor bulk samples described with binary gene expression profiles (Martinez-Garcia, Juan, Rausell *et al*. Genome Medicine 2014). Discretization of expression levels has been previously used to minimize batch effects and reduce noise in microarray data. We hypothesized that an analogous qualitative treatment of gene expression levels can be especially helpful for the analysis of single-cell RNA-seq data where technical and biological noise can lead to significant variation in the observed quantitative values. Furthermore, numerous studies have reported bimodality in gene expression at single-cell level, which may favor such a qualitative approach. Here, different statistical approaches to discretize single-cell RNA-seq levels are evaluated, e.g. a B-spline method, probabilistic mixture models and multi-modal tests. In addition, the possibility of dropout events leading to non-detection of gene expression is treated in a probabilistic way by modeling the relationship between the fraction of cells with non-detection versus the average expression of the cells with detected signal. Two alternative strategies are proposed: a discretization into disjoint bins or a probabilistic assignment into multiple bins. In both cases an MCA treatment is then applied to obtain two related vector spaces: one of individual cells and one of genes. After determining the optimal number of retained dimensions to filter out noise, an automatic unsupervised clustering of cells is performed and the specific sets of genes characterizing each subpopulation is assessed through the MCA pseudo-baricentric relationships. The results of the approach are compared to those led by alternative state-of-the-art methods making use of publicly available single-cell data sets.

Mots clefs : single cell, RNA-seq, immune diseases, statistics, cellular heterogeneity, gene expression, discretization, qualitative

*. Intervenant

Screening of public cancer data reveals RPL5 as a candidate tumor suppressor

Laura Fancello [*1], Kim De Keersmaecker [†1]

[1] Laboratory for disease mechanisms in cancer - KU Leuven – O&N IV Herestraat 49, box 602, 3000 Leuven, Belgique

Poster 5

Several functional categories of proteins show somatic alterations in cancer including transcription factors, signaling molecules and epigenetic regulators. Recently, somatic defects in the ribosome, the cellular machinery in charge of translating mRNA into proteins, were also described in cancer. Indeed, somatic mutations and deletions affecting the ribosomal protein genes RPL5 in glioblastoma, RPL5, RPL10, RPL11 and RPL22 in acute T-cell leukemia and RPS15 in chronic lymphocytic leukemia were discovered. Moreover, patients affected by congenital diseases caused by ribosome defects (ribosomopathies) present a high risk to develop cancer. Finally, some experimental work in zebrafish showed that heterozygous inactivation of ribosomal protein genes promotes tumor development.

These observations suggest a role of ribosome defects in cancer development. However, we still lack an overview of which and how many ribosomal protein genes are most affected, in which cancer types and which role they play in oncogenesis (i.e tumor suppressors or oncogenes).

In order to answer these questions, we performed a systematic screening of the public cancer database TCGA for somatic mutations, copy number alterations, differential expression and/or methylation of 81 ribosomal protein genes in 16 different cancer types. Correlations to significant clinical features and patterns of co-occurring or exclusive mutations were also analysed. This meta-analysis reveals that RPL5, RPL11, RPSA, RPL23A, RPS5 and RPS20 are frequently and significantly targeted by genetic alterations in cancer and therefore represent interesting candidate driver cancer genes. In particular, RPL5 is significantly mutated and/or heterozygously deleted in breast cancer, glioblastoma and melanoma patients, with higher and previously underestimated frequency of alteration in glioblastoma (10.9 %). Interestingly, RPL5 haploinsufficiency is significantly correlated to worse overall survival in glioblastoma patients and cooperates with TP53 to confer worse overall survival in breast cancer patients.

Mots clefs : TCGA, cancer, high throughput sequencing data

*. Intervenant

†. Corresponding author : Kim.DeKeersmaecker@kuleuven.be

Évaluation des outils bioinformatiques dédiés à la caractérisation des sous-clones mutés minoritaires

Poster 6

Benoît Guibert [*] [†1], Thérèse Commes[1], Laurence Lodé[2,3], Anthony Boureux [‡1]

[1] Institut de Recherche en médecine régénératrice, INSERM U1183 (U1183) – Universté Montpellier – Hôpital Saint-Éloi, 80 avenue Augustin Fliche, F-34 295 Montpellier Cedex 5, France – Tél : +33 4 67 33 57 11

[2] CHU de Montpellier – CHRU de Montpellier – Hôpital Saint Éloi, 80 avenue Augustin Fliche, F-34 000 Montpellier, France

[3] UMR CNRS 5235 – CNRS – Hôpital Saint Éloi, 80 avenue Augustin Fliche, F-34 000 Montpellier, France

Problématique biologique

TP53 est un gène retrouvé muté dans 50 % des cancers, toute origine histologique confondue (Leroy et al.,2014). Il code pour la protéine p53 qui est suppresseur de tumeur à l'état sauvage en entraînant l'arrêt du cycle cellulaire et la mort par apoptose des cellules soumises à des dommages de l'ADN susceptibles de devenir tumorales. La protéine p53 mutante peut en revanche acquérir des propriétés oncogéniques à l'état muté et donc favoriser la croissance tumorale (Lane et al., 2010).

Les syndromes myélodysplasiques de bas risque avec délétion 5q (del5q) se manifestent par une anémie non carentielle (réfractaire) due à une hyperexpression de p53 sauvage dans les précurseurs des globules rouges de la moelle osseuse. La découverte récente de l'efficacité remarquable du lenalidomide sur ces cellules tumorales hyperexprimant p53 sauvage a révolutionné la prise en charge thérapeutique de ces patients, guérissant leur anémie en quelques semaines et leur épargnant des transfusions de culots de globules rouges à un rythme soutenu responsables de dégradation de la qualité de vie et de surcharge en fer (hémochromatose secondaire)(Wei et al., 2013).

Des mutations de TP53 sont retrouvées dans 15 à 20 % des syndromes myélodysplasiques de bas risque avec délétion 5q (del5q) et l'existence de protéine p53 mutante dans ces cellules tumorales est associée à un pronostic défavorable par évolution en une forme gravissime de la maladie (leucémie aigüe). Il a été démontré que ces clones mutés pour TP53 n'étaient plus sensibles au traitement par lenalidomide (Jädersten et al., 2011).

L'étude cytogénétique et la recherche rétrospective des mutations de TP53 réalisées de façon longitudinale par technique de deep-NGS (technologie 454 Roche) chez des patients atteints de syndromes myélodysplasiques de bas risque avec del (5q) isolée au diagnostic ont permis de mettre en évidence une forte instabilité génétique évocatrice d'une évolution clonale.

L'analyse des mutations de TP53 par technique de séquençage haut débit (NGS 454 de profondeur > 1000X) montre fréquemment une absence de mutation de TP53 (sensibilité 1-2 %) au diagnostic. L'émergence d'une ou plusieurs mutations survient le plus souvent après mise en route d'un traitement par lénalidomide probablement responsable de l'éradication de clones

*. Intervenant

†. Corresponding author : benoit.guibert@free.fr

‡. Corresponding author : Anthony.Boureux@univ-montp2.fr

indolents porteurs du gène TP53 sauvage et la sélection de clones agressifs porteurs de TP53 muté comme décrit dans la leucémie lymphoïde chronique (Landau et al., 2013).

La question de la pré-existence des mutations TP53 avant traitement se pose avec deux hypothèses envisageables :

— l'émergence de ces clones TP53 est-elle due à une instabilité génétique due à la surcharge en fer et aux ROS ? dans ce cas, il pourrait être intéressant de traiter les patients par lenalidomide à un stade très précoce de la maladie, avant que les patients ne soient transfusés évitant ainsi la surcharge en fer et l'instabilité génétique.

— ces sous-clones TP53 minoritaires sont-ils pré-existants dès les stades précoces de la maladie ? dans ce cas, il serait déconseillé de traiter ces patients par lenalidomide pour éviter l'évolution clonale et la progression de la maladie.

Dix patients ont vu émerger des clones TP53 au cours de leur maladie alors que l'analyse du gène TP53 par NGS de leur prélèvement au diagnostic était négatif au seuil de 1-2 %.

Problématique bioinformatique

Recherche de variants

L'appel de variants, ou « SNV Calling » est l'étape permettant de détecter les variants à partir d'un fichier BAM pour produire un fichier VCF. Il s'agit d'identifier les sites où les données séquencées présentent des variations par rapport à un génome de référence. La principale difficulté est de trouver un compromis entre une bonne sensibilité (afin de minimiser les faux négatifs, non détection des sous-clones), et une bonne spécificité (pour minimiser les faux positifs liés aux artefacts). Cette tâche est confiée à des outils tels que GATK, Samtools, Mutect et développés par le Broad Institute ou l'un des nombreux autres logiciels d'appel de variants dont VarScan2 ou DeepSNV/Shearwater. Selon le logiciel utilisé, différentes approches sont possibles.

Il est également nécessaire de produire des analyses statistiques d'aide à la validation des variants, en particulier les profondeurs de couverture, permettant de connaître le nombre de fois qu'un nucléotide a été séquencé au total, le nombre de fois qu'un nucléotide a été séquencé de façon identique à la référence ou de façon variante.

Les méthodes actuelles pour résoudre la problématique des sous-clones génétiquement distincts dans les échantillons tumoraux nécessitent de regrouper les mutations somatiques par fréquences alléliques. Ainsi, des algorithmes de « variant calling » ont été développés pour distinguer avec précision les vrais polymorphismes SNP et les mutations somatiques SNV/MNV/indels des allèles non référencés dus à des artefacts introduits pendant la préparation des banques d'ADN et par le séquençage.

Recherche des variations dans des sous-clones

La détection des substitutions (SNV) ou petites insertions-délétions (indels) courtes somatiques est une étape clé dans la caractérisation du génome du cancer. Les mutations dans le cancer sont rares (0.1 à 100 par Mb de génome de cancer) et surviennent souvent seulement dans un sous-groupe de cellules séquencées, soit en raison de la contamination par des cellules normales ou soit par le fait de l'hétérogénéité tumorale. En conséquence les méthodes d'appel de mutations ont besoin d'être à la fois spécifiques pour éviter les faux positifs, et sensibles, pour détecter les mutations clonales et sous-clonales (Beerenwinkel et al., 2015). Ainsi, des algorithmes de « variant calling » ont été développés pour distinguer avec précision les vrais polymorphismes SNP et les mutations somatiques SNV/indels des allèles non référencés dus à des artefacts introduits pendant la préparation des librairies et par le séquençage.

Deux méthodes principales d'appel de variants sont utilisées : les méthodes heuristiques, utilisées par des outils comme DeepSNV (Gerstung et al., 2012), VarScan2 (Koboldt et al., 2012),

ou encore Lofreq (Wilm et al., 2012), et les classifieurs bayésiens, implémentés dans les outils Mutect (Cibulskis et al., 2013) ou Shearwater (Beerenwinkel et al., 2015). Les classifieurs bayésiens offrent la flexibilité de prendre en compte la probabilité de survenue d'un événement connu a priori. Ils utilisent à cette fin une base de connaissances comme la base COSMIC (Catalogue of Somatic Mutations in Cancer), leur permettant d'affiner les résultats en fonction d'éléments tels que la position de variants répertoriés en fonction de types de cancer, d'erreurs de séquençages ou d'alignement apparaissants systématiquement sur certaines localisations.

Déterminer précisément les fréquences alléliques de chaque sous-clone permet une meilleure inférence des clones et offre une perspective de détermination de la phylogénie tumorale (Stead et al., 2013). Les performances des outils dépendent de la profondeur de couverture (nombre de reads par position génomique) qui elle-même dépend du nombre d'échantillons passés dans chaque run de séquençage. Il est recommandé en génétique somatique en contexte de routine clinique de dépasser une profondeur de 100X et fréquent d'atteindre des profondeurs de couverture de 500 à 1000X (quand une profondeur de 30X voire moins est acceptable en génétique constitutionnelle).

Résultats

Nous avons utilisé plusieurs types de données :

— Données simulées de reads Illumina d'ADN génomique : elles intègrent des variations (substitutions et insertions/délétions) à différentes profondeurs.
— Données de séquençage d'ADN génomique de patients atteints ou non de syndrome myélodysplasique.
— Données de séquençage d'ADN du gène TP53 de type Illumina ou PacBio.

Dans un premier temps, nous avons évalué les outils sur les données simulées pour déterminer leurs capacités à détecter les différents types de mutations. Les paramètres mesurés ont été la taille mémoire utilisée, le temps de calcul et la sensibilité et la spécificité des résultats.

Nous avons établi des profils d'utilisation optimum en fonction des données issues des grands consortium et des méthodes d'analyses sélectionnées.

En fonction des résultats obtenus, nous avons recherché dans les données de patients la présence ou non de sous-clones tumoraux.

Les analyses bioinformatiques issues de ces travaux permettent au praticien d'adapter les traitements en fonction du profil mutationnel de chaque patient dans une démarche de médecine personnalisée.

Remerciements

Je remercie pour l'accueil et l'encadrement l'équipe de Thèrèse Commes - Bioinformatique et BioMarqueurs (IRMB - U1183).

Je remercie l'Institut de Biologie Computationnelle de Montpellier (CNRS Université de Montpellier) pour le financement du stage.

Et enfin, je remercie le département informatique de la Faculté des Sciences de l'Université de Montpellier (`http://deptinfofds.univ-montp2.fr/`) ainsi que le Labex Numev (`http://www.lirmm.fr/numev/`) pour le financement du congrès.

Références

Leroy B, Girard L, Hollestelle A, Minna JD, Gazdar AF, Soussi T. Analysis of TP53 mutation status in human cancer cell lines: a reassessment. *Hum Mutat*, 35(6):756-65, June 2014.

Lane D, Levine A. p53 Research: the past thirty years and the next thirty years. *Cold Spring Harb Perspect Biol*, 2(12), December 2010.

Wei S, Chen X, McGraw K, Zhang L, Komrokji R, Clark J, Caceres G, Billingsley D, Sokol L, Lancet J, Fortenbery N, Zhou J, Eksioglu EA, Sallman D, Wang H, Epling-Burnette PK, Djeu J, Sekeres M, Maciejewski JP, List A. Lenalidomide promotes p53 degradation by inhibiting MDM2 auto-ubiquitination in myelodysplastic syndrome with chromosome 5q deletion. *Oncogene*, 32(9):1110-20, February 2013.

Jädersten M, Saft L, Smith A, Kulasekararaj A, Pomplun S, Göhring G, Hedlund A, Hast R, Schlegelberger B, Porwit A, Hellström-Lindberg E, Mufti GJ. TP53 mutations in low-risk myelodysplastic syndromes with del(5q) predict disease progression. *J Clin Oncol.*, 20;29(15):1971-9, May 2011.

Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L, Wan Y, Zhang W, Shukla SA, Vartanov A, Fernandes SM, Saksena G, Cibulskis K, Tesar B, Gabriel S, Hacohen N, Meyerson M, Lander ES, Neuberg D, Brown JR, Getz G, Wu CJ. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell.*, 14;152(4):714-26 February 2013.

Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, 43, January 2015.

Niko Beerenwinkel, Roland F. Schwarz, Moritz Gerstung, and Florian Markowetz. Cancer evolution : mathematical models and computational inference. *Systematic Biology*, 64(1):e1–25, January 2015.

Moritz Gerstung, Christian Beisel, Markus Rechsteiner, Peter Wild, Peter Schraml, Holger Moch, and Niko Beerenwinkel. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, 3:811, 2012.

Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2 : somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, March 2012.

Andreas Wilm, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. LoFreq : a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, 40(22):11189–11201, December 2012.

Kristian Cibulskis, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, March 2013.

Lucy F. Stead, Kate M. Sutton, Graham R. Taylor, Philip Quirke, and Pamela Rabbitts. Accurately Identifying Low-Allelic Fraction Variants in Single Samples with Next-Generation Sequencing : Applications in Tumor Subclone Resolution. *Human Mutation*, 34(10):1432–1438, October 2013.

Mots clefs : NGS, TP53, Evolution sous-clonale, appel de variants, variant calling, variant caller, Syndromes myélodysplasiques, classifieurs bayésiens

Galaxy for diagnosis in a multicenter laboratory

Poster 7

Christophe Habib [*1,2], Bruno Francou[2,3], Ivan Sloma[4,5], Alexis Proust[2], Julien Nauroy[6], Anne Spraul Davit[7], Thierry Naas[8], Claire Deback[9], Jacques Young[3,10], Agnès Linglart[1,11], Anne Guiochon-Mantel[2,3], Jérôme Bouligand[2,3]

[1] Plateforme d'Expertises maladies rares Paris-Sud (PEPS) – Assistance publique - Hôpitaux de Paris (AP-HP) – France
[2] Génétique Moléculaire, Pharmacogénétique et Hormonologie, Hôpital Bicêtre (GMPH-HUPS) – Assistance publique - Hôpitaux de Paris (AP-HP) – France
[3] U1185, Faculté de médecine, Univ. Paris-Sud, Inserm, Le Kremlin-Bicêtre – Université Paris-Saclay – France
[4] Unité d'Onco-Hématologie Moléculaire et Cytogénétique, Hôpital Paul Brousse – Assistance publique - Hôpitaux de Paris (AP-HP) – France
[5] U935, Faculté de médecine, Univ. Paris-Sud, Inserm, Villejuif – Université Paris-Saclay – France
[6] Direction Informatique - Pôle Infrastructures Systèmes et Applications Critiques – Université de Paris-Sud Orsay – France
[7] Biochimie, Hôpital Bicêtre (HUPS) – Assistance publique - Hôpitaux de Paris (AP-HP) – France
[8] Microbiologie, Hôpital Bicêtre (HUPS) – Assistance publique - Hôpitaux de Paris (AP-HP) – France
[9] Virologie, Hôpital Paul Brousse (HUPS) – Assistance publique - Hôpitaux de Paris (AP-HP) – France
[10] Endocrinologie Adultes, Hôpital Bicêtre (HUPS) – Assistance publique - Hôpitaux de Paris (AP-HP) – France
[11] Endocrinologie et diabétologie de l'enfant, Hôpital Bicêtre (HUPS) – Assistance publique - Hôpitaux de Paris (AP-HP) – France

Context

Our laboratory of Molecular Genetics, Pharmacogenetics and Hormonology (GMPH) performs diagnosis of several genetic disorders in the fields of Endocrinology, Neurology and Hepatology. Biomolecular investigations for one patient often require hundreds of double stranded Sanger sequencing to eventually find the genetic event leading to the pathology. This process could take several years. We recently acquired a NGS sequencer (Illumina Miseq) to perform targeted exome sequencing and accelerate diagnosis. This new technology raises new challenges such as the processing and the storage of gigabytes of data. Plus, bioinformatics analyses need to achieve high standards according to requirement for quality and competence in medical laboratories (EN ISO15189: 2012). Many laboratories externalized this crucial analytical step to a service provider. This externalization involves questionable responsibility transfer and a supplemental cost. Our laboratory chosed to acquire its own server to master the bioinformatics analyses on a Galaxy instance dedicated to diagnosis. This in-house solution is suitable for a multisite hospital (Kremlin-Bicêtre Hospital, Paul Brousse Hospital and Béclère Hospital) context. Indeed, several departments are performing their analyses on this instance.

Method

We acquired a HP server with 16 cores and 283 Go of RAM running under Debian OS. The Galaxy platform follows the best practices of the galaxy team with uwsgi web server, nqqginx proxy and HTCondor as scheduler. This instance is only accessible and secured through the hospital's intranet. All the sequence files (fastq) are automatically loaded into the data libraries and each

*. Intervenant

lab has access to its own data only. Each pathology has a dedicated workflow in accordance with the GATK best practices of the Broad Institute and department specifications. Workflows were designed and validated by a bioinformatics engineer and concerned biologists.

Results

Our galaxy instance is running since July 2014. It is used by 8 biologists and 6 technicians in 4 departments, dealing with Mendelian diseases but also Cancer and Microbiology. For Orphan diseases, we analyzed the data of more than 500 patients in 9 distinct panels of genes since its launch. Using NGS plus Galaxy for bioinformatics analyses greatly improved the probability to detect a genetic event in relation with the disease. For instance, in GMPH, the percentage of patients with identified pathogenic variants has increased from < 20 % to almost 50 % for congenital hypogonadotropic hypogonadism (CHH) condition. Sensibility was higher due to full transcripts analyses and broader panels of genes investigated. Sanger sequencing confirmed SNVs events identified by NGS and bioinformatics in house process in 100 % of cases. False positive rate is therefore at 0 %. Galaxy gives the great opportunity to keep tools used in our analyses up to date. For instance, the Genome Analyses Toolkit (GATK) wrapper available on the toolshed was updated from version 1.7 to 3.4 within two years. Pathology specific workflow guaranties quality and reproducibility of analyses. Each workflow version can be easily compared to the previous one for validation and improvements assessment. Indeed all workflows parameters and tools versions can be exported and archived with raw data. Plus, the metadata (e.g. time, owner of analysis, and so on) are saved in a postgresql database. The scheduler HTCondor has proved to be necessary to avoid overload of the server and optimize the use of resources in a multi-user context. The sequences generated by one MiSeq run (48 hours on average) for 24 patients are analyzed by batch in 40 minutes.

Conclusion

Galaxy allows us to perform diagnosis with high sensitivity and specificity. The verifications by gold standard methods (e.g. Sanger sequencing) is a strong point to validate our new bioinformatics workflows for ISO 15189 accreditation. Galaxy gives many opportunities to improve constantly the methods thanks to its scalability which guarantees to easily use up-to-date tools. This platform provides the traceability and reproducibility required in a genetic and clinical context over years. Galaxy also demonstrates its capacity to answer quickly in emergency context. Moreover it can take over increasing users since it did not reach its limits in terms of computational power. According to these results, Galaxy proved to be a reliable and valuable platform for in-house bioinformatics dedicated to NGS and medical diagnosis in a multicenter laboratory.

Mots clefs : Galaxy, Diagnosis, Clinical, Medical, NGS, Accreditation, EN ISO15189

Discovery of epigenetically regulated genomic domains in lung cancer

Marugan Jose Carlos [*1], Daniel Jost [†1]

Poster 8

[1] Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG) – CNRS : UMR5525, Université Joseph Fourier - Grenoble I – Domaine de la Merci, F-38 706 La Tronche, France

Cell differentiation is associated with the establishment of specific patterns of cell-type and tissue-specific gene expression, which largely rely on what is called "epigenetic landscaping" of the genome, a process which is itself largely dependent on changes in DNA methylation and post translational histone modifications. In differentiated cells, these epigenetic mechanisms not only help activating and maintaining specific gene expression patterns but also control a genome-wide repression of tissue-specific genes. Failure of preserving the proper epigenetic mark profiles might result in inappropriate gene activity and diseases like cancer [Portela2010].

Although cancer is mainly caused by diverse mutations and genetic alterations, recent investigations have demonstrated a pivotal role also played by epigenetics in cancer initiation, progression and treatment ([Baylin2011, Sandoval2012, Timp2013] for reviews). In particular, global deregulation of epigenetic signaling is an early and recurrent event that occurs during oncogenic cell transformation and, in collaboration with mutation, might provide the phenotypic variation allowing selection of the hallmarks properties of cancer such as rapid proliferation or genomic instability. Aberrant gene activity is a direct consequence of these epigenetic anomalies. For example, epigenetic-associated repression of tumour suppressor genes in cancer cells is well documented and now recognized as an important oncogenic event [Baylin2011]. A less studied consequence of epigenetic deregulation in transformed cells is the illegitimate activation of various cell- and tissue-specific genes [Rousseaux2013, Bert2013]. As it becomes more and more clear that epigenetics is central in cancer, a major challenge of cancer systems biology is to understand how epigenetic deregulation effectively affects gene expression and predicts prognosis. In this project, we propose to address this point by developing an integrated study of genomic, transcriptomic and epigenomic alterations specifically in lung cancers. Lung cancer is the most common form of cancer in the world today (12.6 % of all new cancers, 17.8 % of cancer deaths). The number of new cases is over 1.2 million per year, with 1.1 million deaths [Jemal2011]. Today in Europe it kills more people than the combined total numbers from breast, colon and prostate cancers. Trends over time show that, overall the incidence of lung cancer is steadily increasing in both genders in many recently industrialized countries, including France, hence becoming a major axis of public health policies.

In this poster, our goal is to identify a set of genes whose variability in gene expression between lung cancer and normal lung cells are remarkable, and where this variability cannot be attributed to the copy number variation (CNV) by amplification or deletion. For this purpose, we focus on two datasets: tumorous vs normal lung tissue data taken from The Cancer Genomic Atlas consortium ([TCGA2012]) and lung cancer cell line data taken from [Suzuki2014].

The first step was to build an efficient bioinformatic pipeline in order to analyze uniformly raw RNAseq data from the various datasets. This pipeline is made of 4 units:

— Quality control performed by FastQC [Babraham2015] associated to trimming of low score reads made by fastq-mcf [Aronesty2011].

*. Intervenant

†. Corresponding author : daniel.jost@imag.fr

— Alignment of the paired reads on the human reference genome (GRChr37) using Tophat2 [TopHat2008] and Bowtie [Bowtie2008].
— Association and counting of sequence reads to human transcripts (GTF file from ENSEMBL) using HTSeq [Anders2010].
— Normalization of the counts with two different methods: DESeq [AndWolf2010] and TMM from edgeR [Robinson2010].

We have validated our pipeline in every step with different options and parameters, and we have compared the different results in order to obtain the most reliable ones to our study. Overall, we observe a 15-20 % of reads due to non-unique alignment, absence of features or ambiguity in the transcript reference file.

Using the normalized data, we performed global supervised and unsupervised clustering analysis for the two different datasets (tumorous vs normal tissues, and model cell lines). Strikingly, for tumorous vs normal tissues, we find that obtained clusters reflect almost perfectly the histology of the tumors. For cell lines, we observe two main subgroups that are associated with deregulation of specific gene sets. Identically, performing clustering based on a subset of genes (germline ectopic genes), we observe that resulting clusters are strongly associated with prognosis: expression of many ectopic genes is strongly correlated with low survival rate.

The second step was to performed differential analysis for the two datasets in order to identify genes that are deregulated in lung cancers. Using DESeq and edgeR, we build a robust list of differentially expressed genes. Enrichment analysis using GSEA [Tamayo2005] or ConsensuspathDB [Atanas2011] shows that this list is enriched in immune, signaling, regulatory or cell division pathways or gene ontologies, and also in modules associated with (lung) cancer or ectopic activity. We then concentrate of the location of this gene set along the genome and observe that, while largely spread, some clusters are observed associated with co-regulated genes.

Finally, we identify among this list, genes whose deregulation is not simply product of the variation in the number of copies of the genes. For this we infer the copy number for each genes using Control-FREEC [Boeva2011]. Knowing the copy-number variations (CNV) for each gene in each sample, we were able to identify a small number of genes that always exhibits deregulation in absence of CNV in some dataset subgroups.

In conclusion, we have identified genomic regions that exhibit gene expression deregulation not associated with CNV. These regions are good candidates for epigenetically regulated regions that play an important role in lung cancer. The next step of our study would be to characterize the epigenomic landscape of these regions in normal and tumorous tissues in order to define a epigenomic signature of regions that are susceptible to be affected by epigenetic deregulation mechanisms in cancer.

References

[Baylin2011] S.B. Baylin, P.A. Jones (2011) A decade of exploring the cancer epigenome – biological and translational implications. *Nat. Rev. Cancer* 11:726-734.

[Bert2013] S.A. Bert et al. (2013) Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell* 23:9-22.

[Jemal2011] A. Jemal et al. (2011) Global cancer statistics. *CA Cancer J. Clin.* 61:69-90.

[Portela2010] A. Portela, M. Esteller (2010) Epigenetic modifications and human disease. *Nat. Biotech.* 28:1057-1068.

[Rousseaux2013] S. Rousseaux, et al. (2013) Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med.* 5:186r66.

[Sandoval2012] J. Sandoval, M. Esteller (2012) Cancer epigenomics: beyond genomics. *Curr. Opin. Genet. Dev.* 22:50-55.

[Suzuki2014] A. Suzuki et al. (2014) Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma lung cell lines. *Nucleic Acids Res.* 42:13557-13572.

[Timp2013] W. Timp, A.P. Feinberg (2013) Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat. Rev. Cancer* 13:497-510.

[TCGA2012] The Cancer Genomic Atlas (2012) Research Network: `http://cancergenome.nih.gov/`.

[Babraham2015] Babraham Bioinformatics (2015), Babraham Institute. FastQC, a quality control tool for high throughput sequence data. Retrieved from: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[Aronesty2011] Erik Aronesty (2011). ea-utils : "Command-line tools for processing biological sequencing data"; `http://code.google.com/p/ea-utils/`.

[TopHat2008] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S, Pachter L. (2008). Jonhs Hopkins University, Center for Computational Biology (CCB), "Fast splice junction mapper for RNA-Seq reads"; `https://ccb.jhu.edu/software/tophat/index.shtml`.

[Bowtie2008] Langmead B, Kim D, Antonescu V, Wilks C (2008) Jonhs Hopkins University, "Ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences"; `http://bowtie-bio.sourceforge.net/bowtie2/index.shtml`

[Anders2010] Anders S (2010), at EMBL Heidelberg (Genome Biology Unit), "Python package that provides infrastructure to process data from high-throughput sequencing assays"; `http://www-huber.embl.de/HTSeq/doc/overview.html`

[AndWolf2010] Simon Anders and Wolfgang Huber (2010): Differential expression analysis for sequence count data. *Genome Biology* 11:R106.

[Robinson2010] Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139-140.

[Tamayo2005] Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003). *Nat Genet* 34:267-273.

[Atanas2011] Atanas Kamburov, Konstantin Pentchev, Hanna Galicka, Christoph Wierling, Hans Lehrach, Ralf Herwig (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Research* 39(Database issue):D712-717.

[Boeva2011] Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, Barillot E. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* 27(2):268-9. PMID: 21081509.

Mots clefs : epigenetics, lung cancer, deregulation, RNAseq, clustering, differential analysis

MutaScript : a mutational score for each coding transcript as a new module for exome data filtering

Poster 9

Thomas Karaouzene *[1,2], Nicolas Thierry-Mieg[1], Pierre Ray[2]

[1] Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG) – CNRS : UMR5525, Université Joseph Fourier - Grenoble I – Domaine de la Merci, F-38 706 La Tronche, France

[2] Génétique, Épigénétique et Thérapies de l'Infertilité (GETI) – Inserm, CNRS : UMR5309 – Institut Albert Bonniot, Université Grenoble Alpes, Grenoble, France

Recent technological advances have made it economically feasible to sequence exomes or whole genomes of large numbers of individuals. Whole-exome sequencing (WES) and whole-genome sequencing (WGS) analyses powered by next-generation sequencing (NGS) have become invaluable tools to identify causal mutations responsible for Mendelian diseases. However, with more than 30 000 single nucleotide variants and short insertions/deletions found per exome, interpretation and filtering of extensive variant lists is currently a labor-intensive and error-prone task, and hampers the implementation of WES in genetic research. NGS data analysis can be subdivided into three sequential subtasks. The first task comprises quality control of the raw sequencing reads and the mapping of the trustworthy reads to a reference genome. The second task uses the data obtained during the mapping stage of the analysis to identify or "call" the variants corresponding to the differences between the analyzed genome and the reference genome. The last task is the variants' interpretation in relationship with the patient's phenotype. This step mostly consists of annotating variants, for example predicting their impact on protein sequence and functionality, and filtering those deemed not relevant for the considered analysis. It thus remains a challenge to analyze large numbers of variants to identify causal alleles associated with inherited conditions. As a result, new analysis tools are required to handle the large sets of genetic sequencing data and to increase the proportion of variants flagged as benign that can be automatically filtered out from the geneticist's working list. Some free software is already available like BiERapp or exomeSuite to name a few. These software provide solutions for variant annotation, interpretation and prioritization. However, all of them are focused on annotating variants (and variants only).

Here we are developing an algorithm to annotate transcripts rather than variants. The idea is to calculate a score for each coding transcript based on its "mutational variability": the MutaScript score. Our key hypothesis is that transcripts with a high variability in the general population are unlikely to be involved in critical functions, and have thus a low probability of being causal for a severe mendelian disease. This study is possible due to the existence of large public datasets gathering exome data harvested from numerous individuals. We are using the data provided by the Exome Aggregation Consortium (ExAC), which aggregates exome data from 61,000 individuals. To keep only reliable variants, in addition to ExAC quality filters we removed all the variants having a median coverage < 20x. Remaining variants were annotated using the software Variant Effect Predictor (VEP), which predicts the effect of every variant on transcripts and protein sequence on the Ensembl v75 transcriptome. VEP provide a detailed description of the effect of the variants and classifies them into four categories corresponding to the estimated impact of the variants on the protein. These four impacts are "HIGH", "MODERATE", "LOW" and "MODIFIER": variants with a HIGH impact are expected to be strongly deleterious (e.g. nonsense variants), while the MODIFIER class comprises variants that are likely benign, such as intronic variants.

*. Intervenant

The MutaScript score takes into account the allele counts of all ExAC variants located in the candidate transcript, and each variant is weighted according to its VEP impact. Thus the highest scores are obtained by transcripts with many, frequent, high-impact variants: these transcripts are most likely not involved in severe pathologies. To design and test our scoring model, we use the "Human Phenotype Ontology" (HPO). This ontology provides a standardized vocabulary of phenotypic abnormalities encountered in human diseases. Each HPO term describes a phenotypic abnormality. Thus, for example, we are expecting genes flagged by the "Death" term (HP:0011420) in HPO to have low variability and so a low MutaScript score.

Once finalized, MutaScript should prove useful to annotate coding transcripts and thus to systematically filter high scoring transcripts. It could also be used to prioritize transcripts with the lowest scores, as these are good candidates for causing severe pathologies. Overall we believe that MutaScript score will be very helpful for phenotype / genotype analyses of diseases with mendelian transmission using WES and WGS data.

Mots clefs : Séquençage haut, débit, Analyse bio, informatique, filtrage des données, exome, transcrits

Caractérisation et analyse bio-informatique d'un réseau multi-échelle dans le cancer de la prostate : co-expression génique, mutome, interactome…

Aurélie Martin [*] [†1], Laurent Naudin[1], Sébastien Tourlet [*] [‡1]

Poster 10

[1] IPSEN Innovation – 5 avenue du Canada, F-91 940 Les Ulis, France

Les technologies d'étude du transcriptome à large échelle (*e.g.* puces à ADN, RNA-seq) permettent de générer simultanément un grand nombre d'informations sur les niveaux d'expression génique. L'analyse de grandes quantités de données d'expression obtenues dans différents tissus ou différentes conditions expérimentales permet d'établir des réseaux de relations (*e.g.* co-expression) unissant des groupes de gènes. Un des enjeux principaux réside dans l'analyse de ces réseaux d'expression, tant au niveau topologique (*e.g.* structure générale du réseau, identification des zones fortement connectées), qu'au niveau descriptif (*e.g.* définition des méta-données liées aux expériences et aux échantillons). La méthode présentée ici permet de construire un réseau de co-expression spécifique à une maladie, le cancer de la prostate, par contextualisation d'un réseau global représentatif de l'ensemble des puces à ADN publiées pour l'espèce humaine. L'analyse de ce réseau composé de 6 585 gènes avec 4 mesures de centralité qui sont la « degree centrality », la « betweenness centality », la « closeness centrality » et le « clustering coefficient » permet d'identifier 506 gènes d'intérêts. Dans cette étude, nous nous intéressons plus particulièrement aux gènes codants pour des protéines de type facteurs de transcriptions (TF) ou récepteurs couplés aux protéines G (GPCR). Nous retrouvons ainsi des gènes déjà connus pour jouer un rôle important dans la genèse et le développement du cancer de la prostate.

Mots clefs : Réseau multi échelle, puces à ADN, mesures de centralité, classification, annotation, voies cellulaires, biomarqueurs, gènes cibles, cancer de la prostate, logiciel R, Pipeline Pilot, bioinformatique

*. Intervenant
†. Corresponding author : aurelie.martin@ipsen.com
‡. Corresponding author : sebastien.tourlet@ipsen.com

Evaluation of integrative approaches for the analysis of multi-omics data

Alexei Novoloaca [*] [†1], Frédéric Reynier[1], Jérémie Becker[*1]

[1] BIOASTER – ANR – 40 avenue Tony Garnier, F-69 007 Lyon, France

Poster 11

Recent advances in sequencing and mass spectrometry technologies have enabled biomedical researchers to collect large-scale omics data from the same biological samples. The huge volume of data generated by these technologies represents an opportunity in term of fundamental research by improving the annotation of genome [1] and the identification of coordinated cellular processes at different omics layers [2]. Industrial applications are also to benefit from these technologies by increasing diagnostic and prognostic accuracy for instance [3]. The underlying idea behind these applications is that cellular signals and processes depend on the coordinated interaction among a variety of biomolecules, making their joint analysis crucial to provide a comprehensive view at the system level. In this context, BIOASTER, the first French Technology Innovation Institute in Microbiology, offers multi-omics applications to its industrial partners to help them decipher the mode of action of drugs and characterize biological processes associated with particular phenotype (protein processing, immune response, etc.). However, the complexity of biological processes, the technological biases and the "large p, small n" problem makes integrative analyses not straightforward. To address these challenges, recent statistical approaches have been introduced to capture both data specific variations and joined variations across data types (omics) [4].

To identify the methods that suit best BIOASTER projects, a thorough benchmarking was performed on both simulated and real data. A selection of supervised and unsupervised methods were tested, the latter including Bayesian and multivariate approaches. To avoid favouring one category over the other, two simulation strategies were developed using either one of the studied Bayesian model or one of the multivariate approach. In the first strategy, data were simulated according to multivariate Gaussian distributions. In the second, block diagonal structures were generated by carefully simulating shared latent variables. In both cases, data specific and shared variations were introduced with different signal to noise ratios. The results showed that all the methods recover the shared structure with a high accuracy, but only a handful could identify data specific clusters. Finally, when run on a published dataset, the best of the tested methods led to pathways consistent with the original paper. To our knowledge, no such comparison on integrative approaches has been performed yet.

References

[1] Wang, J., Qi, M., Liu, J., & Zhang, Y. (2015). CARMO: a comprehensive annotation platform for functional exploration of rice multi-omics data. *The Plant Journal*, 83(2):359-374.

[2] Bartel, J., Krumsiek, J., Schramm, K., Adamski, J., Gieger, C., Herder, C., … & Strauch, K. (2015). The human blood metabolome-transcriptome interface. *PLoS Genetics*, 11(6):e1005274.

[3] Bonne, N. J., & Wong, D. T. (2012). Salivary biomarker development using genomic, proteomic and metabolomic approaches. *Genome medicine*, 4(10):1-12.

[4] Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics,* 17(2):167.

*. Intervenant

†. Corresponding author : alexei.novoloaca@etu.univ-lyon1.fr

Mots clefs : Integrative methods, multiomics

Data mining des réseaux de signalisation dans le cancer de la vessie : pipeline automatisé de traitement de données RNA-Seq, et analyse par la théorie des graphes

Poster 12

Mouna Rouass [*] [†1], Aurélie Martin [*] [‡1], Laurent Naudin[1], Sébastien Tourlet [*] [§1]

[1] IPSEN Innovation – 5 avenue du Canada, F-91 940 Les Ulis, France

Les données massives de RNA-Seq permettent une appréhension plus fine du transcriptome et une analyse détaillée des changements moléculaires impliqués dans les cancers. L'étude est basée sur les données publiques disponibles sur la base The Cancer Genome Atlas (TCGA), qui répertorie les informations cliniques et génomiques de 34 cancers. Les données brutes du cancer de la vessie sont traitées par un pipeline automatisé visant à normaliser et filtrer (se dégager du bruit de fond) afin d'obtenir des données d'expression géniques robustes. Cette méthodologie repose sur l'unification, l'optimisation et l'association séquentielle de méthodes de normalisation préexistantes. Ensuite, un sens biologique doit être apposé à ces données. Ainsi, les données normalisées d'expression RNA-Seq doivent être mises en relief dans un réseau multi-échelle pour comprendre les mécanismes moléculaires et biologiques sous-jacents du cancer vessie. Les valeurs normalisées sont traitées par des techniques d'analyse basées sur la théorie des graphes associant les données d'expression RNA-Seq à d'autres données (interactions fonctionnelle, physique). Le pipeline d'analyse et de filtrage permet d'identifier 1 740 gènes qui sont ensuite analysés dans les réseaux biologiques par la théorie des graphes.

Mots clefs : NGS, RNA Seq, Pipeline Pilot, logiciel R, bioinformatique, réseau multi échelle, théorie des graphes, cancer de la vessie

*. Intervenant
†. Corresponding author : mouna.rouass@ipsen.com
‡. Corresponding author : aurelie.martin@ipsen.com
§. Corresponding author : sebastien.tourlet@ipsen.com

iSeGWalker : an easy handling *de novo* genome reconstruction dedicated to small sequence

Benjamin Saintpierre [*1], Johann Beghain [†1], Éric Legrand [‡1], Anncharlott Berglar[2], Deshmukh Gopaul[2], Frédéric Ariey [§1]

Poster 13

[1] Génétique et Génomique des Insectes Vecteurs (GGIV) – Institut Pasteur de Paris, CNRS : URA3012 – 25-28 rue du Docteur Roux, F-75 724 Paris Cedex 15, France
[2] Plasticité du Génome Bactérien (PGB) – CNRS : URA2171, Institut Pasteur de Paris – 25 rue du docteur Roux, F-75 724 Paris Cedex 15, France

Most of "de novo softwares" are global assemblers, meaning they work on the assembling of all reads from a sequencing file. They are not adapted to get a short sequence as the non nuclear DNA from a pathogen. Here we present a Perl software, iSeGWalker (in silico Seeded Genome Walker), developed to accomplish a quick *de novo* reconstruction of a region, by "genome walking" on Next Generation Sequencing (NGS) data.

The first step of the process is to determine an initial seed, which must be unique and specific to the targeted region. The second step is a cycle with the seed search step (an exact-matching reads selection), the alignment of all selected reads and the generation of a consensus sequence. Once the consensus obtained, a new seed, composed by the 30 last consecutive nucleotides, is obtained and a new cycle is performed.

We tested our software using an apicoplaste seed on a Fastq file obtained from Illumina's *Plasmodium falciparum* 3D7 reference strain sequencing. We were able to identify the whole complete genome of the apicoplast including a non-published region harboring a balanced polymorphism that may have a function in the regulation and/or division of the *Falciparum* apicoplast genome.

Mots clefs : NGS, Illumina, Perl, de novo, *Plasmodium*

*. Intervenant
†. Corresponding author : johann.beghain@wanadoo.fr
‡. Corresponding author : eric.legrand@pasteur.fr
§. Corresponding author : frederic.ariey@yahoo.fr

iFiT : an integrative bioinformatics platform for biomarker and target discovery. A case study in neuroendocrine tumors

Poster 14

Sébastien Tourlet [*] [†1], Frederic Scaerou[1], Aurélie Martin [*] [‡1], Arunthi Thiagalingam[2], Isabelle Paty[1], Laurent Naudin[1], Philip Harris[3]

[1] IPSEN Innovation – 5 avenue du Canada, F-91 940 Les Ulis, France
[2] Ipsen Bioscience, Inc. – Cambridge, États-Unis
[3] IPSEN Developments Ltd – Slough/Oxford, Royaume-Uni

iFiT (Ipsen Focused-on-new bIological enTities and biomarkers) is a Bioinformatics platform integrating systems biology functionalities together with semantic & logic-based artificial intelligence within a high-scale computing environment.

Key applications are the discovery of potential therapeutic targets as well as the identification of patient stratification candidate biomarkers.

Given the limited OMICs characterization of neuroendocrine tumors, the identification of driver genes and pathways is challenging: To help circumvent this paucity of molecular information, iFiT was built on the postulate that co-expressed genes participate in the same biological processes. Furthermore, we fed the platform with curated heterogeneous datasets, pre-clinical and clinical, including molecular and phenotypic information.

We focused our search on drugable GPCRs and microRNAs involved in mechanisms such as pancreas islet cells lineage, differentiation, multiplication and hormone secretion. As a result, we identified 42 GPCRs and 10 microRNAs, including well-known NETs-associated genes such as SSTR2 and DRD2. iFiT predicted the driver role of SSTR2 in both proliferation and secretion before the release of the CLARINET study (ESMO 2013). Remarkably, 90 % of candidate genes were validated on tumor tissues from 40 GEP-NET patients.

In conclusion, iFiT achieves an excellent detection rate, and is proving suitable to uncover hidden information and mine translational knowledge in NET.

Mots clefs : target discovery, therapeutic target, systems biology, translational bioinformatics platform, logic, based artificial intelligence, R software, Pipeline Pilot

*. Intervenant
†. Corresponding author : sebastien.tourlet@ipsen.com
‡. Corresponding author : aurelie.martin@ipsen.com

iTox : prediction of toxicity using system's biology approaches on the biological target profile

Sébastien Tourlet [*] [†1], Aurélie Martin [*] [‡1], Laurent Naudin[1]

[1] IPSEN Innovation – 5 avenue du Canada, F-91 940 Les Ulis, France

Poster 15

Adverse effects of synthetic peptides/proteins are mainly due to exacerbated biological target effects. Therefore, it is relevant to develop a tool and a knowledge-driven approach focused on biology systems rather than on chemical structure knowledge data. Classical network computational biology is often based on cell line drug effects. However, with the existing open data strategies, tissue specific drug side effects can be inferred from tissue-specific networks and can emphasize the relevance of exacerbated effects of desired and undesired therapeutic targets. iTox is a technology providing predictive toxicity alerts based on analysis of exacerbated biological effects. It allows to compute potential adverse effects, along the drug development (CA, LI, and due diligences). In conclusion, the technology no need to generate new experimental data to predict targets and off-targets related toxicity alerts.

Mots clefs : predictive toxicology, drug related adverse events, in silico prediction, exacerbated biological effects, text mining, bioinformatics, Pipeline Pilot

*. Intervenant

†. Corresponding author : sebastien.tourlet@ipsen.com

‡. Corresponding author : aurelie.martin@ipsen.com

Single cell profiling of pre-implantation mouse embryos reveals fully Xist-dependent paternal X inactivation and strain-specific differences in gene silencing kinetics

Poster 16

Maud Borensztein[1], Laurène Syx[1,2], Katia Ancelin[1], Patricia Diabangouaya[1], Tao Liu[3], Jun-Bin Liang[3], Ivaylo Vassilev [*1,2], Nicolas Servant[2], Emmanuel Barillot[2], Azim Surani[4], Chong-Jian Chen[3], Edith Heard[1]

[1] Institut Curie, PSL Research University, Genetics and Developmental Biology Unit, INSERM U934/CNRS UMR3215, F-75 005 Paris, France
[2] Institut Curie, PSL Research University, Mines Paris Tech, Bioinformatics and Computational Systems Biology of Cancer, INSERM U900, F-75 005 Paris, France
[3] Annoroad Gene Technology Co., Ltd, Beijing, China – Chine
[4] Wellcome Trust Cancer Research UK Gurdon Institute, Department of Physiology, Development and Neuroscience, Wellcome Trust-MRC Stem Cell Institute, University of Cambridge – Royaume-Uni

During early mammalian development, X-chromosome inactivation (XCI) is established in female embryos. One of the two X chromosomes is converted from an active into an inactive state to ensure X-linked gene dosage compensation between females (XX) and males (XY) [1]. XCI is triggered *in cis* by the non-coding RNA *Xist*, which coats the inactive X. In mice, *Xist* is first expressed only from the paternal allele leading to preferential silencing of the paternal X (Xp). Reactivation followed by random XCI then occurs in the embryo proper, when the extra-embryonic tissues will conserve the Xp silenced [2,3]. However the early Xp inactivation has been extensively investigated, the precise chromosome-wide dynamics is unknown and the importance of *Xist* remains questionable. To address these issues, we have performed 178 single-cell RNA sequencing analysis (scRNAseq) of F1 hybrid embryos (oocytes to 64-cell stages) derived from reciprocal crosses between highly polymorphic strains *Mus musculus castaneus* (Cast/EiJ) and *Mus musculus domesticus* (C57BL6/J). A bioinformatic pipeline was developed to analyse the allele-specific gene expression and thus, follow the Xp transcriptional dynamics through several developmental stages. We have shown that paternal XCI is strictly dependent of *Xist* and its absence leads to genome-wide transcriptional misregulation and lack of extra-embryonic pathway activation. Here, we demonstrate that efficient dosage compensation is a prerequisite to correct female pre-implantation development. This study provides us with important insights into the transcriptional and allelic dynamics of the X chromosome and enhances our knowledge of the very first stages of mammalian development.

References

[1] Lyon M.F. (1961) Gene Action in the X-chromosome of the Mouse (*Mus musculus* L.). *Nature* 190:372–373

[2] Okamoto I et al. (2004) Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science* 303:664-669

[3] Okamoto I et al. (2005) Evidence for *de novo* imprinted X-chromosome inactivation independent of meiotic inactivation in mice. *Nature* 17:369-373.

*. Intervenant

Mots clefs : single, cell, X, embryo, mice, scRNAseq, allele, specific, pipeline, bioinformatic, Xist, inactivation

CloSeRMD : clonal selection for rare motif discovery

Salma Aouled El Haj Mohamed [*1,2,3], Julie Thompson[1], Mourad Elloumi[2]

Poster 17

[1] Laboratoire des sciences de l'ingénieur, de l'informatique et de l'imagerie (ICube) – Université de Strasbourg, CNRS : UMR7357 – 300 boulevard Sébastien Brant, BP 10413, F-67 412 Illkirch Cedex, France (Alsace)
[2] Université de Tunis. Laboratoire des Technologies de l'Information et de la Communication et du génie Electrique (LaTICE) – LR11ES04, 1008, Tunis, Tunisie
[3] Université de Tunis El Manar. Faculté des Sciences de Tunis, École doctorale mathématiques, informatique, sciences et technologie de la matière – Campus de Tunis El Manar. Faculté des Sciences de Tunis. 1068, Tunis, Tunisie

Biological sequence motifs are defined as short, usually fixed length, sequence patterns that may represent important structural or functional features in nucleic acid and protein sequences such as transcription binding sites, splice junctions, active sites or interaction interfaces. They occur in an exact or approximate form within a family or a subfamily of sequences. Motif discovery is therefore an important challenge in bioinformatics and numerous methods have been developed for the identification of motifs shared by a set of functionally related sequences.

Consequently, much effort has been applied to *de novo* motif discovery, for example, in DNA sequences, with a large number of specialized methods. One interesting aspect is the development of nature-inspired algorithms, for example, particle swarm optimization to find gapped motifs in DNA sequences. Unfortunately, far fewer tools have been dedicated to the *de novo* search for protein motifs. This is due to the combinatorial explosion created by the large alphabet size of protein sequences, as well as the degeneracy of the motifs, i.e. the large number of wildcard symbols within the motifs. Some tools can discover motifs in both DNA and protein sequences. Other work has been dedicated to the discovery of specific types of protein motifs, such as patterns containing large irregular gaps, 'eukaryotic linear motifs' or phosphorylation sites. Many studies have been conducted in the comparison of these specific motif discovery tools.

In most cases, *de novo* motif discovery algorithms take as input a set of related sequences and search for patterns that are unlikely to occur by chance and that might represent a biologically important sequence pattern. Since protein motifs are usually short and can be highly variable, a challenging problem for motif discovery algorithms is to distinguish functional motifs from random patterns that are over-represented by chance.

Furthermore, existing motif discovery methods are able to find motifs that are conserved within a complete family, but most of them are still unable to find motifs that are conserved only within a sub-family of the sequences. These sub-family specific motifs, which we will call 'rare' motifs, are often conserved within groups of proteins that perform the same function (specificity groups) and vary between groups with different functions/specificities. These sites generally determine protein specificity either by binding specific substrates/inhibitors or through interaction with other protein. That's why more 'clever' algorithms should be implemented to solve this problem.

We will present a comparative study of some the existing motif discovery methods for protein sequences and their ability to discover biologically important features as well as their limitations for the discovery of new motifs and we will propose our solution to solve this problem using an Artificial Immune System. This algorithm is used to discover both motifs conserved in a

*. Intervenant

set of sequences, and *rare* motifs. The algorithm is based on the Clonal Selection model. The main characteristic of our algorithm consists in its hybrid nature, using both machine learning and statistics to discover motifs. The learning phase consists in the comparison of random antigens with random antibodies, saving both of the pools in memories and extending the pools with mutated individuals. The statistical aspect involves the assignment of fitness scores to the individuals in order to promote the more conserved individuals within subfamilies of sequences

Mots clefs : Motif discovery tools, protein sequences, conserved motifs, rare motifs, Clonal Selection

Associating gene ontology terms with protein domains

Seyed Ziaeddin Alborzi [*1], Marie-Dominique Devignes [†1], Dave Ritchie[1]

Poster 18

[1] Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) – CNRS : UMR7503, Université de Lorraine, Institut National de Recherche en Informatique et en Automatique (INRIA) – Campus Scientifique, BP 239, F-54 506 Vandœuvre-lès-Nancy Cedex, France

Motivation. The fast growing number of protein structures in the protein data bank (PDB) [1] raises new opportunities for studying protein structure-function relationships. In particular, as the biological activity of many proteins often arises from specific domain-domain and domain-ligand interactions, there is a need to provide a direct mapping from structure to function at the domain level.

Many protein entries in PDB and UniProt are annotated to show their component protein domains according to various classifications (Pfam [2], SCOP [3], or CATH [4]), as well as their molecular function through the Gene Ontology (GO) terms [5]. We therefore hypothesize that relevant GO-domain associations are hidden in this complex dataset of annotations.

Method. We use as gold-standard all GO-domain associations available from InterPro database [6] and we define GODomainMiner, a novel recommender-based method to infer associations between GO terms and Pfam domains using SIFTS (Structure Integration with Function, Taxonomy and Sequence [7]) and the UniProt databases.

Recommender systems are a class of information filtering systems which aim to present a list of items that might be of interest to an on-line customer. Content-based filters predict associations between user profiles and description of items by identifying common attributes [8]. Here, we use content-based filtering to associate GO terms with Pfam domains, the PDB chains and UniProt sequences being their common attributes. In other terms, the GODomainMiner approach associates GO terms with Pfam domains based on the structures and sequences that they share.

Results. GODomainMiner finds a total of 43,895 non-redundant GO-Pfam associations for molecular functions in a completely automatic fashion with a Recall of 0.96 with respect to the associations present in the InterPro database (about 1,500 associations).

The novel calculated GO-Pfam associations could add value to the description of structural domains of unknown function in Pfam database. They are currently undergoing comparison with the GO-SCOP and GO-CATH domain associations. Moreover, the GODomainMiner resource could be used to annotate thousands of PDB chains or protein sequences which currently lack any GO annotation although their domain composition is known.

References

[1] Berman, HM, et al. "The protein data bank." *Nucl. Acids Res.* 28:235-242 (2000).

[2] Finn, Robert D., et al. "The Pfam protein families database: towards a more sustainable future." *Nucl. Acids Res.* 44:D279 (2016).

[3] Andreeva, Antonina, et al. "Data growth and its impact on the SCOP database; new developments." *Nucl. Acids Res.* 36:D419 (2008).

*. Corresponding author : Seyed-Ziaeddin.Alborzi@inria.fr

†. Intervenant

[4] Sillitoe, Ian et al. "CATH: comprehensive structural and functional annotations for genome sequences." *Nucl. Acids Res.* 43:D376 (2015).

[5] Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource." *Nucl. Acids Res.* 32:D258-D261 (2004).

[6] Mitchell, Alex, et al. "The InterPro protein families database: the classification resource after 15 years." *Nucl. Acids Res.* 43:D213-D221 (2015).

[7] Velankar, Sameer et al. "SIFTS: Structure Integration with Function, Taxonomy and Sequence resource." *Nucl. Acids Res.* 41:D483 (2013).

[8] Ricci, F Rokach, L and Saphira, B (eds) *Recommender Systems Handbook*. Springer Science +Business Media LLC 2011.

Mots clefs : Gene Ontology, Pfam domain, Content, based filtering, Domain annotation, Structure-function relationships

Towards FFT-accelerated off-grid search with applications to Cryo-EM fitting

Alexandre Hoffmann [*1], Sergei Grudinin [†1]

Poster 19

[1] NANO-D (INRIA Grenoble Rhône-Alpes / LJK Laboratoire Jean Kuntzmann) – Institut polytechnique de Grenoble (Grenoble INP), Université Joseph Fourier - Grenoble I, CNRS : UMR5224, Laboratoire Jean Kuntzmann, INRIA – Inria GIANT DRT/LETI/DACLE Bâtiment 51C - Minatec Campus, 17 rue des Martyrs, F-38 054 Grenoble Cedex, France

We present a novel FFT-based exhaustive search method extended to off-grid translational and rotational degrees of freedom. The method combines the advantages of the FFT-based exhaustive search, which samples all the conformations on a grid, with a local optimization technique that guarantees to find the nearest optimal off-grid conformation. The method is demonstrated on a fitting problem and can be readily applied to a docking problem.

The algorithm first samples a scoring function on a six dimensional grid of size N^6 using a Fast Fourier Transform (FFT). This operation has the asymptotic complexity of $O(N^6 \cdot \log(N))$. Then, the the off-grid search is performed using a local quadratic approximation of the cost function and a trust region optimization algorithm. The computation of the quadratic approximation is also accelerates by FFT, when computed exhaustively at an additional cost of $O(N^6 \cdot \log(N))$. Alternatively, we can compute it only for the top M rigid solutions at an additional cost of $O(M \cdot N^3)$.

We tested our method on fitting atomistic protein models into several synthetic and experimental Cryo-EM maps. Our results demonstrate that spatially proximate rigid poses converge to a single off-grid solution.

Mots clefs : FFT Cryo, EM optimization

*. Intervenant

†. Corresponding author : sergei.grudinin@inria.fr

Knodle – a machine learning-based tool for perception of organic molecules from 3D coordinates

Maria Kadukova [*1], Sergei Grudinin [†2,3]

Poster 20

[1] Moscow Institute of Physics and Technology (MIPT) – 141700, 9, Institutskii per., Dolgoprudny, Moscow Region, Russia/Russie
[2] Laboratoire Jean Kuntzmann (LJK) – CNRS : UMR5224, Université Joseph Fourier – Grenoble I, Université Pierre Mendès-France - Grenoble II, Institut Polytechnique de Grenoble - Grenoble Institute of Technology, Université Pierre-Mendès-France - Grenoble II – Tour IRMA, 51 rue des Mathématiques - 53, F-38 041 GRENOBLE Cedex 9, France
[3] NANO-D (INRIA Grenoble Rhône-Alpes / LJK Laboratoire Jean Kuntzmann) – Institut polytechnique de Grenoble (Grenoble INP), Université Joseph Fourier - Grenoble I, CNRS : UMR5224, Laboratoire Jean Kuntzmann, INRIA – INRIA GIANT DRT/LETI/DACLE, Bâtiment 51C - Minatec Campus, 17 rue des Martyrs, F-38 054 GRENOBLE Cedex, France

Information about the chemical properties of atoms and bonds between them is very important for many computational methods in structural biology, medicine and bioinformatics. For example, the proper assignment of atom types and bond orders is crucial for the success of virtual screening methods in drug-design [1] as well as for the performance of some knowledge-based potentials [2].

We developed a prediction model based on nonlinear Support Vector Machines (SVM), implemented in a KNOwledge-Driven Ligand Extractor called Knodle, a software library for the recognition of atomic types, hybridization states and bond orders in the structures of small molecules. The model was trained using an excessive amount of structural data collected from the PDBbindCN database.

Accuracy of the results and the running time of our method is comparable with other popular methods, such as NAOMI, fconv, and I-interpret. On a set of 3,000 protein-ligand structures collected from the PDBBindCN general data set (v2014), the current version of Knodle along with NAOMI have a comparable accuracy of approximately 3.9 % and 4.7 % of errors, I-interpret made 6.0 % of errors, while fconv produced approximately 12.8 % of errors. On a more general set of 332,974 entries collected from the Ligand Expo database, Knodle made 4.5 % of errors. Overall, our study demonstrates the efficiency and robustness of nonlinear SVM in structure perception tasks.

Knodle is currently used for atom types assignment in our recently developed ConvexPL potential for protein-ligand docking, which shows the best pose prediction results on the Comparative Assessment of Scoring Functions (CASF) benchmark, and was among the top performers in the CSAR 2013–20143 and D3R 2015 Docking Exercises.

Knodle is available at `https://team.inria.fr/nano-d/software/Knodle`, and its GUI will be made available as a part of the SAMSON software platform at `https://www.samson-connect.net/`.

References

[1] Waszkowycz, B.; Clark, D.E.; Gancia, E. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 2011, 2.

[2] Neudert, G.; Klebe, G. *J. Chem. Inf. Model.*, 2011, 51.

*. Intervenant
†. Corresponding author : `sergei.grudinin@inria.fr`

[3] Grudinin, S.; Popov, P., Neveu, E., Cheremovskiy, G. *Journal of Chemical Information and Modeling* 2015.

Mots clefs : Machine learning, Ligand structure perception, ProteinLigand docking

Comparison of RNA suboptimal secondary structure prediction methods including pseudoknots

Audrey Legendre *[1], Fariza Tahi[1], Éric Angel[1]

Poster 21

[1] Informatique, Biologie Intégrative et Systèmes Complexes (IBISC) – Université d'Évry-Val d'Essonne : EA4526 – 23 boulevard de France, F-91 034 Évry Cedex, France

Introduction

RNAs are involved in many processes like the translation or gene's expression and they are involved in numerous pathologies, one can cite for example cancer and neurodegenerative diseases. The RNA secondary structure determination is an essential step in the understanding of its function, in the identification of RNA classes involved in a given biological processes, in the identification of interactions that the given RNA can have with other RNAs or proteins. Secondary structures often include substructures called pseudoknots making them lose their plane conformation, i.e. the addition or deletion of a pseudoknot, completely changes the 3D structure. A lot of softwares were proposed to predict secondary structures with pseudoknots. There are exact methods based on the energetic model [1-3] and heuristic methods (comparative model [4], probabilistic model [5]). The heuristic methods are faster and their predictions are fairly accurate. However, none of these softwares can find suboptimal structures, except Tfold [4] based on the comparative model. Mfold [6], RnaSubopt [7] and Sfold [8] can predict suboptimal structures but without pseudoknots. PknotsRG [9], now known as pKiss, is the only software able to predict suboptimal structures with pseudoknots. It is important to predict an ensemble of possible structures rather than a single one because the real structure can be one of these structures. Indeed, the RNA can have multiple structures (due to an interaction with another molecule or the environment for example), and moreover no method is able to compute with a 100 % accuracy the real structure (approximations in energetic models can result in suboptimal structures that are nearest from the real one than the optimal one). Here we are interested in the comparison of the methods predicting optimal and suboptimal possible secondary structures including pseudoknots from a single sequence. For this purpose, we propose to compare pKiss and an extension of Ipknot [5] using a probabilistic model. In this work, we show that pKiss gives better predictions with a more restricted set of suboptimal structures.

Methods

PKiss. PKiss is the successor of pknotsRG that could predict suboptimal secondary structures of RNA with pseudoknots by computing the minimum free energy structures using dynamic programming. PKiss includes the same functionality and so, the prediction of the canonical simple recursive class of pseudoknots. It also adds the prediction of the canonical simple recursive kissing hairpins, we used the recommended 'A' strategy to compute it. Moreover, pKiss can do shape analysis, computation of probabilities, and can be parameterized by different folding strategies and different models (minimum free energy model, comparative model).

Linear programming. In a linear program we have a set of constraints and an objective function which has to be minimized or maximized. The objective functions and constraints are defined

*. Intervenant

over decision variables. According to the type of these variables, the linear problems are more or less hard to solve. An integer linear problem is harder to solve (exponential time complexity) than a linear problem with only real variables (polynomial time complexity). To solve the linear problem, there exist solvers like GLPK [10], GUROBI [11], and CPLEX [12]. They use several solving algorithms like the simplex algorithm or the interior point algorithm.

Ipknot algorithm. To predict a RNA secondary structure, giving a sequence, Ipknot computes the base pairing probabilities and solve an integer linear problem. The objective function of Ipknot aims to maximizing the expected accuracy of the predicted secondary structure, from the base pair probabilities. Ipknot considers only base pairs with a probability superior to a threshold. Ipknot decomposes the structure in many levels. Each level defines base pairs that can't cross, and therefore are pseudoknot-free. Each base pair i,j is associated with a binary decision variable yp(i,j), which is equal to 1 if the bases i and j at level p are linked, 0 otherwise. The lower level has the most of base pairs and when this level is full or if adding a base pair is not possible because of constraints, a next level is created. Ipknot considers only the base pairs with the highest probabilities and the threshold can be different according to the level. In the linear problem, constraints are defined to prevent base pairs that can't be found in real secondary structure. The constraints allow the prediction of all pseudoknot classes.

Suboptimal solutions. The easier method to find suboptimal solutions in linear programming is to solve the problem many times. At each step, a constraint is added to prevent to find the previous solution. Let assume that we have found an optimal solution x* and let define B={ i | x*i=1 } and N = { i | x*i=0 }. Balas and Jeroslow [13] proposed to add the following constraint to the current integer linear program :

$\sum_{i\in B} xi - \sum_{i\in N} xi \leq |B| - 1$ which is equivalent to

$\sum_{i\in B} (1-xi) + \sum_{i\in N} xi \geq 1$

This constraint ensures that the (Hamming) distance between any feasible solution x and the previous optimal one x* must be at least one, therefore there must be at least one variable xi which takes a different value from x*i.

Implementation. We implemented the extension of Ipknot with the method described precedently, with an IDE for the CPLEX solver, named Oplide. We chosed the CPLEX solver because it is an industrial standard. The base pairing probabilities are computed with the Vienna RNA package [7], with the McCaskill Model [14] and the free energy parameters from the Vienna RNA package. We implemented two levels because three are rarely needed in nature. Also, we used the default thresholds of the Ipknot webserver [5], i.e. 2 for level 1, and 16 for level 2, they give the best predictions.

Mesures. To evaluate the predictions of pKiss and the Ipknot extension, we compute for each sequence, the sensitivity, the specificity and the Matthews correlation coefficient. These statistics are based on the different numbers of base pair with the referenced sequence.

Results and discussion

Dataset. We tested our implementation with the pk168 dataset [15]. This dataset was compiled from Pseudobase++ [16]. It includes 168 RNAs with sixteen pseudoknot classes.

Importance of suboptimal solutions. We predicted five secondary structures for the PKB102, PKB103 and PKB104 RNAs with Ipknot and computed the MCC for each. The goal was to show some examples, where the true solution is not the optimal one. The results are shown in Table 1. We observed that the suboptimal secondary structures often resulted in the addition or deletion of a single base pair in a stack or not. This was expected because of the nature of the constraint we added at each step. At first we may think that the addition or the deletion of one base pair can't have an incidence on the structure, but it can. When we observed the MCC, we see that for the PKB102 and the PKB104 RNAs, the best solutions are the second ones. For the PKB103 RNA,

the best solution is the third one. It really shows that the structure that is nearest to the reality isn't always the optimal according to the Ipknot's model. It is why it is important to predict optimal and suboptimal solutions.

Rank distribution. We determined the rank of the best prediction of pKiss and Ipknot according to the MCC (Figures 1 and 2). The distribution of pKiss (Figure 1) shows that for 79 sequences (47.02 %), the best prediction isn't at the first rank. It is almost the half of the dataset, it confirms that the real solution isn't always the optimal one. We observe that when we cumulate rank 1 and 2, we got 135 sequences. It represents 80.36 % of the dataset for which the best prediction is at these ranks. If we add the ranks 3 and 4, we reach 89.88 % of the dataset. The presence of other ranks are very sporadic, the maximum number of best prediction is at most 2. The distribution of Ipknot shows that the best prediction isn't at the first rank, for 114 sequences (67.86 %). The distribution also shows that the ranks 1, 2 and 3 consist of 77 sequences (45.8 %), i.e., if we choose to predict only two suboptimal solutions, we have a probability of 0.46 to have the best solution among this set. Comparing these results, we would say that pKiss is more interesting because it give best prediction at a small rank. But we need to compare the solutions between the two softwares.

Probability to have the best solution. We represented these probabilities in function of the number of solutions predicted by pKiss and Ipknot (Figure 3). The pKiss curve grows really faster compared to the Ipknot curve. The curves well represent the ranks, for pKiss, we see that the probability is almost at 1.0 with a little set, while for Ipknot the probability of having the best solution is only 0.5 when we have a set of 8 solutions. It shows that the set of suboptimal solutions must be quite large with Ipknot to be sure to have the nearest solution from the reality.

Comparison of the solutions given by pKiss and Ipknot. We represented the sensitivity in function of 1-specificity for each sequence of the pk168 dataset (Figure 4). For each sequence we represented the best solution, according to the MCC, among pKiss and Ipknot (considering an output of three solutions). There are more best solutions from pKiss (116) than Ipknot (52). Moreover, most of the solutions from pKiss have a better accuracy than the ones from Ipknot. Pkiss gives better solutions than our extension of Ipknot.

Time. The two methods exhibit quite similar computation times. To predict the dataset, with a processor i5-3470 3.20 GHz × 4, pKiss needs 36.18 seconds and Ipknot (with twenty suboptimal solutions) 24.67 seconds.

Conclusion

In this work we have shown that predicting suboptimal secondary structures of RNAs is important. PKiss and Ipknot algorithms are based on respectively energetic and probabilistic models. We proposed to extend Ipknot with a simple linear programming method to compare these softwares. We shown that pKiss gives better predictions with a more restricted set of suboptimal structures.

References

[1] Waszkowycz, B.; Clark, D.E.; Gancia, E. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 2011, 2.

[2] Neudert, G.; Klebe, G. *J. Chem. Inf. Model.*, 2011, 51.

[3] Grudinin, S.; Popov, P., Neveu, E., Cheremovskiy, G. *Journal of Chemical Information and Modeling* 2015.

Mots clefs : Machine learning, Ligand structure perception, ProteinLigand docking

Sur l'information apportée par la structure 3D dans la formation des interactions protéines-protéines

Poster 22

Guillaume Launay[1], Nicoletta Ceres[1], Juliette Martin [*1]

[1] MMBS – CNRS : UMR5086, Université Claude Bernard - Lyon I – France

Contexte et question

Les protéines accomplissent leurs fonctions biologiques par le biais d'interactions avec d'autres macro-molécules biologiques (ADN, membranes ou protéines). Le nombre d'interactions fonctionnelles entre protéines au sein d'un organisme n'est qu'un sous-ensemble très réduit de toutes les interactions possibles. Ainsi chez la levure, la taille de l'interactome est estimée à 37 600 interactions binaires, soit 0,2 % de toutes les interactions possibles entre les 6 000 protéines encodées dans son génome.

Comment s'effectue le contrôle des interactions, dans le milieu particulièrement dense en macro-molécules qu'est le cytoplasme, pour parvenir à réprimer 99,8 % des interactions possibles ? Les protéines sont régulées à de multiples niveaux (expression, traduction, localisation et modifications post-traductionnelles). De plus, l'interaction physique entre protéines en rendue possibles par des propriétés structurales telles que la complémentarité de forme et des liaisons non covalentes stabilisantes comme la complémentarité de charges, la présence d'amas hydrophobes, la formation de liaisons hydrogènes ou de ponts salins.

L'information de structure 3D a été utilisée pour la prédiction des interactions protéines-protéines à plusieurs reprises dans la littérature. La plupart des méthodes exploitent la relation d'homologie entre protéines : l'existence d'une interaction entre deux structures permet d'inférer l'interaction entre des structures homologues. En revanche, la démarche qui consiste à considérer les propriétés intrinsèques des structures sans considérer les relations d'homologie pour inférer une interaction a été moins explorée et fournit des résultats plus mitigés.

En fin de compte, il est difficile aujourd'hui d'apprécier la contribution des propriétés intrinsèques des structures 3D dans la capacité à prédire les interactions protéine-protéine, en dehors de la notion d'homologie. Le travail présenté ici a pour but de contribuer à déconvoluer les facteurs qui dictent la formation d'interactions entre protéines, et qui peuvent donc être utilisés en prédiction.

Méthodes et résultats

Dans cette étude, nous avons considéré plusieurs jeux de paires de protéines de *S. cerevisiae* pour lesquelles aucune interaction fonctionnelle n'a été rapportée, dénommées paire négative, et avons comparé leurs structures 3D à celles des complexes disponibles dans la PDB. Nous avons identifié des cas de paires négatives structuralement très similaires à des complexes expérimentaux, ce qui indique que, bien que ces protéines ne forment pas d'interactions fonctionnelles, leurs structures 3D sont compatibles. En tenant compte de la couverture limitée de la PDB, nous estimons que 8,7 % des paires négatives pourraient être ainsi compatibles structuralement. Le

*. Intervenant

nombre d'interactions potentielles correspondant est au moins 40 fois supérieur au nombre d'interactions fonctionnelles (0,2 % des interactions totales).

Nous avons placé ces interactions potentielles dans le réseau d'interactions protéine-protéine natif de la levure, afin d'estimer leur impact fonctionnel en cas de formation effective. Nous avons montré que les interactions formées par les paires négatives avec des structures compatibles sont particulièrement centrales dans le réseau, ce qui suggère des effets particulièrement néfastes.

Structuralement, les modèles de complexes protéine-protéines correspondant aux paires négatives ne sont pas aberrants selon les descripteurs conventionnels de taille et de compacité des interfaces. Certains sont même prédits comme relativement stables par notre champ de force gros grains PaLaCe.

Conclusion

Ces résultats montrent que les outils actuels peinent à séparer les paires fonctionnelles des paires non fonctionnelles et supportent un modèle dans lequel la régulation biologique est capitale pour empêcher des interactions potentiellement très néfastes entre protéines de structures compatibles.

Mots clefs : interactions protéine, protéine, structure 3D, réseaux

Antibiotic resistance : Structural analysis of the (dis)similarities between β-lactamases and penicillin-binding proteins

Poster 23

Mame Ndew Mbaye [*1,2], Dimitri Gilis[1], Marianne Rooman[1]

[1] Université Libre de Bruxelles (ULB) – Avenue Franklin D. Roosevelt 50, 1050 Bruxelles, Belgique
[2] Université Cheikh Anta Diop de Dakar, Sénégal

Since the discovery of penicillin, a lot of antibiotics have been developed against bacteria. *Enterobacteriaceae* are gram-negative bacteria that are at the basis of many diseases. Some are true pathogens whereas others are opportunistic or cause secondary infections of wounds, the urinary and respiratory tracts and the circulatory system. Although several classes of antibiotics have been used against *Enterobacteriaceae* and have been successful for a number of years, these bacteria have developed antibiotic resistance. The major mechanism of *Enterobacteriaceae* resistance is the secretion of β-lactamase enzymes, which inactivate the class of β-lactam antibiotics. In principle, these molecules bind to penicillin-binding proteins (PBPs), thereby blocking the synthesis of cell wall. However, β-lactamases also interact with these antibiotics, cleave them and render them inactive.

The cell wall synthesis in *Enterobacteriaceae* involves several classes of PBP, which share a common enzymatic mechanism based on the presence of an active serine. Their penicillin binding domain harbors three conserved amino acid sequence motifs in the binding cavity: SXXK, (S/Y)XN and (K/H)(S/T)G. The inactivation of the PBPs of type PBP2 or PBP3, or the simultaneous inactivation of both PBP1A and PBP1B, is lethal for the bacteria [1].

β-lactamases can be divided into two groups on the basis of their enzymatic mechanism. The first (class B) consists of metallo-enzymes containing zinc ions that are involved in β-lactams hydrolysis. Enzymes of the second group (containing classes A, C and D) hydrolyze the β-lactams through a serine esterification mechanism [2,3]. We focus in this study only on the latter group that contains active serine β-lactamases, as they have the same catalytic mechanism as PBPs. In particular, the three characteristic motifs of PBPs are common to the serine active β-lactamases [4].

It is therefore interesting to highlight structural similarities and differences between, on the one hand, β-lactamases of class A, C and D and, on the other hand, PBPs – in particular PBPs whose inactivation is lethal. This is the goal of the present work. Such insights can help designing new antibiotics that are specific of either PBPs or β-lactamases and can hence overcome drug resistance.

Methods

Among the different PBPs of which the activation is lethal, only PBP3 has an experimental X-ray structure available in the Protein DataBank (PDB) [5]. We thus restrict our analysis to this type of PBP. Its PDB code is 4BJP.

Each family of A, C and D β-lactamases is divided into several subfamilies, which differ in their sequence, structure and medical relevance. We chose a representative X-ray structure for each of the subfamilies that are known to contribute to antibiotic resistance [6]. The selected PDBs are: 1ZKJ (CMY, class C), 4HBT(CTX, class A), 4GOG (GES, class A), 3C5A (KPC, class

*. Intervenant

A), 1M6K (OXA, class D), 4D2O (PER, class A), 1N9B(SHV, class A), 1DY6 (SME, class A) and 1M40 (TEM, class A).

The representative PBP and b-lactamase structures were superimposed using the DaliLite [7,8] and PDBeFold [9] servers. The active cavities in each structure were assigned using MetaPocket2.0 [10], and refined by visual inspection and comparison.

Results

To answer whether it is possible to design a ligand that specifically binds PBP3 without being recognized and degraded by β-lactamases, we performed a detailed structural comparison of the active cavity in these proteins. We focused on the type of amino acids present in different regions of the cavities and on the distributions of hydrogen bond donors and acceptors. Without surprise, our results show that the three motifs SXXK, (S/Y)XN and (K/H)(S/T)G, which are located in the central region of the cavity and are responsible for the catalytic activity of the proteins, are highly conserved in PBP3 and in the selected β-lactamases (see Figure 1).

In contrast, some other regions in the cavity were shown to be occupied by amino acids of different types and/or show differences in the H-bond donor/acceptor distributions between on the one hand PBP3 and on the other hand all the β-lactamases. For instance, Figure 2 shows a specific region of the cavity, with a positively charged residue that is only present in PBP3 and is very conserved. The β-lactamases have negatively charged, aromatic or polar amino acids in this region. Moreover, the H-bond donor/acceptors are differently distributed (Figure 2b). We found also other regions where differences are detected. We also observed that the cavity of PBP3 is shaped differently: it is more flat in a region and thus slightly larger than the cavity found in the β-lactamases.

In summary, our structural analysis reveals several features of the PBP3 cavity that distinguish it from the β-lactamase cavities. These differences can be exploited for the design of a ligand that targets specifically these particular regions, in view of overcoming the resistance mechanism.

References

[1] Ghuysen J-M. Serine β-lactamases and penicillin-binding proteins. *Annual review Microbiology* 1991; 45:37–67.

[2] Rice L. B. and Bonomo R.A. Genetic and biochemical mechanisms of bacterial resistance to antimicrobial agents. In: Loridan V, editor. *Antibiotics in laboratory medicine. 5th ed Baltimore, USA: Williams & Wilkins*; 2005, p. 485.

[3] Bebrone C. Metallo-β-lactamases (classification, activity, genetic, organization, structure, zinc coordination) and their superfamily. *Biochemical pharmacology* 2007: 4:1686-1707.

[4] Pfeifle D, Janas E, and Wiedeman B. Role of Penicillin-Binding Proteins in the Initiation of the AmpC β-Lactamase Expression in *Enterobacter cloacae. Antimicrobial agents and chemotherapy* 2000; 44:169-172.

[5] Berman H.M., Westbrook Z, Feng G., Gilliland T.N., Bhat T.N., Weissig I.N., Shindyalov P.E. and Bourne P.E. The Protein Data Bank. *Nucleic Acids Research* 2000; 28:235-242.

[6] Bush K. and Jacoby G.A. Updated functional Classification of β-Lactamases. *Antimicrobial agents and chemotherapy* 2010; 54:969-976.

[7] Holm L. and Sander C. Mapping the protein universe. *Science* 1996; 273:595-603.

[8] Holm L. and Park J. DaliLite workbench for protein structure comparison. *Bioinformatics* 2000; 6:566-577.

[9] Krissinel E. and Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta crystallographica* 2004; D60:2256-2268.

[10] Zhang Z, Li Y, Lin B, Schroeder M and Huang B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* 2011; 27:2083-2088.

Mots clefs : Enterobacteriaceae, Antibiotics resistance, Penicillin binding proteins, β-lactamases

Une nouvelle méthode de clustering avec incertitude de données de séquençage

Alexandre Bazin *[1], Didier Debroas[2], Engelbert Mephu Nguifo[1]

Poster 24

[1] Laboratoire d'Informatique, de Modélisation et d'optimisation des Systèmes (LIMOS) – Institut Français de Mécanique Avancée, Université Blaise Pascal - Clermont-Ferrand II, Université d'Auvergne - Clermont-Ferrand I, CNRS : UMR6158 – Bâtiment ISIMA, Campus des Cézeaux, BP 10025, F-63 173 Aubière Cedex, France

[2] Microorganismes : génome et environnement (LMGE) – Université d'Auvergne - Clermont-Ferrand I, CNRS : UMR6023, Université Blaise Pascal - Clermont-Ferrand II – Université Blaise Pascal, Campus des Cézeaux, 24 avenue des Landais BP 80026, F-63 170 Aubière, France

La détermination de la structure des communautés (abondance, richesse, composition) d'un écosystème est un enjeu central en écologie et donc en écologie microbienne. De récents travaux effectués à l'aide des nouvelles techniques de séquençage (NGS) ont mis en évidence que cette richesse spécifique était certainement sous-estimée ; la plus grande partie de la biodiversité est représentée par des OTUs (i.e. espèces microbiennes) faiblement abondants : la biosphère rare. La détermination de la structure (richesse, abondance, diversité, composition) de la biosphère rare repose sur la détermination des OTUs. Or, cette détermination varie en fonction des méthodes de classification (clustering). En effet, les méthodes de classification habituellement utilisées (UCLUST [1], Swarm [2]…) attribuent généralement un cluster unique à un objet de façon certaine. Sachant qu'une séquence peut se trouver à la frontière de plusieurs clusters, ou plus précisément avoir des ressemblances avec des séquences de différents clusters, les clusters obtenues par ces méthodes ne sont pas toujours fiables et en outre ils dépendent de l'ordre dans lequel les données sont traitées.

Afin d'améliorer la qualité des clusters, nous proposons d'utiliser une version floue des algorithmes habituellement utilisés. Ainsi, les séquences ne sont plus associées à un seul cluster de façon certaine mais peuvent l'être à plusieurs clusters avec différents degrés de d'appartenance (0, 0.1, …, 0.9, 1). Cela permet non seulement d'identifier efficacement les séquences problématiques qui pourraient bénéficier d'une attention particulière mais aussi d'évaluer la « qualité » des clusters. En effet, un cluster idéal comprendrait beaucoup de séquences lui appartenant certainement et peu de séquences lui appartenant de façon moins sûre. Au contraire, un cluster comprenant beaucoup de séquences lui appartenant faiblement et aucune lui appartenant certainement serait une erreur de l'algorithme de classification. En mesurant la différence entre la répartition des degrés d'appartenance obtenus et le cluster idéal, nous pouvons attribuer à chaque cluster une valeur représentant sa qualité. Les clusters de faible qualité peuvent alors être automatiquement modifiées, fusionnées ou supprimées tandis que les plus ambigües peuvent être soumises à l'expertise de l'utilisateur.

Cette méthode est en cours d'implémentation et sera testée sur plusieurs jeux de données, et comparée aux méthodes standard en utilisant des indices de qualité du clustering. Cette méthode permettra donc d'accroître la fiabilité des clusters proposés, et en outre nous espérons que cela aidera aussi à réduire l'influence de l'ordre sur le résultat final.

Références

[1] Edgar, Robert C. "Search and clustering orders of magnitude faster than BLAST." *Bioinformatics* 26.19 (2010).

*. Intervenant

[2] Mahé, Frédéric, et al. “Swarm: robust and fast clustering method for amplicon-based studies.” *PeerJ* (2015).

Mots clefs : Clustering

Minimal perfect hash functions in large scale bioinformatics

Antoine Limasset [*1]

[1] GENSCALE (INRIA - IRISA) – École normale supérieure (ENS) - Cachan, Universite de Rennes 1, CNRS : UMR6074, INRIA – Campus de Beaulieu, F-35 042 RENNES Cedex, France

Poster 25

Indexing objects has always been a fondamental task in bioinformatics. But indexing has been proven extremely expensive or impossible for large scale problems.

Recently, Minimal Perfect Hash Functions (MPHF) has been used to tackle this problem in various fields such as assembly, read mapping or RNA quantification. The problem is that the construction of such functions is based on complex structures as hypergraphs and requires significant computational ressources. Existing MPHF construction tools typically use more memory than the function itself and can be prohibitively slow.

For adressing this problem, we present here BBhash, a scalable library to construct Minimal Perfect Hash Functions for very large datasets and show some example applications of MPHF in bioinformatics. Our library BBhash is designed to be pratical and does not aim to achieve the theoritical bound of memory usage of 1.44 bits by elements. It is to our knowledge the only tool able to construct a MPHF for up to 100 billions element in hours, with limited memory fingerprint (< 30GB).

BBhash is available at `https://github.com/rizkg/BooPHF`.

Mots clefs : Minimal perfect hash function, high performance computing, Scalable method

*. Intervenant

Paraload : un programme de répartition de charge à large échelle pour les calculs en bioinformatique

Poster 26

Dominique Guyot[1], Simon Penel[2], Vincent Navratil[1], Daniel Kahn[3], Guy Perrière [*] [†1,2]

[1] Pôle Rhône-Alpes de Bioinformatique (PRABI) – Université Claude Bernard - Lyon I (UCBL) – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[3] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – Université Claude Bernard - Lyon I (UCBL), Institut national de la recherche agronomique (INRA) – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France

Paraload est un programme permettant de distribuer des tâches de calculs indépendantes sur un très grand nombre de processeurs, que ceux ci soient sur la même machine ou non. Cet outil permet ainsi d'accélerer de façon très significative la vitesse de traitement de tout calcul parallélisable par les données. Ses domaines d'application en bioinformatique sont donc particulièrement nombreux : recherche de similarités, alignements d'un grand nombre de familles, construction de profils, rééchantillonnage de type bootstrap ou jackknife, etc.

Paraload est une application de type client/serveur qui utilise des connexions TCP/IP afin de distribuer des données ainsi que la commande à exécuter en parallèle sur de multiples serveurs. La particularité de son fonctionnement réside dans le fait que c'est au client de demander quels sont les calculs à effectuer et non le serveur. En effet, la fonction du serveur consiste seulement en la distribution des tâches, ce qui est extrêmement rapide. De ce fait, les clients sont utilisés à leur maximum, même si la granularité du parallélisme est très fine.

Afin d'optimiser le temps de calcul, un système de gestion de la taille des tâches utilisant une fonction de complexité algorithmique est implémenté. Cinq options sont disponibles : $O(1)$, $O(n)$, $O(n^2)$, $O(n^3)$ et $O(n \cdot \log(n))$. Dans le cas d'une tâche impliquant la construction d'un ensemble d'alignements multiples on choisira $O(n^2)$ car la plupart des programmes disponibles implémentent des algorithmes quadratiques. Par ailleurs, ce système de gestion récolte les données de temps de calculs effectif pour pouvoir modifier cette fonction de complexité si besoin est. Paraload implémente également un système de contrôle d'intégrité des calculs, un système de contrôle des transmissions des données et permet la reprise sur point en cas de panne critique.

Concernant l'utilisation du programme, l'accent a été mis sur la simplicité. Le fonctionnement de Paraload nécessite un fichier d'entrée et un fichier de configuration. Le serveur lit le fichier de configuration, découpe le fichier d'entrée en morceaux qu'il envoie au clients de calcul. Le serveur récupère ensuite le résultat pour l'empiler sur un fichier de sortie puis il inscrit les informations de chaque calcul dans un fichier de log.

Ce programme a notamment été utilisé pour la production de la version 7 d'HOGENOM. En particulier, il a permis d'achever la phase de comparaison globale des homologies (132 320 heures de calculs soit 15 années de temps CPU) en une quinzaine de jours. Il a également été utilisé dans la procédure de construction de la banque ProDom pour un total avoisinant les 1 000 000 d'heures soit 115 années de temps CPU.

Le source de Paraload est disponible au téléchargement à l'adresse :
ftp ://pbil.univ-lyon1.fr/pub/logiciel/paraload/

*. Intervenant
†. Corresponding author : guy.perriere@univ-lyon1.fr

Paraload est compilable et utilisable sur toute machine Linux disposant d'un compilateur C/C++ ainsi que des bibliothèques C/C++ standard.

Mots clefs : Calcul parallèle, répartition de charge

TOGGLE-3 : a tool for on the fly pipelines creation and performing robust large-scale NGS analyses

Poster 27

Sébastien Ravel [*] [†1], Christine Tranchan-Dubreuil [‡2], Cécile Monat[2], Gautier Sarah[3], Julie Orjuela-Bouniol[4], François Sabot [§2]

[1] Biologie et génétique des interactions plantes-parasites pour la protection intégrée (BGPI) – Institut national de la recherche agronomique (INRA) : UR0385, Centre de coopération internationale en recherche agronomique pour le développement [CIRAD] : UMR54 – Campus International de Baillarguet - TA 41 / K, F-34 398 Montpellier Cedex 05, France

[2] Diversité, adaptation, développement des plantes (DIADE) – Institut de recherche pour le développement [IRD] : UR232, Université Montpellier II - Sciences et techniques – Centre IRD de Montpellier 911 av Agropolis BP 604501, F-34 394 Montpellier Cedex 5, France

[3] Amélioration Génétique et Adaptation des Plantes Méditerranéeennes et Tropicales (AGAP) – Montpellier SupAgro, Institut national de la recherche agronomique (INRA) : UMR1334, CIRAD-BIOS – TA A-108/03 Avenue Agropolis, F-34 398 Montpellier Cedex 5, France

[4] ADNid company – 830 avenue du Campus Agropolis Baillarguet, F-34 980 Montferriez sur Lez, France

Web site:
`https://github.com/SouthGreenPlatform/TOGGLE`

Dear biologist, have you ever dreamed of using the whole power of those numerous NGS tools that your bioinformatician colleagues use through this awful list of command lines ?

Dear bioinformatician, have you ever wished for a really quick way of designing a new NGS pipeline without having to retype again dozens of code lines to readapt your scripts or starting from scratch ?

So, be happy ! TOGGLE is for you !

With TOGGLE (TOolbox for Generic nGs anaLysEs), you can create your own pipeline through an easy and user-friendly approach. Indeed, TOGGLE integrate a large set of NGS softwares and utilities to easily design pipelines able to handle hundreds of samples. The pipelines can start from Fastq (plain or compressed), SAM, BAM or VCF (plain or compressed) files, with parallel (by sample) and global analyses (by multi samples).

Moreover, TOGGLE offers an easy way to configure your pipeline with a single configuration file:

- organizing the different steps of workflow,
- setting the parameters for the different softwares,
- managing storage space through compressing/deleting intermediate data,
- determining the way the jobs are managed (serial or parallel jobs through scheduler SGE and SLURM).

TOGGLE can work on your laptop, on a single machine server as well as on a HPC system, as a local instance or in a Docker machine. The only limit will be your available space on the storage system, not the amount of samples to be treated or the number of steps.

*. Intervenant
†. Corresponding author : `sebastien.ravel@cirad.fr`
‡. Corresponding author : `christine.tranchant@ird.fr`
§. Corresponding author : `francois.sabot@ird.fr`

TOGGLE was used on different organisms, from a single sample to more than one hundred at a time, in RNAseq, DNAreseq/SNP discovery and GBS analyses.

List of bioinformatics tools included:
- BWA : bwaAln, bwaSampe, bwaSamse, bwaIndex, bwaMem
- SamTools : samToolsFaidx, samToolsIndex, samToolsView, samToolsSort, mergeHeader, samToolsMerge, samToolsIdxstats, samToolsDepth, SamToolsFlagstat, samToolsMpileUp
- PicardTools : picardToolsMarkDuplicates, picardToolsCreateSequenceDictionary, picardToolsSortSam, picardToolsAddOrReplaceReadGroup, picardToolsValidateSamFile, picardToolsCleanSam, picardToolsSamFormatConverter
- GATK : gatkBaseRecalibrator, gatkRealignerTargetCreator, gatkIndelRealigner, gatkHaplotypeCaller, gatkSelectVariants, gatkVariantFiltration, gatkReadBackedPhasing, gatkUnifiedGenotyper, gatkBaseRecalibrator, gatkPrintReads
- Fastqc : fastqc
- FastxToolkit : fastxTrimmer
- Tophat : bowtie-build, bowtie2-build, tophat2
- Snpeff : snpeffAnnotation
- Cutadapt : cutadapt
- Graphviz v2.xx (optional)

References

[1] TOGGLE: Toolbox for generic NGS analyses. Cécile Monat, Christine Tranchant-Dubreuil, Ayité Kougbeadjo, Cédric Farcy, Enrique Ortega-Abboud, Souhila Amanzougarene, Sébastien Ravel, Mawussé Agbessi, Julie Orjuela-Bouniol, Maryline Summo and François Sabot. *BMC Bioinformatics* 2015, 16:374 doi:10.1186/s12859-015-0795-6

Mots clefs : NGS, Toolbox, On the fly pipeline, Flexible, RNASeq, GBS, SNP, Scheduler

Simulating the surface diffusion and concentration of receptors in cells membrane

Pascal Bochet [*] [†1,2], Thierry Rose[1]

Poster 28

[1] Groupe Réseaux et Signalisation (CITECH) – Institut Pasteur – 25-28 rue du Docteur Roux, F-75 724 Paris, France
[2] Polarité cellulaire, Migration et Cancer – CNRS : UMR3691 – 25-28 rue du Dr Roux, F-75 724 Paris, France

Helper lymphocytes (CD4+ T cells) regulate the immune response and are among the targets of HIV-1 virus. They patrol the organism and, when activated, bind to the wall of blood vessels before they cross it in the extravasation process to penetrate adjacent tissues.

This binding is due to the concentration of integrin molecules in specific areas of the cell membrane with a different lipid composition, the lipid rafts. The activation of integrins, which them to bind to fibronectin molecules present in the extracellular matrix in the vessel walls and immobilize CD4+ T cells.

We ran molecular diffusion models, treating each integrin molecule as a independant entity, to test the hypothesis that the accumulation of integrin molecules in the lipid rafts is partially due to their slower diffusion rate in the different lipids which constitute the rafts.

Based on the total amount of integrin on the cell surface and the total number of lipid rafts we were able to estimate the number of integrin molecules in each raft. By comparing these results with measures of the force required to tear off activated CD4+ T cells bound to fibronectin-coated surfaces we are able to check the consistency of these numbers with the known affinity of integrin-fibronectin interaction.

References

[1] T. Rose, A. Pillet, V. Lavergne, B. Tamarit, P. Lenormand, J.C. Rousselle, A. Namane and J. Thèze, Interleukin-7 compartmentalizes its receptor signaling complex to initiate CD4 T lymphocyte response. *J Biol Chem*, 285:1489814908, 2010.

[2] B. Tamarit, F. Bugault, A. Pillet, V. Lavergne, P. Bochet, N. Garin, U. Schwarz, J. Thèze and T. Rose. Membrane microdomains and cytoskeleton organization shape and regulate the IL7-receptor signalosome in human CD4 T cells. *J Biol Chem*, 288:8691-8701, 2013.

Mots clefs : diffusion, cell membrane, receptors, lipid raft, binding

*. Intervenant
†. Corresponding author : pascal.bochet@pasteur.fr

How to visualize and uniquely identify bound fatty acids during their biosynthesis ?

Olivier Clerc [*1], Éric Maréchal[1], Sylvaine Roy [*†1]

Poster 29

[1] Laboratoire Physiologie Cellulaire & Végétale Institut de Biosciences et de Biotechnologie de Grenoble (PCV) – CEA, Université Joseph Fourier - Grenoble I, Centre national de la recherche scientifique (CNRS), Institut National de la Recherche Agronomique (INRA) – 17 rue des Martyrs, F-38 054 Grenoble Cedex 9, France

In the lipid metabolism, the fatty acids biosynthesis involves Free (non-esterified) Fatty Acids (FFA) and Bound Fatty Acids (BFA). The latter are bound to what we call a "co-substance" which can be more or less complex; it can be for instance, the Coenzyme A (CoA), the Glycerol (in the case of triglycerids), an Acyl Carrier Protein (ACP) or a Serine Protein (SerProt) whose detailed sequence may be unknown.

A biologist, who studies the lipid metabolic pathways, needs to handle and to uniquely identify the Fatty Acids even when they are bound to such a co-substance. In the case of triglycerids, he or she must know the position on the glycerol backbone (sn1, sn2 or sn3). This biologist needs also to visualize clearly each Fatty Acid, be it free or bound, with its possible co-substance.

It is easy to handle, uniquely identify and correctly create an image for visualization when the fatty acid is free, since it is a relatively small chemical molecule. Its structure can be clearly described thanks to a classical chemical format such as SDF [1], Smiles [2], CML [3],... It can be uniquely identified thanks to the InChI identifier [4] which is a textual identifier for chemical substances. Every classical tool used for small molecule allows to display correctly the fatty acid in any image format.

For BFA, by contrast, the situation is much more complex, since the co-substance can be a protein whose sequence is unknown or even can be a position concept. Their visualization cannot be realized with the classical tools that deal with small chemical molecules or those that handle macromolecules, and their identification raises some issues of conceptualization and implementation.

In this poster, we will present the software that we have realized; it allows the automatic creation of BFA images exactly with the same shapes the biologists are used to, when they draw them manually. This software will be soon freely available for the scientific community. We will also give a brief overview of the issues raised by the identification of BFA as well as the first solutions that we are considering for our internal database.

Acknowledgments

This work is supported by the Oceanomics Consortium which has received funding from the French "Investissements d'Avenir" Program under the grant agreement ANR-11BTBR-0008. The authors thank the ChemAxon company (`http://www.chemaxon.com/`) for allowing academics to freely use their software, here Molecule File Converter, version 14.7.7.0, under the FreeWeb license.

*. Intervenant

†. Corresponding author : `sylvaine.roy@cea.fr`

References

[1] Dalby et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited *J. Chem. Inf. Comput. Sci.* 1992 32(3):244-255.

[2] Weininger. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28(1):31-36.

[3] Murray-Rust and Rzepa. Chemical markup, XML, and the world-wide web. 1. Basic principles *J. Chem. Inf. Comput. Sci.* 1999, 39:928–942.

[4] `http://www.iupac.org/inchi/`.

Mots clefs : lipids, fatty acid, biosynthesis, lipidomics, metabolism, image, visualization, acide gras, lipide, lipidomique, biosynthèse, métabolisme, visualisation

Generation of gamma rhythms in a modelling of the hippocampus CA1 area

Jean de Montigny [*1,2], Bruce Graham[2]

[1] Parcours « Bioinformatique, Connaissances, Données » du Master Sciences & Numérique pour la Santé de l'Université de Montpellier – Université Montpellier II - Sciences et Techniques du Languedoc – France
[2] University of Stirling, Computing Science & Mathematics division – Royaume-Uni

Poster 30

After being almost ignored during decades, gamma rhythms (GR, 30 to 120 Hz oscillations) are nowadays the topic of many studies, investigating both their origins and purpose. Due to its high neuron density and its implication in several cognitive tasks, the hippocampus and its CA1 area, which exhibits GR, represents a key structure for those GR studies (Colgin et. al, 2009).

A widespread hypothesis is that the gamma frequency range can be split into two rhythms, slow (30 to 50 Hz) and fast (60 to 120Hz), having distinct neural source and underlying distinct memory process (Colgin, 2015). Slow GR of the CA1 could be generated by the stratum radiatum, a layer known to receive input from the CA3 area that tends to give a slow gamma input to the CA1 area. On the contrary, fast GR seems to be generated by the stratum lacunosummoleculare, a layer of the CA1 receiving input from the entorhinal cortex that seems to communicate with fast gamma (Colgin, 2016). Yet, even if the stratum radiatum and the stratum lacunosummoleculare layers show clear predispositions respectively for slow and fast GR, other layers of CA1 could be able to generate both GR.

However, this way of slow/fast GR generation is still controversial and need further studies (Colgin, 2015). By reproducing neural network and oscillations, computational neurosciences help investigating those questions and improve our comprehension of the GR.

The pyramidal interneuronal network gamma (PING) described by Kopell and al. in the chapter *Gamma and Theta Rhythms in Biophysical Models of Hippocampal Circuits* (from the book *Hippocampal Microcircuits: A Computational Modeler's Resource Book*, 2010) use Hodgkin-Huxley model, that produce realistic neurons behaviours, in order to model slow GR. Using the Kopell and al. PING network we show that an Integrate-and-fire model (IF) is enough to reproduce the GR. By substituting the Hodgkin-Huxley model by a basic leak IF model we reduce strongly the computational cost. More, we show that the important information to produce GR is the action potential presence itself rather than its biological verisimilitude.

We also show that it is possible to generate either slow or fast GR in a simple PING network receiving stochastic input, depending on the neurons characteristics (time constants). Using the Python module Brian (Goodman and Brette 2009), we created a PING network made of 80 pyramidal cells (PC, the main excitatory neuron in the hippocampus) and 20 interneurons (inhibitory neurons) modelled by quadratic IF. A synaptic connection (alpha synaptic model) is created from every interneuron to every PC and from every PC to every interneuron. Only the PC received a stochastic input from 20 neurons generating independent Poisson spike trains. With the same network specifications, it is possible to generate either slow or fast gamma only by changing the times constants describing the neurons. This result shows that the GR frequency depend on the neurons characteristics (membrane time and synaptic time constant, which can vary over time) rather than the network configuration (connexions and action potential weight, which are more stable over time).

By giving the output of the precedent network as an input to a second structurally identical neural network, we show that the second network, respectively with a slow or fast gamma as an

*. Intervenant

input, is able to generate either slow or fast GR. It is important to note that the generation of the two GR by the second network is achieved without a change of its time constant. This result can be an explanation step of slow/fast GR generation in the CA1 area following the hypothesis exposed at the beginning of this abstract.

In this work, we have shown that a single neural network model can exhibit both slow and fast GR, depending on the input rhythm and not on the network structure. This result can be a contribution to the understanding of gamma rhythms generation in the hippocampus CA1 field. The gain in computational cost archived by using simple model (like IF model) could permit modelling of greater and more elaborate neural networks, which could be a way to increase physiological relevance. More, it could permit the modelling of several communicating networks, and so, the modelling of information processing in the hippocampus.

I would like to thank the computer science department of the Faculty of Science of the University of Montpellier (`http://deptinfofds.univ-montp2.fr/`) and the Numev Labex (`http://www.lirmm.fr/numev/`) as well as the Computing Science & Mathematics department of the University of Stirling for accepting to finance my participation in JOBIM.

References

- Colgin LL., Denninger T., Fyhn M., Hafting T., Bonnevie T., Jensen O., et al. (2009). Frequency of gamma oscillations routes flow of information in the hippocampus. *Nature* 462:353–357.

- Colgin LL (2015). Do slow and fast gamma rhythms correspond to distinct functional states in the hippocampal network? *Brain Research* 1621:309–315.

- Colgin LL (2016). Rhythms of the hippocampal network. *Nat Rev Neurosci* 17(4):239-49.

- Goodman DF and Brette R (2009). The brian simulator. *Front Neurosci* doi:10.3389.

- Kopell N., C. Borgers, D. Pervouchine, P. Malerba, and A. Tort, "Gamma and Theta Rhythms in Biophysical Models of Hippocampal Circuits" in *Hippocampal Microcircuits: A Computational Modeler's Resource Book*. Springer Series in Computational Neuroscience 5, 2010.

Mots clefs : neuroscience, modelling, gamma rhythms

Metabolic investigation of the mycoplasmas from the swine respiratory tract

Mariana Galvao Ferrarini [*1], Scheila Gabriele Mucha[2], Marie-France Sagot[3], Arnaldo Zaha[2]

Poster 31

[1] ERABLE (LBBE Lyon / INRIA Grenoble Rhône-Alpes) – CNRS : UMR5558, INRIA, Université Claude Bernard - Lyon I (UCBL), Laboratoire de Biométrie et Biologie Évolutive. UMR CNRS 5558 Campus de La Doua - Université Claude Bernard - Lyon 1 Bâtiment Grégoire Mendel - 16 rue Raphaël Dubois, F-69 100 Villeurbanne – INRIA Grenoble - Rhône-Alpes Inovallée, 655 avenue de l'Europe, Montbonnot, 38 334 Saint Ismier Cedex, France
[2] Universidade Federal do Rio Grande do Sul - UFRGS – Brazil/Brésil
[3] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France

The swine respiratory tract is colonised by several bacteria, among which are three common *Mycoplasma* species: *Mycoplasma hyopneumoniae*, *Mycoplasma flocculare* and *Mycoplasma hyorhinis* [Mare and Switzer, 1965; Meyling and Friis, 1972; Rose *et al.*, 1979]. Even though little information is available concerning the prevalence of bacteria in healthy lungs, these three *Mycoplasma* species have been isolated from the respiratory tract from both healthy and diseased pigs [Ciprian *et al.*, 1988; Fano *et al.*, 2005; Pieters *et al.*, 2010]. While colonization by *M. flocculare* is virtually asymptomatic, *M. hyopneumoniae* is the causative agent of enzootic pneumonia and *M. hyorhinis* is present in cases of pneumonia, polyserositis and arthritis [Kobisch and Friis, 1971]. Several studies also suggest a synergistic role for *M. hyopneumoniae* in the initiation of a variety of other bacterial and viral infections [Ciprian *et al.*, 1994], which explains why this pathogen is considered to be a major cause of economic loss in the pig industry [Maes *et al.*, 1996]. In addition to these mycoplasmas, *Mycoplasma hyosynoviae*, the primary agent of non-purulent arthritis, can occasionally be found in the lower respiratory tract as an opportunistic bacteria of pre-existing pneumonic lesions [Friis, 1971].

Besides mycoplasmal pneumonia, the Porcine Respiratory Disease Complex (PRDC) has emerged as an economically signigicant respiratory disorder characterized by the slow growth, fever, cough, loss of appetite, lethargy and dyspnea in pigs [Thacker *et al.*, 2001; Choi *et al.*, 2003]. Even though many species are related to PRDC, it is essential to note that enzootic pneumonia caused by *M. hyopneumoniae* is by far the most costly disease in pig industry, and this bacteria is usually seen as an essential component to the successful establishment of a pathogenic community in the host [Sorenson *et al.*, 2013]. Also, *M. hyopneumoniae* infections take longer to cause lesions and take longer to be successfully eliminated than infections from other pathogens [Thacker *et al.*, 2001].

While mycoplasma diseases in swine have been extensively studied, they have not been explored from a mathematical/computational point of view, mostly because their genome sequences were not available until recently [Minion *et al.*, 2004; Vasconcelos *et al.*, 2005; Liu *et al.*, 2010; Liu *et al.*, 2011; Calcutt *et al.*, 2012; Siqueira *et al.*, 2013; Liu *et al., 2013; Calcutt et al.*, 2015]. Moreover, although recent studies have placed *M. hyopneumoniae*, *M. flocculare* and *M. hyorhinis* in the hyopneumoniae clade by phylogenomic analysis [Siqueira *et al.*, 2013], which corroborates with their high 16S rRNA sequence similarity [Stemke *et al.*, 1992], it is not yet clear what causes the specific pathogenicity or lack thereof in each of them. This elevated genomic resemblance combined with their different levels of pathogenicity is an indication that these species, as for most mycoplasmas, have unknown mechanisms of virulence and of differential expression.

*. Intervenant

Moreover, establishing which are the virulence and pathogenicity factors is generally seen as an open problem in *M. hyopneumoniae*, mostly because non-pathogenic strains are extremely similar to the pathogenic ones. And even though pathogenic determinants such as adhesion to the host cells and evasion from the immune response have already been well-described in the literature for both *M. hyopneumoniae* and *M. hyorhinis* [Citti *et al.*, 1997; Djordjevic *et al.*, 2004; Whittlestone, 2012; Xiong *et al.*, 2016], no link between metabolism and pathogenicity has been made for either species up to date; mostly because the lack of experimental information hinders the formulation of hypotheses concerning the specific pathologies of each species.

The genome sizes of *Mycoplasma* spp. range from 580 kb (*Mycoplasma genitalium*) to more than 1,358 kb (*Mycoplasma penetrans*), representing an important example of genome reduction (and minimal metabolism) during the evolutionary process. It is possible that in an initial symbiotic phase, the host provided a broad range of metabolites for these bacteria. This, together with the ability of the bacteria to uptake such compounds, made several activities dispensable for the bacterial life. Over the course of evolution, these bacteria would have lost some of the genes that became unnecessary for life in an environment conditioned by another genome [Andersson and Kurland, 1998].

In this way, we aimed at studying the reduced metabolism of *M. hyorhinis*, *M. hyopneumoniae* and *M. flocculare* to try to understand what could influence their different life-styles and pathogenicity. We reconstructed the metabolic models for several strains of each of the species with the intention of comparing the networks and finding possible explanations for the different levels of pathogenicity observed among and within species. We were able to detect slight differences in the metabolic networks reconstructed that can partially explain the incapacity of *M. flocculare* to cause disease or the ability of *M. hyorhinis* to grow faster than the other two species. The models for *M. hyorhinis* could uptake a wider range of carbohydrates which, in turn, might reflect the overall better growth rates obtained for this species *in vitro*. This may also explain why this species is considered a common contaminant of cell cultures [Nikfarjam and Farzaneh, 2012].

As for the lack of pathogenicity of *M. flocculare*, the enzyme responsible for the production of the highly toxic hydrogen peroxide (a well-characterized virulence factor in the lung pathogens *Mycoplasma mycoides* and *Mycoplasma pneumoniae* [Vilei *et al.*, 2001; Hames *et al.*, 2009]), is absent in this species. On the other hand, *M. hyopneumoniae* and *M. hyorhinis* harbor in their genomes this specific enzyme and can use glycerol-sn-3-phosphate as carbon source with the production of hydrogen peroxide as a byproduct of the reaction. Moreover and as mentioned before, *M. hyopneumoniae* and *M. flocculare* are closely related genetically, and both species have been shown to adhere to cilia in a similar way [Young *et al.*, 2000], indicating that the inability of *M. flocculare* to cause disease (or the ability of *M. hyopneumoniae* to cause disease) might not be directly related to adhesion.

Nevertheless, there is little metabolic experimental data available for the three species, which makes the reconstruction of a reliable metabolic model an extremely time-consuming work. Together with the fact that these mycoplasmas are generally grown in complex media, with high serum concentrations, we needed additional experimental data in order to compare the *in silico* reconstructed metabolic networks with the *in vivo* metabolic characteristics of the three species. For this reason, we also performed nuclear magnetic resonance spectroscopy (NMR) analyses to detect metabolites consumed and produced in both complex and defined media. These experiments corroborated with the reconstructed models and suggested two new features in particular: (i) the uptake of myo-inositol in *M. hyopneumoniae* might be related to a higher acetate production, and (ii) *M. hyorhinis* showed a surprisingly reduced ability to convert pyruvate to acetate in the growth conditions used in this study.

All these *in silico* and *in vivo* metabolic differences might influence the different levels of pathogenicity in each of the species studied here. Furthermore, this work will serve as a basis for the study of the differential metabolism and pathologies caused by the swine respiratory tract mycoplasmas and may help to propose ways to prevent disease development in the future.

Mots clefs : Mycoplasma, Systems Biology, Metabolic Networks, Pathogenicity, Swine

Unravelling the transcriptome architecture of a non-model bacterium : *Flavobacterium psychrophilum*

Poster 32

Cyprien Guérin [*] [†1]

[1] INRA, UR1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut national de la recherche agronomique (INRA) – Bâtiment 210-233 Domaine de Vilvert, F-78 350 Jouy en Josas Cedex, France

Tatiana Rochat (1), Cyprien Guérin (2), Erwin Vandijk (3), Brigitte Kerouault (1), Bogdan Mirauta (2,6), Claude Thermes (3), Francis Repoila (4,5), Éric Duchaud (1), Pierre Nicolas (2)

(1) INRA, UR892, Virologie et Immunologie Moléculaires, 78352 Jouy-en-Josas, France

(2) INRA, Mathématiques et Informatique Appliquées du Génome à l'Environnement, 78352 Jouy-en-Josas, France

(3) Next-Generation Sequencing Facility, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Sud, Université Paris-Saclay, F-91198 Gif-sur-Yvette Cedex, France.

(4) INRA, UMR1319 Micalis, F78350 Jouy-en-Josas, France

(5) AgroParisTech, UMR Micalis, F78350 Jouy-en-Josas, France

(6) Biologie Computationnelle et Quantitative, UPMC and CNRS UMR7238, Paris

The fish-pathogen Flavobacterium psychrophilum, a member of the family Flavobacteriaceae in the phylum Bacteroidetes, is a major sanitary concern for the salmonid farming industry worldwide. Juvenile salmonids are particularly susceptible to this Gram-negative aerobic bacterium, which is able to survival outside the fish during long periods of starvation in freshwater environments. The genome of a strain of F. psychrophilum has been sequenced, assembled and annotated (Duchaud et al., 2007) and data on the genetic diversity of the species has been accumulated (Nicolas et al., 2008; `http://pubmlst.org/fpsychrophilum/`). However, the function of many genes and the regulatory networks of this bacterium belonging to a phylum remote from most well studied bacteria remain almost completely unknown. In this context, trancriptome analyses can considerably increase our knowledge of the genetic mechanisms that underlie the physiological adaptation of this non-model bacterium to its different life-styles.

Our objective in this work was to combine in a cost effective manner several genome-wide experimental approaches to unravel the transcriptome architecture of F. psychrophilum. Different techniques were used to address several objectives: i) characterize transcription start sites (TSS) and 5' and 3' regions with base resolution, ii) discover new transcribed regions, iii) study expression levels across a large diversity of biological conditions.

A virulent strain isolated from a Coho salmon was selected. Global and 5'-targeted RNA-seq analyses were applied to a pool of bacterial RNAs extracted from cells undergoing various conditions to enrich the structural annotation of genome by establishing a repertoire of transcribed regions and TSS. In silico analyses involved: the delineation of transcription units from the global RNA-seq profiles using Parseq, an approach based on state-space models (Mirauta et al., 2014); identification of the TSSs based on a custom analysis framework of 5'-targeted RNA-Seq data; clustering of the promoters based on sigma factor binding sites using treemm a dedicated

*. Intervenant

†. Corresponding author : `cyprien.guerin@jouy.inra.fr`

motif finding algorithm (Nicolas et al., 2012). Based on this enriched structural annotation, a transcription array was designed to analyse transcriptome changes across conditions.

Expressed genes as well as potential regulatory elements were listed (i.e. TSS, 5'-untranslated regions, non-coding and antisense RNAs). In silico searches for consensus promoter sequences upstream of identified TSSs revealed more than one thousand sigma-70 dependent promoters as well as three promoters groups responding likely to alternative sigma factors. Data on the condition-dependent transcriptome obtained with this array are already available for twenty-four conditions and new experiments aiming at maximizing the coverage of the life-styles of this bacterium are currently under development.

References

E. Duchaud, M. Boussaha, V. Loux, J.-F. Bernardet, C. Michel, B. Kerouault, S. Mondot, P. Nicolas, R. Bossy, C. Caron, P. Bessiéres, J.-F. Gibrat, S. Claverol, F. Dumetz, M. Le Hénaff and A. Benmansour. (2007) Complete genome sequence of the fish pathogen Flavobacterium psychrophilum. Nature Biotechnology. 25. 763-9.

P. Nicolas, S. Mondot, G. Achaz, C. Bouchenot, J.-F. Bernardet and E. Duchaud. (2008) Population structure of the fish-pathogenic bacterium Flavobacterium psychrophilum. Appl. Environ. Microbiol. 74. 3702-9.

P. Nicolas, U. Mäder, E. Dervyn, T. Rochat, A. Leduc, N. Pigeonneau, E. Bidnenko, E. Marchadier, M. Hoebeke, (41 authors), and P. Noirot. (2012) Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis. Science. 335. 1099-1103.

B. Mirauta, P. Nicolas, and H. Richard (2014) Parseq: reconstruction of microbial transcription landscape from RNA-Seq read counts using state-space models. Bioinformatics. 30:1409-16.

Mots clefs : transcriptome, transcriptomic, non model, bacterium, RNA seq, array, transcription array, genome wide

Multipus : conception de communautés microbiennes pour la production de composés d'intérêt

Poster 33

Alice Julien-Laferrière [*1,2], Laurent Bulteau[3], Delphine Parrot[1,2], Alberto Marchetti-Spaccamela[4], Leen Stougie[5], Susana Vinga[6], Arnaud Mary[1,2], Marie-France Sagot [†1,2]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Équipe Érable (INRIA Grenoble Rhône-Alpes) – L'Institut National de Recherche en Informatique et en Automatique (INRIA) – France
[3] Laboratoire d'Informatique Gaspard-Monge (LIGM) – Fédération de Recherche Bézout, École des Ponts ParisTech (ENPC), ESIEE, Université Paris-Est Marne-la-Vallée (UPEMLV), CNRS : UMR8049 – Cité Descartes, Bâtiment Copernic, 5 boulevard Descartes, F-77 454 Marne-la-Vallée Cedex 2, France
[4] Dipartimento di Informatica e Sistemistica [Rome] – University of Rome "La Sapienza" via Salaria 113, I-00198 Rome, Italie
[5] Centrum voor Wiskunde en Informatica (CWI) – Kruislaan 413 P.O. Box 94079 1090 GB Amsterdam, Pays-Bas
[6] Departamento de Engenharia Mecanica Instituta Superior Technico [Lisboa] (IDMEC/IST) – Av Rovisco Pais 1049-001 Lisboa, Portugal

Depuis le début des années 2000, la biologie synthétique a permis de produire des composés innovants et d'intérêt pour différentes industries telles que la pharmacologie (pénicilline, insuline, artemésinine), les énergies (biogaz) ou les polymères (1,3-propanediol). Ces productions sont généralement effectuées en cultures pures. Mais dernièrement, l'usage de communautés microbiennes a été proposé afin de permettre par exemple : la division des voies de synthèses entre différentes souches, l'exploitation de co-produits, la levée de certaines inhibitions dues à l'accumulation de métabolites toxiques ainsi que la co-production de composés.

Ici nous proposons un algorithme permettant la sélection de communautés artificielles (composées d'espèces dont l'interaction n'est pas observée dans la nature) ou synthétiques (composées d'organismes modifiés génétiquement pour acquérir de nouvelles capacités métaboliques). L'algorithme permet aussi d'obtenir les voies de synthèses utlisées.

Cet algorithme utilise la topologie des réseaux métaboliques, en représentant les modèles métaboliques comme des hypergraphes dirigés. Dans ce cas, les nœuds sont les métabolites du réseau tandis que les hyperarcs représentent les réactions allant d'un ensemble de substrats à un ensemble de produits. Nous fusionnons les réseaux métaboliques de tous les microorganismes du consortium (que nous appellerons *travailleurs*). Pour cela, le réseau métabolique de chaque travailleur est modélisé tel un hypergraphe. Dans chacun de ces réseaux, des composés et réactions peuvent être ajoutés si ils sont existants dans d'autres organismes que les travailleurs (il s'agit alors de l'insertion de métabolites et/ou de réactions exogènes). Enfin, le transport entre les organismes du consortium est permis en ajoutant un arc simple reliant un composé dans un organisme au même composé dans un autre organisme. Les hyperarcs sont ensuite pondérés en utilisant un *a-priori* sur le coût d'une réaction vis-à-vis du procédé de production d'un composé d'intérêt. Dans un premier temps, nous avons divisé les hyperarcs en trois catégories : endogènes (présents initialement chez les travailleurs), exogènes (nécessitant l'insertion de gènes codant pour des

*. Intervenant
†. Corresponding author : marie-france.sagot@inria.fr

enzymes dans le réseau) et transports (permettant l'échange d'un composé entre deux travailleurs). Les deux dernières catégories auront des poids supérieurs. En effet l'insertion d'un ou plusieurs gènes codants pour une enzyme dans un organisme peut s'avérer coûteux. Cela nécessite des manipulations génétiques et de l'énergie pour l'organisme afin d'exprimer les nouveaux gènes introduits. Enfin, exporter et importer des composés demande l'utilisation de transporteurs qui peuvent être demandeur en énergie ou nécessiter de grands gradients de concentrations.

Nous essayons ensuite de résoudre le problème de l'hyperarbre de Steiner de poids minimum. Une instance du problème est composée d'un hypergraphe dirigé et pondéré, d'un ensemble de composés sources et d'un ensemble de composés cibles. Nous désirons trouver la solution la plus légère, c'est à dire un ensemble d'hyperarcs (réactions) tel que tous les composés cibles soient atteints à partir des sources. Cependant, un hyperarc ne peut être sélectionné que si tous ses substrats ont été produits auparavant ou font partie de l'ensemble de sources initial. Nous nommons les sources d'un hyperarc des tentacules, un hyperarc est tentaculaire si il a plus d'un tentacule.

Nous montrons que ce problème est NP-difficile mais qu'il existe un algorithme FPT (*fixed-parameter tractable*) avec comme paramètres le nombre de cibles et le nombre d'hyperarcs tentaculaires. L'algorithme énumère tout d'abord toutes les combinaisons possibles des hyperarcs tentaculaires où une combinaison est un sous-ensemble des hyperarcs ordonnés selon l'ordre topologique de la solution. Pour chaque combinaison, en utilisant le sous-réseau composé uniquement d'arc simples (non tentaculaires), l'algorithme calcule la manière optimale (de poids minimum) pour relier les hyperarcs tentaculaires choisis. Cela est fait en utilisant une routine de programmation dynamique qui généralise un algorithme FPT classique pour le problème d'arbre de Steiner dirigé (en utilisant le nombre de cibles comme paramètre).

Nous avons assigné des poids uniformes pour les hyperarcs au sein des catégories définies précédemment (endogène = 0.01, exogène = 1, transport = 1) et appliqué l'algorithme à deux consortiums en fixant le nombre d'hyperarcs tentaculaires dans les solutions à k=3 au maximum.

Tout d'abord nous proposons une communauté synthétique pour la production de deux antibiotiques : la pénicilline et la céphalosporine C. Nous avons sélectionné comme possibles travailleurs trois actinobacteries (*Streptomyces cattleya*, *Rhodococcus jostii* RAH 1, *Rhodococcus erythropolis* BG43) et une archée méthanogène (*Methanosarcina barkeri*) et proposé comme seule source la cellulose qui est un substrat peu coûteux et facilement disponible. Les réactions insérées proviennent de deux organismes : un champignon (*Aspergillus Nidulans*) et une actinobactérie (*Streptomyces rapamycinicus*) qui possèdent la capacité de synthétiser les deux antibiotiques bêta-lactame visés. La solution obtenue montre que le meilleur consortium microbien pour la production des deux bêta-lactamines est constitué de *Streptomyces cattleya* et *Methanosarcina barkeri*. Multipus nous permet donc d'obtenir les voies métaboliques nécessaires à la synthèse des produits mais aussi de sélectionner le meilleur consortium possible dans un plus large ensemble d'espèces.

Nous avons ensuite testé un consortium artificiel constitué de *Clostridium butyricum* qui produit naturellement du 1,3-propanediol (PDO) et une archée méthanogène *Methanosarcina mazei*. Le 1,3-propanediol est un polymère d'intérêt pour l'industrie, il est utilisé pour la fabrication de nombreux produits (peintures, composites, etc.). Lors de la production de PDO, de l'acétate est aussi synthétisé. Or l'acétate exerce une action inhibitrice sur la production de PDO et la croissance de C. butyricum car il est toxique lorsque présent en grandes concentrations. Cependant *M. mazei* peut pousser sur de l'acétate et produit du méthane qui peut être récupéré pour créer du biogaz. Puisque les deux organismes peuvent produire naturellement les deux produits choisis (PDO et méthane), nous n'avons pas introduit de nouvelles réactions. Nous avons proposé d'utiliser comme source de carbone unique du glycérol, un sous-produit de la production de biodiesel et donc un substrat de choix pour des procédés biotechnologiques. Dans un premier temps, les poids des réactions (endogène, exogène et transport) ont été définis comme pour l'exemple précédent. Afin de produire du PDO et du méthane avec du glycérol, les deux organismes échangent de l'acétyl-CoA. Or l'acétyl-CoA est essentiel pour *C. butyricum*, celui-ci n'aurait donc pas d'intêret à

diminuer sa concentration interne d'acétyl-CoA. Cependant, puisque l'acétate est toxique pour *C. butyricum*, nous avons supposé qu'un processus d'excrétion de l'acétate pouvait exister. Aussi dans un second temps, nous avons diminué le poids du transport de l'acétate de *Clostridium butyricum* à *Methanosarcina mazei* (de 1 à 0.5). Dans ce cas, nous obtenons un ensemble de réactions permettant effectivement de produire à la fois du PDO et du méthane en partant uniquement du glycérol grâce à la consommation de l'acétate par *M. mazei*. Cette consommation pourrait permettre d'obtenir une production plus importante de 1,3-propanediol puisque l'acétate ne serait alors plus présent en grande quantité dans le milieu. Nous pouvons donc, à l'aide d'un paramétrage plus fin des poids au sein des catégories de réactions, obtenir des sous-réseaux de production plus réalistes.

Pour conclure, nous proposons Multipus, un programme contenant un algorithme d'énumération qui permet d'inférer quelles espèces inclure au sein d'une communauté ainsi que les réactions à utiliser pour une production de composés d'interêt. Nous avons appliqué Multipus à deux cas : tout d'abord à la production jointe de deux antibiotiques (la pénicilline et la céphalosporine C) puis à la production d'un polymère (1,3-propanediol) associée à celle du méthane afin de consommer un sous-produit toxique (l'acétate) et de permettre un meilleur rendement.

Multipus (MULTIple species for the synthetic Production of Useful biochemical Substances) est disponible à l'adresse : `http://multipus.gforge.inria.fr/`.

Mots clefs : microbial consortia, synthetic biology, combinatorial algorithm

Étude d'un réseau toxico-génomique pondéré en vue de la prédiction *in silico* de la toxicité des substances chimiques

Poster 34

Estelle Lecluze [*1], Thomas Darde[1], Frédéric Chalmel[1], Antoine Rolland[1], Emmanuelle Becker [* †1]

[1] Institut de recherche, santé, environnement et travail [Rennes] (Irset) – Inserm : U1085, École des Hautes Études en Santé Publique [EHESP], Universite de Rennes 1 – 9 avenue du Pr. Léon Bernard, F-35 000 Rennes, France

Introduction et contexte international

L'Union Européenne a coordonné une campagne de recensement ayant dénombré plus de 100 000 substances dont seulement 3 % ont fait l'objet d'analyses approfondies pour en évaluer la toxicité et établir des liens avec des pathologies humaines et des phénotypes délétères. Suite à ce constat, elle a depuis 2007 mis en place le programme *Registration, Evaluation, Authorization of CHemicals* (REACH), qui impose aux industriels produisant ou important des produits chimiques sur le territoire européen d'évaluer leur toxicité et de démontrer l'innocuité de ceux-ci.

Parmi les différentes approches qui existent pour évaluer la toxicité d'une substance, l'utilisation de modèles animaux (*in vivo* ou *ex vivo*) reste la plus utilisée. Cependant, cette approche n'est pas optimale pour plusieurs raisons : (i) il existe des différences entre ces organismes modèles et l'homme, constituant donc une barrière à l'interprétation des résultats ; (ii) ces techniques sont onéreuses et relativement complexes à mettre en place ; et enfin (iii) le nombre d'animaux à sacrifier pour mener à bien ces études est extrêmement important. Des modèles *in vitro* sont d'ores et déjà utilisés, sous la forme de lignées cellulaires ou de cultures organotypiques dont la mise en place est complexe. Cependant, les conditions *in vivo* de ces cellules ne sont pas nécessairement reproduites, et les résultats de ce type d'approches peuvent être différents d'une réponse physiologique.

La prise en compte simultanée de (i) l'ampleur du défi que représente l'investigation toxicologique de milliers de composés, et de leurs potentiels mélanges, ainsi que (ii) la complexité et le coût des approches *in vivo*, *ex vivo* et *in vitro* existantes, pousse les législateurs et scientifiques à considérer les approches prédictives *in silico* comme des compléments pertinents aux approches expérimentales, permettant à la fois d'identifier les molécules ou les associations de molécules à étudier en priorité, de déterminer les tissus potentiellement impactés et pour certaines approches d'identifier les mécanismes sous-jacents.

L'objectif serait ici de développer une méthode permettant une caractérisation fine de la toxicité, qui permette de différencier les différents types de toxicité (nephrotoxicité, neurotoxicité, reprotoxicité…) en se basant sur la signature toxico-génomique du composé. La signature toxico-génomique se définit par l'ensemble des gènes différentiellement exprimés suite à une exposition. La toxicité d'un composé c étant fonction de sa dose d, de la durée d'exposition d'exposition t, ainsi que de l'organe étudié o, nous considèrerons chaque condition C définie par un quadruplet $\overline{C}$ comme indépendante des autres, et chaque quadruplet $\overline{C}$ aura sa propre signature toxico-génomique.

*. Intervenant

†. Corresponding author : emmanuelle.becker@univ-rennes1.fr

Dans un premier temps, nous décrirons les données utilisées ainsi que leur traitement en vue d'aboutir à une matrice de signatures toxico-génomiques. Nous présenterons ensuite la construction d'un réseau pondéré de signatures toxico-génomiques, et étudierons quelques propriétés de celui-ci. Enfin, nous nous intéresserons à la modularité au sein de ce réseau, avant de tester la pertinence du transfert d'annotations de toxicité via les modules de ce réseau.

Traitement des données pour aboutir à une matrice de données toxico-génomique

Les données proviennent des banques de données DrugMatrix [1] et TgGate [2], et contiennent les résultats d'expériences toxicologiques à grande échelle menées chez le rat, testant les effets de produits chimiques variés (par exemple, 638 composés thérapeutiques, industriels ou environnementaux dans la banque DrugMatrix).

Après un contrôle qualité manuel et une normalisation globale (méthode RMA), chaque condition a été comparée à son contrôle, et les gènes différentiellement exprimés ont été identifiés pour chaque condition (changement de ratio d'expression > 1.5, test statistique avec modèle linéaire d'estimation de la variance, et correction pour les tests multiples). Seuls les composés ayant une signature toxico-génomique claire (plus de 10 gènes différentiellement exprimés) ont été conservés, et les gènes ne présentant aucune variation d'expression ont été retirés du jeu de données.

Au final, on dispose d'une matrice de 3 022 conditions d'exposition (colonnes) sur 11 434 gènes (lignes). Les 3 022 conditions correspondent à différentes conditions d'expositions à 410 composés chimiques distincts. Les organes les plus étudiés sont deux organes clés du point de vue toxicologique : le foie, organe de la détoxication (2 246 conditions d'expositions testées), et le rein, organe d'élimination des métabolites (573 conditions). La plupart des composés n'ont été testés que dans une seule condition, alors que les composés les plus étudiés ont été testés dans plusieurs dizaines de conditions : 20 conditions et 5 doses différentes pour la gentamicine (antibiotique), 29 conditions et 7 doses différentes pour le cisplatine (chimiothérapie), et jusqu'à 56 conditions et 5 doses différentes pour l'ethinylestradiol (œstrogène actif par voie orale le plus utilisé au monde, notamment dans la plupart des pilules contraceptives combinées).

Un réseau toxico-génomique pondéré

Construction du réseau toxico-génomique pondéré

Les réseaux biologiques ont été étudiés à différentes échelles : réseaux d'interactions protéine-protéine, réseaux de maladies, réseau de composés, etc... Néanmoins, si ces graphes de composés ont été mis en évidence, l'établissement d'un réseau toxicologique connectant des conditions toxicologiques précises définies par des quadruplets $\overline{C}$ n'a jamais été entrepris. Nous proposons donc de définir un tel graphe, dont les sommets sont les quadruplets $\overline{C}$, et dont les sommets sont reliés entre eux si leurs signatures toxico-génomiques sont proches.

Afin de pondérer les arêtes par une mesure reflétant la similarité des signatures toxico-génomiques, nous proposons d'utiliser un indice de Kappa de Cohen pondéré. Le Kappa de Cohen est habituellement utilisée afin de mesurer l'accord de deux variables quantitatives ayant les mêmes modalités, et cherche à estimer la concordance des deux variables [3]. Pour transposer cet indice à notre problématique, les signatures toxico-génomiques des composés seront assimilées aux variables aléatoires, les modalités observées étant : (a) gène sur-exprimé, (b) pas de variation d'expression du gène, et (c) gène sous-exprimé. L'indice du Kappa de Cohen pondéré varie de 1 (accord parfait) à —1 (désaccord total).

Pour construire un graphe à partir de ces mesures, chaque condition correspond à un sommet, et deux conditions sont reliées entre elles si leur indice Kappa de Cohen pondéré est supérieur ou égal à 0. Les arêtes sont pondérées par la valeur du Kappa de Cohen correspondante ; les poids

varient donc entre 0 et 1. Dans un premier temps, les réseaux des deux organes les plus représentés ont été étudiés séparément : deux réseaux ont donc été construits : l'un à partir des 573 conditions du rein, l'autre à partir des 2 246 conditions du foie.

Étude de la centralité et des poids

Le graphe étant pondéré, il est intéressant de comparer la distribution des degrés des sommets et celle des forces des sommets ; la force d'un sommet se définissant comme la somme des poids des arêtes incidentes au sommet [4].

Dans le graphe du rein (573 conditions), le poids moyen des arêtes est de 0,067, la force moyenne est de 29,10 et varie entre 4,29 et 61,52. Dans le graphe du foie (2 246 conditions), le poids moyen des arêtes est de 0,055, la force moyenne est de est de 88,94 et varie entre 3,13 et 217,99. Au sein des 2 graphes, on observe une synergie entre degré des sommets et poids des arêtes adjacentes ; la force des sommets dépend du degré sous la forme $s(k) \approx k^b$: elle n'est pas proportionnelle à leur degré, mais inférieure à l'hypothèse proportionnelle pour les faibles degrés, et supérieure à l'hypothèse proportionnelle pour les forts degrés. Cette propriété est liée à la profondeur d'analyse des composés, le foie et le rein étant les organes au sein duquel le plus grand nombre de conditions ont été testées ; et s'explique par (i) un plus grand nombre de conditions testées par composé, lesquelles ont des interactions de poids fort ; (ii) aux conditions d'expositions les plus longues, qui s'agglomèrent entre elles ; et (iii) par les composés azolés, qui présentent des similarités de signatures importantes ; (iv) un grand nombre d'interactions non significatives, c'est-à-dire dont la signature toxico-génomique n'est pas clairement définie.

Modularité dans le réseau toxico-génomique pondéré

En nous basant sur des indicateurs topologiques de modularité (coefficients de clustering pondéré par la méthode de Barrat et al [4], modularité de Newman [5]), nous proposons trois axes d'analyse principaux :

(1) La proximité des signatures toxico-génomique se traduit-elle fréquemment par une co-annotation de toxicité ? Il s'agit de valider l'hypothèse que des composés ayant des signatures toxico-génomiques proches ont des toxicités proches ; et nos résultats permettent de valider cette hypothèse.

(2) Inversement, les composés connus pour présenter une certaine toxicité montrent –ils une tendance à s'agglomérer dans le graphe obtenu ? Cette assertion est la réciproque de la précédente, et nos résultats obtenus en comparant le graphe toxico-génomique avec des graphes toxico-génomiques randomisés montrent que cette proposition est vérifiée pour certaines toxicités (par exemple cardiotoxicité **, perturbation endocrinienne **, maladie de Parkinson*) et pas d'autres. Il est intéressant de noter que les toxicités montrant une forte modularité ne sont pas particulièrement liées à l'organe sur lequel ont été obtenues les signatures toxico-génomiques, ce qui peut être une propriété très intéressante pour des approches prédictives.

(3) Peut-on discriminer des sous-ensembles ou classes de conditions densément connectées, en cherchant par exemple à maximiser un surcroît d'arêtes interne tenant compte des pondérations via un critère de modularité de Newman pondéré ? N'ayant pas d'a priori sur la taille des classes, ni sur les seuils de similarités intéressants, nous proposons de ne pas nous limiter à la production d'un système de classe, mais d'étudier aussi les relations entre ces ensembles, en organisant notre système de classes sous la forme d'un arbre (une approche similaire ayant été employée pour les graphes d'interactions protéine-protéine [6,7]).

Transfert d'annotations de toxicité au sein du réseau toxico-génomique pondéré

À partir des classes discriminées précédemment (système de classes chevauchantes généré de façon à maximiser la modularité de Newmann pondérée), nous étudions enfin la pertinence

d'un transfert d'annotations de toxicité au sein de ces classes. Le principe mis en œuvre est similaire à celui exploitant les graphes d'interactions protéine-protéine afin d'inférer la fonction des protéines [6] : après avoir annotés les modules obtenus (quelles toxicités sont enrichies dans ce modules ? lesquelles sont fortement représentées ?), les composés n'ayant pas d'annotations présents dans le modules se voient « transférer » les annotations du module auxquels ils appartiennent.

La pertinence de cette approche a été évaluée par une méthode de bootstrap consistant à répéter l'opération suivante : supprimer les annotations de toxicité connues pour un composé, et comparer celles-ci avec les annotations de toxicités transférées à ce composé via les modules auxquels il appartient. Les résultats soulignent que les annotations volontairement supprimées sont fréquemment re-transférées via l'annotation des modules, et que les annotations transférées sont majoritairement pertinentes.

Références

1. Auerbach, S. S. *et al.* Predicting the hepatocarcinogenic potential of alkenylbenzene flavoring agents using toxicogenomics and machine learning. *Toxicol. Appl. Pharmacol.* 243:300–314 (2010).

2. Igarashi, Y. *et al.* Open TG-GATEs: A large-scale toxicogenomics database. *Nucleic Acids Res.* 43:D921–D927 (2015).

3. Cohen, J. A coefficient of agreement of nominal scales. *Educ. Psychol. Meas.* 20:37–46 (1960).

4. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.* 101:3747–52 (2004).

5. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* 103:8577–82 (2006).

6. Brun, C. *et al.* Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol.* 5:R6 (2003).

7. Becker, E., Robisson, B., Chapple, C. E., Guénoche, A. & Brun, C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics* 28:84–90 (2012).

Mots clefs : toxico, génomique, toxicologie, réseau, modularité, transfert d'annotations

VirHostPred, une méthode d'inférence des interactions protéine-protéine virus/host basée sur l'homologie de séquences protéiques

Poster 35

Justine Picarle [*1], Dominique Guyot[1], Vincent Navratil [†1]

[1] Pôle Rhône-Alpes de Bioinformatique (PRABI) – Université Claude Bernard - Lyon I (UCBL) – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France

Les maladies infectieuses, causées par des virus ou des bactéries, sont une des principales causes de mortalité dans le monde (`http://www.who.int/fr/`). Les virus, contrairement à la majorité des bactéries, sont des parasites intracellulaires obligatoires et utilisent par le biais d'interactions moléculaires complexes et finement régulées, la machinerie de réplication de leurs hôtes afin de se répliquer. Plus particulièrement, les interactions protéine-protéine entre les protéines virales et les protéines de l'hôte jouent un rôle fonctionnel clef à tous les niveaux du cycle infectieux ainsi que dans les processus d'échappement aux réponses antivirales orchestrés par les cellules hôtes (V Navratil et al., 2010). Dans certaines conditions, ces mêmes interactions protéine-protéine sont susceptibles de participer aux perturbations moléculaires impliquées dans la mise en place de maladies (Vincent Navratil, de Chassey, Combe, & Lotteau, 2011). Le traitement des infections virales et des maladies qui leurs sont associées demeure l'un des principaux défis de la santé publique. Il est donc important de mieux caractériser les interactions protéine-protéine virus/hôte puisqu'elles peuvent nous fournir des cibles thérapeutiques potentielles dans le but de produire de nouveaux médicaments (de Chassey, Meyniel-Schicklin, Vonderscher, André, & Lotteau, 2014).

Les technologies expérimentales dites à « haut débit » (crible double hybrides, purification par affinité couplée à la spectrométrie de masse) ont permis une première caractérisation des réseaux d'interaction protéine-protéine virus/hôtes et de leurs propriétés systémiques (Vincent Navratil et al., 2011). La base de connaissance VirHostNet 2.0 a pour mission de collecter, par une approche de biocuration experte, l'ensemble de ces données d'interactions virus/hôtes produites et publiées dans la littérature (Guirimand, Delmotte, & Navratil, 2015). Ainsi, dans la version de janvier 2016, plus de 23 000 interactions protéine-protéine virus/hôte ont été annotées. Ces données sont devenues en quelques années un *gold standard* pour la communauté scientifique. Pour en faciliter l'accès, elles ont été mises à disposition à travers une interface web d'interrogation et de visualisation (`http://virhostnet.prabi.fr/`). Elles sont par ailleurs diffusées au standard international PSI-MI à travers le service web PSICQUIC (Proteomics Standard Initiative Common QUery InterfaCe) (Aranda et al., 2011).

Néanmoins, les ressources pour l'étude des interactions protéine-protéine entre les virus et les hôtes restent encore très limitées. Ces dernières sont souvent restreintes aux quelques virus majeurs en santé humaine tel que les virus de la grippe, HIV, HCV, les herpesvirus et sont ainsi indisponibles pour les virus infectant les espèces d'hôte autres que l'humain. En se basant sur l'ensemble des 18 530 interactions virus/hôtes connues et annotées à ce jour à l'échelle de l'espèce, notre estimation de l'espace totale de recherche de l'interactome virus/hôte atteindrait plus de 6 x 10e11 d'interactions protéine-protéine (communication personnelle). Compte tenu de l'étendu de cet espace de recherche et du coût élevé des techniques expérimentales « haut débit », les approches de prédiction *in silico* d'interactions protéine-protéine en bio-informatique basée sur l'homologie de séquences et de structures sembleraient être un bon point de départ pour accélérer

*. Intervenant

†. Corresponding author : `navratil@prabi.fr`

la recherche dans le domaine (Lewis, Jones, Porter, & Deane, 2012) (de Chassey et al., 2013). Alors que ce type d'approche est devenu un classique pour la prédiction d'interactions protéine-protéine intra-espèces (Garcia-Garcia, Schleker, Klein-Seetharaman, & Oliva, 2012), encore peu d'études se sont penchés sur la prédiction inter-espèces impliquant des virus. Par ailleurs, à notre connaissance, aucun outil bioinformatique n'est disponible et/ou adapté à un passage des prédictions à l'échelle de l'ensemble des 18 530 couples virus/hôte connus.

Pour répondre à cette problématique, nous avons développé VirHostPred, une méthode d'inférence d'interactions protéine-protéine interespèces virus/hôte basée sur l'information d'homologie de séquences protéiques. Notre hypothèse de départ repose sur l'observation qu'une interaction protéine-protéine virus/hôte (VH) puisse être conservée au cours de l'évolution soit par spéciation ou convergence évolutive, on parle alors d'interactions interologues. Ainsi, si la protéine A du virus V est connue pour interagir avec la protéine B de l'hôte H et que la protéine A' du virus V' est homologue de A, alors on peut inférer l'interaction interologue entre la protéine A' et B. De même, si la protéine A du virus V est connue pour interagir avec la protéine B de l'hôte H et que la protéine A' du virus V' et la protéine B' de l'hôte H' sont des homologues respectifs de A et B, alors on peut inférer l'interaction interologue entre la protéine A' et B'. La méthode de prédiction VirHostPred, a été appliquée à un jeu de données d'interactions protéine-protéine connues (n=855 012 interactions) extraites de Virhostnet 2.0. La première étape de VirHostPred repose sur l'utilisation intensive de blastp afin de mesurer la similarité entre la totalité des séquences protéiques d'UniProt (n=60 560 696, version de janvier 2016) et une base de séquences de partenaires protéiques (n=94 260). Les résultats de blastp sont ensuite pré-filtrés selon les critères d'e-valeur, de pourcentage d'identité et de pourcentage de couverture des alignements. Dans la phase d'évaluation de notre méthode, nous avons utilisé un seuil minimal de 80 % pour la couverture globale d'alignement, un seuil minimal de 50 % pour l'identité et de 10^{-4} pour la e-valeur. L'algorithme de recherche des interologues virus/hôte a été entièrement parallélisé et nous a permis de parcourir 20 % de l'espace total de recherche (communication personnelle). Le nombre total de prédictions obtenues avoisine les 3×10^8 interactions protéine-protéine et concerne plus de 250 nouvelles espèces d'hôte échantillonnées. Une validation et un *scoring* des interactions sont en cours et seront diffusés au format PSI-MI à la communauté scientifique pour des validations expérimentales ultérieures et cela après publication de la méthode.

Références

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S. L., Ceol, A., Chautard, E., … Hermjakob, H. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature Methods*, 8(7):528–9. doi:10.1038/nmeth.1637

de Chassey, B., Meyniel-Schicklin, L., Aublin-Gex, A., Navratil, V., Chantier, T., André, P., & Lotteau, V. (2013). Structure homology and interaction redundancy for discovering virus-host protein interactions. *EMBO Reports*, 14(10):938–44. doi:10.1038/embor.2013.130

de Chassey, B., Meyniel-Schicklin, L., Vonderscher, J., André, P., & Lotteau, V. (2014). Virus-host interactomics: new insights and opportunities for antiviral drug discovery. *Genome Medicine*, 6(11):115. doi:10.1186/s13073-014-0115-1

Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J., & Oliva, B. (2012). BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic Acids Research*, 40(Web Server issue):W147–51. doi:10.1093/nar/gks553

Guirimand, T., Delmotte, S., & Navratil, V. (2015). VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Research*, 43(Database issue):D583–7. doi:10.1093/nar/gku1121

Lewis, A. C. F., Jones, N. S., Porter, M. A., & Deane, C. M. (2012). What evidence is there for the homology of protein-protein interactions? *PLoS Computational Biology*, 8(9):e1002645. doi:10.1371/journal.pcbi.1002645

Navratil, V., de Chassey, B., Combe, C. R., & Lotteau, V. (2011). When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. *BMC Systems Biology*, 5:13. doi:10.1186/1752-0509-5-13

Navratil, V., de Chassey, B., Meyniel, L., Pradezynski, F., André, P., Rabourdin-Combe, C., & Lotteau, V. (2010). System-level comparison of protein-protein interactions between viruses and the human type I interferon system network. *Journal of Proteome Research*, 9(7):3527–36. doi:10.1021/pr100326j

Mots clefs : Virus/Host, interaction protéine-protéine, inférence, interologie

BRANE Cut : inférence de réseaux de gènes par post-traitement, application à l'étude transcriptomique de données modèles et d'un champignon filamenteux

Poster 36

Aurélie Pirayre [*1], Frédérique Bidard[2], Laurent Duval[1]

[1] Direction Technologie, Informatique et Mathématiques appliquées – IFP Énergies Nouvelles – France
[2] Direction Chimie et Physico-chimie appliquées (DCPA) – IFP Énergies Nouvelles – 1-4 avenue Bois-Préau, F-92 852 Rueil-Malmaison, Cedex, France

Introduction

Idéalement, les réseaux de gènes sont des outils adéquats pour visualiser des interactions entre gènes et permettant d'en déduire des voies de régulations correspondant à un phénotype donné. Cependant, malgré une effervescence de méthodes d'inférence de Réseaux de Régulation de Gènes (RRG), la fiabilité des résultats sur données réelles reste encore un défi majeur [9]. Ceci est en partie dû à la structure des données elles-mêmes, où le nombre de variables (les gènes) à traiter est souvent très supérieur à la quantité d'observations (les conditions expérimentales). Les données en question sont issues de techniques de type puces à ADN ou plus récemment de type séquençage haut-débit tel que le RNAseq. De telles données reflètent, pour chaque gène i, un niveau d'expression (relatif ou absolu) dans différentes conditions expérimentales. Le gène i est donc caractérisé par son profil d'expression.

À partir de ces données, l'inférence de réseaux peut être traitée sous deux points de vues : le premier est basé sur des mesures statistiques entre toutes les paires de profils d'expression [4, 8, 2] tandis que le second se base sur des modèles [5, 12, 3]. La plupart de ces méthodes permettent d'attribuer à chaque paire de gènes (i,j) un poids $\omega(i,j)$ reflétant le niveau d'interaction entre les gènes i et j. L'ensemble de ces poids peut être collecté dans une matrice carrée de taille $G \times G$, où G est le nombre total de gènes. Cette matrice peut alors être vue comme la matrice d'adjacence pondérée du réseau de gènes G sous-jacent. Ce réseau G est donc défini par un ensemble de nœuds V (correspondant aux gènes) et un ensemble d'arêtes E, où l'arête $e(i,j)$ liant les nœuds i et j est pondérée par le poids $\omega(i,j)$. À partir de ce réseau complet (tous les nœuds sont reliés entre eux) et pondéré, un seuillage est classiquement opéré pour ne sélectionner que les liens les plus forts. Ainsi, en définissant un seuil λ, le réseau final G* ne contient que le sous-ensemble d'arêtes E* pour lesquelles $\omega(i,j) > \lambda$. Dans cet article, nous présentons la méthode BRANE Cut [10] qui raffine le seuillage classique et ses performances sur des données modèles et réelles de transcriptome.

Modélisation mathématique

En définissant pour chaque arête $e(i,j)$, un label binaire $x(i,j)$ traduisant la présence ($x(i,j) = 1$) ou l'absence ($x(i,j) = 0$) de l'arête dans le graphe, le seuillage classique peut être obtenu en minimisant $\omega|x - 1| + \lambda x$. En effet, la solution explicite de ce problème d'optimisation donne $x(i,j) = 1$ si $\omega(i,j) > \lambda$ et 0 sinon. À partir d'un réseau de gènes, pondéré et complètement connecté, la méthode BRANE Cut (Biologically Related A priori Network Enhancement with Graph cuts) [10] que nous proposons permet de raffiner le seuillage classique par ajout de contraintes reflétant

*. Intervenant

des *a priori* structuraux ainsi que des *a priori* biologiques sur les mécanismes de régulation des gènes. Ces différents *a priori* sont, entre autres, basés sur la connaissance des gènes codant pour des facteurs de transcription (FTs), l'ensemble de ces gènes étant noté T. À l'inverse, les gènes n'ayant pas été renseignés comme codant pour un facteur de transcription seront notés nonFTs.

Seuillage double

À la différence du seuillage classique, la valeur du seuil λ dans BRANE Cut dépend de la nature du lien entre les nœuds i et j et ce seuil adaptatif est alors noté $\lambda(i,j)$. Afin de n'inférer que des arêtes impliquant au moins un FT, on pourra prendre $\lambda(i,j) = 2^*\max\{\omega(i,j)\}$ si (i,j) n'appartient pas à l'ensemble T^2. Une contrainte additionnelle peut être ajoutée pour inférer préférentiellement les liens FT-nonFT par rapport aux liens FT-FT. Pour cela, le paramètre $\lambda(i,j)$ peut être défini comme la somme de deux seuils λi et λj, chacun agissant au voisinage du nœud i et j respectivement. Ainsi si le nœud i est un FT, $\lambda i = \lambda FT$ et $\lambda i = \lambda nonFT$ sinon. Afin de respecter complètement la contrainte additionnelle, il est nécessaire que $\lambda FT \geq \lambda nonFT$.

Co-régulation

En complément d'un seuillage double, nous intégrons également un *a priori* biologique exploitant la co- régulation des gènes. En effet, si au moins un gène k est identifié comme co-régulé par le couple de FTs (j,j') alors, ce même couple de FTs peut co-réguler d'autres gènes i. D'un point de vue du réseau, cette co-régulation se traduit en favorisant l'assignation d'une même valeur aux labels $x(i,j)$ et $x(i,j')$ si des gènes sont identifiés comme co-régulés par le couple de FTs (j,j'). D'un point de vue mathématique, cela se traduit par l'ajout d'un terme variationnel évaluant la différence entre les labels et pondéré par une probabilité de co-régulation notée $\rho(i,j,j')$. Dans ce travail, un gène k est dit co-régulé par le couple de FTs (j,j') si les poids $\omega(k,j)$, $\omega(k,j')$ et $\omega(j,j')$ sont tous les trois supérieurs au $(G-1)$e quantile de tous les poids. Ainsi, pour un gène i différent de k et un couple de FTs (j,j'), nous pouvons définir une probabilité de co-régulation (notée $\rho(i,j,j')$) proportionnelle au nombre de gènes identifiés comme co-régulés par (j,j') par rapport au nombre de gènes que le couple (j,j') peut co-réguler.

Stratégie d'optimisation

En intégrant le double seuillage ainsi que l'*a priori* de co-régulation, une nouvelle fonction de coût à minimiser peut être définie. Ce nouveau problème d'optimisation permettant de retrouver x^* (la labélisation optimale des arêtes) correspond à un problème de coupe minimale dans un graphe. Grâce à la dualité coupe minimale / flow maximal, la résolution de ce problème revient à maximiser un flux dans un graphe dit de transport et noté Gf. Dans un tel graphe, le flux est une fonction qui assigne à chaque arête une valeur réelle (un poids) sous deux contraintes : *i)* une contrainte de capacité limite : le flux de chaque arête ne doit pas dépasser le poids de l'arête et *ii)* une contrainte de conservation du flux : pour chaque nœud du graphe de transport (sauf pour les nœuds source s et puits t), le flux entrant doit être égal au flux sortant. Sous ces deux contraintes, maximiser le flux de la source au puits équivaut à trouver la coupe minimale. Dans ce travail, nous construisons donc un graphe de transport Gf qui correspond au problème proposé de coupe minimale [6] et obtenons la solution grâce à l'algorithme max-flow [1].

Résultats établis

BRANE Cut a été testé avec des poids issus des méthodes CLR [4] et GENIE3 [5] sur des données de *benchmark* simulées comprenant cinq réseaux du challenge DREAM4 (*Dialogue for Reverse Engineering Assessments and Methods*) [7] ainsi que sur des données réelles de *Escherichia coli* [4]. Pour chacune de ces données, les performances du seuillage par BRANE Cut (utilisé en

post-traitement des poids CLR et GENIE3) ont été comparées à celles obtenues par un seuillage classique, en terme d'AUPR (*Area Under Precision-Recall curve*). Les résultats montrent une amélioration atteignant 6% (pour CLR) et 11% (pour GENIE3) sur données simulées (DREAM4) et une amélioration d'environ 12% (pour CLR) et 3% (pour GENIE3) sur les données réelles d'*Escherichia coli*.

Nouveaux résultats pour un cas d'application réel

Dans ce papier, nous présentons également de nouveaux résultats biologiques obtenus grâce à l'utilisation de BRANE Cut [10] sur des données transcriptomiques issues du champignon filamenteux *Trichoderma reesei*. Ce champignon est connu pour produire et sécréter des enzymes (appelées cellulases) capable de dégrader les molécules de cellulose et hémicellulose. Il est donc un acteur de choix dans le procédé de fabrication des biocarburants de seconde génération. Cependant, à une échelle plus industrielle, la souche sauvage de *T. reesei* présente des rendements de production de cellulases encore trop faibles. Il est donc nécessaire de générer de nouvelles souches hyper-productrices. Jusqu'à présent, des techniques de mutagénèses aléatoires étaient utilisées mais arrivent maintenant à leurs limites et des techniques de mutagénèses dirigées semblent plus appropriées. Ce champignon étant peu connu, il est nécessaire de comprendre plus en détail son fonctionnement, et plus particulièrement comment se déroule la régulation des gènes lors de la production de cellulases. À cette fin, des expériences en *fed-batch* ont été menées afin d'étudier le transcriptome du champignon à 24h et 48h pour différentes sources de carbone : 100 % glucose (non inducteur), 100 %lactose (inducteur) et deux mélanges de glucose/lactose en proportions différentes. À partir de ces données transcriptomiques, 650 gènes ont été sélectionnés du fait de leur expression différentielle en présence de lactose par rapport au glucose. En se basant sur ces 650 gènes, nous avons calculé les similarités entre les profils d'expression de gènes par la méthode CLR et inféré un réseau de gènes avec BRANE Cut. Nous avons obtenu un réseau contenant 161 gènes et 205 arêtes, duquel on peut extraire trois sous-réseaux. Un premier sous-réseau fait apparaître des gènes induits sur lactose avec une dépendance à la concentration de lactose. On retrouve dans ce sous-réseau les cellulases les plus connues, associées à des facteurs de transcription préalablement identifiés pour participer à la production de cellulases. Ce sous-réseau permet de retrouver les connaissances actuelles [11] et ainsi d'apporter une certaine fiabilité à l'ensemble du réseau inféré par BRANE Cut. Le second sous-réseau fait intervenir des gènes induits sur lactose, quelle que soit sa concentration. En d'autres termes, seul le signal lactose modifie le comportement de ces gènes. Dans ce cas également, nous retrouvons des gènes plutôt liés à la production de cellulases. Enfin, le troisième sous-réseau présente en majorité des gènes réprimés en présence de lactose avec une influence de sa concentration. Ces gènes sont essentiellement liés au développement, mais on y retrouve également des gènes induits et liés à la production de cellulases. Ces trois sous-réseaux semblent indiquer un lien entre le développement et la production de cellulases, lien qui, pour l'heure, n'a pas encore été révélé par d'autres études sur *Trichoderma reesei*.

Conclusions

Nous présentons ici une application originale de l'outil BRANE Cut. Cet outil permet, à partir d'un réseau pondéré complètement connecté, de sélectionner les arêtes les plus pertinentes au sens d'un critère basé sur la sélection d'arêtes *i)* de fort poids en privilégiant les arêtes liées à un facteur de transcription ainsi que *ii)* traduisant une co-régulation des gènes. Sa formulation originale sous une fonction de coût à minimiser permet de résoudre ce problème par des algorithmes rapides tels que l'algorithme du max-flow. Cette méthode a permis, sur des données réelles de transcriptome, d'établir des liens nouveaux entre les mécanismes de développement et de production de cellulases chez le champignon filamenteux *Trichoderma reesei*. Suite à l'utilisation de BRANE Cut, une étude prédictive est en cours afin de valider les hypothèses formulées à partir du réseaux inférés.

Références

[1] Y. Boykov *et al.*, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004.

[2] A. J. Butte *et al.*, *Pac. Symp. Biocomput.*, 2000.

[3] C. Charbonnier *et al.*, *Stat. Appl. Genet. Mol. Biol.*, 2010.

[4] J. J. Faith *et al.*, *PLoS Biol.*, 2007.

[5] V. A. Huynh-Thu *et al.*, *PLoS One*, 2010.

[6] V. Kolmogorov *et al.*, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004.

[7] D. Marbach *et al.*, *Proc. Nat. Acad. Sci. U.S.A.*, 2010.

[8] A. A. Margolin *et al.*, *BMC Bioinformatics*, 2006.

[9] F. Picard *et al.*, *Gaz. Math.*, 2011.

[10] A. Pirayre *et al.*, *BMC Bioinformatics*, 2015.

[11] D. Poggi-Parodi *et al.*, *Biotechnol. Biofuels*, 2014.

[12] X. Zhang *et al.*, *Bioinformatics*, 2013.

Mots clefs : Inférence, Réseaux de gènes, optimisation discrètes, coupe minimale, données d'expression de gènes

La reconstruction de réseaux métaboliques, une manière d'étudier globalement le métabolisme secondaire du genre *Penicillium*

Poster 37

Sylvain Prigent [*] [†1], Jens Christian Nielsen[1], Jens Nielsen[1]

[1] Sysbio, Department of Biology and Biological Engineering, Chalmers University of Technology – Suède

Depuis de nombreuses années la découverte de nouveaux antibiotiques décroît alors que les résistances aux antibiotiques disponibles augmentent. Cela amène à une crise majeure de résistance aux antibiotiques qui pose une menace majeure pour la santé publique et les efforts de recherche de nouveaux antibiotiques doivent être renouvelés. Les champignons filamenteux sont bien connus en tant que producteurs naturels de nombreuses molécules possédant une forte activité biologique. L'exemple le plus célèbre de molécule produite par ces espèces est peut-être la pénicilline, produite par *Penicillium rubens,* dont la découverte en 1928 a permis de décroître énormément la mortalité due aux infections bactériennes. Mettre l'accent sur la recherche autour de la production de métabolites secondaires par les espèces du genre *Penicillium* pourrait amener à la découverte de nouvelles molécules biologiquement actives et/ou de nouvelles méthodes de productions à grande échelle de ces molécules.

Récemment la recherche autour du métabolisme secondaire des champignons filamenteux a grandement bénéficié du développement d'outils bioinformatiques permettant de rechercher et d'identifier, dans des génomes, les gènes impliqués dans la synthèse de métabolites secondaires. Ces outils (tels qu'antiSMASH) exploitent le fait que ces gènes se regroupent en clusters dans le génome, les clusters de gènes biosynthétiques. En couplant ces outils à des données de transcriptomique et de métabolomique il est aujourd'hui possible de prédire la capacité de production et les voies de synthèse de métabolites secondaires.

Dans le cadre de ce projet nous avons séquencé dix espèces du genre *Penicillium*, auxquels s'ajoutent 14 génomes disponibles publiquement. Ces séquences couplées à des analyses métabolomiques et transcriptomiques nous permettent d'identifier les clusters de gènes impliqués dans la production de métabolites secondaires d'intérêt. Néanmoins, si savoir qu'une espèce possède la capacité génétique de produire une molécule, très souvent les champignons filamenteux ne produisent ces molécules que dans des environnements spécifiques. Reconstruire les réseaux métaboliques de ces espèces permettrait d'aider à comprendre globalement le métabolisme de ces espèces et en se concentrant sur leur métabolisme secondaire nous devrions être en mesure de mieux comprendre les conditions d'activation de ces voies métaboliques.

Nous avons donc reconstruit 24 réseaux métaboliques en se basant sur les dix génomes nouvellement séquencés et sur les 14 génomes publiquement disponibles. Deux méthodes différentes ont été utilisées pour reconstruire ces réseaux métaboliques. La première reconstruction est basée sur trois réseaux métaboliques existants chez des espèces proches. Celui de *Penicillium rubens* (précédemment appelé *Penicillium chrysogenum*), Aspergillus niger et Aspergillus nidulans. Une recherche d'orthologues entre les enzymes catalysant des réactions métaboliques chez ces trois espèces et le protéome prédit chez nos 24 espèces d'intérêt a permis d'inférer la présence de réactions métaboliques chez ces dernières. D'autre part une reconstruction de-novo a été réalisée en se basant sur MetaCyc, une base de données de réactions métaboliques contenant de nombreuses informations sur le métabolisme secondaire de manière générale. Enfin les voies métaboliques

*. Intervenant

†. Corresponding author : prigent@chalmers.se

dérivées des prédictions de clusters de gènes biosynthétiques seront intégrées manuellement au réseau afin d'avoir une description aussi précise que possible du métabolisme secondaire.

Dans un premier temps, les reconstructions de réseaux ont permis d'affiner l'annotation fonctionnelle du génome des 10 espèces nouvellement séquencées. D'autre part l'étude des réseaux obtenues nous apporte de premiers indices sur l'évolution du métabolisme de ces espèces et plus particulièrement son métabolisme secondaire. Ainsi, il semblerait que le métabolisme primaire de ces espèces ne soit pas parfaitement corrélé à leur histoire phylogénétique. Ces espèces auraient en effet développé des capacités métaboliques spéciales basées sur leur habitat. Il en va de même pour le métabolisme secondaire, pour lequel il ne semble pas toujours y avoir de corrélation entre la capacité de production de certaines molécules et l'histoire évolutive des espèces. Là encore notre hypothèse est que ces espèces auraient développé des capacités de production de molécules spéciales en se basant sur les organismes rencontrés dans leur habitat naturel.

La prochaine étape de se projet sera le recueillement d'informations sur la biomasse des différentes espèces pour permettre le gap-filling puis une modélisation numérique des réseaux obtenus. Les simulations réalisées devraient ainsi permettre, par exemple, de prédire des inactivations de gènes permettant la redirection de flux de molécules vers la production de composés d'intérêt. Il devrait également être possible de prédire différents transferts de gènes entre espèce du genre *Penicillium* afin d'optimiser, encore une fois, la production de molécules biologiquement actives et intéressantes d'un point de vue médical.

Mots clefs : Réseaux métaboliques, métabolisme secondaire

Exploration of the enzymatic diversity of protein families : data integration and genomic context

Poster 38

Guillaume Reboul [*] [†1,2,3], Mark Stam[1,3], Benjamin Viart[1,3], David Vallenet [‡1,3], Claudine Médigue[1,3]

[1] Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme (LABGeM) – Direction de la Recherche Fondamentale, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA), Institut de Génomique, Genoscope – 2 rue Gaston Crémieux, F-91 057 Évry, France
[2] Université de Rouen - Master Bioinformatique – Normandie Université – France
[3] Génomique métabolique (UMR 8030) – CEA, CNRS : UMR8030, Université d'Évry-Val d'Essonne – Genoscope, 2 rue Gaston Crémieux, F-91 057 Évry Cedex, France

Next Generation Sequencing technologies have dramatically increased the number of available sequences in public databanks. In contrast, many enzymatic activities ($\approx 20\,\%$) are orphans of protein sequence which highlight the lack of knowledge in metabolism (Sorokina *et al.*, 2014). The growing amount of available protein sequences is a great opportunity to fill gaps in many metabolic pathways as well as to discover novel enzymatic activities. Multiple ways to reduce the number of orphan enzymes and elucidate pathway holes have already been investigated and proposed by our team. One of these is based on combining genomic and metabolic contexts to predict candidate genes for orphan enzymes (Smith *et al.*, 2012). This method has been recently extended by a new network representation of the metabolism to detect conserved chemical transformation modules (Sorokina *et al.*, 2015). In parallel, other methods were developed using structural features like ASMC (Active Site Modelling and Clustering) which classifies active site pockets of an enzyme family and detects important residues for substrate specificity (de Melo-Minardi *et al.*, 2010). These methods were successfully combined to elucidate the enzymatic diversity of a protein family of unknown function (Bastard *et al.*, 2014).

Here, we present improvements in the detection of protein families of interest and in their analysis using a genomic context clustering method based on conserved syntenies. In order to highlight enzyme families of interest, we have integrated knowledge from several public databases on proteins, families and metabolism (*i.e.* UniProtKB, InterPro, Pfam, PIRSF and MetaCyc). Then, a strategy combining multiple criteria was established with the objective to select protein families of unknown function linked to some experimental evidences illustrating a potential enzymatic activity. Beside, we have developed a method to classify proteins of a family based on their conserved genomic contexts. First, each protein is compared against all others to determine if there corresponding genes share a conserved synteny. Second, a graph is established using these synteny results. Nodes are proteins of the family and are connected by an edge when a conserved synteny is observed. Weights on edges represent the average number of genes in synteny. To investigate the enzymatic diversity of the family, clustering algorithms are then applied on this weighted graph to define protein groups which are supposed iso-functional. Furthermore, functions of conserved neighbor genes within each group may give clues to predict the precise function for yet uncharacterized proteins of the family. As an illustration, we applied this method on the PF07071 family of the Pfam database (v29.0, 12/2015). This family is described as a "Domain of Unknown Function" (DUF1341) but contains few members annotated as 2-dehydro-3-deoxy-phosphogluconate aldolase (DgaF protein).

*. Intervenant
†. Corresponding author : reboul_guillaume@yahoo.fr
‡. Corresponding author : vallenet@genoscope.cns.fr

References

Bastard, K. *et al.* Revealing the hidden functional diversity of an enzyme family. *Nat. Chem. Biol.* 10:42–49 (2014).

de Melo-Minardi, R. C., Bastard, K. & Artiguenave, F. Identification of subfamily-specific sites based on active sites modeling and clustering. *Bioinformatics* 26:3075–3082 (2010).

Smith, A. A. T., Belda, E., Viari, A., Médigue, C. & Vallenet, D. The CanOE strategy: Integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. *PLoS Comput. Biol.* 8, (2012).

Sorokina, M., Stam, M., Médigue, C., Lespinet, O. & Vallenet, D. Profiling the orphan enzymes. *Biol. Direct* 9, 10 (2014).

Sorokina, M., Médigue, C. & Vallenet, D. A new network representation of the metabolism to detect chemical transformation modules. *BMC Bioinformatics* 16:385 (2015).

Mots clefs : metabolism, enzymatic diversity, protein families, genomic contexts, synteny, clustering

Investigating long non-coding RNA's role in nucleating protein complexes

Poster 39

Diogo Ribeiro [*] [†1,2], Andreas Zanzoni[1], Lionel Spinelli[1], Davide Cirillo[2], Gian Tartaglia[2], Christine Brun[1]

[1] UMR S1090 Technological Advances for Genomics and Clinics (TAGC) – Institut National de la Santé et de la Recherche Médicale - INSERM, Aix-Marseille Université - AMU – Parc Scientifique de Luminy, Case 928, F-13 288 Marseille, France
[2] Centro de Regulación Genómica (CRG) – Barcelone Espagne

At least 50 % of the mammalian genome is predicted to be transcribed but not translated. Whereas most of this pervasive transcription is still thought to have little or no impact in the cell, evidence is mounting for at least a portion of this transcription to be functional in a variety of ways.

Recently, a significant part of long non-coding RNAs (lncRNAs) – transcripts broadly defined as having > 200 nt and low or nonexistent coding-potential – has been found to affect several cellular functions, including regulation of transcription, mRNA processing and protein activity. This action can occur through binding of DNA, RNA or protein, often in a regulated fashion. Examples of such action are demonstrated by the binding of the HOTAIR and MEG3 lncRNAs to proteins of the Polycomb repressive complex 2 (PRC2), creating functional protein-RNA complexes responsible for targeted gene expression regulation [1]. An increasing amount of other lncRNAs shows to interact with diverse protein complexes, and to be differentially expressed or regulated in a cell-type or condition-specific manner, providing indication for their functionality. As cellular functions are performed by interactions between macromolecules, we can expect that non-coding transcripts, and specially lncRNAs, may play a central role in piecing together protein components, for example by acting as scaffolding molecules and/or promoting the assembly of protein complexes. Despite this, the potential function or impact of the vast majority of the > 16,000 curated human lncRNAs is yet to be deciphered, as research in this relatively new topic has been done mostly on case by case basis, being hindered by the lack of large-scale analysis methods.

Here, to assess the binding potential of human lncRNAs to RNA-binding proteins (RBPs) on a large-scale, we use the catRAPID "omics" algorithm [2,3], which predicts their binding potential based on physico-chemical features (e.g. secondary structure, hydrogen bonding, van der Waals forces). To explore the potential prevalence and importance of lncRNAs being a part of macromolecular complexes, we integrate RNA molecules and their protein interactions into the human protein-protein interaction network, thereby creating a bipartite interaction network. By investigating the ability of lncRNAs to bind proteins of the same functional network module (i.e. a group of interacting proteins involved in the same biological process), protein complex or pathway, we can pinpoint lncRNAs that may play a role in those units.

Using protein-RNA interaction predictions between the majority of known human lncRNAs and RBPs (> 13 million interactions), should allow us to discover novel protein-RNA complexes and understand the general role of RNA in nucleating protein-protein interaction networks, as well as categorize groups of lncRNAs through their protein interaction profiles. Moreover, addition of extra information, such as tissue expression and conservation data, will allow determining the specificity and regulation of candidate protein-RNA complexes and prioritize interactions to be

*. Intervenant
†. Corresponding author : diogo.ribeiro@inserm.fr

validated in vivo.

References

[1] Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., … Kanduri, C. (2015). MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. *Nature Communications*, 6:7743. http://doi.org/10.1038/ncomms8743.

[2] Bellucci, M., Agostini, F., Masin, M., & Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nature Methods*, 8(6):444–445. http://doi.org/10.1038/nmeth.1611.

[3] Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., & Tartaglia, G. G. (2013). catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, 29(22):2928–2930. http://doi.org/10.1093/bioinformatics/btt495.

Mots clefs : protein, RNA interactions, interaction networks, long non, coding RNA

Functional genetic diversity of the yeast galactose network

Poster 40

Magali Richard [*1], Daniel Jost[2], Hélène Duplus-Bottin[1], Florent Chuffart[1], Étienne Fulcrand[1], Gaël Yvert[1]

[1] Laboratoire de Biologie Moléculaire de la Cellule (LBMC) – CNRS : UMR5239, Institut national de la recherche agronomique (INRA) : UR5239, Université Claude Bernard - Lyon I (UCBL), École Normale Supérieure (ENS) - Lyon – ENS de Lyon, 46 allée d'Italie, F-69 364 Lyon Cedex 07, France
[2] Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG) – CNRS : UMR5525, Université Joseph Fourier - Grenoble I – Domaine de la Merci, F-38 706 La Tronche, France

Optimizing growth and survival in face of an environmental signal is a major challenge for single-cell organisms. The yeast galactose regulatory network is a standard model system for studying both cellular and molecular responses to an extracellular input. This model offers the possibility to investigate the effects of genetic diversity on a regulatory response at a very fine scale. In order to better understand how natural diversity affects single-cell fates in response to an external stimuli, we are studying natural strains and artificial mutants of *Saccharomyces cerevisiae* to i) relate genetic diversities to changes in network parameters and to ii) identify novel genetic variations involved in the network properties. A Green Fluorescent Protein (GFP) reporter is used to monitor by flow cytometry the dynamics of the network upon galactose induction on a population of cells. This allowed us to unravel a large diversity of responses among natural strains, and to identify natural alleles that can radically change network properties. In order to interpret the functional properties of genetic variants in terms of molecular processes, we are currently putting in perspective *in vivo* experiments with *in silico* modelling, based on a mathematical model inspired from the literature. Next, we intend to experimentally validate our predictions on the involved molecular features, such as transcription rate, mRNA stability or protein-protein interaction affinities. In parallel, we are also investigating natural wild isolates that display bimodality in the network response. These different – yet complementary – strategies should allow us to characterize natural genetic routes that change a gradual response to a binary response.

This project is supported by grant SiGHT StG-281359 from the E.U.

Mots clefs : galactose network, yeast, bimodality, modelisation, quantitative genetic

*. Intervenant

Méthode d'apprentissage supervisé à partir d'un réseau de coexpression pour l'annotation fonctionnelle des gènes d'*Arabidopsis thaliana*

Poster 41

Rim Zaag [*1,2], Cécile Guichard [†1,2], Étienne Delannoy [‡1,2], Marie-Laure Martin-Magniette [§1,2,3]

[1] Institute of Plant Sciences Paris Saclay (IPS2) – Institut national de la recherche agronomique (INRA) : UMR1403, CNRS : UMR9213, Université Paris Sud - Paris XI, Université d'Évry-Val d'Essonne, Université Paris Saclay – Bâtiment 630, Rue Noetzlin, F-91 405 Orsay, France

[2] Institute of Plant Sciences Paris Saclay (IPS2) – Université Paris Diderot - Paris 7, PRES Sorbonne Paris Cité – Bâtiment 630, Rue Noetzlin, F-91 405 Orsay, France

[3] Unité de recherche Mathématiques et Informatique Appliquées UMR518 (MIA) – AgroParisTech, Institut national de la recherche agronomique (INRA), Université Paris-Saclay – 16 rue Claude Bernard, F-75 231 Paris Cedex 05, France

Introduction

Le processus d'annotation fonctionnelle des gènes inconnus reste un défi majeur de la biologie. Selon différentes estimations, 20 à 40 % des gènes prédits des organismes eucaryotes dont le génome est complètement séquencé, n'ont aucune fonction attribuée (Wortman *et al.* 2003 ; Gollery *et al.* 2006, 2007 ; Hanson *et al.* 2010). Concernant la plante modèle *Arabidopsis thaliana,* séquencée en 2000, la fonction d'environ 5 000 gènes soit environ 19 % des gènes annotés (Zaag *et al.* 2015) reste totalement inconnue selon la dernière version de l'annotation officielle TAIR10 (`http://www.arabidopsis.org/`) (Garcia-Hernandez *et al.* 2002). Ces gènes sans aucun indice sur leur fonction potentielle sont appelés gènes orphelins (Domazet-Lozo and Tautz, 2003 ; Fukushi and Nishikawa, 2003) et sont systématiquement mis de côté dans les analyses de génétique inverse ou autres approches fonctionnelles dirigées. À côté de ces gènes orphelins il existe aussi un nombre très important de gènes partiellement annotés. Les méthodes d'annotation fonctionnelle automatiques récentes sont fondées sur la recherche de partenaires fonctionnels. La fonction des gènes inconnus est alors déduite à partir de celles de leurs partenaires en s'appuyant sur l'hypothèse d'association par culpabilité. Pour la recherche de partenaires, la majorité de ces méthodes s'appuient sur l'intégration de plusieurs données omiques en particulier des données transcriptomiques pour la construction de réseaux d'interactions moléculaires. Des évaluations de ces méthodes ont montré un taux élevé de faux positifs parmi les prédictions, qui pourrait s'expliquer par le manque de spécificité du contexte biologique et l'hétérogénéité des données qui sont souvent issues de différentes expériences et différents laboratoires. De plus, la plupart de ces méthodes sont centrées sur une prédiction par gène alors que la performance de prédiction varie en fonction de l'ontologie et même des termes analysés (Radivojac *et al.* 2013 ; Ryngajllo *et al.* 2011).

L'objectif de notre travail a été de proposer une nouvelle méthode d'annotation fonctionnelle par terme contrôlant l'hétérogénéité des données utilisées. Cette méthode utilise les résultats d'une analyse de coexpression à large échelle de 18 catégories de stress chez *Arabidopsis thaliana*. Pour chaque terme des trois ontologies de la GOSlim, la méthode prédit les gènes qui y sont associés

*. Intervenant
†. Corresponding author : `guichard@versailles.inra.fr`
‡. Corresponding author : `delannoy@evry.inra.fr`
§. Corresponding author : `marie_laure.martin-magniette@agroparistech.fr`

avec un contrôle du FDR à 20 %. Le travail a consisté à identifier les paramètres importants pour la prédiction et d'évaluer la robustesse des prédictions par validation croisée. Cette méthode a permis d'identifier une règle de prédiction pour 16 termes et de proposer une annotation pour 4 349 gènes orphelins ou partiellement annotés.

Méthodes

Construction du réseau de corégulation

Une analyse de coexpression à large échelle de 18 catégories de stress chez *Arabidopsis* a permis d'identifier 681 groupes de gènes ayant le même profil d'expression transcriptomique. La caractérisation fonctionnelle et relationnelle de ces groupes a montré leur pertinence biologique (Zaag et al. 2015, frei-dit-Frey et al. 2014). Cependant la coexpression peut être insuffisante pour supposer leur implication dans les mêmes fonctions biologiques (D'haeseleer et al. 2000). Nous avons construit un réseau de gènes corégulés à partir de l'intégration horizontale des analyses de coexpression des 18 stress en identifiant les couples de gènes présents dans un même cluster de coexpression pour plusieurs catégories de stress.

Méthode d'annotation

La méthode de prédiction consiste en une méthode binaire d'apprentissage supervisée centrée sur les termes par ontologie. Cette méthode permet pour chaque terme de prédire si un gène doit être annoté ou pas avec ce terme. Le principe de la méthode est de considérer une collection de classifieurs dépendants d'un ensemble de paramètres et qui calculent pour chaque gène un score représentatif de la présence du terme analysé dans son voisinage au sein du réseau de corégulation. Chaque classifieur appliqué à un ensemble de gènes permet d'obtenir une liste décroissante de scores. Une collection de règles de décision est ensuite créée en définissant un score seuil associé à chacune de ces listes : si le score d'un gène est supérieur à ce score seuil, le terme est attribué au gène et si le score est inférieur, le terme ne lui est pas attribué.

Pour la détermination du score seuil associé à chaque classifieur, j'ai décidé de le définir de manière à ce que la liste de gènes ayant un score supérieur à ce score seuil comporte au maximum 20 % de faux-positifs. Cette assurance de qualité est importante car elle garantit aux biologistes de limiter leurs efforts de validation expérimentale qui sont souvent coûteux en temps et en ressources. Cette valeur de FDR autorisée à 20 % maximum peut paraître élevée, mais en l'état de l'annotation fonctionnelle actuelle, il peut être considéré comme étant un taux très stringent. En effet, Wang *et al.* (2013) montrent que le FDR de plusieurs méthodes d'annotation actuelles varie entre 85 % et 46 %, en particulier celui de l'approche du vote majoritaire est de 70 %.

L'objectif est de déterminer le score seuil de chaque règle et de sélectionner la meilleure règle par terme. Pour cela, la méthode a été mise au point en analysant tous les gènes ayant une annotation pour l'ontologie considérée constituant un jeu de travail. Afin de mesurer la performance de la méthode, le jeu de travail est découpé en un jeu d'apprentissage pour définir les listes des scores et les scores seuils associés et un jeu test pour mesurer la performance de ces règles. De plus, j'ai utilisé une procédure de cross-validation (CV) afin de considérer plusieurs jeux d'apprentissage et de test à partir d'un même jeu de travail. La meilleure règle par terme est alors déterminée par l'évaluation du FDR et du F-measure sur les jeux tests. Cette règle est ensuite appliquée aux gènes orphelins ou partiellement annotés.

Résultats

L'intégration horizontale des 18 catégories de stress a permis de mettre en évidence l'existence de nombreux couples de gènes coexprimés dans plusieurs catégories de stress allant jusqu'à 14 catégories. Ce réseau est constitué de 12 373 gènes dont 1778 orphelins et 4322 partiellement annotés

et la valeur de la coexpression varie entre 1 et 14. Afin d'exploiter ce réseau pour la prédiction des fonctions des gènes orphelins ou partiellement annotés dans cet espace, plusieurs paramètres ont été considérés. Ils correspondent à la valeur de coexpression, à la prise en compte de cette valeur ou seulement de l'existence de l'arête, au score estimant la présence du terme dans le voisinage du gène et à la manière de classer les gènes selon leur score. Ces paramètres peuvent prendre différentes valeurs et leurs différentes combinaisons ont permis de définir 198 classifieurs pour chacun des 43 termes GO Slim analysés.

Afin d'évaluer l'impact de ces différents paramètres et des termes GO Slim sur la qualité de la classification mesurée par l'AUC, j'ai réalisé une analyse de sensibilité. Ceci a permis de montrer que la valeur de la coexpression sur les 18 catégories n'apporte pas plus d'information que l'existence de cette coexpression. L'ensemble des autres paramètres ainsi que les termes sont quant à eux pertinents. Cette analyse a révélé également l'existence d'interactions entre ces paramètres et les termes appuyant ainsi l'importance d'identifier une règle spécifique pour chaque terme.

Pour l'assignation de chaque gène à un terme, j'ai déterminé un score seuil associé à chaque classifieur qui devait contrôler le FDR à 20 % au maximum sur chaque jeu d'apprentissage. Sur les 100 jeux d'apprentissages considérés, il y a plusieurs classifieurs pour lesquels, le score seuil n'a pas pu être défini car le FDR n'était jamais inférieur à 20 %. J'ai également constaté la difficulté de plusieurs autres règles à contrôler le FDR sur les jeux tests au seuil demandé. Par ces deux critères de définition des règles de prédiction, 16 termes ont été retenus sur les 43 termes initiaux qui définissent les 3 ontologies et il existait plusieurs règles par terme pour la majorité d'entre eux.

Afin de les départager j'ai étudié le Fmeas, une métrique communément utilisée pour l'évaluation des modèles prédictifs et qui correspond à une moyenne harmonique entre la précision et la sensibilité. En comparant la performance des règles sélectionnées en fonction de cette métrique, nous avons constaté que les valeurs de Fmeas variaient entre 0,97 et 0,002. Pour les termes ayant une seule règle, celle-ci a été naturellement sélectionnée pour être appliquée aux gènes orphelins ou partiellement annotés. Les règles correspondantes ayant un FDR moyen de 0,137 ont cependant une faible valeur de Fmeas dont la moyenne est de 0,017. Étant donné que lorsque le FDR est contrôlé cela garantit une bonne précision et sachant que le Fmeas est un compromis entre la sensibilité et la précision, ce résultat indique que pour ces règles contrôlant la précision, la sensibilité est faible. Pour les autres termes ayant plusieurs règles, sélectionner la règle avec le Fmeas maximal limite le nombre de gènes à annoter car ils sont obtenus à partir de l'analyse des paires de gènes très souvent coexprimés. J'ai donc opté pour un compromis entre la performance des règles retenues mesurée avec le Fmeas et le nombre de gènes à prédire avec ces règles. Ce compromis est trouvé en imposant une valeur de Fmeas minimale de 0,2 et en maximisant le nombre de gènes annotés prédits positifs. Ce choix a été effectué pour 12 termes qui sont prédictibles avec une moyenne Fmeas de 0,421 et un FDR de 0,199 soit une précision de 0,80 ce qui correspond à une meilleure performance que celle des méthodes comparées dans Wang et al. (2013).

Au final 16 termes correspondant à 11 termes GO Slim spécifiques et 5 termes non spécifiques ayant le mot « other » sont prédictibles. La règle de prédiction sélectionnée pour chacun de ces 16 termes a été appliquée pour l'annotation des 1 778 gènes orphelins et 5 846 gènes partiellement annotés. La prédiction a été réalisée à l'aide des scores seuils déterminés par la procédure de validation croisée pour chaque terme. Cela a permis de proposer un indice de confiance de la prédiction défini par le nombre de fois où le gène est prédit positif parmi 100 prédictions. Ces indices permettent ainsi d'augmenter la confiance accordée à la prédiction de chaque gène. L'approche a déclaré 5 803 prédictions positives avec un indice de confiance supérieur à 80, et a permis de caractériser 4 349 gènes. Si nous nous intéressons uniquement aux 11 termes spécifiques, 61 prédictions positives ont été effectuées permettant de caractériser 49 gènes dont 23 orphelins avec un indice de confiance supérieur à 80.

Cette analyse a également permis de souligner la sensibilité de la valeur du FDR obtenu pour chaque terme en fonction du nombre de gènes annotés par ce terme dans le réseau et de sa représentativité. De plus, l'utilisation de la technique de validation croisée avec tirage aléatoire des

blocs accentue le déséquilibre entre exemples positifs et exemples négatifs pour l'apprentissage surtout pour les termes les moins représentés dans le réseau. Cela explique probablement en grande partie la difficulté de la méthode à caractériser ces termes et indique ainsi la nécessité d'adapter la valeur maximale autorisée pour le FDR à leur représentativité et donc à chaque terme.

D'un autre côté, la difficulté des règles à contrôler le FDR pour plusieurs termes reflète probablement aussi l'insuffisance des données transcriptomiques pour la caractérisation de ces termes notamment ceux de l'ontologie MF. En effet, il a été démontré dans la littérature que les méthodes qui reposent sur le principe de culpabilité par association permettent d'améliorer la performance de prédiction des termes de l'ontologie CC et BP mais ils sont moins performants pour la prédiction des termes de l'ontologie MF qui sont mieux prédits par les méthodes fondées sur la similarité de séquence. Cela est dû au fait que les partenaires fonctionnels interagissent ensemble dans l'objectif d'effectuer une même fonction biologique sans pour autant avoir nécessairement la même fonction moléculaire. Cependant certains termes de l'ontologie BP et CC ne peuvent pas être prédits à partir du réseau de corégulation uniquement. L'intégration de ce réseau avec d'autres données omiques permettrait probablement d'améliorer la performance de prédiction de ces termes.

Mots clefs : annotation fonctionnelle, apprentissage supervisé, intégration, transcriptome, réseaux de gènes

Exploring the functional landscape of RNA-binding proteins through predicted protein-RNA interactions

Poster 42

Andreas Zanzoni [*] [†1], Davide Cirillo[2], Diogo Ribeiro[1,2], Lionel Spinelli[1], Elisa Micarelli[1], Gian Gaetano Tartaglia[2], Christine Brun [‡1]

[1] UMR_S1090 Technological Advances for Genomics and Clinics (TAGC) – Institut National de la Santé et de la Recherche Médicale (INSERM), Aix-Marseille Université (AMU) – Parc Scientifique de Luminy, Case 928, F-13 288 Marseille, France
[2] Centro de Regulación Genómica (CRG) – Barcelone, Espagne

RNA-binding proteins (RBPs) play a fundamental role in all the aspects of RNA fate. In recent years, several large-scale experiments identified hundreds of mRNA-binding proteins (mRBPs) from yeast and human cells, including not only well-known RBPs but also many novel and unexpected candidates. For instance, among the 1,200 mRBPs identified in human cells, more than 300 proteins (*e.g.*, metabolic enzymes and structural proteins) have no RNA-related function and lack known RNA-binding domains. Nevertheless, the functional role of many mRBPs in the cell is still unknown and, for the vast majority of them, their physiological mRNA targets have not been identified yet.

In order to characterize the functional landscape of the experimentally identified human mRBPs, we investigated whether they bind to the mRNAs encoding proteins involved in the same pathways, thereby suggesting a post-transcriptional regulation. For this, we have implemented a computational strategy that combines protein-RNA interaction prediction and statistical enrichment analysis.

First, we have predicted the interactions between 1,156 mRBP protein sequences and more than 35,000 mRNAs using the catRAPID *omics* algorithm (Agostini F., Zanzoni A. *et al.*, Bioinformatics, 2013), leading to ≈37 millions mRNA-protein interaction predictions. Second, for each mRBP, KEGG and Reactome pathways have been assessed for enrichment/depletion of mRBP's targets using both Fisher's Exact test and GSEA.

Our preliminary analysis outlines an interesting global pattern: some pathways are likely to be frequently targeted by mRBPs, whereas others seem to avoid such type of interactions. In particular, pathways related to metabolism seem to be extensively regulated at the post-transcriptional level. We are currently exploiting these results to classify mRBPs based on the pathways they can regulate.

Overall, this analysis should bring new insights on the pervasiveness of the post-transcriptional regulation mediated by mRBPs.

Mots clefs : RNA, binding proteins, protein, RNA interactions, post-transcriptional regulation

*. Intervenant
†. Corresponding author : zanzoni@tagc.univ-mrs.fr
‡. Corresponding author : christine-g.brun@inserm.fr

NaS : une méthode hybride de correction des erreurs du séquençage Nanopore

François-Xavier Babin *1,2

Poster 43

[1] Genoscope - Centre national de séquençage – CEA, Genoscope – 2 rue Gaston Crémieux CP5706, F-91 057 Évry Cedex, France
[2] Université de Rouen – Normandie Université – Rue Thomas Becket, F-76 130 Mont-Saint-Aignan, France

Les technologies de séquençage longues lectures ont fait leur apparition il y a quelques années. Ces dernières, contrairement aux technologies de séquençage courtes lectures, permettent de régler des problèmes liés à l'assemblage des régions répétées et au phasage des haplotypes.

L'année dernière, Oxford Nanopore a annoncé un nouveau séquenceur de longues lectures appelé « MinION ». Les deux brins d'ADN préalablement coupés en fragments de plusieurs Kb sont reliés entre eux par une molécule en épingle à cheveux à leur extrémité, puis guidés jusqu'à un nanopore. Le nanopore est parcouru en permanence par un courant électrique qui lors du passage d'une molécule va être perturbé. Cette perturbation est enregistrée sous forme d'évènements qui résument chacun le passage d'un 6-mer. Les brins d'ADN sont lus l'un après l'autre ce qui permet de constituer une séquence consensus (également appelée lecture « 2D »), dans le cas de la lecture d'un seul des deux brins une lecture dites « 1D » est générée. Les séquences produites en sortie du nanopore possèdent un taux d'erreur entre 20 % et 30 % ce qui complique l'assemblage avec les outils actuellement disponibles.

Pour pallier à ce défaut de qualité des séquences, nous proposons une approche hybride utilisant des lectures illumina pour corriger les lectures Nanopore. Cette stratégie, appliquée sur le génome d'*Acinetobacter baylyi ADP1*, a généré des lectures synthétiques NaS (Nanopore Synthetic-long) dépassant les 100 Kb s'alignant parfaitement et sans erreur sur le génome de référence.

Mots clefs : Oxford Nanopore, longue lecture, correction, assemblage

*. Intervenant

Sentinel, a Tilling NGS analysis tool. Detection and identification of EMS mutations in a Tilling crop population

Poster 44

Guillaume Beaumont [*1], Brahim Mania [†1], Joseph Tran [*‡2], Fabien Marcel [§1], Marion Dalmais [¶1], Abdelhafid Bendahmane [1]

[1] Unité de recherche en génomique végétale (URGV) – CNRS : UMR8114, Institut national de la recherche agronomique (INRA) : UR1165, Université d'Évry-Val d'Essonne – 2 rue Gaston Crémieux, BP 5708, F-91 057 Évry Cedex, France

[2] Institut Jean-Pierre Bourgin (IJPB) – Institut national de la recherche agronomique (INRA) : UMR1318, AgroParisTech – INRA Centre de Versailles-Grignon, Route de St-Cyr (RD10), F-78 026 Versailles Cedex, France

Sentinel, as a computational analysis tool, was developped since January 2012 to tackle new sequencing technology challenges (NGS) along with the improvements of the Tilling platform of INRA URGV at Évry (known as Institute of Plant Science of Paris-Saclay, IPS2, Gif-sur-Yvette since January 1st, 2015).

Sentinel allows rapid identification of rare mutations in a wide variety of individuals, from different origins and species, in a Tilling crop population sequenced by NGS. Therefore, this tool is essential in translational research to help researchers transfer biological knowledge obtained on model plants to agronomic important crops.

Without an efficient homolog recombination non-GM tools in crops, targeted mutagenesis is difficult to achieve. Though, the Tilling (Targeted Induced Local Lesion IN Genome) reverse genetic approach was developped to obtain an EMS-mutagenized population with the phenotype of interest. The main challenge was to find an effective method to identify rare mutations (< 1 %) in this large set of individuals. The Sentinel pipeline achieves this last goal focusing on known candidate genes and can effectively handle massive sequencing data produced by NGS. But this task and the post-processing analysis are time consuming depending on the Tilling population size and the number of candidate genes studied.

Therefore, a "user friendly" graphical interface was also developped to help Tilling users manage their NGS data, perform the analysis until the identification of their mutations. It has been validated upon 38 Tilling populations collections, representing 11 species of agronomic interest, and over more than 600 amplicons from gene of interest. This tool offers an easy and rapid method for the analysis of Tilling by NGS data, with a useful interface providing graphical plots to explore the results along with a relational database to manage the experiments and the results.

Mots clefs : Variant detection, Tilling population, NGS, EMS, mutagenised crop

*. Intervenant
†. Corresponding author : brahim.mania@u-psud.fr
‡. Corresponding author : joseph.tran@u-psud.fr
§. Corresponding author : fabien.marcel@u-psud.fr
¶. Corresponding author : marion.dalmais@u-psud.fr
Corresponding author : bendahm@evry.inra.fr

Development of a bioinformatics pipeline for differential CLIP-seq analysis

Poster 45

Mandy Cadix [*1,2], Galina Boldina[2], Jérôme Saulière[3], Pierre Gestraud[1], Rym Sfaxi[2], Aurélie Teissandier[4,5], Leila Bastianelli[3], Hervé Le Hir[3], Stéphan Vagner[2], Nicolas Servant[1], Martin Dutertre[2]

[1] Bioinformatics and Computational Systems Biology of Cancer (U900) – Institut Curie, PSL Research University, MINES ParisTech - École nationale supérieure des Mines de Paris, Inserm : U900 – F-75 005 Paris, France

[2] Institut Curie (UMR 3348) – Institut National de la Santé et de la Recherche Médicale - INSERM, CNRS : UMR3348, PSL Research University – Orsay, France

[3] Institut de Biologie de l'ENS (Paris - Ulm) (IBENS) – École Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Inserm – 46 rue d'Ulm, 75 005 Paris, France

[4] U900 – Inserm : U900, Institut Curie, MINES ParisTech - École nationale supérieure des Mines de Paris – France

[5] U934 – Inserm : U934, CNRS : UMR3215, Institut Curie – France

CLIP-seq (Cross-Linking, Immunoprecipitation and Sequencing) is a method allowing to map the binding sites of a specific RNA-binding protein throughout the transcriptome at (or near) nucleotide resolution. However, very few studies have used differential CLIP-seq analysis, which aims at comparing binding maps of a given protein in two different conditions (e.g., with and without a stimulus) and at identifying regulated binding sites. For this, CLIP-seq data need to be normalized to RNA-seq data in each condition. The aim of this study was to develop a bioinformatics pipeline for differential CLIP-seq analysis. Our pipeline includes five steps: 1) CLIP-seq data pre-processing; 2) CLIP-seq reads mapping to genome; 3) identification of binding sites (peaks) in CLIP-seq data; 4) quantification of gene expression in RNA-seq data; and 5) differential analysis of each binding site (normalized to matched gene expression level) between two conditions. We applied this approach to a set of experimental data obtained in our lab to determine the impact of DNA damage on RNA binding of two related RNA binding proteins, hnRNP H and F, in human cells. Our method allowed us to map RNA binding sites of hnRNP H/F on a genome-wide scale and to identify regulated binding sites of hnRNP H/F upon DNA damage.

Mots clefs : CLIP seq, RNA binding proteins, post transcriptional regulations, RNA processing

*. Intervenant

Genome-wide analysis of transpositional activity in the cultivated rice *Oryza sativa*

Marie-Christine Carpentier [*] [†1], Fu-Jin Wei[2], Brigitte Courtois[3], Yue-Ie Hsing[2], Olivier Panaud[1]

Poster 46

[1] Laboratoire Génome et Développement des plantes (LGDP) – CNRS : UMR5096, Université de Perpignan – Bât. T, 58 avenue Paul Alduy, 66 860 Perpignan Cedex, France
[2] Institut of Plant and Microbial Biology, Academia Sinica – No. 128,Section 2, Yien-Chu-Yuan Road Taipei, 11529, Taïwan
[3] CIRAD – TA A-108/3, Avenue Agropolis, 34 398 Montpellier Cedex 05, France

Rice, *Oryza sativa*, is the staple foold for half the world population. It is the first crop the genome of which was sequenced ten years ago (IRGSP 2005). This model species benefits from various genomic ressources among which genome sequences of 3,000 rice varieties that were publicly released in 2014. This provides a unique opportunity to unravel the genetic diversity of the crop with accessions from 89 countries, distributed into 5 varietal groups – *indica, japonica, aus/boro, basmati/sandri* and *intermediate*. Rice (*O. sativa*) was domesticated ten thousand years ago in Asia. There were two independent domestications from the same wild ancestor *O. rufipogon*, from the North and the South side of the Himalayas. This double origin explains the structure of the genetic diversity of the cultivated gene pools into two major groups which has been evidenced by various molecular markers, including the most recent set of 185,000 SNPs. Transposable elements (TEs) can provide a more precise picture of rice genome dynamics on a shorter evolutionary scale (posterior to the domestication) because of transposition rate is higher than base substitutions. We developped a pipeline to detect all retrotransposons insertions in the 3,000 genomes dataset. We present the results of a preliminary analysis of 19 LTR-retrotransposon families. A total of 23,939 polymorphic insertions were detected throughout the 12 rice chromosomes. Interstingly, a majority of these insertions are very recent, as evidenced by their presence in only few accessions. Given that 35 % of the rice genome is composed of TEs, our results suggest that as much as one third of the genome of the crop is labile and of very recent origin. It also provides a strong evidence that transposition in rice occurs *in planta* and in the fields, thus providing a source of genomic diversity that remains to be exploited.

Mots clefs : transposable elements, rice, genomic, NGS

*. Intervenant
†. Corresponding author : marie-christine.carpentier@univ-perp.fr

Developing tools to classify polymorphism in the oyster genome *Crassostrea gigas*

Cristian Chaparro [*1]

Poster 47

[1] Laboratoire Interactions Hôtes-Pathogènes-Environnements UMR5244 (IHPE) – Centre national de la recherche scientifique - CNRS (France), Université de Perpignan – 58 avenue Paul Alduy, F-66 860 Perpignan, France

The pacific oyster (*Crassostrea gigas*) is a very important economic activity. The biggest producer is China followed by Japan and France. The genome has been sequenced (Zhang et al., 2012) and is composed of approximately 12,000 scaffolds and an automatic annotation layer representing 28,027 genes. Due to a high degree of polymorphism in *C. gigas* of about 1.3 %, an inbred line of 4 generations of brother-sister mating was produced for the sequencing. Unfortunately most research groups do not work with the same genome as the one that was sequenced, as animals are often taken from the natural area or from the population used to reproduce oysters in reproduction centers. This material is important as it is closer to the reality of culture conditions and thus will produce answers that are more applicable to the industry.

When re-annotating RNA-Seq assemblies we observe that in average 10 % of genes have significant differences with those described in the reference genome due to this inherent polymorphism. Furthermore, transposable elements are abundant and actively transcribe which might lead to effective transposition and reshaping of the genomic landscape. We decided to investigate the possibility of developing methods that would help us on the re-annotation process by discriminating the source of the genomic variation with respect to the reference genome.

We will present preliminary results on the analysis of the re-annotation of RNA-Seq data like that from GIGATON (Rivière et al., 2015) against the reference genome while discriminating the origin of the differences, whether they are due to the natural divergence of a polymorphic population, speciation due to selection or due to the action of transposable elements.

References

Rivière, G., Klopp, C., Ibouniyamine, N., Huvet, A., Boudry, P., Favrel, P., 2015. GigaTON: an extensive publicly searchable database providing a new reference transcriptome in the pacific oyster *Crassostrea gigas*. *BMC Bioinformatics* 16:401. doi:10.1186/s12859-015-0833-4

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Xie, Y., Holland, P.W.H., Paps, J., Zhu, Y., Wu, F., Chen, Y., Wang, J., Peng, C., Meng, J., Yang, L., Liu, J., Wen, B., Zhang, N., Huang, Z., Zhu, Q., Feng, Y., Mount, A., Hedgecock, D., Xu, Z., Liu, Y., Domazet-Lošo, T., Du, Y., Sun, X., Zhang, S., Liu, B., Cheng, P., Jiang, X., Li, J., Fan, D., Wang, W., Fu, W., Wang, T., Wang, B., Zhang, J., Peng, Z., Li, Y., Li, N., Wang, J., Chen, M., He, Y., Tan, F., Song, X., Zheng, Q., Huang, R., Yang, H., Du, X., Chen, L., Yang, M., Gaffney, P.M., Wang, S., Luo, L., She, Z., Ming, Y., Huang, W., Zhang, S., Huang, B., Zhang, Y., Qu, T., Ni, P., Miao, G., Wang, J., Wang, Q., Steinberg, C.E.W., Wang, H., Li, N., Qian, L., Zhang, G., Li, Y., Yang, H., Liu, X., Wang, J., Yin, Y., Wang, J., 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490:49–54. doi:10.1038/nature11413

Mots clefs : comparative genomics, RNAseq, oyster

*. Intervenant

BAdabouM, un outils rapide de détection des variants structuraux génomiques

Tristan Cumer *[1], Frédéric Boyer[1]

Poster 48

[1] Laboratoire d'écologie alpine (LECA) – CNRS : UMR5553, Université Joseph Fourier - Grenoble I, Université de Savoie – Bâtiment D - Biologie, 2233 rue de la piscine, BP 53, F-38 041 Grenoble Cedex 9, France

Contexte

Le développement des technologies de séquençages NGS a entraîné un essor des quantités de données génomiques à l'échelle populationnelle (ex. : projet 1000 génomes humains). L'analyse de ces données réalignées sur un génome de référence, outre la détection de polymorphismes de type SNP, permet l'étude des variations dans la structure génomique des individus [1]. Ces variants structuraux (insertions, délétions, variations du nombre de copies, inversions, réarrangements chromosomiques) sont connus pour avoir un fort impact sur les individus et restent encore aujourd'hui peu étudiés dû notamment au challenge que représente leur détection.

État de l'art. Avec la disponibilité de ces données, de nombreux outils ont été développés pour relever ce challenge et visent à détecter et étudier ces variations. Cependant, les études comparatives montrent la très faible concordance entre ces différentes méthodes et la nécessité de multiplier les logiciels pour couvrir l'ensemble des types d'évènements et réduire les faux positifs ce qui peut s'avérer lourd en temps de calcul [2]. De plus l'installation, le paramétrage et la fusion des fichiers de sortie peut se révéler fastidieux.

BAbabouM

Le logiciel BAdabouM tente de palier aux principaux problèmes listés ci-dessus. Facile d'installation, rapide d'usage, il détecte tous les types de variants structuraux. BAdabouM parcourt le fichier d'alignement (`.bam`) en recherchant des configurations d'alignement anormales et caractéristiques de chaque type de variant. Le principe se base sur la détection d'anomalies dans des fenêtres glissantes adjacentes le long des chromosomes. Cela permet de repérer et de caractériser les variants, de préciser leur nature ainsi que leur position à la base près, le tout avec une faible emprunte mémoire et un temps d'exécution réduit.

Paramétrage et utilisation

Le programme peut s'utiliser de deux façons différentes : soit en utilisant un jeu de paramètres peu stringent pour filtrer rapidement des zones suspectes qui seront ensuite soumises à l'examen d'autres logiciels, soit de façon autonome. Le logiciel est entièrement paramétrable mais peut aussi s'utiliser sans donner de paramètres, dans ce cas, les paramètres de base (taille de librairie et profondeur de séquençage, seuils de détection) sont inférés. L'ensemble des paramètres peut être fixé par l'utilisateur afin de pouvoir garantir la détection la mieux adaptée avec le type de données de re-séquençage.

*. Intervenant

Tests et Comparaison

BAdabouM a été testé et comparé à d'autres logiciels sur des données de séquençage haut débit de génomes complets de moutons (voir Tableau 1). Le nombre de variants détectés par BAdabouM est faible par rapport aux autres logiciels et ces différences de détections sont en cours d'investigation.

Implémentation et disponibilité

Le logiciel est implémenté en langage C. BAdabouM est encore en cours de test, il sera prochainement distribué sous licence CeCILL.

Références

[1] Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., ... & Konkel, M. K. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75-81.

[2] Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., ... & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2):256-278.

Mots clefs : Variants Structuraux, BAdabouM, Resequencing

A complete genome of the domesticated apple (Golden Delicious)

Nicolas Daccord [*] [†1], Jean-Marc Celton[1], Elio Schijlen[2], Nathalie Choisne[3], Hadi Quesneville[3], Étienne Bucher[1]

Poster 49

[1] Institut de Recherche en Horticulture et Semences (IRHS) – Institut national de la recherche agronomique (INRA) : UMR1345 – Agrocampus Ouest, UMR1345 IRHS, F-49 045 Angers, France
[2] Wageningen UR – Pays-Bas
[3] Unité de Recherche Génomique Info (URGI) – Institut national de la recherche agronomique (INRA) : UR1164 – INRA, Centre de recherche de Versailles, Bâtiment 18, Route de Saint Cyr, F-78 000 Versailles, France

Apple is one of the most commonly cultivated fruit for which numerous genetic marker-assisted breeding programs exist. Thus, the demand for a complete, annotated apple genome is great.

A first draft of the genome was published before [1]. However, the quality of this genome is quite poor due to several reasons: The variety that was selected for the sequencing project was a heterozygous (Golden Delicious) and the apple genome is rich in repetitive elements that could not be resolved by the NGS technologies available at that time.

The domesticated apple genome is highly heterozygous and has a size of approximately 700 Mb. In this work, to simplify the assembly step, we used a doubled haploid (thus completely homozygous) derivative of Golden Delicious to create a high-contiguity annotated apple genome.

First, we performed an assembly of the genome using the PacBio sequencing technology (35X coverage) which produces long reads (mean length > 8 kb). The main advantage of such long reads is that they can cover very large repetitive sequences which will be correctly resolved by the assembler and retrieved in the final contigs. The high error-rate of PacBio reads (10-15 %) was resolved by a self-correction of the reads (using the PBcR pipeline [2] which uses the MHAP aligner [3]) prior to the assembly with Celera assembler [4]. A gap-closing step was performed after the assembly using PBJelly [5] to extend or merge contigs.

After the assembly and the gap-closing step, the contig were corrected one more time, using high-confidence illumina paired-end reads from the very same apple line. This second correction was done in order to correct the remaining errors coming from the high-error rate of PacBio reads which weren't corrected by the self-correction step before the assembly. This step was important to correct systematic errors (deletions) resulting from the PacBio sequencing that can be observed at homopolymers.

The assembly gave us 7983 contigs and a N50* of 216 kb, which is more than 13 times greater than the first version of the genome. The total assembly length was 700 Mb which is 100 Mb more than the first version. This difference comes probably from missing long repetitive regions that couldn't be detected before due to the short reads that were used.

To provide a complete package of the genome, a new annotation of the new genome is being performed. First, the genes were annotated using RNA-seq data and EuGene [6] on the bases of a 600X coverage (on genes) of RNA-seq data originating from 10 libraries. The RNA-seq reads were de-novo assembled using Trinity [7]. The EuGEne annotation tool will then be used to map these RNA-seq contigs on the genome to predict genomic features. EuGene will also include

*. Intervenant

†. Corresponding author : nicolas.daccord@angers.inra.fr

information resulting from Blastx searches against all known proteins on the apple genome to strengthen the predictions. Transposable elements will be annotated in a collaborative effort with the group of Hadi Quesneville (URGI, Versailles).

In order to also document the epigenetic level, methylome (methylated cytosines and histones marks) and the small RNAs were mapped to the annotated genome.

Therefore our work will provide a complete atlas of the apple genome and epigenome, constituted of five layers : long contiguous sequences, genes annotation, DNA and histone modifications and small RNAs.

*The N50 value is the length of the sequence for which all the shorter sequences contain 50 % of all bases.

References

[1] Velasco, Riccardo, et al. "The genome of the domesticated apple (*Malus* × *domestica* Borkh.)." *Nature genetics* 42.10 (2010):833-839.

[2] PBcR pipeline : `http://wgs-assembler.sourceforge.net/wiki/index.php/PBcR`.

[3] Berlin, Konstantin, et al. "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing." *Nature biotechnology* 33.6 (2015):623-630.

[4] Myers, Eugene W., et al. "A whole-genome assembly of Drosophila." *Science* 287.5461 (2000):2196-2204.

[5] English, Adam C., et al. "Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology." *PloS one* 7.11 (2012):e47768.

[6] Foissac, Sylvain, et al. "Genome annotation in plants and fungi: EuGène as a model platform." *Current Bioinformatics* 3.2 (2008):87-97.

[7] Grabherr, Manfred G., et al. "Full-length transcriptome assembly from RNA-Seq data without a reference genome." *Nature biotechnology* 29.7 (2011):644-652.

Mots clefs : genomics, genome assembly, genome annotation, methylome, apple, long reads

Development of SNPs markers from pooled Rad-seq data and high-throughput genotyping, applied to the study of several harvested population genetics in the Upper-Maroni (French guiana)

Poster 50

Chrystelle Delord [*1,2], Gilles Lassalle[*1], Sophie Launey[1], Pierre-Yves Le Bail[1]

[1] INRA Rennes (INRA) – Institut national de la recherche agronomique (INRA) – Domaine de la Motte au Vicomte, F-35 653 Le Rheu, France
[2] Hydréco Guyane – Laboratoire Environnement de Petit Saut, BP 823, F-97 388 Kourou Cedex, France

Dans une perspective de conservation, l'appréhension du bon état des ressources génétiques d'une espèce se révèle souvent incontournable. Dans cette optique, un des défis consiste à pouvoir génotyper à moindre coût un grand nombre de marqueurs moléculaires et sur un grand nombre d'individus afin de renforcer la pertinence des interprétations biologiques qui s'ensuivent. L'essor des NGS a permis le développement récent de méthodes de génotypage haut-débit, comme entre autres le « génotypage par milliers » (*Genotyping-in-thousands by sequencing*, Campbell et al. 2015) ou par génotypage sur puce (ex. technologie SmartChip WaferGen, Wilde & al. 2013). Ceci implique de disposer au préalable d'un catalogue de marqueurs moléculaires pertinents et adéquats, challenge particulièrement complexe lorsqu'il s'agit d'espèces non-modèles pour lesquelles peu de connaissances génétiques sont disponibles. Nous nous proposons d'utiliser une stratégie innovante pour développer rapidement et à moindre coût des marqueurs de type *Single-Nucleotide-Polymorphism* (SNP) sur des espèces non-modèles, sur la base du séquençage type RAD-Seq (*Restriction-site Associated DNA*) d'un pool d'individus. Pour cela, il est nécessaire de se poser les questions adaptées, portant à la fois sur la nature même des marqueurs que l'on souhaite utiliser, puis sur la procédure (*pipeline*) informatique à mobiliser pour extraire ces marqueurs des données issues du RAD-Seq. Nous sommes en train de mettre au point cette méthode sur des données de séquençage RAD de truite fario (*S. trutta*). L'objectif est de caractériser ensuite ces marqueurs via la méthode de « génotypage par milliers ». Cette procédure nous permettra d'obtenir les génotypes de très nombreux individus sur l'ensemble des marqueurs sélectionnés pour leur pertinence. Nous appliquerons cette démarche dans son ensemble pour étudier l'impact de la pression de pêche et du paysage sur la diversité et la structuration génétique des populations chez plusieurs espèces de poissons, exploitées et non-modèles, du Haut-Maroni (Guyane française, en collaboration avec le Parc Amazonien). Notre objectif final est de déterminer les parts respectives des deux facteurs susmentionnés : pêche et paysage, conjugués au potentiel dispersif des espèces, sur leur ressources génétiques afin de soutenir une gestion durable des pêcheries artisanales associées.

Mots clefs : Genotyping, by, sequencing, GT, Seq, RAD, Seq, population genomics, PCR multiplexe, STACKS

*. Intervenant

Sequana : a set of flexible genomic pipelines for processing and reporting NGS analysis

Dimitri Desvillechabrol [*1], Christiane Bouchier[1], Thomas Cokelaer [*†2]

Poster 51

[1] Institut Pasteur – Département de Génomes et Génétique, CITECH, Pôle Biomics – Institut Pasteur de Paris – 26-28 rue du Docteur Roux, F-75 015 Paris, France
[2] Institut Pasteur – Hub Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS – Institut Pasteur de Paris – 26-28 rue du Docteur Roux, F-75 015 Paris, France

With an ever-increasing speed and scalability, Next-Generation Sequencing (NGS) enables research laboratories to study complex biological questions. Even though the large amount of data generated is overwhelming, the analysis of such data benefits from a large community of developers. Yet, the heterogeneity of this community led to a fragmented set of tools, which needs to be assemble into specialised pipelines (e.g. quality control, variant detection, de-novo assembly...).

Sequencing platforms that produces sequencing data on a daily-basis must developed NGS pipelines to guarantee that good-quality sequencing data are delivered to research labs. For example, the Biomics pole (Institut Pasteur, CITECH), produces about 100 runs a year combining MiSeq and HiSeq 2500 technologies.

Here, we describe Sequana (`http://sequana.readthedocs.io/`), a Python-based software dedicated to the development of NGS pipelines, which is used within the Biomics pole to analyse various samples (viruses, bacteria, fungi, yeast). We will present the methodology used within the Sequana software to develop and deploy pipelines. We will highlight 3 major aspects. First, pipelines embedded within Sequana are based on the Snakemake framework that ease the decomposition of pipelines into modular sub-units. Although widely used, some bottlenecks have been identified and solved within Sequana (e.g. re-usability of a module within the same pipeline). Secondly, with each pipeline developed, we associate an HTML report. Based on JINJA templating and Javascript, this report summarizes the results of a pipeline and also provides all material to reproduce it. Third, we provide a set of Python tools that (1) bridge the gap between external tools that have been developed independently and (2) alleviate the need for a plethora of small external scripts.

Currently, Sequana provides pipelines to:

1. analyse sequence data after demultiplexing to check the sequence quality,
2. check the taxonomic content of the data,
3. discover variants(SNPs, indels),
4. detect anomalies in the mapping coverage.

Sequana is developed with the aim of simplifying the development of new tools (for developers) and the deployment of the pipelines (for users). The extended documentation and test suite provides a high-quality software that will be used in production within the Biomics NGS platforms. This should also be of interest to help researchers or other NGS platforms to build on Sequana to re-use or extend existing pipelines.

Mots clefs : snakemake, python, NGS, variant, coverage, quality assessment

*. Intervenant
†. Corresponding author : `thomas.cokelaer@pasteur.fr`

Analyses en transcriptomique de la domestication du palmier pacaya, exploité pour son inflorescence comestible en Amérique latine

Abdoulaye Diallo *[1,2,3]

Poster 52

[1] Étudiant en première année du parcours « Bioinformatique, Connaissances, Données (BCD) » du master Sciences et Numérique pour la Santé (SNS) – Université Montpellier II - Sciences et Techniques du Languedoc – France
[2] UMR DIADE IRD/UM/CIRAD,IRD - France Sud – Christine Tranchant-Dubreuil, James Tregear – 911 avenue Agropolis, BP 64501, F-34 394 Montpellier Cedex 5, France
[3] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université Montpellier II - Sciences et techniques, Alban Mancheron – 860 rue de St Priest, CC 05016, F-34 095 Montpellier Cedex 5, France

La floraison des plantes est un élément clé qui influe sur leurs adaptations écologiques mais aussi, dans le cas des espèces cultivés, sur leur productivité. Les arécacées (palmiers) forment un groupe d'environ 2 600 espèces considéré comme l'une des familles de plantes les plus utilisées par l'homme. Le palmier pacaya (*Chamaedorea tepejilote*) est une espèce dioïque dont la distribution naturelle s'étend du sud du Mexique jusqu'au nord de la Colombie en passant par l'Amérique Centrale. Il est exploité depuis plus de deux millénaires, principalement pour son inflorescence mâle qui ressemble à un épi de maïs et est consommé cuit ou cru (le mot tepejilote se traduit par « maïs de montagne »). Dans certains pays, notamment au Guatemala, les palmiers pacaya cultivés se distinguent par des caractères particuliers, recherchés par les consommateurs. Ils produisent notamment de plus grosses inflorescences avec une ramification accrue.

Le but de l'étude est de mieux caractériser les changements morphologiques liés à la domestication du palmier pacaya et de rechercher, par une approche NGS, les modifications moléculaires correspondantes à travers des études du transcriptome. Pour ce faire, un échantillonnage d'inflorescences issues de palmiers cultivés et sauvages de plusieurs pays (Guatemala, Colombie, Belize, Honduras) a été réalisé ; et séquencé par la technologie Illumina HiSeq.

Pour mener à bien les analyses bioinformatiques, il y a plusieurs étapes clés :

— Il n'y a aucun génome de référence ni transcriptome de référence pour cette espèce de palmier. La première étape du projet a été de réaliser un transcriptome de référence *in silico*.
— L'étape de mapping est cruciale dépendant de l'algorithme utilisé pour l'étape d'assemblage.
— Le développement de pipeline d'analyse de données transcriptomiques.

Je vous présenterai d'une part un benchmarking d'outils d'assemblage pour obtenir un transcriptome de référence *de novo*, et de logiciels de mapping des données RNAseq (tophat, CRAC, STAR).

Pour finir, je présenterai les premiers résultats obtenus suite aux analyses différentielles entre les accessions sauvages et domestiquées du palmier pacaya.

Je tiens à remercier le département informatique de la Faculté des Sciences de l'Université de Montpellier (`http://deptinfofds.univ-montp2.fr/`) pour la formation, ainsi que le Labex Numev (`http://www.lirmm.fr/numev/`) pour avoir accepté de financer ma participation à JOBIM2016. Je remercie également les équipes EDI et RICE de l'IRD (UMR DIADE) pour leur encadrement tout au long de mon stage.

*. Intervenant

Mots clefs : RNA-seq, Transcriptomique, palmier pacaya, Illumina, NGS, assemblage de novo, mapping, benchmarking

Une méthode d'optimisation pour la reconstruction des haplotypes

Thomas Dias Alves *[1], Michael Blum[1], Julien Mairal[2]

Poster 53

[1] Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG) – CNRS : UMR5525, Université Joseph Fourier - Grenoble I – Domaine de la Merci, F-38 706 La Tronche, France

[2] LEAR (INRIA Grenoble Rhône-Alpes / LJK Laboratoire Jean Kuntzmann) – CNRS : UMR5224, Institut polytechnique de Grenoble (Grenoble INP), Université Joseph Fourier - Grenoble I, Laboratoire Jean Kuntzmann, INRIA – INRIA Grenoble Rhône-Alpes, 655 avenue de l'Europe, F-38 330 Montbonnot, France

La reconstruction d'haplotypes consiste à estimer les haplotypes à partir des génotypes. Un haplotype est une suite d'allèles qui sont transmis ensemble et qui sont situés sur différents loci du même chromosome. La connaissance des haplotypes est un outil important pour la génomique des populations. C'est une étape préalable pour réaliser l'imputation des données manquantes, détecter les gènes impliqués dans la sélection naturelle ou bien détecter la structure des populations à une échelle fine.

Nous présentons une méthode rapide et précise permettant de reconstruire les haplotypes à partir des génotypes. La méthode est une méthode populationnelle qui tire de l'information provenant d'une population de génotypes. La méthode repose sur un modèle de clustering local des haplotypes qui est similaire au modèle implémenté dans le logiciel fastPHASE. Dans ce modèle, l'haplotype d'un individu est représenté par une mosaïque d'haplotypes types appelés clusters. La méthode repose sur une contrainte de parcimonie qui limite le nombre de sauts entre clusters.

Notre méthode consiste à moyenner des solutions approchées d'un problème d'optimisation avec contraintes. L'algorithme proposé est un algorithme de moindres carrés alternés. L'étape de moindres carrés qui revient à trouver les clusters dans lesquels piocher repose sur une recherche de plus court chemin qui se fait avec un algorithme de programmation dynamique. La mise à jour des haplotypes à reconstruire est faite au sein de l'algorithme en utilisant un échantillonnage aléatoire. Différentes solutions du problème d'optimisation produisent différents haplotypes qui sont ensuite moyennés pour produire une solution plus optimale.

L'avantage de l'algorithme proposé est que la complexité de calcul est linéaire par rapport au nombre de clusters alors que la complexité de fastPHASE augmente de manière quadratique avec le nombre de clusters. De plus, puisque la méthode repose sur le calcul de plusieurs optima locaux qui sont ensuite moyennés, nous avons implémenté l'algorithme en parallèle pour profiter du fait que les ordinateurs modernes ont des processeurs multi-cœurs.

La méthode proposée a été évaluée avec des critères de précision et de vitesse. Pour construire la vérité-terrain, des données génomiques pour des trios (parents et un enfant) permettent d'élaborer des haplotypes avec une grande fiabilité. En comparant les haplotypes reconstruits avec notre méthode et ceux reconstruits en utilisant l'information provenant des trios, il est possible de calculer une mesure d'erreur. Pour les données de génomique humaine HAPMAP, notre méthode obtient des erreurs similaires à fastPHASE mais avec un temps de calcul réduit d'au moins un facteur 10. En reconstruisant les haplotypes du chromosome 1 des individus mormons de HAPMAP (CEU), la mesure de l'erreur est de 4.7 % pour notre algorithme, de 3.7 % pour fastPHASE et de 4.3 % pour Beagle qui est un autre logiciel de phasage. Les tests ont aussi été effectués sur

*. Intervenant

les individus mexicains et africains de HAPMAP et nous avons obtenu des résultats similaires en terme de mesure d'erreur et de vitesse.

L'algorithme que nous proposons pour reconstruire des haplotypes a l'avantage d'être rapide, précis et de fournir une modélisation des données, sous forme de clusters locaux, qui peut être utilisée dans des analyses de génomique des populations. La méthode présentée est implémentée en C++ et est disponible dans un package Python.

Mots clefs : Haplotype, Phasing, Optimization, Reconstructions, Big Data

Exploring the mystery leading to a specific parasite infection for marine blooming dinoflagellates by gene expression screening

Poster 54

Sarah Farhat [*] [†1,2], Benjamin Noël[1], Corinne Da Silva[1], Jean-Marc Aury[1], Laure Guillou[3], Betina Porcel[1,2], Patrick Wincker[1,2]

[1] Genoscope - Centre national de séquençage – CEA, Genoscope – 2 rue Gaston Crémieux CP5706, F-91 057 Évry Cedex, France
[2] Génomique métabolique (UMR 8030) – CEA, CNRS : UMR8030, Université d'Évry-Val d'Essonne – Genoscope, 2 rue Gaston Crémieux, F-91 057 Évry Cedex, France
[3] UMR 7144 Adaptation et Diversité en milieu marin – CNRS : UMR7144, Université Pierre et Marie Curie [UPMC] - Paris VI – France

Alveolata is one of the major eukaryotic lineages. It is composed of protist classes. One of them, Syndiniales is a diverse and highly widespread group of ubiquitous marine parasites infecting many planktonic species. Many Syndiniales kills obligatorily their host to complete their life cycle. This is the case for organisms belonging to the *Amœbophrya* genus, which are able to infect toxic micro-algae (dinoflagellates) responsible of toxic algal blooms. Because of their virulence and abundant offspring, such parasites have the potential to control dinoflagellate populations. It has been shown that two categorizes of this genus exists depending on the range of hosts they could colonize: the generalists and the specialists' parasites. In culture, specialists' parasites exhibit a narrow host spectrum, generally infecting no more than 1-2 dinoflagellate species. In the field, the same parasitic species may infect the same host species, years after years. Coastal planktonic ecosystems are by nature characterized by strong environmental fluctuations. This should theoretically lead to the natural selection of generalist parasites at the expense of specialists. Thus, the persistence and ecological success of specialists among marine planktonic parasites is an intriguing paradox and a potential limitation to dinoflagellate blooms. *Amœbophrya* life cycle alternate between a free-living stage (dinospore) and the infection phase; the infection starts when a dinospore attaches to his host cell's surface, and then enter into the host's cytosol. Once inside, the parasite infests the host's nucleus to become a trophont. Then, the trophont increases in size by having a series of nuclear divisions. Once fully matured, the trophont expands through the cell wall of the host and transforms into vermiform. Finally, this vermiform structure separated into many individual dinospore.

The overarching goal of this project is to unveil the molecular components, mechanisms and evolutionary forces that determine the ability of specialized parasites to infect their primary host and adapt to a novel host, and how frequent host-range variations could be in nature. This is performed by the genomic analysis of two distinct *Amœbophrya* strains in terms of impact on the host. We sequenced the genome of these two strains and did the gene annotation. With a global analysis of their genomes, we had revealed unusual characteristics about these strains. In fact, *Amœbophrya*'s gene structure does not follow the standard GT-AG intron borders. Moreover, these parasites seem to be phylogenetically distant to other known organisms: 48 % and 42 % of the annotated protein-coding genes have no known domains (Interproscan), and 36 % and 32 % of these proteins have a significant match to the UniProtKB database. We are interested to understand different aspect of *Amœbophrya* mechanisms: How the infection affects the parasite life cycle inside the host at a molecular level? What genes and/or metabolic pathways are triggered

*. Intervenant
†. Corresponding author : sfarhat@genoscope.cns.fr

along the infection? And how does gene expression change between the free living stage and the life cycle inside the host? Using RNAseq data, at different time of the infection of *Amœbophrya* strain on its host, we were able to have an overview of the gene expression during the parasite's life cycle. In order to understand each step of the infection from a transcriptome point of view, we mainly used DESeq2, an R package tool that allow studying gene expression. We used the free-living stage as a reference of gene expression and screened all gene expression at different times of the infection. Then with a clustering method, we categorized gene expression by similar profile. In parallel, we identified the different molecular interactions of each organism with using different tools.

Results will be shown starting with the global expression analysis of each strain and a comparative analysis between the two organisms in order to understand differences and similarities of the gene expression between a generalist and a specialist parasitoid.

Mots clefs : Parasite, transcriptomic, gene expression

Selection in humans, the lost signal

Elsa Guillot [*1,2], Jérôme Goudet[1,2], Marc Robinson-Réchavi[1,2]

[1] Université de Lausanne (UNIL) – Suisse
[2] Swiss Institute of Bioinformatics (SIB) – Suisse

Poster 55

Selection has been acting at multiple scale of evolutionary history. Typically, population genetics studies have found very recent signal of selection in humans by comparing samples from different population. On a larger temporal scale, methods from molecular evolution permit to identify positive selection by comparing different species. However the closer these species, the weaker the signal.

Hence, it remains a challenge to find selection signal in the human genome, older that the human expansion out of Africa (> 100kya) but younger than the split with chimpanzees (> 6,500kya), which would be specific to human as a species.

The aim of this study is to find new methods for detecting signals of selection in humans by using both inter-species and intra-species data. Built upon previous methods (Keightley et al. 2007, Boyko et al. 2008, Eyre-Walker et al. 2009) this approach infers selection alongside demographic history. Particularly challenging is the scalability of this computations relying on large matrix operations. In this context we built a new inference framework, optimizing this computations to apply on large human dataset.

Mots clefs : Population genetics, selection, humans

*. Intervenant

Oxford Nanopore Technologies : données et applications

Benjamin Istace [*1]

[1] Genoscope - Centre national de séquençage – CEA, Genoscope – 2 rue Gaston Crémieux CP5706, F-91 057 ÉVRY Cedex, France

Poster 56

Les régions génomiques complexes, telles que les répétitions ou les régions polymorphes, peuvent être difficiles à assembler en utilisant des lectures courtes. Ces problèmes sont partiellement résolus par l'avènement des longues lectures, puisqu'elles permettent de traverser les zones répétées et de phaser les haplotypes.

L'année dernière, Oxford Nanopore a révélé un séquenceur de très petite taille nommé le MinION. Celui-ci peut être branché directement à un ordinateur via un port USB. Les deux brins de l'ADN, liés par une structure en épingle à cheveux (hairpin), sont guidés à travers un nanopore, parcouru par un courant électrique, par une enzyme. Les bases sont ensuite lues par 6-mers, chaque variation du courant électrique correspondant à un 6-mer spécifique. Si les deux brins sont lus avec succès, une séquence consensus appelée 'lecture 2D' est générée, sinon, seule la séquence forward (aussi appelée 'lecture 1D') est conservée. Cette technologie offre plusieurs avantages, le premier étant que le MinION est très petit et très peu cher. De plus, la préparation des librairies ne requiert pas d'amplification par PCR, ce qui élimine tous les biais liés à cette technique.

Des tests ont été menés par tous les utilisateurs, à la base sur le génome du phage lambda, et ont montré l'obtention de lectures mesurant en moyenne 5 kb. Cependant, ces tests ont aussi montré un haut taux d'erreur, de l'ordre de 15 %. Malgré ce taux d'erreur élevé, plusieurs génomes bactériens et eucaryotes ont pu être assemblés.

Cette présentation orale aura pour but de récapituler les propriétés du MinION, autant du point de vue de la technique utilisée que des statistiques concernant les lectures générées. Dans une seconde partie, la méthode NaS, permettant de corriger efficacement les longues lectures bruitées grâce à des données de séquençage Illumina sera présentée. Enfin, les résultats d'assemblage du génome de la levure S288C séquencée par le MinION seront dévoilés.

Mots clefs : Oxford Nanopore, longues lectures, assemblage

*. Intervenant

Intraspecific epigenetic conservation of duplicate genes associated to their transposable element neighborhood in human

Romain Lannes [*1], Emmanuelle Lerat [†1]

Poster 57

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France

Epigenetic modifications correspond to heritable changes caused by modifications in the chromatin structure rather than in the DNA molecule itself. Three intertwined mechanisms including DNA methylation, histone modifications and RNA interference are currently considered to initiate and sustain epigenetic changes [1]. They can modulate the gene expression such as differential expression between tissues or in response to environmental changes [2]. DNA methylation usually occurs in the context of CpG dinucleotides and is associated with transcription silencing [3–6]. RNA interference involves small noncoding RNAs, which target messenger RNAs and trigger their degradation [7,8]. Histone modifications correspond to post-translational changes occurring at particular amino acid residues of these proteins [6,9,10]. According to the type of histone modification, the effect can be either activating or repressing the gene expression [11,12]. Understanding the mechanisms involved in the initiation, maintenance and heritability of epigenetic states is an important aspect to better understand the cell functioning and to evaluate their implication in the genome evolution. Functionally important regions of genomes are conserved throughout evolution. Since epigenetic modifications may have crucial impact on both gene and cell functioning, we can ask whether some selective constraints are acting on these modifications and whether there is a possible evolutionary conservation of the epigenome linked to the conservation of DNA sequence.

Changes in the gene regulation may play a role in the adaptation and evolution of organisms [13]. Since part of the variation of gene expression can be explained by variation in epigenetic modifications, they could be important implied factors. Moreover, the epigenetic conservation or divergence can be linked to the DNA sequence conservation. In human, hypomethylated CpG islands are under selective constraints [14]. These CpG islands were also shown to be more enriched in particular histone modifications [15]. The acquisition of hypermethylated DNA in human is coupled to a very rapid nucleotidic evolution near CpG sites [16]. A comparison of three histone modifications among several cell types from human and mouse showed a strong association between the stability (intraspecies) and the conservation (interspecies) of these modifications against both genetic and environmental changes [17]. At an intraspecific level, epigenetic modifications may be implicated in functional divergence by facilitating tissue-specific regulation. For example, human duplicate genes are initially highly methylated, then gradually loose DNA methylation as they age. Within each pair of genes, DNA methylation divergence increases with time. Moreover, tissue-specific DNA methylation of duplicates correlates with tissue-specific expression, implying that DNA methylation could be a causative factor for functional divergence of duplicate genes [18]. However, epigenetic modifications may also play a role in the functional conservation. For example, in some plants, paralogous genes associated to a particular histone modification showed the highest coding sequence divergence but the highest similarity in expression patterns and in regulatory regions when compared to paralogous genes in which only one gene was the target of this histone modification [19]. By comparing recent segmental duplications

*. Intervenant

†. Corresponding author : emmanuelle.lerat@univ-lyon1.fr

regions in human, a widespread conservation of DNA methylation and some histone modifications was observed [20].

Eukaryotic genomes are formed from a variety of elements among which protein-coding genes are a minority. The human genome is indeed constituted by only < 2 % of protein-coding genes, whereas repeated sequences represent more than the half [21]. While the non-coding part was first thought to bare no function [22], it is now known to be composed of a mixture of repetitive DNA and non-coding sequences crucial for transcriptional and post-transcriptional gene regulation [23,24]. The greater part of repeated DNA is classified as transposable elements (TEs), with several millions of them inserted throughout the human genome. TEs are middle-repeated DNA sequences that have the ability to move from one position to another along chromosomes. They typically encode for all the proteins necessary for their movement and possess internal regulatory regions, allowing for their independent expression. Different categories of TEs have been identified [25,26]. Because of their presence in genomes, TEs have a significant impact on genome evolution [27,28]. TEs can promote various types of mutations, which are expected to be mostly deleterious when affecting functional regions. The mammalian genomes harbor million of insertions of TEs that have not been counter-selected during evolution or that are too young to have been eliminated since they did not have any deleterious impacts or because they were on the contrary selected for some advantages they provided to the organism. Then, fixed TE insertions can still influence the genome evolution, expression or regulation [29]. To counteract their deleterious effects, TEs are regulated by the host genome via epigenetic mechanisms to suppress or silence the TE activity [30,31]. In normal mammalian cells, TEs are usually methylated, therefore transcriptionally silenced. In cancer cells where DNA methylation is abolished, TEs can be mobilized, resulting in a potential impact on the integrity of the cell [32,33]. A change in the local epigenetic landscape associated with the presence of TEs is expected to affect the expression of the neighboring genes since these modifications occurring at TE sequences can spread to neighboring sequences [34–36]. The spreading of TE histone modifications to adjacent regions has also been observed [37–39] indicating that the presence of TEs can potentially influence the epigenetic state of neighboring genes. When comparing segmental duplications in human, Alu elements were observed to be enriched around methylated sites of discordant paralogous regions [20].

The question we want to address here is how the epigenetic modifications are conserved and what is the role of TEs in this conservation. For that, we have studied the conservation of the epigenome at an intraspecific level in human. By measuring, in a given environmental condition, the divergence of epigenetic modifications associated to duplicated genes and linked to the differential presence of TEs near the genes, we have determined the impact of TEs on epigenetic changes associated with the time since duplication, and the function divergence of these genes. For that, we have retrieved functional pairs of duplicate genes from the HOGENOM database [40] and have crossed their position in the genome with that of TE insertion to determine the TE environment richness of each gene. The time since gene duplication inside each pair was determined by computing the synonymous substitutions rates. Functional divergence was estimated using the Gene Ontology annotation. Finally, we have determined the histone enrichment of each gene for four different histone modifications in several tissues as well as the gene expression level. Because of the negative regulation of TEs by epigenetic mechanisms, we could expect a greater variation in the epigenetic environment in gene pair displaying different TE environment.

References

[1] Egger, G., Liang, G., Aparicio, A. & Jones, P. a. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 429, 457–463 (2004).

[2] Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* 33 Suppl, 245–254 (2003).

[3] Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* 16, 6–21 (2002).

[4] Weber, M. & Schübeler, D. Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr. Opin. Cell Biol.* 19, 273–80 (2007).

[5] Jones, P. A. & Liang, G. Rethinking how DNA methylation patterns are maintained. *Nat. Rev. Genet.* 10, 805–11 (2009).

[6] Bernstein, B. E., Meissner, A. & Lander, E. S. The Mammalian Epigenome. *Cell* 128, 669–681 (2007).

[7] Carthew, R. W. & Sontheimer, E. J. Origins and Mechanisms of miRNAs and siRNAs. *Cell* 136, 642–55 (2009).

[8] Ghildiyal, M. & Zamore, P. D. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* 10, 94–108 (2009).

[9] Grant, P. A. A tale of histone modifications. *Genome Biol.* 2, REVIEWS0003 (2001).

[10] Peterson, C. L. & Laniel, M.-A. Histones and histone modifications. *Curr. Biol.* 14, R546–51 (2004).

[11] Ha, M., Ng, D. W.-K., Li, W.-H. & Chen, Z. J. Coordinated histone modifications are associated with gene expression variation within and between species. *Genome Res.* 21, 590–598 (2011).

[12] Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* 128, 707–19 (2007).

[13] Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science (80-.).* 165, 349–357 (1969).

[14] Cocozza, S., Akhtar, M. M., Miele, G. & Monticelli, A. CpG islands undermethylation in human genomic regions under selective pressure. *PLoS One* 6, e23156 (2011).

[15] Akhtar, M. M., Scala, G., Cocozza, S., Miele, G. & Monticelli, A. CpG islands under selective pressure are enriched with H3K4me3, H3K27ac and H3K36me3 histone modifications. *BMC Evol. Biol.* 13, 145 (2013).

[16] Hernando-Herraez, I. *et al.* The interplay between DNA methylation and sequence divergence in recent human evolution. *Nucleic Acids Res.* 43, 8204–8214 (2015).

[17] Woo, Y. H. & Li, W. H. Evolutionary conservation of histone modifications in mammals. *Mol. Biol. Evol.* 29, 1757–1767 (2012).

[18] Keller, T. E. & Yi, S. V. DNA methylation and evolution of duplicate genes. *Proc. Natl. Acad. Sci.* 111, 5932–5937 (2014).

[19] Berke, L., Sanchez-Perez, G. F. & Snel, B. Contribution of the epigenetic mark H3K27me3 to functional divergence after whole genome duplication in Arabidopsis. *Genome Biol.* 13, R94 (2012).

[20] Prendergast, J. G. D., Chambers, E. V. & Semple, C. a. M. Sequence-Level Mechanisms of Human Epigenome Evolution. *Genome Biol. Evol.* 6, 1758–1771 (2014).

[21] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–45 (2004).

[22] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).

[23] Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10, 691–703 (2009).

[24] Ludwig, M. Functional evolution of noncoding DNA. *Curr. Opin. Genet. Dev.* 12, 634–639 (2002).

[25] Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8, 973–982 (2007).

[26] Kapitonov, V. V & Jurka, J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* 9, 411–412; author reply 414 (2008).

[27] Kidwell, M. G. & Lisch, D. R. Transposable elements and host genome evolution. *Trends Ecol. Evol.* 15, 95–99 (2000).

[28] Biémont, C. & Vieira, C. Genetics: junk DNA as an evolutionary force. *Nature* 443, 521–524 (2006).

[29] Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science (80-.).* 351, aac7247–aac7247 (2016).

[30] Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 8, 272–285 (2007).

[31] Huda, A. & Jordan, I. K. Epigenetic regulation of mammalian genomes by transposable elements. *Ann. N. Y. Acad. Sci.* 1178, 276–284 (2009).

[32] Kulis, M. & Esteller, M. DNA methylation and cancer. *Adv. Genet.* 70, 27–56 (2010).

[33] Ross, J. P., Rand, K. N. & Molloy, P. L. Hypomethylation of repeated DNA sequences in cancer. *Epigenomics* 2, 245–269 (2010).

[34] Morgan, H. D., Sutherland, H. G., Martin, D. I. & Whitelaw, E. Epigenetic inheritance at the agouti locus in the mouse. *Nat. Genet.* 23, 314–8 (1999).

[35] Rebollo, R. *et al.* Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS Genet.* 7, (2011).

[36] Eichten, S. R. *et al.* Spreading of Heterochromatin Is Limited to Specific Families of Maize Retrotransposons. *PLoS Genet.* 8, (2012).

[37] Gendrel, A.-V., Lippman, Z., Yordan, C., Colot, V. & Martienssen, R. A. Dependence of heterochromatic histone H3 methylation patterns on the Arabidopsis gene DDM1. *Science* 297, 1871–3 (2002).

[38] Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* 297, 1833–7 (2002).

[39] Lippman, Z. *et al.* Role of transposable elements in heterochromatin and epigenetic control. *Nature* 430, 471–476 (2004).

[40] Penel, S. *et al.* Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 10 Suppl 6, S3 (2009).

Mots clefs : transposable element, epigenetics, genome evolution

The red queen dynamic in the kingdom of recombination

Thibault Latrille *[1]

[1] École Normale Supérieure de Lyon (ENSL) – 46 allée d'Italie, F-69 007 Lyon, France

Poster 58

In humans and many other species, recombination events cluster into narrow hotspots within the genome. Given the vital role recombination plays in meiosis, we might expect that the positions of these hotspots would be tightly conserved over evolutionary time. However, there is now strong evidence that hotspots of meiotic recombination in humans are transient features of the genome. For example, hotspot locations are not shared between human and chimpanzee. Biased gene conversion in favor of alleles that locally disrupt hotspots is a possible explanation of the short lifespan of hotspots.

Remarkably, Prdm9 has been proposed to be a key determinant of the positioning of recombination hotspots during meiosis, and the most rapidly evolving gene in human. Prdm9 genes often exhibit substantial variation in their numbers of encoded zincfingers, not only between closely related species but also among individuals of a species.

Here, we propose a population genetic model which exhibits hotspots transience while reflecting the PRDM9 features, resulting in a intragenomic red queen dynamic. Our model account for empirical observations regarding the molecular mechanisms of recombination hotspots and the nonrandom targeting of the recombination by PRDM9. We further investigate and compare to known data the diversity of PRDM9, the hotspots turnover and the genome wide disruption of hotspots.

Mots clefs : Recombination, hotspots transience, red queen, population genetic model, PRDM9

*. Intervenant

Étude comparative des génomes et transcriptomes de *Mucor* spp.

Poster 59

Annie Lebreton [*1], Laurence Meslet-Cladière[1], Jean-Luc Jany[1], Georges Barbier[1], Erwan Corre [†2]

[1] Laboratoire Universitaire de Biodiversité et Écologie Microbienne (LUBEM) – IFR148 ScInBioS, PRES Université Européenne de Bretagne (UEB), Université de Bretagne Occidentale (UBO) : EA3882 – ESIAB - Parvis Blaise Pascal - Technopôle Brest-Iroise, F-29 280 Plouzané, France
[2] Station biologique de Roscoff (SBR) – Université Pierre et Marie Curie (UPMC) - Paris VI, CNRS : FR2424 – Place Georges Teissier - BP 74, F-29 682 Roscoff Cedex, France

Parmi les champignons d'intérêt technologique en agroalimentaire, sont retrouvées, en dehors des ascomycètes tels que *Penicillium camemberti* ou *P. roqueforti*, plusieurs espèces de champignons appartenant au sous phylum des Mucoromycotina. Au sein de ce groupe, le genre *Mucor* comprend des espèces dites technologiques utilisées notamment dans l'affinage de certains fromages tels que les Tommes, le Saint-Nectaire ou encore le Cantal, dans lesquels elles sont à l'origine du développement de caractères organoleptiques typiques. A contrario, d'autres fromages sont, de manière récurrente, contaminés par certaines espèces de *Mucor* (dites d'altération) qui engendrent des accidents de fabrication comme celui dit du « poil de chat » ou sont possiblement à l'origine d'intoxications alimentaires. Ces accidents sont responsables de pertes de production qui peuvent être économiquement significatives. Afin d'améliorer les connaissances sur les *Mucor*, une étude de génomique et transcriptomique comparatives est actuellement menée au sein du laboratoire. Un intérêt particulier est porté aux éléments du génome (gènes, éléments répétés, systèmes de régulation etc.) pouvant être liés à une adaptation à la matrice fromagère.

Cette étude a ciblé cinq espèces de *Mucor* : *M. fuscus* et *M. lanceolatus*, espèces technologiques prééminentes dans les fromages, *M. racemosus*, contaminant fréquent des productions fromagères, *M. circinelloides*, contaminant occasionnel et *M. endophyticus*, espèce endophyte à la croissance ralentie sur matrice fromagère. Deux espèces appartenant au sous-phylum Mucoromycotina et non retrouvées dans les fromages ont également été intégrées à l'étude : *Phycomyces blakesleeanus* et *Rhizopus oryzae*. Cette dernière étant utilisée dans la production traditionnelle d'aliments asiatiques fermentés.

Les souches *M. fuscus* UBOCC 1.09.160, *M. lanceolatus* UBOCC 1.09.153, *M. racemosus* UBOCC 1.09.155 et *M. lanceolatus* CBS 385-95 ont été séquencées par le laboratoire avec une approche mate-pair (MP) et paired-end (PE). Les données MP de *M. lanceolatus* ainsi que MP et PE *M. endophyticus*, sont en cours de séquençage et devraient être accessibles prochainement. Les génomes de *M. circinelloides* CBS 277.49, *P. blakesleeanus* NRRL1555 et *R. oryzae* 99-880 étaient disponibles dans les bases de données publiques. En parallèle, les transcriptomes des quatre souches *M. fuscus* UBOCC 1.09.160, *M. lanceolatus* UBOCC 1.09.153, *M. racemosus* UBOCC 1.09.155 et *M. lanceolatus* CBS 385-95 ont été séquencés à partir de banques de cDNAs totaux.

Des assemblages ont été réalisés avec les assembleurs SOAPdenovo2, clc-assembly cell, Velvet et spades. À cette date, des recherches d'orthologies avec *M. circinelloides*, *P. blakesleeanus* et *R. oryzae* ont été effectuées pour les deux génomes *M. fuscus* et *M. racemosus*. Les données transcriptomiques ont été alignées sur leurs génomes respectifs avec STAR, les transcrits ont été reconstruits avec Cufflinks et une structure génique a été prédite avec TransDecodeur. Des prédictions géniques ont également été menées avec Glean et Augustus. Des annotations géniques fonctionnelles ainsi

*. Intervenant
†. Corresponding author : corre@sb-roscoff.fr

qu'une prédiction des éléments répétés ont été réalisées pour ces mêmes espèces. Toutes ces informations ont été formatées pour être consultées sur le « genome browser » JBrowse.

L'assemblage de *M. fuscus* est constitué de 4000 contigs pour une taille de l'ordre de 40Mb. Sur ces contigs, environ 7500 gènes ont été prédits. L'assemblage de *M. racemosus* est constitué de 3 500 contigs pour une taille de 47 Mb. Sur ces contigs environ 12 500 gènes ont été prédits. La taille moyenne de ces gènes, de leur CDS et introns sont équivalents entre les deux espèces. Sur ces deux génomes environ 22 000 transcrits, incluant les isoformes, ont été reconstruits par Cufflinks.

Le meilleur assemblage de *M. lanceolatus*, obtenu avec spades, contient environ 12 000 contigs pour une taille de l'ordre de 50 Mb. L'assemblage de *M. lanceolatus* est actuellement trop fragmenté pour réaliser une étude de comparaison génomique. Celui-ci sera amélioré par la disponibilité prochaine de données MP.

Le nombre et la structure des gènes prédits pour *M. racemosus* sont similaires à ceux trouvés dans les bases de données pour *M. circinelloides*. La sensibilité de détection des gènes n'est cependant pas aussi bonne pour *M. fuscus*. Cela serait dû à une moindre qualité d'assemblage et probablement un manque de robustesse du modèle de gènes utilisé pour les prédictions. On notera que la pertinence des prédictions n'a pas été vérifiée, cette étape devra être réalisée lorsque tous les génomes seront disponibles.

Dans l'attente du jeu de données génomique complet, les transcriptomes ont été analysés. Les transcrits ont été reconstruits *de novo* avec Trinity puis annotés avec le pipeline Trinotate. Les termes GO des transcrits ont été déduit des domaines PFAM-A et des orthologies avec Uniref90 et SwissProt. Des recherches d'orthogroupes ont été réalisées avec OrthoFinder.

L'assemblage *de novo* des données RNAseq avec Trinity a conduit aux reconstructions d'environ 21 000 transcrits répartis en 14 000 gènes pour chacune les quatre espèces étudiées. Parmi ces gènes environ 10 000 ont été prédits comme codant des protéines.

L'analyse d'orthologie réalisée sur les transcrits reconstruits par Trinity a permis de générer environ 11 000 orthogroupes dont 6 644 sont communs aux cinq espèces. Parmi ces orthogroupes, 439 d'entre eux sont constitués uniquement des gènes de *M. fuscus* et *M. lanceolatus*.

Le nombre et la taille moyenne des transcrits reconstruits *de novo* sont cohérents entre les espèces. L'étude transcriptomique a permis d'estimer la taille et la composition du « *core transcriptome* ». Elle a également permis d'identifier un premier groupe de gènes qui serait conservé entre les espèces technologiques et absentes des autres espèces étudiées.

Les techniques testées dans le cadre de cette étude transcriptomique pourront être appliquées lors de la comparaison des génomes. D'une part afin de déterminer un « core génome » qui pourrait délivrer des informations importantes concernant les métabolismes au niveau du genre *Mucor*. D'autres part pour découvrir des gènes spécifiquement retrouvés ou absents des espèces technologiques. Ces groupes de gènes permettraient de détecter de possibles marqueurs d'adaptation, notamment au milieu fromager, ainsi que des métabolites à impact positif et négatif de différentes espèces dans un contexte fromager. Outre le contexte fromager, l'étude des génomes du genre *Mucor* apporte un éclairage original sur la l'histoire évolutive de ce groupe du fait de sa position basale par rapport aux champignons dits supérieurs. À titre d'exemple, les gènes impliqués dans la synthèse de métabolites secondaires des champignons supérieurs sont organisés en clusters, l'analyse des génomes de *Mucor* permettra d'étudier également l'évolution de l'organisation génique (dispersion ou clusterisation) de ces gènes.

Mots clefs : assemblage, annotation, adaptation, fromage

pcadapt : an R package to perform genome scans for selection based on principal component analysis

Poster 60

Keurcien Luu [*1], Michael Blum [†1]

[1] Techniques de l'Ingénierie Médicale et de la Complexité – Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG) – CNRS : UMR5525, Université Joseph Fourier - Grenoble I – Domaine de la Merci, F-38 706 La Tronche, France

We introduce the R package pcadapt that provides lists of candidate genes under selection based on population genomic data. The package is fast and can handle large-scale data generated with next-generation technologies. It works at the individual scale and can handle admixed individuals as it does not assume that individuals should be grouped into populations. It returns a list of P-values, which provides the opportunity to control for the false discovery rate. The statistical method implemented in pcadapt assumes that markers that are excessively related with population structure, as ascertained with principal component analysis, are candidates for local adaptation. The package computes vectors that measure associations between genetic markers andprincipal components. For outlier detection, a vector of associations is then transformed into a test statistic using Mahalanobis distance. Using simulated data, we compared the false discovery rate and statistical power of pcadapt to the ones obtained with FLK, OutFlank and BayeScan. For data simulated under an island model, we find that all software provide comparable results. However, in a model of divergence between populations, we find that BayeScan is too liberal, Outflank and BayeScan are too conservative, and pcadapt provide intermediate results. In terms of running time, we find that pcadapt is the fastest software of all included in the comparison. Because pcadapt can handle molecular data generated with next sequencing technologies, we anticipate that it will be a valuable tool for modern analysis in molecular ecology.

Mots clefs : population genetics, outlier detection, Mahalanobis distance, principal component analysis

*. Intervenant

†. Corresponding author : michael.blum@imag.fr

Association genetics to identify genes involved in aggressiveness traits in the plant pathogenic fungus *Mycosphaerella fijiensis*

Poster 61

Léa Picard [*1], Marie-Françoise Zapater[1], Daniel Bieysse[1], Françoise Carreel[1], Sébastien Ravel[2], François Bonnot[1], Yanetsy Montero[3], Véronique Roussel[1], Remy Habas[1], Luis Perez-Vicente[3], Catherine Abadie[1], Jean Carlier [†1]

[1] Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) – 42 rue Scheffer, F-75 116 Paris, France
[2] Biologie et génétique des interactions plantes-parasites pour la protection intégrée (BGPI) – Institut national de la recherche agronomique (INRA) : UR0385, Centre de coopération internationale en recherche agronomique pour le développement [CIRAD] : UMR54 – Campus International de Baillarguet - TA 41 / K, F-34 398 Montpellier Cedex 05, France
[3] Instituto de Investigaciones de Sanidad Vegetal (INISAV) – Cuba

The fungal pathogen *Mycosphaerella fijiensis* causing black leaf streak disease of banana may be able to adapt to quantitatively resistant cultivars through changes in its aggressiveness. To understand this adaptation, it is necessary to determine the genetic basis of the aggressiveness traits involved. This study aims to identify genes of aggressiveness traits using a genome wide association studies (GWAS) approach.

About 130 *M. fijiensis* isolates were collected on a susceptible and a quantitatively resistant banana cultivar in three different locations in Cuba. The genome of these isolates was then sequenced (Illumina sequencing, paired-end, 100 bp) and mapped against the reference genome using bwa software. SNP (single nucleotide polymorphism) calling was performed with GATK unified genotyper. SNPs were kept if they had a minimum coverage of 10X, alleles with a ratio of at least 90 % and complete data for all isolates, leading to a total of 154,612 validated SNPs.

Population structure was evaluated using the STRUCTURE software on 10,000 randomly selected SNPs across the whole genome and including data from 50 supplementary isolates from outside Cuba as outgroups. Amova was also conducted with the Cuban isolates with these SNPs. There was low genetic structure between the three different Cuban populations (Fst=0.055).

Linkage disequilibrium (LD) was calculated using the Hap-r2 option from vcftools. It declined to 50 % within about 6 kb. The population mutation rate θ was calculated using the egglib python library. The first estimate gave θ at a value of 0.06 %, which is quite low. Associated with the rather elevated LD, this might reflect the recent bottleneck event that accompanied the introduction of *M. fijiensis* in Cuba about 20 years ago.

The θ value was then used to run the Interval program from the LDhat package in order to calculate the population recombination rate ρ for each chromosome of *M. fijiensis* core genome. The mean ρ ranged from 0.98 Morgans/kb at scaffold 12 to 6.47 M/kb at scaffold 2, with an overall mean ρ for the core genome of 2.49 Morgans/kb. The recombination rate *r* was estimated at 0.00017 Morgans/kb on the core genome. From the effective population size calculated from ρ ($N\rho$) and from θ ($N\theta$) the frequency of sex was estimated to occur once out of 4 to 40 generations for a mutation rate of 10^{-8} to 10^{-9}, respectively.

A phenotype study was conducted to assess two aggressiveness traits using inoculations under controlled conditions: the number of symptoms and their total surface. About 100 isolates were

*. Intervenant

†. Corresponding author : jean.carlier@cirad.fr

inoculated on three cultivars (highly susceptible, susceptible and quantitatively resistant). We conducted a variance analysis using a linear model on the three traits. This revealed a strong cultivar effect and strong isolates in origin (resistant or susceptible) and origin in locality effects. Based on these results, we decided to use a simplified linear model taking only the cultivars and isolates effects into account, to calculate least-square means.

These preliminary analyses showed that the sample analyzed satisfy conditions to conduct a GWAS analysis (reasonable LD, limited population structure and phenotypic variability for the quantitative traits considered). Such a study is currently being performed using the GAP-IT R package using a multi-locus mixed model (MLMM).

Mots clefs : Banana, *Mycosphaerella fijiensis*, plant pathogenic fungus, aggressiveness traits, population genomics, GWAS

Screening of transposable element insertions in the mosquito *Anopheles gambiae*, the main malaria vector in Africa

Poster 62

Quentin Testard [*] [†1], Julien Veyssier[1], Anna-Sophie Fiston-Lavier [‡1]

[1] Institut des Sciences de l'Évolution - Montpellier (ISEM) – CNRS : UMR5554, Institut de recherche pour le développement [IRD] : UMR226, Université Montpellier II - Sciences et techniques – Place E. Bataillon CC 064, F-34 095 Montpellier Cedex 05, France

Anopheles gambiae is the main vector of malaria in Africa, a human disease that infects up to 500 million people and responsible for almost three million deaths per year. As traditional control strategies such as pesticides failed to halt the spread of these diseases, genetic studies of mosquito genomes has been investigated. Previous studies highlighted transposable element (TE) insertions that can induce transgene fixation in populations (Gould and Schliekelman, 2004). Other studies suggested that effector molecules (*i.e.*, genes affecting the ability of insects to transmit pathogens) are linked to TEs (Boete and Koella, 2002). TEs that can spread through populations in spite of fitness costs (Gould and Schliekelman, 2004) can be consider as good genetic elements to control the human disease spreading. Understanding the dynamics of TEs in *An. gambiae* will help the biological control efforts. Unfortunately, much less is known about TE population dynamics in *An. gambiae*.

Here, we investigate the TE population dynamics in *An. gambiae* taking advantage of the recent resequencing of more than 800 *Anopheles* individuals from the MalariaGEN project (`http://www.malariagen.net/`). This sequencing project consists in the sequencing of 765 wild-caught *Anopheles* specimens from eight countries in sub-Saharan Africa and also 80 *Anopheles* specimens comprising parents and progeny of four crosses. This dataset offers the unprecedented opportunity to explore the TE population dynamics in *An. gambiae but also* to assess the impact of TEs on genome structure and evolution in the *Anopheles* complex.

The first and major challenge of this study is to identify the TE insertions in each single individual. We recently developed a broadly applicable and flexible tool called T-lex2 (Fiston-Lavier et al 2011, 2014; `http://sourceforge.net/projects/tlex`). T-lex2 allows automatic and accurate genotyping of individual known TE insertions (*i.e.*, annotated in a reference genome). T-lex2 is composed of five modules allowing investigating the analysis of the flanking regions of the known TE insertions before detecting the presence and/or absence of these known TEs in individuals. The detection approaches are based on the analysis of the mapping and the read-depth coverage at the flanking regions of the TEs. Then T-lex2 combines the detection results to genotype the TEs and/or to estimate the TE frequencies in populations. As the TE detection depends on the TE annotation quality, T-lex2 offers the possibility to analyze the flanking regions of each TE to identify TEs putatively miss-annotated. For that, T-lex2 allows annotating TSDs (Target Site Duplications). The TSDs are short tandem repeats induced by the transposition mechanisms expected to be located at the flanking regions of the most of the TE insertions. The T-lex2 TSD detection is an unbiased and accurate approach that can be used for all types of TEs (Fiston-Lavier et la 2014). T-lex2 also now includes a new module to discover the *de novo* TE insertions (*i.e.*, novel TE insertions not present in the reference sequence). We will present the upgrade version of T-lex2 including this "*de novo*" module.

*. Intervenant
†. Corresponding author : `quentintest@gmail.com`
‡. Corresponding author : `anna-sophie.fiston-lavier@umontpellier.fr`

After the screening of the individual TE insertions in the *An. gambiae* populations, we will be able to analyze the full TE spectrum in *An. gambiae* and hope to propose a TE population dynamic model in *An. gambiae*. Focusing on TEs fixed in the population, active and close to genes, such analysis should help highlighting TEs putatively involved in the control of the spreading of malaria.

Mots clefs : Transposable element, mosquitoes, Anopheles, malaria, population genomics, sequencing, NGS, T, lex2

Mapping-Based Microbiome Analysis (MBMA) for diagnostic and molecular epidemiology

Anita Annamalé *[1], Amine Ghozlane[2,3], Philippe Glaser[4]

[1] Institut Pasteur – Institut Pasteur de Paris, CNRS : UMR3525, Université Paris Diderot - Paris 7 – France
[2] Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur de Paris – France
[3] Institut Pasteur – Institut Pasteur de Paris, CNRS : USR3756 – France
[4] Institut Pasteur – Institut Pasteur de Paris, CNRS : UMR3525 – France

Poster 63

A main challenge in the field of metagenomics is to quickly and precisely determine the composition and the relative abundance of constituents of a microbial community. State-of-the-art metagenomic methods rely on mapping- [1,2], kmer- [3,4] or assembly- [5] based approaches for identification and quantification. However, currently available methods are either time-consuming or fail to provide an accurate quantification at gene and strain levels.

To overcome these limitations, we developed Mapping Based Microbiome Analysis (MBMA), a mapping-based approach for identification and quantification of species, strains and resistance genes, with two main innovations: the use of a more efficient and discriminatory database for rapid quantification, and an optimized counting method for an accurate abundance prediction.

MBMA, implemented in Python, identifies which genes are present within a sample based on the mapping of Illumina-reads against publicly available databases. RefMG [6] was used for taxonomical classification and ResFinder [7] for identification of resistance genes. However, MBMA is compatible with any other database and is adapted to commonly-used mapping tools (Bowtie2 [8], BWA [9], NovoAlign [10]). It also includes different counting methods for quantification ("unique", "ex-aequo", "shared"), while providing a counting matrix directly compatible with statistical and visualization softwares (phyloseqR or shaman). Each step in MBMA was benchmarked using seven simulated metagenomes with closely related species, generated using ART ngs read simulator [11].

To determine the best mapping tool and the most accurate method for abundance estimation, we compared the observed and expected composition of the simulated metagenomes, by calculating the Kullback Leibler divergence and the Pearson and Spearman correlation. The best mapping tool was determined to be Bowtie2, with "shared" as the most accurate counting method. Finally, MBMA was compared to assembly- and k-mer-based (KmerFinder [3]) strategies for species identification.

We obtained a Kullback Leibler divergence of 18.64 %, 43.49 % and 2.48 % for KmerFinder, assembly-based approach and MBMA, respectively. MBMA successfully and accurately estimated the composition in all seven datasets within 15 to 20 minutes (with 4 CPUs).

In this work, we developed an accurate and reliable strategy for determining the composition of microbial communities containing limited diversity. We are currently testing this method on biological samples, while optimizing the detection and quantification of resistance genes. This tool will represent a promising alternative for the rapid identification and quantification of infectious agents in clinical microbiology laboratories.

References

[1] Luo, Chengwei et al. "ConStrains identifies microbial strains in metagenomic datasets."

*. Intervenant

Nature biotechnology 33.10 (2015):1045-1052.

[2] Scholz, Matthias et al. "Strain-level microbial epidemiology and population genomics from shotgun metagenomics." *Nature Methods* (2016).

[3] Hasman, Henrik et al. "Rapid whole genome sequencing for the detection and characterization of microorganisms directly from clinical samples." *Journal of clinical microbiology* (2013): JCM. 02452-13.

[4] Wood, Derrick E, and Steven L Salzberg. "Kraken: ultrafast metagenomic sequence classification using exact alignments." *Genome Biol* 15.3 (2014):R46.

[5] Kultima, Jens Roat et al. "MOCAT: a metagenomics assembly and gene prediction toolkit." *PLoS One* 7.10 (2012):e47656.

[6] Ciccarelli, Francesca D et al. "Toward automatic reconstruction of a highly resolved tree of life." *Science* 311.5765 (2006):1283-1287.

[7] Zankari, Ea et al. "Identification of acquired antimicrobial resistance genes." *Journal of antimicrobial chemotherapy* 67.11 (2012) 2640-2644.

[8] Langmead, Ben, and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2." *Nature methods* 9.4 (2012):357-359.

[9] Li, Heng. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." *arXiv preprint* arXiv:1303.3997 (2013).

[10] Hercus, C. "Novoalign." *Selangor: Novocraft Technologies* (2012).

[11] Huang, Weichun et al. "ART: a next-generation sequencing read simulator." *Bioinformatics* 28.4 (2012):593-594.

Mots clefs : metagenomics, microbiome, species identification, abundance estimation, antibiotic resistance, mapping

Taxonomic profiling and comparison of infectious metagenomics samples

Anaïs Barray [*] [†1], Mathias Vandenbogaert[1], Syria Laperche[2], Virginie Sauvage[2], Jean-Claude Manuguerra[3], Valérie Caro[1]

Poster 64

[1] Environment and Infectious Risks Unit, Genotyping of Pathogens Pole (ERI PGP) – Institut Pasteur de Paris – France
[2] Département d'études des agents transmissibles par le sang, Centre National de Référence des hépatites B et C et du VIH en transfusion – Institut National de la Transfusion Sanguine – France
[3] Environment and Infectious Risks Unit (ERI) – Institut Pasteur de Paris – France

Introduction

High-Throughput Sequencing (HTS) has gained in throughput and cost-efficiency, strongly affecting public health and biomedical research and enabling the conduct of large scale genomic projects. In this context, metagenomics has become a fast developing field for characterizing microbial communities in environmental and/or clinical samples, at the genomic level in order to reach functional and taxonomic conclusions.

In infectious metagenomics studies, it is critical to detect rapidly and sensibly potential life-threatening pathogens. With this poster, we want to introduce solutions to characterize the bacterial and viral composition suitable to clinical samples. It first requires a fast and accurate method for recovering the known microbial biodiversity (i.e listed in databases), followed with phylogenetic approaches to study the remaining species. The resulting taxonomic profiles will enable thorough sample comparisons in order to pinpoint infectious agents, giving leads in the establishment of infectious diagnostics.

Necessity of a new phylogenetic assessment strategy for infectious metagenomics samples

A community profile (the identification and quantification of the present species in a sample) is usually obtained by assigning the sequencing reads to different set of taxa. In metagenomics projects where huge amounts of data are produced, this is still a challenging problem which raises a number of computational issues :

1/ Similarity based methods use algorithms such as BLAST [1] are considered the most accurate methods for read assignment and classification. The relatively short read lengths of existing deep sequencing technologies and the homology across viral and bacterial species is part of the reason the current state-of-the-art methods struggle with ambiguous matches and false positives. This common problem causes available workflows to ignore the low abundance organisms as these may be false positives. Although this can be part of a logical approach to bypass false positives, it can also be troublesome in situations where the detection of a potentially disease-causing pathogen depends on identifying traces (very low number of reads) in the dataset.

2/ Accurate viral identification and quantification in complex mixtures containing other eukaryotic and prokaryotic sequences, has been relatively unattended despite the wealth of profiling and binning methods. Similarity based approaches using nucleotide similarities for inference are suboptimal for virus detection. This is especially critical when species are absent from the database, or when some closely related strains exist altogether in the sample.

*. Intervenant
†. Corresponding author : anais.barray@pasteur.fr

3/ Phylogenetic analysis of each read in a HTS dataset, by accurately placing an organism in the sample according to a ground-truth taxonomic tree, can be computationally challenging for large datasets. Individual reads can often only be placed inaccurately, and reporting on this uncertainty is still difficult.

4/ Using traditional (alignment and homology based) approaches, the task of assigning taxonomic labels to metagenomic DNA sequences has been relatively slow and computationally expensive, an approach that ultimately shifted in favor of using faster abundance estimation programs, which only classify small subsets of metagenomic data.

Using exact alignments of k-mers, novel approaches achieve classification accuracies comparable to BLAST. These approaches have the benefit of being several orders of magnitude faster, and provide powerful comparison tools to distinguish different organisms present in metagenomics HTS read datasets.

Aligning several sets of multi-million reads against public databases or a smaller subset of target reference genomes, remains computationally difficult even with optimized tools. A large fraction of reads may remain unmapped, as read alignment algorithms usually compute high-score alignments only, and fail to report low-quality alignments.

As such, alignment-free methods have come into use. The latter do not compute read alignments, but perform metagenomic classification of NGS reads based on the analysis of shared k-mers between input reads and the genomes from pre-compiled databases. The processing pipeline used in this work has been designed with viral sequence identification from RNA-sequencing as a main goal, yet being scalable to fit whole genome shotgun bacterial profiling.

A first study aimed to test the performance of a pipeline based on a k-mer alignment approach (Kraken [2]). Our pipeline has been assessed on clinical datasets from blood samples, its performance compared to more traditional methods addressing the community profiling task. The samples were pre-treated such that the majority of human (host) genetic material is removed and thus, in the end, leaves a manageable amount (up to a few million reads) for community profiling. For such clinical samples, the ground truth is naturally not known. However, the metagenomic landscape of the samples was artificially modified using spike-ins of potential viral pathogens, for the purpose of assessing the sensitivity and specificity of upstream DNA/RNA extraction/purification methods, while enhancing the speed/efficiency of our pipeline for the sake of its usability on larger scale data sets. Each sample was spiked with 7 viral pathogens at varying titers, and pipeline performances were compared on the profiling and quantification of these viruses. The Kraken pipeline was designed with both viral and bacterial community identification in mind, and several databases are being tested. In order to find species with high precision, k-mers' length was set to 31 bp, as longer exact alignment lowers false positive rate.

Early results are encouraging, as the Kraken pipeline can effectively recover the majority of the diversity found with BLASTs, while being approximately 6,000 to 10,000 times faster and less CPU consuming. Notably, the k-mer alignment approach performs effectively on unassembled reads, giving hints for a faster analysis solution. Further work is necessary to exhaust the known microbiome diversity remaining in a sample, for example with fast protein alignment tools alternative to BLASTx.

After the reduction of the known background in a metagenomic sample, analyzes of the remaining and potentially pathogenic diversity can be led without a priori. Current study shows that only 5-13% of the metagenomics reads can be classified accurately to the genus taxonomic level. Assembly of reads to generate more informative sequences followed by phylogenetic placement in a tree of reference microbes would be a strategy to harness the unannotated diversity within blood samples.

Comparison of several samples

Metagenomic data usually consists of very large collections of short anonymous sequences, rendering the comparison of two metagenomes notoriously difficult. Sequence-by-sequence based analysisis to identify all pairwise similarities between two metagenomic data sets represents a computational cost that is prohibitively expensive. Instead of handling complete sequences it has become common practice to compare feature profiles that can represent relevant aspects of the functional and taxonomic composition of metagenomic sequence data.

Comparison of samples recovers the shared background diversity, and gives lights to the microbiome individuality of samples, easing the pathogen identification task. Furthermore, inspection of closely related microbiomes and their associated annotations can be used as a quality control of the dataset and may reveal unexpected flaws of the sampling, sequencing or data processing procedures: (1) Neighboring microbiomes with unexpected habitat labels may indicate contamination of the sample. (2) Related metagenome datasets in the neighborhood can be used as additional data sources for comparative analyses.

K-mer distributions of a set of metagenomic samples give an indication of the presence, abundance and evolutionary relatedness of novel organisms present in the samples. K-mer frequencies are used either for taxonomic binning of individual reads or for computing the overall composition. Depending on the k-mer size used, this method is only suitable for higher level phylogenetic analysis (using small k-mers), or is highly dependent on the training database used (using larger k-mers). Using longer k-mers allows for higher specificity but using k-mers that are unique to specific taxa in the dataset ignores a great deal of information about evolutionary relatedness provided by other k-mers.

Our own experience : in a preliminary study using Kraken's taxonomic results, we could rapidly compare multiple human clinical samples. Further optimizations are needed to increase the significativity of the attested homologies, like (1) the filtering parameters for quality and host sequences, (2) the confidence scoring implemented in Kraken, (3) the k-mer size. An alternative to be explored would be the use of Bloom filters to achieve fast bitwise querying between samples' k-mers, in the manner of the COMMET [3] tool.

Conclusion

A limitation of our pipeline which is universal for similarity based methods, is its reliance on the content and completeness of public reference databases. Public databases hardly represent the real biological diversity, especially pertinent for the viruses that are mostly undiscovered. Additionally their content is biased towards cultivable organisms and human pathogens. Therefore reads from novel microorganisms that are sufficiently divergent from known species will remain unclassified. Unclassified reads need to be identified for follow up investigations: (1) Further assess the sequences assigned to the "unknown category", looking for nucleotide similarity with NR-NT using BLASTn. A large number of reads originated from an untranslated region of the virus genomes, happen to not be captured by the protein reference database. (2) Handle the absence of closely related sequences in the database and mis-assignments (to species that share very low levels of similarity) of these reads resulting in the great number of false positives, for example with phylogenetic approaches.

The classifications and resulting profiles may differ depending on the choice of the database. The RefSeq database is limited, while in GenBank contains several representatives for each bacteria/virus, capturing the high mutation rates of a given species. Informed database selection will improve results, however users should always bear in mind the inherent limitations, biases and potential errors in order to avoid misinterpretations of unlikely findings.

References

[1] Altschul et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

[2] Wood & Salzberg (2014). Kraken : Ultrafast metagenomics sequence classification using exact alignments. *Genome Biology* 15:R46.

[3] Maillet et al. (2014). "COMMET: comparing and combining multiple metagenomic datasets". *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2014.

Mots clefs : HTS analysis, Infectious metagenomics, Taxonomic profiling, k-mer clustering

Analyze your microbiota sequencing data using a Galaxy-based framework

Bérénice Batut [*] [†1], Clémence Defois[1], Kévin Gravouil[1], Cyrielle Gasc[1], Thomas Eymard[1], Éric Peyretaillade[1], Jean-François Brugère[1], Pierre Peyret [‡1]

Poster 65

[1] Conception, Ingénierie et Développement de l'Aliment et du Médicament (CIDAM) – Université d'Auvergne - Clermont-Ferrand I – Facultés de Médecine et de Pharmacie CBRV - RdC et 5e étage BP 38, 28, Place Henri-Dunant, F-63 001 Clermont-Ferrand, France – Tél : +33 4 73 17 79 52.

Nowadays, complex communities of microorganisms can be studied in depth with metagenomics and metatranscriptomics driven by evolution of sequencing techniques. Indeed, these meta'omics techniques offer insight concerning structure and functions of the studied communities. Notwithstanding, raw microbiota sequencing data are difficult to analyze due to their size and the numerous available tools [1,2] for analytical workflow. Indeed, sequences analysis requests successive bioinformatics tasks (e.g. quality control, sequence sorting, taxonomic analysis, functional analysis, statistical analysis). Choosing the best tools with correct parameter values and combining tools together in an analysis chain is a complex and error-prone process. Therefore, there is an urgent need for modular, accessible and sharable user-friendly tools.

Several solutions are proposed in the form of bioinformatic pipelines to exploit metagenomic data. Tools such as *QIIME* [3], *Mothur* [4], *MEGAN* [5], *IMG/M* [6,7] or *CloVR-metagenomics* [8] are useful tools, but they do not provide a complete analytical workflow or are limited to amplicon sequences. Other interesting solutions such as *SmashCommunity* [9], *RAMMCAP* [10], *MetAMOS* [11] propose complete analysis pipelines. However, these command-line tools without proper graphical interface are not user-friendly solutions for researchers with low programming or informatics expertise. For their part, web-based services (*MG-RAST* [12,13], *EBI metagenomics* [14]) analyze microbiota data with automated pipelines working as a black box that also prevents the mastering and monitoring of the data process. Reproducibility can be questioned by this lack of transparency and flexibility. Alternative approaches to improve usability, transparency and modularity can be found in open-source workflow systems (Taverna [15], Wings [16], Galaxy [17,18] or Kepler [19]). They can be used to build a complete modular, accessible, sharable and user-friendly framework to analyze sequence data using chosen tools and parameters.

To overcome these limitations, we have developed ASaiM, an open-source opinionated Galaxy-based framework. It is accessible via a web-based interface to ease microbiota sequence data analyses for biologists. With a custom collection of tools, workflows and databases, the framework is dedicated to microbiota sequence analyses (particularly extraction of taxonomic and metabolic information from raw sequences). By its intrinsic modularity, ASaiM allows adjustment of workflows, tools parameters and used databases for general or specific analyses and provides therefore a powerful framework to easily and rapidly analyze microbiota data in a reproducible and transparent environment.

Implementation

Based on a custom Galaxy instance, ASaiM framework is integrating tools, specifically chosen for metagenomic and metatranscriptomic studies and hierarchically organized to orient user choice

*. Intervenant
†. Corresponding author : berenice.batut@gmail.com
‡. Corresponding author : pierre.peyret@udamail.fr

toward the best tool for a given task. We integrate tools such as *FastQ-Join* [20] for paired-end assembly, *FastQC* [21] and *PRINSEQ* [22] to control quality, *vsearch* [23] to dereplicate, *CD-HIT* [24,25] to cluster sequences, *SortMeRNA* [26] to sort rDNA or rRNA sequences, *NCBI Blast+* [27,28] and *Diamond* [29] for similarity search, *MetaPhlAn2* [30] to assign taxonomy of sequence type, *HUMAnN2* [31] to analyze metabolism, *Group HUMAnN2 to GO slim term* [32] to get a broad overview of metabolism functions, *GraPhlAn*, *KRONA* [33] and *R* scripts to get graphical outputs and statistical tests. All available tools, their versions are described in ASaiM documentation (`http://asaim.readthedocs.org/`). These tools can be used alone or orchestrated inside workflows.

To guide users and then ease exploitation of raw metagenomic or metatranscriptomic sequences, we propose a standard but customizable workflow. This workflow is composed of the following steps (i) quality control, (ii) rRNA/rDNA sequence sorting, (iii) taxonomic assignation, (iv) functional assignation to *UniRef50* gene families, *MetaCyc* pathways and GO slim term, and (v) linking of taxonomic and functional analyses. Graphical outputs are generated at the different stages of the workflow (with custom scripts or using dedicated tools). All choices (tools, version, parameters, execution order) are described in ASaiM documentation (`http://asaim.readthedocs.org/`). Other workflows are also available for comparative analyses of taxonomic and/or functional results obtained with the previously described worklow.

ASaiM framework source code is available under Apache 2 license at `https://github.com/asaim/framework`. The code source includes customization of the Galaxy instance: automatic configuration, launch and provision with chosen tools, workflows and needed databases.

Validation

ASaiM framework was tested on two mock metagenomic datasets. These datasets are metagenomic shotgun sequences (> 1,200,000 single-end sequences from a 454 GS FLX Titanium) from a controlled microbiota community (with 22 known microbial species), available on EBI metagenomic database (`https://www.ebi.ac.uk/metagenomics/projects/SRP004311`). Taxonomic and functional results from ASaiM workflow were compared to the ones obtained with EBI metagenomics pipeline [14] (version 1.0). Details about these analyses (scripts, results, extended report, input data) are available on a dedicated GitHub repository (`https://github.com/ASaiM/hmp_mock_tests`).

For these tests, an ASaiM customed Galaxy instance was deployed on a Debian GNU/Linux System with 8 cores Intel(R) Xeon(R) at 2.40 GHz and 32 Go of RAM. On both datasets, execution of whole main workflow (from quality control to taxonomic and functional analyses integration) lasts less than 5h30, uses maximum of 1.52 Go of RAM.

In both EBI metagenomic pipeline and ASaiM main workflow, raw sequences are pre-processed with first a quality control to remove low quality, small and duplicated sequences and then a rRNA/rDNA sorting. After quality control, more sequences are conserved (> 96 %) with ASaiM workflow than with EBI metagenomics pipeline (< 87 %), despite identical input sequences. The different tools (*PRINSEQ* [22] and *vsearch* [23] for ASaiM workflow, *Trimmomatic* [34], *UCLUST* [35] and *RepeatMasker* [36] for EBI metagenomics pipeline) and parameters in EBI and ASaiM workflows may explain the observed differences. The quality-controlled sequences are afterwards sorted in rRNA or non rRNA sequences (with *SortMeRNA* [26] and *rRNASelector* [37] for respectively ASaiM and EBI metagenomics pipelines). With both pipelines, the proportion of rRNA sequences is low (< 1.5 %) as expected for any metagenomic datasets.

These rRNA sequences (< 9,500 sequences) are used in EBI metagenomics pipeline for taxonomic assignation with *QIIME* [3]. ASaiM workflow uses *MetaPhlAn2* [30] on all quality controlled sequences (> 1,100,000 sequences), leading to more accurate and statically supported taxonomic assignations, which are also more precise (until species levels, against family levels with EBI metagenomics pipeline). Contrary to taxonomic results from EBI metagenomics pipelines, all

genera found with ASaiM workflow correspond to expected genera. Despite previous differences, relative abundances of found all families are similar for both pipelines and are correlated with expected abundances of these mock datasets.

For functional analyses, ASaiM workflow uses *HUMAnN2* [31] to extract relative abundance of gene families (*UniRef50*) and metabolic pathways (*MetaCyc*). In EBI metagenomics pipeline, functional analyses are only based on *InterPro* similarity search. In both pipelines, gene families or proteins are afterwards grouped into slim Gene Ontology term (cellular components, biological processes and molecular functions), with relative abundance. In ASaiM, this step based on UniRef50 gene families is performed by specially developed tool [32]. The variability of relative abundances of GO slim terms between both pipelines explains at most 35 % of overall variability of relative abundances of GO slim terms. Therefore, GO slim term differences between both pipelines exist, but they are negligible.

In addition, in *HUMAnN2* results, abundances of gene families and pathways are stratified at community level. The functional information (abundance of given pathway or gene families for a given taxon) can be related to global taxonomic information and, particularly, species abundances, in ASaiM workflow. This type of information is not available with EBI metagenomic pipeline. A strong correlation is observed between gene family or pathway mean abundances and related species abundances inside both datasets.

Conclusion

With an easy-to-install custom Galaxy instance, ASaiM provides a simple and intuitive pre-configured environment to study microbiota. The build workflow with chosen tools, parameters and databases allows a rapid (few hours on a standard computer), easy, reproducible, complete and accurate analysis of a metagenomic or metatranscriptomic dataset from raw sequences to taxonomic and functional analysis. This workflow as well as workflows for comparative analyses are proposed, but in each case their tools with their parameters and execution order are customizable by the user. To help in tool, parameter and workflow choices, a documentation is available at `http://asaim.readthedocs.org/`.

ASaiM framework is then a powerful framework to analyze shotgun raw sequence data from complex communities of microorganisms. This open-source and biologist-oriented solution enhances usability, reproducibility and transparency of such studies.

References

1. M.L.Z. Mendoza et al, *Brief Bioinform* 16 (2015) 745–758.
2. N. Segata et al, *Molecular Systems Biology* 9 (2013) 666.
3. J.G. Caporaso et al, *Nature Methods* 7 (2010) 335–336.
4. P.D. Schloss et al, *Appl. Environ. Microbiol.* 75 (2009) 7537–7541.
5. D.H. Huson et al, *Genome Res.* 17 (2007) 377–386.
6. V.M. Markowitz et al, *Nucl. Acids Res.* 42 (2014) D568–D573.
7. V.M. Markowitz et al, *Nucl. Acids Res.* 36 (2008) D534–D538.
8. S.V. Angiuoli et al, *BMC Bioinformatics* 12 (2011) 356.
9. M. Arumugam et al, *Bioinformatics* 26 (2010) 2977–2978.
10. W. Li, *BMC Bioinformatics* 10 (2009) 359.
11. T.J. Treangen et al, *Genome Biol.* 14 (2013) R2.
12. R.K. Aziz et al, *BMC Genomics* 9 (2008) 75.

13. F. Meyer et al, *BMC Bioinformatics* 9 (2008) 386.

14. S. Hunter et al, *Nucl. Acids Res.* 42 (2014) D600–D606.

15. K. Wolstencroft et al, *Nucl. Acids Res.* 41 (2013) W557–W561.

16. Y. Gil et al, *IEEE Intelligent Systems* 26 (2011) 62–72.

17. D. Blankenberg et al, *Curr Protoc Mol Biol* 0 19 (2010) Unit–19.1021.

18. J. Goecks et al, *Genome Biology* 11 (2010) R86.

19. B. Ludäscher et al, *Concurrency Computat.: Pract. Exper.* 18 (2006) 1039–1065.

20. E. Aronesty, Ea-Utils: "Command-Line Tools for Processing Biological Sequencing Data," (2011).

21. S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data (2010).

22. R. Schmieder, R. Edwards, *Bioinformatics* 27 (2011) 863–864.

23. T. Rognes et al, Vsearch: VSEARCH 1.4.0 (2015).

24. L. Fu et al, *Bioinformatics* 28 (2012) 3150–3152.

25. W. Li, A. Godzik, *Bioinformatics* 22 (2006) 1658–1659.

26. E. Kopylova et al, *Bioinformatics* 28 (2012) 3211–3217.

27. P.J. Cock et al, *GigaScience* 4 (2015) 39.

28. C. Camacho et al, *BMC Bioinformatics* 10 (2009) 421.

29. B. Buchfink et al, *Nat Meth* 12 (2015) 59–60.

30. D.T. Truong et al, *Nat Meth* 12 (2015) 902–903.

31. S. Abubucker et al, *PLoS Comput Biol* 8 (2012) e1002358.

32. B. Batut, Group abundances of UniRef50 gene families obtained with HUMAnN2 to Gene Ontology (GO) slim terms with relative abundances: release v1.2.0 (2016).

33. B.D. Ondov et al, *BMC Bioinformatics* 12 (2011) 385.

34. A.M. Bolger et al, *Bioinformatics* 30 (2014) 2114–2120.

35. R.C. Edgar, Bioinformatics 26 (2010) 2460–2461.

36. A. Smit et al, RepeatMasker Open-3.0, 1996.

37. J.-H. Lee et al, *J Microbiol.* 49 (2011) 689–691.

Mots clefs : metagenomics, metatranscriptomics, Galaxy, intergrated analysis

SkIf (Specific k-mers Identification) : un outil d'identification rapide de gènes ou de régulateurs de gènes d'intérêt

Poster 66

Martial Briand [*1], Romain Gaborieau[1], Marie-Agnès Jacques[1], Tristan Bourreau[1], Sylvain Gaillard[*1], Nicolas Chen[1]

[1] Institut de recherche en Horticulture et Semences (IRHS) – Institut national de la recherche agronomique (INRA) : UMR1345, Agrocampus Ouest, Université d'Angers – IRHS 42 rue Georges Morel, F-49 071 Beaucouzé Cedex, France

La génomique comparative est devenue un outil de base pour les biologistes afin d'identifier des gènes cibles d'intérêt fondamental, biotechnologique ou biomédical. De nombreuses approches existent pour effectuer ces comparaisons : la recherche et la visualisation de synténies, la recherche d'orthologues, l'analyse de SNPs… Cependant, ces approches ne permettent pas toujours d'identifier rapidement et simplement les gènes ou régulateurs de gènes potentiellement responsables d'un caractère d'intérêt. Dans le but d'optimiser l'identification rapide de gènes ou régulateurs d'intérêt, nous avons développé le logiciel SkIf, basé sur la recherche de séquences spécifiquement présentes ou absentes chez les organismes vivants présentant un caractère d'intérêt.

SkIf peut prendre en entrée des groupes de séquences nucléiques ou protéiques (e.g. génomes, transcriptomes, ou protéomes). Dans un premier temps, SkIf construit une matrice contenant la totalité des oligomères de taille k (k-mers) présents dans chacune des séquences d'entrée. Cette matrice est ensuite utilisée pour identifier les k-mers spécifiquement présents ou absents dans un groupe de séquences d'intérêt défini par l'utilisateur. Les k-mers chevauchants sont ensuite concaténés afin de permettre l'identification rapide de séquences plus larges que k (long-mers). SkIf fournit en sortie les séquences des k-mers et long-mers spécifiquement trouvés ainsi que leurs coordonnées sur une séquence de référence définie par l'utilisateur.

Les résultats obtenus comprennent à la fois des SNPs et des ilots génomiques plus ou moins larges. SkIf est donc particulièrement efficace pour l'étude d'organismes polyphylétiques pour un caractère donné, car il permet de détecter à la fois les signaux d'évolution convergente et les transferts horizontaux. Les gènes ou régulateurs de gènes ainsi trouvés seront des candidats privilégiés pour l'étude du caractère d'intérêt. Ces régions spécifiques pourront également être utilisées comme cibles pour élaborer des tests d'identification ou de détection spécifiques des organismes qui présentent ce caractère.

SkIf a été développé en C++, la lecture des séquences se fait à l'aide de la libraire bpp-seq de Bio++. L'outil s'utilise en ligne de commande ou à travers un environnement Galaxy.

Mots clefs : Génomique comparative, kmer, Identification gènes d'intérêt

*. Intervenant

Study of microbial diversity and plant cell wall-degrading enzymes during flax dew-retting by using targeted-metagenomics

Poster 67

Christophe Djemiel [*] [†] [1], Sébastien Grec [‡] [1], Simon Hawkins [§] [1]

[1] Unité de Glycobiologie Structurale et Fonctionnelle (UGSF) – CNRS : UMR8576, Université Lille 1, Sciences et Technologies - Lille 1 – Université de Lille - Sciences et Technologies, Bâtiment C9, F-59 655 Villeneuve d'Ascq Cedex, France

The first step in the industrial transformation of flax stems into industrial fibers used in textiles or composite materials is achieved by dew-retting on the soil surface. At harvest, the flax stems are 'pulled' (up-rooted) and laid down directly on the soil in swaths (strips). Subsequent exposure to alternating periods of rain and heat combined with the development and the action of soil microflora favor the separation of cellulosic bast fibers from the stems via the hydrolysis of cell wall polymers (hemicelluloses, pectins). Despite many studies of this process (Akin, 2013; Henriksson, & Akin, 1997; Martin *et al.,* 2013) relatively little is known about i) the composition of the microflora population, ii) the kinetics of microbial colonization of plant material and iii) the evolution of the microbial population in the soil. As an example of biological processing involving microbial activities, dew retting represents a convenient model to study the microbial dynamics of plant colonization and degradation of cell wall components.

In order to study this biological process, we undertook a targeted-metagenomics approach. A metagenomic pipeline including sample preparation, marker amplification, raw sequence and data analyses was built. We evaluated and compared different protocols and tools such as mothur (Schloss *et al.,* 2009), qiime (Caporaso *et al.,* 2010), PIPITS (Gweon *et al.,* 2015) and clustering tools such as swarm (Mahé *et al.,* 2014). The resulting pipeline analyses and homemade tools used to redesign primers so as to specifically amplify bacterial rRNA genes from complex samples including plants are detailed. Analyses are completed by functional predictions of enzymatic activity using an adaptation of PiCRUSt (Langille *et al.,* 2013) software for the CAZy database (Lombard *et al.,* 2014).

As a conclusion and proof of pipeline efficiency, we report the first exhaustive microbial inventory of dew-retting and its evolution during the process obtained by using a targeted-metagenomics approach. Potential enzymatic functions related to cell wall degradation based on functional prediction using the bioinformatic software PICRUSt are listed and used to establish a potential chronology of cell wall polymer degradation.

This work is funded within the framework of the collaborative French 'Future project' SINFONI. Christophe Djemiel thanks the 'Haut de France region', and OSÉO for their financial support. The authors thank the Genomic platform of Genopole Toulouse Midi Pyrénées where sequencing was performed.

References

Akin, D. E. (2013). Linen most useful: Perspectives on structure, chemistry, and enzymes for retting flax. ISRN Biotechnology, 2013, 23. http://doi.org/http://dx.doi.org/10.5402/

*. Intervenant
†. Corresponding author : christophe.djemiel@gmail.com
‡. Corresponding author : sebastien.grec@univ-lille1.fr
§. Corresponding author : simon.hawkins@univ-lille1.fr

2013/186534.

Henriksson, G., & Akin, D. (1997). Identification and retting efficiencies of fungi isolated from dew-retted flax in the United States and europe. Applied and Environmental Microbiology. Retrieved from http://aem.asm.org/content/63/10/3950.short.

Martin, N., Mouret, N., Davies, P., & Baley, C. (2013). Influence of the degree of retting of flax fibers on the tensile properties of single fibers and short fiber/polypropylene composites. Industrial Crops and Products, 49, 755–767. http://doi.org/10.1016/j.indcrop.2013.06.012.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology, 75(23), 7537–41. http://doi.org/10.1128/AEM.01541-09.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature Publishing Group, 7(5), 335–336. http://doi.org/10.1038/nmeth0510-335.

Gweon, H. S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D. S., … Schonrogge, K. (2015). PIPITS: An automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. Methods in Ecology and Evolution, 6(8), 973–980. http://doi.org/10.1111/2041-210X.12399.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. PeerJ, 2, e593. http://doi.org/10.7717/peerj.593.

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., … Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol, 31(9), 814–821. http://doi.org/10.1038/nbt.2676.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research, 42(Database issue), D490–5. http://doi.org/10.1093/nar/gkt1178.

Mots clefs : Targeted, metagenomics, NGS, Pipelines, Flax, Dew, retting, Soil

Environmental metatranscriptomics of marine eukaryote plankton

Poster 68

Marion Dupouy [*1], Éric Pelletier [†1], Quentin Carradec[1], Adriana Alberti[1], Olivier Jaillon[1], Corinne Da Silva[1], Amos Kirilovsky[1], Karine Labadie[2], Bonora Mario Neou[1], Marc Wessner[1], Patrick Wincker[1,3]

[1] Genoscope - Centre national de séquençage – CEA, Genoscope – 2 rue Gaston Crémieux, CP5706, F-91 057 Évry Cedex, France
[2] Commissariat à l'Énergie Atomique (CEA) – France
[3] Génomique métabolique (UMR 8030) – CEA, CNRS : UMR8030, Université d'Évry-Val d'Essonne – Genoscope, 2 rue Gaston Crémieux, F-91 057 Évry Cedex, France

150 worldwide open ocean stations were sampled during the Tara Oceans expedition that covered during 3 years most of the open ocean regions across the world. Physicochemical paramaters as well as imaging data complete the collection. For each station, sea water for two depths (sub-surface and Deep Chlorophyll Maximum) was filtered to aggregate five ranges of organisms sizes fro 0 to 2000 µm. DNA and RNA were extracted and sequenced using Illumina HiSeq methodology. mRNA reads were cleaned up and assembled with Velvet and Oases resulting in more than 116 millions sequence clusters (called "unigenes") representing the largest collection of marine eukaryotic genes. This collection was then annotated with taxonomic (from Uniref and the MMETSP) and functional (from PFAM) information. Individual unigenes occurrences and expression in each individual filter were then computed. Analysis of those data allowed us characterize the taxonomic composition of the Tara Oceans eukaryote gene catalog and to unveil the main genetic function expressed by eukaryotic plankton across their distribution in the world oceans. Novel bioinformatics approaches developed in the context of this project, concerning both gene catalog reconstruction and massive global analysis will be presented and discussed.

Mots clefs : Métagénomique, Métatranscriptomique, plancton, Tara Oceans

*. Intervenant
†. Corresponding author : eric.pelletier@genoscope.cns.fr

Picome : un workflow de production d'inventaire de communautés en metabarcoding : méthodes exactes et utilisation du calcul HTC

Poster 69

Jean-Marc Frigerio[1,2], Agnès Bouchez[3], Philippe Chaumeil[1,2], Maria Kahlert[4], Frédéric Rimet[3], Franck Salin[1,2], Sylvie Thérond[5], Alain Franc [*] [†1,2]

[1] Biodiversité, Gènes & Communautés (BioGeCo) – Université de Bordeaux, Institut national de la recherche agronomique (INRA) : UMR1202 – Site de recherche Forêt, Bois de Pierroton, 69, route d'Arcachon, F-33 612 Cestas Cedex, France
[2] Équipe Pleiades – INRIA – Centre de recherche INRIA Bordeaux - Sud-Ouest, France
[3] UMR CARRTEL – Institut National de la Recherche Agronomique - INRA (FRANCE) – 75 avenue de Corzent, F-74 203 Thonon-les-Bains, France
[4] Swedish University of Agricultural Sciences (SLU) – S-750 07 Uppsala, Suède
[5] Institut du développement et des ressources en informatique scientifique (IDRIS) – CNRS : UPS851 – Bâtiment 506 BP 167, F-91 403 Orsay Cedex, France

Enjeu

Depuis plusieurs années, la qualité des eaux douces en rivières et lacs est estimée en routine par des bioindicateurs. La théorie écologique, et l'observation, nous enseignent qu'il est possible d'estimer la composition des communautés connaissant les habitats (notion de profil écologiques). La bioindication revient à inverser la variable estimée et le conditionnement, comme une sorte d'application étendue du théorème de Bayes : inférer les habitats connaissant les communautés. Les diatomées sont des microalgues utilisées en routine pour estimer la qualité des eaux douces. Un indice de polluo-sensibilité est attaché à chaque espèce, et la qualité des eaux est estimée par l'indice moyen d'une communauté, modéré par les abondance. La partie « fastidieuse » du travail est de construire un inventaire, actuellement sur une base naturaliste. Notre travail consiste à proposer la construction d'un inventaire sur une base moléculaire, connaissant un jeu de donnée de reads, d'un même marqueur (approche de type « amplicons »).

Choix du modèle biologique et des méthodes

Le choix des diatomées comme modèle biologique repose sur deux critères : (i) ces organismes sont utilisés en routine en bioindication, au niveau européen, donc un résultat en ce domaine est utile aux services environnementaux qui réalisent les inventaires et (ii) ce sont des micro-organismes dont la systématique et la taxonomie peuvent être étudiés tant par une approche moléculaire (barcoding) que naturaliste (observation des frustules sous microscope). Ainsi, ces organismes permettent une comparaison entre d'une part une construction d'OTUs moléculaires supervisés par une classification sur base phénotypique, et d'autres parts une construction d'OTUs sur le même jeu de données par classifications non supervisées. Le marqueur choisi pour cette étude est rbcL, issu du chloroplaste, connu comme étant résolutif au niveau de l'espèce et universel chez les diatomées. La taille des jeux de reads ne permettent (pas encore…) de réaliser des phylogénies par maximum de vraisemblance ou (encore moins) par approches bayésiennes. Aussi, nous nous sommes orientés vers des méthodes à partir de distances. La distance entre deux reads ou deux

*. Intervenant
†. Corresponding author : Alain.Franc@pierroton.inra.fr

séquences est calculée exactement à partir du score de l'algorithme de Smith-Waterman. C'est en fait un dissimilarité, rigoureusement. Nous avons ensuite procédé en deux temps : une étude de la base de référence, et un mapping des reads sur cette même base.

Étude de la base de référence

Nous avons d'abord vérifié que le marqueur choisi est bien résolutif, et qu'il y a accord entre les approches moléculaires et optiques sur la base de référence (issue d'une collection de diatomées maintenue à l'INRA de Thonon et d'une curation en continu de bases de références publiques). Chaque souche de diatomée de la collection est identifiée par voie optique et séquencée pour le marqueur considéré. Nous avons calculé toutes les distances entre deux séquences de la base, et construit une image de la base dans un espace euclidien par isométrie, ou autant que faire se peut (multidimensional scaling, version « metric »). Puis, nous avons utilisé des outils de classification supervisée du « machine learning », comme les SVM, ou non (communautés sur graphes, spectral clustering, clustering). Nous avons comparé les classifications supervisés et non supervisées (qui mènent à des OTU). Chacune de ces méthodes a ses avantages et limites. C'est la convergence (ou non) des résultats qui est informative. Nous en déduisons que le marqueur est fiable au niveau du genre, et souvent mais pas toujours au niveau de l'espèce. Il s'agit là d'un travail de modélisation statistique.

Mapping de chaque read sur la base

Nous avons ensuite comparé chaque read d'un jeu de données NGS à l'ensemble de la base, c'est à dire réalisé 10**5 x (2.10**3) comparaisons. Nous avons pour cela construit un programme C qui calcule ces valeurs de distances, que nous avons ensuite parallélisé avec MPI pour distribuer le calcul par paquets de reads versus l'ensemble de la base. Nous affectons un read a un taxon de la base de référence par une technique du type k-NN (nearest neighbors). On choisit un seuil (ici 9, car il représente 3 % de la longueur des reads en moyenne) et affectons un read à un taxon si toutes les séquences de la base à une distance inférieure au seuil sont du même taxon. Nous avons séquencé 171 échantillons de rivières de Mayotte, par multiplexage, sur un Ion Torrent de la plateforme Génome-Transcriptome de Bordeaux. Le nombre de librairies par run a été ajusté par une estimation de la profondeur nécessaire. Nous avons réalisé ainsi 171 inventaires de communauté. L'ensemble des calculs pour comparaison de chaque read avec la base a été réalisé sur 512 cœurs d'un Blue Gene Q d'IBM, à l'IDRIS (Turing), qui est une machine hyperparallèle. Nous avons utilisé la programmation MPI comme du map-reduce. Le même travail est en cours en distribuant le travail sur 2**14 = 16384 CPU. Nous nous attendons à ce que le temps d'exécution soit effectivement divisé par 120 (plus précisément 2**7=128), car l'équilibre des charges d'un tel programme est de très bonne qualité. Sur 512 cœurs, le temps d'exécution été de 236 heures wall time (11 lots de 20h, un lot de 10h et un lot de 6 heures). Cela représente environ 1h20 wall time par échantillon, soit environ 700 heures (un mois) si le calcul était effectué séquentiellement sur un PC. Ces même échantillons ont été inventoriés par voie optique. La comparaison des deux approches (moléculaire et optique) est en cours, et s'avère prometteuse, étant donné le nombre des échantillons. Une seconde expérience de même nature est en cours avec plus de cent échantillons de communautés de diatomées de rivières suédoises, dans d'autres conditions environnementales (conditions boréales).

Conclusions et perspectives

:

Nous avons montré la faisabilité d'inventaires en routine de communautés de diatomées par voie moléculaire (amplicons), en construisant pour cela un « workflow » qui permet d'estimer la qualité de la base de référence et d'assigner chaque read d'une communauté, en un temps court via un

accès à une machine de type HTC. Cependant, et c'est un intérêt de cette étude, bien des questions méthodologiques restent à étudier. Nous en discuterons quelques unes :

(i) Un premier verrou non négligeable est la non congruence entre les systématiques moléculaires, à base de phylogénies, et naturalistes, à base phénotypique, pour un nombre significatif de taxons. Si tel reste le cas, il ne sera pas possible de réconcilier des inventaires à base naturaliste et moléculaire. La travail de systématique doit donc se poursuivre, pour une convergence entre les deux « référentiels », comme cela est le cas sur les angiospermes (APG).

(ii) Un second verrou est que, pour être reconnu, un taxon doit être présent dans la base. Ainsi, un critère essentiel est la couverture taxonomique de la base. Il est irréaliste d'espérer qu'à terme l'ensemble des espèces y soit représenté (il existe plus de 100 000 espèces de diatomées). Ce verrou peut être contourné par deux voies en parallèles : concevoir un plan d'échantillonnage de la diversité des diatomées d'intérêt en bioindication ou pour la systématique pour compléter la base, et mettre en œuvre des outils comme Pplacer qui permet de placer un read en position optimale (profondeur du nœud) dans une phylogénie, donc effectuer une phylogénie moléculaire de la base de référence, et placer chaque read de façon optimale dans cette phylogénie (le temps est alors linéaire avce le nombre de reads).

(iii) L'assignation taxonomique par une méthode du type k-NN est de bonne qualité (cette approche est très étudiée en machine learning), mais probablement non optimale. Des travaux sont en cours pour améliorer la qualité de cette étape. Par exemple, on peut observer que la phase intensive du calcul est le calcul des distances de toutes les paires reads x références. Nous travaillons sur l'idée suivante : il est possible de placer un read (une query) au sein d'un nuage de points dans un espace euclidien de dimension d en connaissant les distances à kd points considérés (rigoureusement, d+1) comme des ancres (une triangulation), et choisis de façon optimale connaissant la forme du nuage de points associé aux références. Il est possible également de définir une fois pour toutes un « champ de probabilité » de tel ou tel taxon en toute position au sein du nuage de points associé aux références. La position d'une « query » étant connue, il est alors possible de lui d'assigner, ou non, un taxon de la base de référence.

Remerciement

Nous remercions pour la réalisation de ce projet l'ONEMA, qui a financé cette étude, le labex CEBA (Centre d'Études de la Biodiversité Amazonienne) qui a également soutenu le développement des outils, le projet e-Biothon qui a permis un premier déploiement des calculs sur une machine hyperparallèle, et le projet DARI i2015037360 qui a permis le déploiement des calculs en production sur Turing.

Publication de l'équipe sur ce projet

Kermarrec, L.; Franc, A.; Rimet, F., Chaumeil, Ph., Humbert J.-F. & Bouchez, A. – 2013 - Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, 13(4):607-619

Kermarrec L., Franc A., Rimet F., Chaumeil P., Frigerio J.M., Humbert J.F., Bouchez A., - 2014 - A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, 33:349-364.

Duarte A. M. S. & al. (+ 11 auteurs, ordre alphabétique, dont A. Franc) - 2015 - Future opportunities and trends for e-infrastructures and life sciences: going beyond the grid to enable life science data analysis – *Frontiers in genetics*, 6: `http://dx.doi.org/10.3389/fgene.2015.00197`

Daydé, M., Depardon, B., Franc, A., Gibrat, J.-F., Guillier, R., Karami, Y., Sutter, F., Taddese B. & Thérond, S. – 2015 – E-Biothon : an experimental plaform for BioInformatics – *IEEE conference CSIT'15 (10th International Conference on Computer Science and information technologies)*.

Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A. & Bouchez, A. - 2016 - R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. *Database*, 1–21, doi: 10.1093/database/baw016

Mots clefs : metabarcoding, bioindication, inventaires de communautés, amplicons, OTU, écologie microbienne

Functional metagenomics of microbial communities inhabiting deep and shallow serpentinite-hosted ecosystems

Poster 70

Éléonore Frouin *[1], Fabrice Armougom[1], Gaël Erauso[1]

[1] Institut méditerranéen d'océanologie (MIO) – CNRS : UMR7294, Université du Sud Toulon - Var, Institut de recherche pour le développement [IRD] : UMR235, Aix Marseille Université – Campus de Luminy, Case 901, F-13 288 Marseille Cedex 09, France

Introduction

During the last decade, serpentinite-hosted hydrothermal ecosystems have generated considerable interest since these extreme environments may contribute to a better understanding of the origins of life [1]. The ultramafic rocks, a major component of the earth's mantle, can undergo a widespread exothermic and abiotic geochemical process known as serpentinization. The serpentinization involves the hydration and transformation of the ultramafic rocks by percolating seawater that result in the liberation of high amount of hydrogen gas, methane and small organic molecules (acids and amino-acids) carried out by the end-member alkaline fluids. The alkaline and anoxic fluids emanating from these hydrothermal systems precipitate upon mixing with cold seawater, resulting in the formation of impressive submarine chimneys. For a long time, the only one known example of such very peculiar marine hydrothermal system was the Lost-City Hydrothermal Field (LCHF) discovered in 2000, near axis of the Mid-Atlantic Ridge at around 800 m depth and which since have been object of many multi-disciplinary studies. More recently, our group has initiated such studies on a shallow analog of LCHF, re-discovered in the Bay of Prony in New-Caledonia and that we thus called the Prony Hydrothermal Field (PHF). There, like in LCHF, tall (up to 38 m) carbonates chimneys discharge alkaline (pH > 11) fluids at moderate temperatures (< 40°C) at maximal 50 m depth. Contrary to LCHF, the fluids in PHF are low-salted as they originated from terrestrial run-off waters but like in LCHF they are enriched in dihydrogen, methane and small organic molecules produced abiotically that might constitute potential sources of energy for microbial growth in chemolithoautotrophy.

Indeed, abundant and complex microbial communities (microbiome) were found living inside the chimney walls, fueled by contact with the alkaline and anoxic fluids. Recent 16S rRNA surveys of the LCHF and PHF serpentinite-hosted microbiomes revealed a huge bacterial diversity, consisted mostly of Firmicutes, Chloroflexi, and Proteobacteria [2], but a low archaeal one with the predominance of the Methanosarcinales order represented by only two major phylotypes that seemed to be specific to these serpentinite-hosted ecosystems [3]. Although specific hydrogen utilization was reported by Brazelton and colleagues [4] at the LCHF, the metabolic capabilities of such serpentinite-hosted microbiomes remain largely unknown. In this way, metagenomic investigations could reveal the potential core metabolic pathways as well as microbiome-specific functionality. We therefore reported in this study an overview of the functional capability of the microbiome inhabiting the LCHF and PHF hydrothermal systems, through a comparative metagenomics approach.

*. Intervenant

Materials and methods

Four samples of active chimneys were collected for comparative metagenomic analyses: two from the LCHF and two other from the PHF. Total DNA of each microbiome was extracted and a whole genome shotgun sequencing was performed using Illumina technology with a 150 pb paired-end MiSeq strategy. The raw metagenomic data were trimmed using the Trimommatic software v0.32. The trimmed paired-end reads were merged with PandaSeq v2.8 and finally assembled into contigs with IDBA-UD version 1.1.0. The open reading frames (ORFs) were predicted by the Prodigal software v47. The functional annotation of the predicted ORFs was performed using RPS-Blast search with the Clusters of Othologous Group (COG) database and using Blastp search with the Kyoto Encyclopedia of Gene and Genome (KEGG) database, with an E-value threshold of 10^{-6}. The number of hits to the functional categories from COG and KEGG were then normalized. A statistical procedure was used to highlight the major differences between functional profiles of the four chimney microbiomes. Reconstructions of metabolic pathways were carried out using the KEGG annotations. The predicted functions were grouped into pathways with MinPath software v1.2, using a parsimony approach to find a minimum number of pathways that can explain all identified functions. The pathway graphs were produced with a slightly modified R package: "Pathview".

Results

On average, 11.5 million reads were assembled from each sample across 44,629 contigs. Approximately 61 % and 42 % of the total predicted genes showed matches against COG and KEGG databases, respectively. The functional profiles were quite similar in both environments. The functions identified in all microbiomes allowed us to define a "functional core" which could be representative of marine serpentinized ecosystems.

However, some rare differences are observed between deep and shallow environment. Indeed, the comparison of the LCHF and PHF functional microbiome profiles showed that the PHF metagenomes were enriched in functional COG categories associated to Energy production and conversion, as well as to Carbohydrate transport and metabolism while the LCHF metagenomes were enriched in the COG category corresponding to Inorganic ion transport and metabolism. In this COG category, more genes associated with outer membrane receptor proteins, mostly for Fe transport (COG1629), were found in LCHF metagenomes. In the 'cellular processes' category, a larger proportion of predicted genes of LCHF metagenomes were assigned to the Cell motility and signal transduction mechanisms sub-category. In particular, we noticed that the genes associated with chemotaxis are more abundant in LCHF. It has been reported by Lertsethtakarn and colleagues [5] that motility and chemotaxis-related functions could assist bacteria in their responds to dynamic environments. Over-representation of predicted genes in this peculiar category in LCHF metagenome may reflect more challenging environmental conditions in LCHF vs PHF regarding the energy and carbon sources supply as PHF microbial ecosystem also benefits from the photosynthetic primary production input in the photic zone, which is not the case of LCHF ecosystem which depends almost exclusively on chemosynthesis based on abiotic products from serpentinization reactions.

Because chemosynthesis is the driving force characterizing these ecosystems, we also attempted identify and to reconstruct the main pathways of carbon dioxide assimilation in prokaryotic autotrophs and thus decipher which ones were preferentially used. The key enzymes of the Calvin Besson Bassham (CBB) cycle and the reverse TriCarboxylic Acid (rTCA) cycle were identified in our four metagenomes, whereas none of the enzymes involved in Wood-Ljungdahl pathway and in the 3-hydroxypropionate cycle were found in our data. Methane, dihydrogen, sulfur and nitrogen pathways were also reconstructed to identify which oxydo-reductions reactions fuel with energy the microbial communities to perform carbon fixation. These analyses highlighted two anaerobic respiration processes: the dissimilatory sulfate reduction and the denitrification. All of

the enzymes involved in denitrification pathway were found in the four metagenomes, indicating that denitrification pathway was utilized by microbial communities inhabiting the two studied sites, contrary to the results previously found in LCHF metagenome by Xie et al [6].

Conclusion

These results advance our understanding of the functioning of microbial communities in a deep (LCHF) and shallow (PHF) serpentinite-hosted ecosystems. The comparative metagenomic analysis allowed us to identify a core of functional capabilities, shared by the microbiomes of these two environments but also to rise new hypothesis on the environmental factors that may shape this peculiar microbial ecosystem. This pioneering study will serve as a basis to link potential metabolisms to the microbial taxonomic groups identified in serpentinite-hosted hydrothermal field and to propose new models of functioning of these extreme microbial ecosystems.

References

[1] W Martin et al. *Nature Reviews* (2008) 6:805-814

[2] W Brazelton et al. *Applied and Environmental Microbiology* (2006) 72:6257–6270

[3] M Quéméneur et al. *Environmental Microbiology Reports* (2014) 6:665-74

[4] W Brazelton et al. *Frontiers in Microbiology* (2012) 2:268

[5] P Lertsethtakarn et al. *Annu Rev Microbiol* 65:389-410

[6] W Xie et al. *The ISME Journal* (2011) 5:414-4426

Mots clefs : metagenomics, hydrothermal, serpentinization, functional profiles

Stratégies de reconstruction de génomes microbiens à partir de métagenomes

Poster 71

Kévin Gravouil [*] [†1,2,3], Corentin Hochart[2], Bérénice Batut[1], Clémence Defois[1], Cyrielle Gasc[1], Pierre Peyret[1], Didier Debroas [‡2], Marie Pailloux [§3], Éric Peyretaillade [¶1]

[1] Conception, Ingénierie et Développement de l'Aliment et du Médicament (CIDAM) – Université d'Auvergne - Clermont-Ferrand I – Facultés de Médecine et de Pharmacie CBRV, RdC et 5e étage BP 38, 28 place Henri-Dunant, F-63 001 Clermont-Ferrand, France – Tél : +33 4 73 17 79 52

[2] Microorganismes : génome et environnement (LMGE) – Université d'Auvergne - Clermont-Ferrand I, CNRS : UMR6023, Université Blaise Pascal - Clermont-Ferrand II – Université Blaise Pascal, Campus des Cézeaux, 24 avenue des Landais, BP 80026, F-63 170 Aubière, France

[3] Laboratoire d'Informatique, de Modélisation et d'optimisation des Systèmes (LIMOS) – Institut Français de Mécanique Avancée, Université Blaise Pascal - Clermont-Ferrand II, Université d'Auvergne - Clermont-Ferrand I, CNRS : UMR6158 – Bâtiment ISIMA, Campus des Cézeaux, BP 10025, F-63 173 Aubière Cedex, France

Les limites de la reconstruction des génomes complets

La métagénomique permet aujourd'hui d'étudier les micro-organismes non cultivables et ainsi d'appréhender le fonctionnement biologique de tout type d'écosystème. Néanmoins, le séquençage massif (ciblé ou non) ne permet pas toujours un inventaire exhaustif des micro-organismes et encore moins de relier la structure de la communauté et les fonctions biologiques qu'elle assure. Pour pallier à cela, il est nécessaire de reconstruire les génomes complets des populations présentes dans les environnements étudiés. En revanche, la faible taille des séquences générées par le séquençage haut-débit (< 500 pb) et l'incroyable diversité pouvant être retrouvée (1 g de sol peut contenir 109 cellules bactériennes représentant 106 espèces) rendent l'assemblage des séquences d'autant plus compliqué. De plus, les approches développées pour l'étude de génomes isolés, maintenant bien rodées s'avèrent inadaptées pour la caractérisation d'environnements complexes. Les algorithmes de traitement des données doivent donc être repensés pour pouvoir prendre en compte les caractéristiques propres à cette approche ultra haut-débit : masse considérable de données, multitudes d'organismes aux abondances inégales, profondeur de séquençage souvent insuffisante, organismes encore inconnus donc sans génomes de référence…

Ainsi, un des freins à l'exploitation de ces données, notamment leur assemblage, est la puissance de calcul disponible. En effet, pour être aussi exhaustif que possible, il convient de ré-analyser l'ensemble des données issues d'un environnement (plusieurs dizaines de téra-octets pour le seul microbiote humain). Cela passe nécessairement par l'utilisation de machines aussi performantes que possibles mais aussi de méthodes adaptées. Finalement, la grande part d'inconnu du monde microbien impose d'employer des méthodes non ciblées et *de novo* afin de s'affranchir des biais dus aux connaissances *a priori*.

Plusieurs stratégies ont été développées pour tenter de reconstruire des génomes à partir de données métagénomiques, notamment par l'utilisation du *binning*. Cette méthode consiste à établir le profil des fragments de génomes selon leur composition en nucléotides et/ou leur abondance au

*. Intervenant
†. Corresponding author : kevin.gravouil@udamail.fr
‡. Corresponding author : didier.debroas@univ-bpclermont.fr
§. Corresponding author : pailloux@isima.fr
¶. Corresponding author : eric.peyretaillade@udamail.fr

sein d'un ou plusieurs métagénomes. Deux contigs aux profils similaires appartiendraient ainsi au génome d'un même organisme. Il est cependant nécessaire d'assembler préalablement les données brutes pour augmenter la robustesse du *binning*.

Cette approche a été utilisée avec succès pour reconstuire 83 génomes, dont 29 complètement nouveaux, à partir de 32 millions de *reads* de métagénomes d'eaux saumâtres de la mer Baltique (Hugerth *et al.*, 2015). Cette étude a permis la description des gènes, de la dynamique temporelle de la population et de la biogéographie d'un nouveau bactérioplancton. Evans *et al.*, (2015) ont également utilisé ce type d'approche pour mettre en évidence deux archées méthanogènes ne faisant pas partie du *phylum* des Euryarchaeota à partir d'un aquifère profond, apportant ainsi un regard nouveau sur le cycle de méthane.

Le développement des méthodes de *binning* appliquées aux données métagénomiques est en plein essor. De nombreuses approches ont été envisagées (Sangwan *et al.*, 2016 ; Soueidan *et al.*, 2015), mais à l'heure actuelle aucun consensus ne s'est dégagé. Ce manque nous a conduit à évaluer les méthodes de *binning* existantes et à imaginer une nouvelle stratégie. Nous présenterons donc dans un premier temps une étude comparative à partir de jeux de données obtenues sur des communautés microbiennes artificielles de composition connue (e.g. : SRR072232, SRR606249). Dans un second temps, nous proposerons une stratégie originale de *binning*. Cette dernière intègre une caractérisation des séquences basée sur les « *k-mers* à trous » (*spaced seed*), des techniques d'extraction de données et de *clustering* issues du *machine learning* tout en exploitant les masses des données brutes et les connaissances existantes.

Étude comparative des méthodes de *binning*

Plusieurs grandes catégories de méthodes de *binning* existent : les méthodes basées sur la composition de séquences, les méthodes basées sur la co-abondance des séquences et les méthodes hybrides. Il est également nécessaire d'évaluer la fiabilité et les performances des différentes stratégies de *binning*. La robustesse du *binning* repose en partie sur la taille des séquences fournies. Les données sont préalablement assemblées en contigs à l'aide d'algorithmes éprouvés. Ces contigs sont alors caractérisés par l'un des critères suivants : (i) la fréquence de k-mers (avec généralement k=4), (ii) la co-abondance des séquences au travers de différents réplicats, (iii) d'une métrique construite pour intégrer (i) et (ii) à la fois. Optionnellement, on peut extraire de ces profils les informations pertinentes et ainsi réduire la complexité des données à traiter. Ces profils sont ensuite clusterisés afin de constituer les *bins*. Pour finir, les *bins* sont soumis à une validation sur des critères biologiques.

TETRA (Teeling *et al.*, 2004), pionnier dans l'approche basée sur la composition des séquences, se base sur la corrélation des motifs d'utilisation des tétranucléotides pour caractériser des séquences. Dick *et al.* (2009) ont exploité les travaux de Teeling *et al.* (2004) définir des clusters à l'aide d'une *emergent self-organizing map*. Par la suite, Nielsen *et al.* (2014) se basent sur la co-abondance des gènes dans de nombreux réplicats afin de regouper les séquences ayant des abondances similaires. D'autres méthodes hybrides ont émergé comme MetaBAT (Kang *et al.*, 2015). Cette dernière exploite à la fois les fréquences des tétranucléotides et la co-abondance des séquences *via* une métrique empirique construite à l'aide de 1414 génomes bactériens complets. Les profils obtenus sont ensuite clusterisés avec un algorithme des *k-medoids* modifié afin de former les *bins*. Ding *et al.* (2015) ont proposé une approche basée sur des outils radicalement différents. Elle consiste à calculer la corrélation intrinsèque des nucléotides, à extraire les informations pertinentes de ces profils par une régression des moindres carrés partiels (*kernel partial least squares*) puis à clustériser ces profils à l'aide d'une méthode de *machine learning* (machine à vecteurs de support, ou *SVM*).

La diversité des approches rend les évaluations et les comparaisons nécessaires. Pour ce faire, des métagénomes artificiels servent de données de référence dans l'évaluation des approches. Des outils dédiés à la validation du *binning* ont également été développés. C'est notamment le cas de

CheckM (Parks *et al.*, 2015) qui recherchent au sein de chaque *bin* des gènes-marqueurs présents en une seule copie au sein des génomes. D'autres part, le challenge CAMI (*Critical Assessment of Metagenomic Interpretation* ; McHardy *et al.*, 2014) propose des protocoles pour une évaluation standardisée des différents outils.

Bien que les méthodes de *binning* permettent de reconstruire des *drafts* de génomes à partir de données métagénomiques, le postulat initial repose sur l'homogénéité de la composition en nucléotide et/ou sur une profondeur de séquençage uniforme au sein des génomes. Cependant, cette homogénéité peut être perturbée par les nombreux transferts horizontaux entre micro-organismes potentiellement très éloignés phylogénétiquement (Garcia-Vallvé *et al.*, 2000), voire également entre procaryote et eucaryote (Hotopp *et al.*, 2007). Ces échanges de matériel génétique vont modifier la composition locale en nucléotides des génomes. Les séquences ainsi obtenues ne répondront plus au premier postulat des méthodes de *binning*. Il en est de même au sujet de la non-uniformité de la profondeur de séquençage (Peng *et al.*, 2012). De plus, des portions de génomes hautement conservées (e.g. familles de gènes) peuvent être partagées par plusieurs individus. Il est alors difficile de les attribuer à un individu plutôt qu'à un autre. Il pourrait ainsi être pertinent de les attribuer aux deux en même temps. Nous avons conduit une comparaison approfondie de différents outils de *binning* couvrant un large panel de techniques. L'évaluation des outils existants repose sur les métagénomes artificiels et les méthodes d'évaluation précédemment décrits. Ces outils sont alors jugés sur la pertinence des résultats obtenus mais aussi sur les ressources informatiques nécessaires au traitement des données. Ce travail permet alors d'identifier les atouts et inconvénients de chaque méthode et de proposer une nouvelle approche de reconstruction de génomes microbiens à partir de métagénomes.

Approche proposée : processus itératif de reconstruction de génomes

Les assemblages pré- et post-*binning* sont assurés par des algorithmes robustes et éprouvés, à savoir les graphes de De Bruijn et l'Overlap-Layout-Consensus.

Pour reconstruire des génomes microbiens à partir de métagénomes, la stratégie itérative suivante est proposée : (i) Les séquences sont caractérisées selon de multiples critères (e.g. : fréquences des *k-mers*, co-abondances des séquences, *etc.*) ; (ii) La complexité des données générées pourra être réduite par des techniques d'extraction de données ; (iii) Les profils établis sont clusterisés pour former des *bins* ; (iv) Chaque *bin* ainsi formé est assemblé indépendamment des autres en vue de reconstruire de plus grands fragments de composition homogène ; (v) Un second assemblage *cross-bins* permet ensuite d'assembler des séquences chevauchantes mais dont la composition diffère localement (*e.g.* bordure d'une portion provenant d'un transfert horizontal). Les *bins* dont deux séquences ont été assemblées sont alors fusionnés ; (vi) Chaque *bin* est évalué selon les modalités des outils existants. Si un *bin* rempli les critères de validation, il est alors extrait et ne participera pas à l'itération suivante ; (vii) Les séquences ayant été utilisées pour générer des contigs sont supprimées du *pool* initial de séquences et ces contigs sont placés dans ce *pool* afin de d'entamer un nouveau cycle.

Ce cycle se poursuit tant qu'il reste des séquences à traiter ou quand l'évaluation des *bins* (étape vi) produit de meilleurs résultats par rapport à l'itération précédente. Les régions uniformes en terme de composition ou de couverture sont alors correctement assemblées entre elles et des régions pourtant différentes en terme de composition et couverture peuvent aussi être assemblées. La caractérisation de séquences combine les différentes approches couramment utilisées pour la métagénomique non ciblée et *de novo* ainsi que l'approche basée sur les *k-mers* à trous (ou « *spaced seed* ») encore jamais appliquée pour ces études *de novo*. D'autres types de *clustering* seront testés comme le *clustering* « flou » permettant d'attribuer une même séquence à plusieurs *bins*. Les séquences communes à plusieurs espèces sont donc correctement attribuées. L'évaluation rigoureuse de ce *clustering* se basera sur des communautés artificielles précédemment décrites et se fera de la même manière que les outils précédemment testés. Le volume de données disponible imposera également l'utilisation de méthodes et architectures issues du Big Data, notamment avec

l'utilisation d'Apache Spark. Ces outils permettent d'envisager la ré-analyse de grandes masses de données existantes.

Références

Blainey, P.C. (2013) The future is now: single-cell genomics of bacteria and archaea. *FEMS microbiology reviews*, 37:407–427.

Dick, G.J. et al. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome biology*, 10:1–16.

Ding, X. *et al.* (2015) DectICO: an alignment-free supervised metagenomic classification method based on feature extraction and dynamic selection. *BMC bioinformatics*, 16:323.

Evans, P.N. *et al.* (2015) Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science*, 350:434–438.

Garcia-Vallvé, S. *et al.* (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Research*, 10:1719–1725.

Gasc, C. *et al.* (2015) Capturing prokaryotic dark matter genomes. *Research in microbiology*, 166:814–830.

Hotopp, J.C.D. *et al.* (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, 317:1753–1756.

Hugerth, L.W. *et al.* (2015) Metagenome-assembled genomes uncover a global brackish microbiome. *Genome biology*, 16:1–18.

Kang, D.D. *et al.* (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165.

McHardy, A. *et al.* (2014) Critical Assessment of Metagenomic Interpretation.

Nielsen, H.B. *et al.* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature biotechnology*, 32:822–828.

Parks, D.H. *et al.* (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25:1043–1055.

Peng, Y. *et al.* (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28:1420–1428.

Qin, J. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464:59–65.

Sangwan, N. *et al.* (2016) Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, 4:1.

Soueidan, H. and Nikolski,M. (2015) Machine learning for metagenomics: methods and tools. *arXiv preprint* arXiv:1510.06621.

Teeling, H. *et al.* (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC bioinformatics*, 5:163.

Mots clefs : assemblage, reconstruction de génomes

Cheese ecosystems insights with shotgun metagenomics and a metadata extended genomics database

Poster 72

Thibaut Guirimand [*] [†1], Charlie Pauvert [*] [‡1,2], Sandra Dérozier[3], Anne-Laure Abraham[1], Mahendra Mariadassou[4], Valentin Loux[3], Pierre Renault[1]

[1] MICrobiologie de l'ALImentation au Service de la Santé humaine (MICALIS) – AgroParisTech, Institut national de la recherche agronomique (INRA) : UMR1319, F-78350 Jouy-en-Josas, France
[2] Master de Bioinformatique – Université de Rouen – France
[3] Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut national de la recherche agronomique (INRA) : UR1404 – Bâtiment 210-233, Domaine de Vilvert, F-78 350 Jouy en Josas Cedex, France
[4] Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut National de la Recherche Agronomique – INRA, Université Paris-Saclay – INRA Unité MaIAGE, Bâtiment 233, Domaine de Vilvert, F-78 352 Jouy-en-Josas Cedex, France

The manufacturing process of cheeses, as for most fermented food, involves a complex flora, which is composed of bacteria, yeast and filamentous fungi. They can be directly inoculated as starter culture or develop from the food-chain environment (raw milk, cheese factory...). Therefore the exact composition of most cheeses is not completely known.

Further understandings of cheeses ecosystems and control of cheese quality both need a better characterization of the cheese flora with a precise taxonomic identification. The FoodMicrobiome-Transfert project aims to address these challenges.

In the framework of this project, we are developing a tool to facilitate metagenomics analysis. This tool is composed of read alignment wrapper tool, a database and a web interface to run analysis.

GeDI : an in-house metagenomics analysis tool

Shotgun metagenomics sequencing data brings information about the studied ecosystem, but also yields noise signal. Hence retrieving the link between sequence and organism is not trivial and require different strategies.

Several current metagenomics tools are based on a set of gene markers, or on the k-mer composition of the reads, but few are able to identify species up to the strain level. We are developing an in-house software to wrap read alignments on reference genomes and extract information from these processes. It relies on the intersection between features (CDS) and alignments data (BAM) to infer species presence or absence.

A web application and a database to exploit GeDI possibilities

The application will allow the users to submit metagenomes and personal genomes. They will be able to choose a list of genomes from our public database and from their personal genome library. They will finally be able to execute GeDI to analyze their metagenome data.

The database, currently in development, will store (i) genomics data from food-related microorganisms that will be used for metagenomics data analysis, (ii) metadata associated with the

*. Intervenant
†. Corresponding author : thibaut.guirimand@jouy.inra.fr
‡. Corresponding author : charlie.pauvert@jouy.inra.fr

ecology of these microorganisms and (iii) metagenomics analysis results. The database genomics part will be enriched with expert annotations, with a focus on genes of technological interest.

The application will permit to visualize and compare analysis results and cheese environments metadata.

Technical specificities

The tools will be hosted on the Migale platform. The GeDI software will be run transparently on the Migale Galaxy portal using our specific web interface. The Python 3 Django web communicates with Galaxy using the bioblend library and allow us to easily manage datasets and outputs. Information are exchanged through bioinformatics standard files (GFF, BAM, etc.), thus easing the use or the export to others tools.

Mots clefs : Database, Metagenomics, NGS, Cheese, Shotgun metagenomics sequencing

Biogeography of predatory protists in neotropical forest soils

Poster 73

Guillaume Lentendu [*] [†1,2], Frédéric Mahé[3], Micah Dunthorn[1]

[1] Technical University of Kaiserslautern (TU Kaiserslautern) – Erwin-Schroedinger Street, 67663 Kaiserslautern, Allemagne
[2] Helmholtz Centre for Environmental Research (UFZ) – Theodor-Lieser-Str. 4, 06120 Halle/Saale, Allemagne
[3] CIRAD – Centre de coopération internationale en recherche agronomique pour le développement : UPR39 – France

Metabarcoding have greatly enhance our understanding of soil protists and provided an uncomparable amount of information on their diversity and assemblages. Though, most of the studies were carried out in temperate soils, while information is missing from tropical forest soils, in which diversity and biogeographic patterns are unknown. In particular, it was previously described that despite a high local diversity, eukaryotic species turnover from arid soil was not high at the regional scale (Green *et al.* 2004). This is in line with further analyses which supported the 'moderate endemicity model' for protist (Foissner 2006; Bass *et al.* 2007), but this was not yet challenged with next generation sequencing data from highly diverse tropical forest soils. To fill this gap, soils in three central America lowland Neotropical forests were sampled at a total of 279 positions. The V4 region of 18S rRNA was tag amplified from the soil extracted DNA and further sequenced with Illumina MiSeq. A first analysis revealed a hyperdiverse protistan community dominated by Apicomplexan parasites, and an extreme high OTU turnover within and between forests (Mahé et al., submitted). In the current study, we further analyzed the spatial turnover of the two other dominant protist groups found in those soils, ciliates and cercozoa, by using the taxa-area relationship and the distance-decay relationship. We also explored the correlations of the previous turnover patterns with the soil and climatic parameters using multivariate statistical approaches. This will thus permit us to predict the expected ciliates and cercozoa diversity at the continental scale by using spatial scaling.

References

Bass D, Richards TA, Matthai L, Marsh V, Cavalier-Smith T (2007) DNA evidence for global dispersal and probable endemicity of protozoa. *BMC Evolutionary Biology*, 7:162.

Foissner W (2006) Biogeography and dispersal of micro-organisms: a review emphasizing protists. *Acta Protozoologica*, 45:111–136.

Green JL, Holmes AJ, Westoby M *et al.* (2004) Spatial scaling of microbial eukaryote diversity. *Nature*, 432:747–750.

Mahé F, de Vargas C, Bass D, Czech L, Stamatakis A *et al.* (submitted) Soil Protists in Three Neotropical Rainforests are Hyperdiverse and Dominated by Parasites.

Mots clefs : protist, ciliates, cercozoa, soil, biogeography, neotropical forest

*. Intervenant
†. Corresponding author : guillaume.lentendu@ufz.de

Housekeeping gene targets and NGS to investigate *Vibrio* spp. communities in coastal environments

Laura Leroi [*] [†1], Joëlle Cozien[2], Fanny Marquer[1], Laure Quintric[1], Marie Agnès Travers[3], Dominique Hervio-Heath [‡2]

Poster 74

[1] Service Ressources Informatiques et Communications (IMN/IDM/RIC) – Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) – Centre de Brest, Pointe du Diable, F-29 280 Plouzané, France
[2] Laboratoire Santé Environnement et Microbiologie (RBE/SG2M/LSEM Brest) – Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) – Pointe du Diable, F-29 280 Plouzané, France
[3] Laboratoire de Génétique et Pathologie des Mollusques Marins (RBE/SG2M/LGPMM La Tremblade) – Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) – Station de La Tremblade, Avenue de Mus de Loup, F-17 390 La Tremblade, France

Bacteria of the genus *Vibrio* are widespread in marine and estuarine waters worldwide and present a high degree of genomic and phylogenetic diversity. Vibrios are one of the most diverse and important marine heterotrophic bacterial groups. To date, more than 130 *Vibrio* species have been described, among which 12 were classified as human pathogens [1]. Climatic changes and human pressure have been linked to pathogens proliferations in costal environment [2]and have led to outbreaks of *Vibrio* diseases in humans and marine organisms and to mass mortalities in oysters [3]. Studies investigating *Vibrio* population dynamics have predominately focused on pathogens in the environment or within their hosts. Very little is known regarding involvement and dynamics of other microorganism communities in the emergence of these pathogens or in the disease process. The ecological dynamics that led to this diversity is a key factor to understand how pathogens evolve in environmental reservoirs.

Due to the limitations of the 16S rRNA phylogeny to elucidate an "Integrated Vibrio Biology" [4], we cannot assess the diversity of closely related but distinct bacterial organisms. In order to monitor key taxonomic groups, *Vibrio* spp. and other bacteria, in aquatic ecosystems, we propose a DNA metabarcoding approach using 16S rRNA and non-conventional markers to directly access the genetic content of environmental samples. The sequencing of the DNA already presents in a water sample using specie target group specific primers makes it a non-invasive method. It is an ecological [5] powerful tool to detect all species of a target group present on the study site. It enables to improve detection of rare species, in comparison with conventional isolation methods [6].

In this study, we designed new primers targeting the main *Vibrio* spp. phylogenetic clades and possibly the major human and animal pathogen species with Next Generation Sequencing.

In silico analysis of sequences from Genbank and Silva databases available for a high number of *Vibrio* isolates and species resulted in the selection of six housekeeping genes with the highest taxonomic coverages. To refine our selection, we assessed the taxonomic resolution and the chimera operational taxonomy unit (OTU) percentage of each potentially discriminant gene by clustering using UCLUST [7]. The alignment of the more discriminant genes sequences by Clustal Omega allowed to identify conserved and variable regions. Regions that could better differentiate vibrios were chosen to design primers. KASpOD [8] was used to design specific primers by comparing target and non-target sequences.

Specificity and sensitivity of the primers and of the yielded amplicons are being validated by PCR and sequencing on several bacterial species (*Vibrio* and other marine bacteria) from

*. Intervenant
†. Corresponding author : laura.leroi@ifremer.fr
‡. Corresponding author : dominique.hervio.heath@ifremer.fr

collection or isolated from the environmental studies. These new primers will be used for NGS on mixed bacterial populations in natural seawater and shellfish samples. The comparison of these multigenics sources OTUs table will allow us to strengthen the taxonomic coverage.

Investigation of the dynamics and diversity of bacterial and *Vibrio* spp. populations at different periods and in different environmental conditions will be useful to the understanding of the disease process and spreading. These informations could be used to develop a risk prediction model.

References

[1] Association of Vibrio Biologists website [Internet]. Available from: `http://www.vibriobiology.net/`

[2] Baker-Austin C, Trinanes JA, Taylor NGH, Hartnell R, Siitonen A, Martinez-Urtaza J. Emerging Vibrio risk at high latitudes in response to ocean warming. *Nat. Clim. Change.* (2013) 3:73–7.

[3] Le Roux F, Wegner KM, Baker-Austin C, Vezzulli L, Osorio CR, Amaro C, et al. The emergence of Vibrio pathogens in Europe: ecology, evolution, and pathogenesis (Paris, 11-12th March 2015). *Front. Microbiol.* (2015) 6:830.

[4] Gomez-Gil B., Thompson C. C., Matsumura Y., Sawabe T., Iida T., Christen R., et al. (2014). Family Vibrionaceae (Chapter 225), in *The Prokaryotes*, 4th Edn. eds Rosenberg E., DeLong E., Thonpson F. L., Lory S., Stackebrandt E., editors. (New York, NY: Springer;), 88

[5] Valentini A, Pompanon F, Taberlet P. DNA barcoding for ecologists. *Trends Ecol. Evol.* (2009) 24:110–7.

[6] Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, et al. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Mol. Ecol.* (2016) 25:929–42.

[7] Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* (2010) 26:2460–1.

[8] Parisot N, Denonfoux J, Dugat-Bony E, Peyret P, Peyretaillade E. KASpOD–a web service for highly specific and explorative oligonucleotide design. *Bioinforma. Oxf. Engl.* (2012) 28:3161–2.

Mots clefs : metabarcoding, primers, *Vibrio*

OBITools3 : Une suite logicielle pour la gestion des analyses et des données de DNA metabarcoding

Poster 75

Céline Mercier * †[1], Frédéric Boyer[1], Éric Coissac[1]

[1] Laboratoire d'écologie alpine (LECA) – CNRS : UMR5553, Université Joseph Fourier - Grenoble I, Université de Savoie – Bâtiment D - Biologie, 2233 rue de la piscine - BP 53, F-38 041 Grenoble Cedex 9, France

Introduction. Le DNA metabarcoding offre de nouvelles perspectives dans le domaine de la recherche sur la biodiversité [1]. Cette approche de l'étude des écosystèmes repose largement sur l'utilisation du séquençage nouvelle génération (NGS), et par conséquent requiert la capacité de traiter d'importants volumes de données. La suite logicielle OBITools satisfait cette exigence grâce à un ensemble de programmes spécifiquement conçus pour l'analyse de données NGS dans le contexte du DNA metabarcoding [2] – `http://metabarcoding.org/obitools/`. Leur capacité à filtrer et éditer les séquences en prenant éventuellement en compte l'information taxinomique permet de mettre en place facilement des pipelines d'analyse couvrant un large champ d'applications du DNA metabarcoding.

Les OBITools3. Cette nouvelle version des OBITools cherche à améliorer significativement l'efficacité du stockage et la rapidité de traitement des données. Pour cela, les OBITools3 reposent sur un système de base de données *ad hoc* à l'intérieur duquel toutes les données qu'une expérience de DNA metabarcoding doit considérer sont stockées : les séquences, les métadonnées décrivant notamment les échantillons, les bases de données de séquences de référence utilisées pour l'annotation taxinomique, ainsi que les bases de données taxinomiques. L'avantage de cette nouvelle structure intégrée, outre les gains d'efficacité qu'elle apporte aux OBITools, est de permettre un échange facile de la totalité des données associées à une expérience.

Un stockage colonne-centré. Un pipeline d'analyse correspond à l'enchaînement d'une série de commandes, réalisant chacune un calcul élémentaire et où le résultat de la commande *n* est utilisé comme donnée par la commande *n+1*. Les données de DNA metabarcoding pouvant aisément être représentées sous la forme de tables, chaque commande peut donc être considérée comme une opération transformant une ou plusieurs tables « entrées » en une ou plusieurs tables « résultats », qui pourront à leur tour être utilisées par la commande suivante. En pratique, beaucoup des opérations élémentaires d'un pipeline reportent dans les tables résultats une grande parties des données incluses dans les tables entrées sans les modifier, et n'utilisent pour leur calculs que quelques unes des informations incluses dans les tables entrées. Actuellement, les OBITools stockent ces tables de données sous la forme de fichiers de séquences annotées au format FASTA ou FASTQ. Cela a deux conséquences : i) si l'on souhaite garder les résultats intermédiaires du processus de traitement, le stockage des fichiers intermédiaires induit le stockage d'une grande quantité d'information redondante, ii) Le décodage et codage de toutes les informations incluses dans les fichiers de séquences finalement non utilisées et/ou non modifiées par la commande représentent une part prépondérante du processus de traitement des OBITools actuels. Le nouveau système de gestion de données des OBITools3 (DMS pour Data Management System) repose sur une organisation colonne-centrée. Les colonnes sont non-mutables et peuvent être associées en vues représentant les tables de données. Les données non modifiées par une commande dans une

*. Intervenant

†. Corresponding author : `celine.mercier@metabarcoding.org`

table entrée peuvent ainsi aisément être associées à la vue résultat sans dupliquer l'information. Des données non utilisées par une commande mais devant être associées au résultat pourront même être insérées dans la vue résultats sans être lues. Il résulte de cette stratégie un gain d'efficacité en espace disque en limitant la redondance des données ainsi qu'un gain en temps de traitement en limitant les opérations de lecture, conversion et d'écriture des données. Enfin, pour optimiser l'accès aux données, chaque colonne est stockée dans un fichier binaire directement mappé en mémoire pour les opérations de lecture et d'écriture.

Optimisation du stockage. Les données de DNA metabarcoding sont intrinsèquement très redondantes. Par exemple, la même séquence correspondant à une espèce sera produite des milliers de fois au sein d'un échantillon et entre les différents échantillons. Pour limiter l'espace de stockage et augmenter l'efficacité des opérateurs de comparaison, l'ensemble des données de type chaîne de caractères est stocké dans les colonnes en utilisant une structure d'index complexe, efficace sur des millions de valeurs, couplant clés de hachage, filtres de Bloom et arbres AVL. Enfin, les séquences d'ADN sont aussi compressées en encodant chaque nucléotide sur 2 ou 4 bits suivant que les séquences ne contiennent que les quatre nucléotides (A,C,G,T) ou utilisent les codes d'incertitude IUPAC.

Stockage de l'historique du traitement des données. L'ensemble des informations manipulées par les OBITools3 sont stockées dans des structures de données non-mutables du DMS. Si une commande a besoin de modifier une colonne utilisée en entrée pour produire son résultat, une nouvelle version de cette colonne est produite, laissant intacte la version initiale. Ce système de stockage permet de conserver à moindre coût l'ensemble des données intermédiaires produites par le pipeline d'analyse. Le stockage de métadonnées décrivant l'opération ayant produit une vue dans le DMS permet de décrire un hypergraphe orienté où les nœuds sont les vues et les arcs représentent les opérations de transformation. En remontant les relations de dépendance au sein de cet hypergraphe, il est possible de reconstruire a posteriori le processus d'analyse complet ayant permis d'obtenir une table résultat.

Outils. Les OBITools3 proposent les mêmes outils que les OBITools originaux. À terme, de nouvelles versions des programmes ecoPrimers (conception d'amorces de PCR) [3], ecoPCR (PCR *in silico*) [4], ainsi que Sumatra (alignement de séquences) et Sumaclust (alignement et clustering de séquences) [5] tirant parti de la structure de gestion des données propre aux OBITools3 seront ajoutés.

Implémentation et disponibilité. Les couches basses de gestion du DMS ainsi que toutes les fonctions de calcul intensif sont codées en C99 pour des raisons d'efficacité. Une couche objet Cython (`http://www.cython.org/`) permet une implémentation simple mais efficace des commandes OBITools3 en Python 3.5. Les OBITools3 sont encore en développement, la version actuelle du code est disponible sur le site git (`https://git.metabarcoding.org/obitools/obitools3/`). Les premières versions fonctionnelles sont prévues pour l'automne 2016.

Références

[1] Taberlet P, Coissac É, Hajibabaei M, Rieseberg LH: Environmental DNA. *Mol Ecol* 2012:1789–1793.

[2] Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac É: OBITools: a Unix-inspired software package for DNA metabarcoding. *Mol Ecol Resour* 2015, 16: 176–182.

[3] Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac É: ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Res* 2011, 39:e145.

[4] Ficetola GF, Coissac É, Zundel S, Riaz T, Shehzad W, Bessière J, Taberlet P, Pompanon F: An in silico approach for the evaluation of DNA barcodes. *BMC Genomics* 2010, 11:434.

[5] Mercier C, Boyer F, Bonin A, Coissac É (2013) SUMATRA and SUMACLUST: fast and exact comparison and clustering of sequences. Available: `http://metabarcoding.org/`

sumatra/ and http://metabarcoding.org/sumaclust/

Mots clefs : métabarcoding

Predicting DNA methylation by means of CpG o/e ratio in the case of a pan-species study

Benoît Aliaga [*1]

Poster 76

[1] Interactions Hôtes Pathogènes et Environnement (IHPE) – CNRS : UMR5244, Université de Perpignan – Université de Perpignan, 52 avenue Paul Alduy, F-66 860 Perpignan Cedex, France

Epigenetic mechanisms contribute to generate heritable phenotypic variability and could therefore be involved in evolution. The most studied of these mechanisms is DNA methylation, i.e. the addition of a methyl group to cytosines leading to 5-methyl-cytosine (5mC). Methylation is predominantly (but not exclusively) located in CpG dinucleotides. Whereas in vertebrates and plants, DNA methylation is known to be involved in control of gene transcription, its role in invertebrates is far less understood. In spite of being one of the most analysed epigenetic information carriers, DNA methylation has not been exhaustively studied in many species, on account of its high cost and the expertise it requires. Since 5mC is spontaneously deaminated to thymine, CpG underrepresentation can be used as a proxy to estimate 5mC levels and distribution patterns. Our work consisted in predicting the methylation status in transcriptomes for 605 species from CpG observed to expected ratios (CpG o/e) by means of a statistical method based on kernel densities estimations. We will present this new software called Notos (standalone and galaxy based) that allows to predict DNA methylation by only using mRNA data.

Mots clefs : DNA methylation prediction, 5, methyl, cytosine, CpG o/e, database, evolution

*. Intervenant

Origin and evolution of extreme halophilic archaeal lineages

Monique Aouad [*1], Najwa Taib[1], Anne Oudart[1], Manolo Gouy[1], Simonetta Gribaldo[2], Céline Brochier-Armanet[1]

Poster 77

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE) – Institut Pasteur de Paris – 28 rue du Docteur Roux, F-75 724 Paris Cedex 15, France

The recent advances in sequencing technologies have provided a valuable material to study the evolutionary history of Life. This rainfall of genomic data offers a unique opportunity to improve phylogenetic inferences, by allowing the simultaneous analysis of hundreds of markers and to combine the weak phylogenetic signal carried by each individual marker toward a stronger signal. In the case of *Archaea*, the analysis of conserved core genes has the ancient evolutionary history of this Domain. While significant advances emerged from these analyses, some highly debated nodes remain unresolved, such as the phylogenetic position of extreme halophilic lineages (i.e. *Halobacteriales* and nano-sized *Nanohaloarchaea*). These extremophiles thrive in high salt concentrations and require salt for growth. Pinpointing the phylogenetic position of *Halobacteriales* and *Nanohaloarchaea* is however crucial to understand the evolutionary patterns and the adaptive molecular processes at the origin of these extremophilic organisms.

Here we present an in-depth phylogenomic analysis aiming at tackling this issue. Using comparative genomics approaches, we identified more than 250 protein markers carrying a reliable phylogenetic signal to study the relationships among *Nanohaloarchaea*, *Halobacteria* and other archaeal lineages. Accurate analyses combining noise-removing and amino-acid recoding approaches indicates that *Nanohaloarchaea* and *Halobacteria* represent two independent lineages. This implies that adaption to high salt concentrations emerged twice independently in *Archaea*. This result provides a robust frame to decipher the evolutionary paths at the origin of extreme halophily in *Archaea*.

Mots clefs : evolutionary history, genomic data, phylogenetic signal, *Archaea*, methanogens, *Halobacteriales*, *Nanohaloarchaea*

*. Intervenant

Détection sans a priori et étude des communautés bactériennes au cours des différents stades de développement de la tique

Poster 78

Émilie Bard [*1], Patrick Gasqui[1], Maria Bernard[2], Séverine Bord[1], Valérie Poux[1], Maggy Jouglin[3], Olivier Duron[4], Jean-Louis Chapuis[5], Gwenaël Vourc'h[1], Suzanne Bastian[3], Laurence Malandrin[3], Olivier Plantard[3], Xavier Bailly[1]

[1] Unité d'Épidémiologie Animale (UR346), INRA (INRA) – Institut National de la Recherche Agronomique - INRA (FRANCE) – Rue de Theix, F-63 122 Saint Genès Champanelle, France
[2] Système d'Information des GÈNomes des Animaux d'Elevage (SIGENAE) – Institut national de la recherche agronomique (INRA) : UMR1313 – France
[3] UMR 1300 INRA/ONIRIS Bioagression, Épidémiologie et Analyse de Risques (BIOEPAR) – INRA, Oniris – École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation, Nantes-Atlantique, Atlanpole - La Chantrerie, B.P. 40706, F-44 307 Nantes Cedex 03, France, France
[4] Institut des Sciences de l'Évolution, Université Montpellier 2 (ISEM) – Université Montpellier II - Sciences et Techniques du Languedoc : EA34095 – Université Montpellier 2, CNRS, UMR5554, F-34 095 Montpellier Cedex 05, France, France
[5] Muséum national d'histoire naturelle (MNHN) – Ministère de l'Écologie, du Développement Durable et de l'Énergie, Ministère de l'Enseignement Supérieur et de la Recherche, Muséum National d'Histoire Naturelle (MNHN) – 57 rue Cuvier, F-75 231 Paris Cedex 05, France

La surveillance des maladies transmises par les tiques dépend de méthodes de détection qui se limitent généralement à un nombre réduit de pathogènes et à la détection de ceux qui sont déjà connus. Les études de communautés bactériennes sans *a priori* permettent de s'affranchir de cette limite en obtenant une vision globale de la diversité des taxons portés par les tiques au cours de leurs différents stades de développement. Afin de valider cette hypothèse, nous avons procédé à un séquençage d'ARN ribosomique 16S pour des tiques à différents stades de développement.

434 tiques, appartenant à l'espèce *Ixodes ricinus*, ont ainsi été échantillonnées dans la forêt de Sénart (France). Afin d'avoir une vision d'ensemble des communautés bactériennes présentes chez ces tiques, une amplification PCR a été réalisée à l'aide d'amorces universelles de la région codant pour l'ARN 16S (Claesson et al. 2010) des bactéries. Les produits PCR obtenus ont ensuite été séquencés par pyroséquençage (454). Les données issues du séquençage ont été analysées à l'aide de Mothur (Schloss et al. 2011), un outil de traitement et d'analyse de séquences 16S permettant de réduire les artefacts liés à l'amplification et au séquençage et d'assigner les séquences à des taxonomies. L'analyse des données obtenues suggère un changement majeur des communautés bactériennes portées par les tiques entre le stade larvaire et les stades suivants. Un test de Fisher a permis de mettre en évidence les différences de fréquences entre stades pour chaque taxonomie. Des représentations graphiques ont également permis de visualiser les proportions occupées par chaque taxonomie en fonction des différents stades de développement. Ces données permettent également de suivre plus en détail les cycles d'agents pathogènes d'intérêts, malgré différents problèmes liés à la faible fréquence des pathogènes dans les communautés bactériennes ciblées ou leur faible représentation dans les alignements de référence généralement utilisés pour l'analyse.

*. Intervenant

Références

Schloss PD, Gevers D, Westcott SL. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS ONE*. 6:e27310.

Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*. 38:e200

Mots clefs : *Ixodes ricinus*, ARN 16S, Communautés bactériennes, Mothur, Phylogénie

Unexpected genome inflation and streamlining in variable environments

Bérénice Batut [*1], Guillaume Beslon[2], Carole Knibbe[2]

Poster 79

[1] Conception, Ingénierie et Développement de l'Aliment et du Médicament (CIDAM) – Université d'Auvergne - Clermont-Ferrand I – Facultés de Médecine et de Pharmacie CBRV - RdC et 5e étage BP 38, 28 Place Henri-Dunant, F-63 001 Clermont-Ferrand, France – Tél : +33 4 73 17 79 52
[2] BEAGLE (Insa Lyon / INRIA Grenoble Rhône-Alpes / UCBL) – INRIA, Institut National des Sciences Appliquées [INSA] - Lyon, Université Claude Bernard - Lyon I (UCBL) – Antenne INRIA Lyon la Doua Bâtiment CEI-1, 66 boulevard Niels Bohr, F-69 603 Villeurbanne, France

Natural ecosystems undergo different levels of environmental perturbations from seasonal fluctuations to less frequent natural disasters and global changes. These different levels impact differently ecosystems, their organisms and adaptive ability. Here, we focus on stochastic environmental fluctuations and their effect on genome evolution. In bacterial genomes, there is some empirical evidence for a correlation between the modularity of the metabolic network and environmental variability (Parter et al., 2007). Moreover, theoretical studies suggest that temporal variability of the environment can impact genome structure (Bentkowski et al., 2015), modularity (Lipson et al., 2002), network structure (Kashtan and Alon, 2005), evolution speed (Kashtan et al., 2007) and evolvability (Earl and Deem, 2004; Crombach and Hogeweg, 2008). It was also predicted that with greater variability of the environment, genomes should have more genes (Bentkowski et al., 2015).

Most of these studies rely on abstract or population genetic models of evolution, without explicitly representing the genome sequence of each individual. A few other studies (Crombach and Hogeweg, 2008; Bentkowski et al., 2015) use individual-based evolutionary models where each individual has genes, but non coding regions are not taken into account. Non coding DNA, even when it has no direct impact on the phenotype, is a source of variability through mutations and rearrangements, possibly increasing evolvability (Knibbe et al., 2007). This ability to generate adaptive genetic diversity (evolvability) is important to face environmental changes (Crombach and Hogeweg, 2008). Thus non coding DNA is an important factor to take into account to study the links between temporal environmental variability, evolution of evolvability and evolution of genome organization.

Here, we used Aevol, an individual-based evolutionary model (Batut et al., 2013; Knibbe et al., 2007), to investigate the impact of environmental variation speed and amplitude on genome organization and evolvability. Our simulations show that evolved genome size and gene density strongly depend on the speed of environmental fluctuations. As discussed below, this behavior can be explained by indirect selection for evolvability when the speed is low to moderate, and by indirect selection for robustness when the speed is so high that the selection becomes stabilizing rather than directional.

Material and methods

Aevol (Batut et al., 2013; Knibbe et al., 2007) model simulates the evolution of a population of N artificial organisms using a variation-reproduction cycle. The population has a constant size over time and is completely renewed at each time step.

Each artificial organism owns a circular, double-stranded chromosome. The chromosome is a string of binary nucleotides, 0 being complementary to 1. This chromosome contains cod-

*. Intervenant

ing sequences (genes) separated by non-coding regions. Each coding sequence is detected by a transcription-translation process and decoded into a "protein" that contributes positively or negatively to a subset of abstract phenotypic characters. Adaptation is then measured by comparing the net values of the phenotypic characters to target values.

At each time step, N new individuals are created by reproducing preferentially the fittest individuals from the parental generation. Afterwards, all individuals from the parental population die. In the experiments presented here, reproduction was strictly asexual.

When a chromosome is replicated, it can undergo point mutations, small insertions and small deletions, but also large chromosomal rearrangements: duplications, large deletions, inversions, translocations. Thus mutations can modify existing genes, but also create new genes, delete some existing genes, modify the length of the intergenic regions, modify gene order, etc.

In this model, the environment is indirectly modelled as the set of target values for the characters, *i.e.* the set of processes needed for an individual to survive. These target values fluctuate over time, according to an autoregressive process of order 1 (an Ohrstein-Uhlenbeck process in discrete time) with parameters sigma and tau, where *sigma* controls the amplitude of the fluctuations and *tau* controls the speed at which the values tends to return to their mean.

To estimate the impact of environmental variation on genome organization, simulations were run with 5 different *sigma* values and 21 different *tau* values during 300,000 generations. Each (*sigma*,*tau*) couple was tested with 5 independent populations.

Results and discussion

For all *sigma* values, variation speed (inverse of *tau*) has a non-linear impact on genome structure. Indeed, as variation speed increases, the evolved genome size first increases and then decreases, with a maximum size observed for mild *tau* values. Hence the relation of environmental variation speed and genome size is bell-shaped, and it is mostly due to variation in the amount of non-coding bases. Intermediate speeds of environmental fluctuations yield the genomes with the lowest gene densities. This suggests that non-coding DNA could play a role in adapting to environmental fluctuations when these fluctuations occur at intermediate speed.

To investigate this hypothesis, we conducted complementary experiments on a subset of the evolved populations (those with the highest sigma values and 5 different tau values). We removed all the non-coding DNA inside the evolved genomes and let them evolve for 100,000 additional generations. We observed that after a few thousand generations, about as much non-coding DNA was regained as the quantity that was removed. The bell-shaped relation between the amount of non-coding DNA and environmental variation speed was quickly restored.

Variation amplitude (*sigma*) does not qualitatively change the shape of the relation between *tau* and genome size. A higher *sigma* intensifies the relation by making the bell more peaked and shifting the peak towards smaller values of *tau*.

This relation between the speed of environmental variation and genome size may be driven by evolvability and indirect selection of mutational variability level. Indeed, as spontaneous rates of local mutations and rearrangements are per base, a larger genome undergoes more local mutations and more rearrangements. Changes of genome size can then modulate the number of local mutations and rearrangements. Increasing the fraction of non-coding DNA only marginally impact the mutational variability of point mutations and indels. However, it has been shown that impact of large duplications and large deletions on coding sequences increases as the proportion of non-coding DNA increases (Knibbe et al., 2007; Fischer et al., 2014). When a beneficial local mutation or rearrangement is selected, the genome where the mutation/rearrangement occured and its size are selected. Genome size can then be indirectly selected by "hitchhiking".

In an environment that fluctuates very slowly, the need for mutational variability is low, hence there is no indirect selective pressure to maintain non-coding DNA, a source of deleterious

mutations. If fluctuations occur faster, then more beneficial mutations or rearrangements are needed to adapt. And they are more likely to occur in large genomes, which indirectly selects for genomes with much non-coding DNA. However, above a certain fluctuation speed, the target changes so frequently that beneficial mutations at a given generation are no longer beneficial at next generation. Selection evolves from stabilizing to directional. The phenotype will stabilize on the temporal mean of the target. In these conditions, large genomes are conter-selected: they mutate too much and there is indirect selection for robustness instead of evolvability. This explains the observed genome streamlining in the fast changing environments.

This work shows the strong non-intuitive influence of environmental variability on genome architecture. As environments varying at mild speeds require more phenotypic variability, they promote indirect selection of variability and then genome inflation through the accumulation of non-coding sequences. Indeed, without effet on the phenotype, these non-coding sequences are directly selected. However, they increase global genetic variability and thus help organisms to face environmental variations. On the other hand, in too quickly or too slowly varying environment, genetic variability is more deleterious than beneficial. Non-coding sequences are then washed-out from the genome and genome streamlining is observed.

References

Batut, B. et al. (2013) In silico experimental evolution: a tool to test evolutionary scenarios. *BMC Bioinformatics*, 14:S11.

Bentkowski, P. et al. (2015) A Model of Genome Size Evolution for Prokaryotes in Stable and Fluctuating Environments. G*enome Biology and Evolution*, 7:2344–2351.

Crombach, A. and Hogeweg, P. (2008) Evolution of Evolvability in Gene Regulatory Networks. *PLoS Comput Biol*, 4:e1000112.

Earl, D.J. and Deem, M.W. (2004) Evolvability is a selectable trait. *Proceedings of the National Academy of Sciences of the United States of America*, 101:11531–11536.

Fischer, S. et al. (2014) A Model for Genome Size Evolution. *Bulletin of Mathematical Biology*, 76:2249–2291.

Kashtan, N. and Alon, U. (2005) Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102:13773–13778.

Kashtan, N. et al. (2007) Varying environments can speed up evolution. Proceedings of the National Academy of Sciences *of the United States of America*, 104:13711–13716.

Knibbe, C. et al. (2007) A Long-Term Evolutionary Pressure on the Amount of Noncoding DNA. *Molecular Biology and Evolution*, 24:2344–2353.

Lipson, H. et al. (2002) On the origin of modular variation. *Evolution*, 56(8):1549–1556.

Parter, M. et al. (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biology*, 7:169.

Mots clefs : genome size evolution, environmental variation

La visualisation d'arbres phylogénétiques sur le web

Jérôme Bourret [*1,2], Anne-Muriel Arigon Chifolleau[1], Vincent Lefort [†1]

Poster 80

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – CNRS : UMR5506, Université de Montpellier – CC 477, 161 rue Ada, F-34 095 Montpellier Cedex 5, France
[2] Parcours Bioinformatique, Connaissances et Données du Master Sciences et Numériques pour la Santé (Master SNS - Parcours BCD) – Université de Montpellier – 163 rue Auguste Broussonnet, F-34 090 Montpellier, France

Au milieu des années 2000, l'apparition des techniques de séquençage à haut-débit a provoqué une augmentation fulgurante de la production de données génomiques et transcriptomiques. Depuis, la taille et l'abondance de ces données ne cessent d'augmenter, ce qui requiert le développement de nouveaux outils bioinformatiques pour les analyser efficacement. Ces grands jeux de données permettent aussi d'envisager des études phylogénétiques de grande ampleur, ce qui nécessite une amélioration constante des outils bioinformatiques dédiés.

Parmi ces outils, les logiciels de visualisation d'arbres permettent de représenter graphiquement les phylogénies obtenues à partir de ces jeux de données. Ces dernières années, les innovations en matière d'outils de visualisation d'arbres ont étés principalement portées sur leur capacité à supporter et à afficher de grandes phylogénies, sur les possibilités d'affichage pour faciliter l'interprétation et sur la gestion des données d'annotation. Les outils proposant les fonctionnalités les plus avancées sont des logiciels lourds que l'utilisateur doit installer localement sur un ordinateur. Cependant, afin de rendre ces outils accessibles au plus grand nombre, des logiciels permettant une utilisation à distance ont été développés. Ils sont souvent mis en oeuvre au sein des plate-formes de bioinformatique.

La plate-forme ATGC (`http://www.atgc-montpellier.fr/`) de l'IFB (Institut Français de Bioinformatique), adossée à l'équipe MAB du LIRMM (Méthodes et Algorithmes pour la Bioinformatique), est dédiée à la bioinformatique pour la génomique évolutive comparative et fonctionnelle. Parmi les outils disponibles sur ATGC, l'applet java Archaeopteryx est utilisée pour la visualisation d'arbres phylogénétiques (Han et Zmasek, 2009). Néanmoins, pour des raisons de sécurité, les navigateurs modernes ont tendance à refuser ou à limiter l'exécution de ce type d'applet. Pour assurer le bon fonctionnement de la plate-forme ATGC, il est donc primordial de remplacer Archaeopteryx.

Notre objectif est de trouver un outil proposant une interface simple et agréable d'utilisation, tout en permettant de visualiser et de manipuler aisément les arbres. Cet outil devra constituer une alternative à Archaeopteryx. Nous avons tout d'abord réalisé une étude bibliographique et une veille technologique sur les outils de visualisation d'arbres phylogénétiques. Ainsi, 25 outils ont été identifiés et analysés selon différents critères discriminants. Ces derniers portent sur la manipulation des arbres ainsi que sur certaines caractéristiques des outils :

- La représentation de la phylogénie sous plusieurs formes (axe, type de diagramme…).
- La modification de l'arbre : regroupement de nœuds (« collapsing »), permutation de branches (« swapping »), permutation automatique de branches dans un ordre de profondeur croissant ou décroissant (« ladderizing »), enracinement de l'arbre (« re-root »), coloration des nœuds et branches…
- L'ajout d'annotations graphiques ou textuelles sur les arbres.

*. Intervenant
†. Corresponding author : `Vincent.Lefort@lirmm.fr`

— La prise en charge des formats standards Newick (Olsen, 1990), Nexus (Maddison et Maddison, 1997) et PhyloXML (Han et Zmasek, 2009).
— La production de fichiers de sortie au format vectoriel (PDF, SVG) et bitmap (PNG, JPEG).
— L'utilisation de l'outil doit rester fluide même lors de la visualisation de grands arbres.
— L'existence d'une communauté active de développeurs.

Ce premier filtre nous a permis de sélectionner 8 outils candidats pour une intégration sur ATGC : The Newick Utilities (Junier et Evgeny, 2010), BioJS (Yachdav et al., 2015), JsPhyloSVG (Smits et Ouverney, 2010), FigTree (Rambaut, 2014), Mesquite (Maddison et Maddison, 2015), T-REX (Boc et al., 2012), ETE Toolkit (Huerta-Cepas et al., 2010) et Phylo.io (Robinson et al., 2016).

Ces outils ont étés par la suite étudiés en détail afin de déterminer lequel était le plus adapté à nos besoins. Afin d'assurer une bonne interactivité avec les utilisateurs, nous avons choisi de favoriser les outils de visualisation s'exécutant côté client, et donc basés sur la technologie Javascript. Nous souhaitons par ailleurs pouvoir personnaliser l'interface graphique afin d'assurer une intégration cohérente au sein de la plate-forme ATGC. Cette contrainte nous a permis de privilégier les outils mettant à disposition des développeurs un ensemble de modules ou bibliothèques dédiés aux données biologiques.

Ainsi, notre étude nous a conduit à identifier l'outil communautaire BioJS comme étant le plus adapté pour une intégration sur ATGC (Yachdav et al., 2015). Il s'agit d'une suite de bibliothèques Javascript dédiées à la visualisation et à la manipulation de données biologiques, notamment les arbres phylogénétiques. L'une des particularités de BioJS réside dans la facilité de créer et de partager des bibliothèques par le biais du Node Package Manager (npm) de Node.js (Tilkov et Vinoski, 2010). Ce procédé reflète l'existence d'une communauté active où chaque bibliothèque est régulièrement mise à jour tout en étant vérifiée par les utilisateurs. Ce sont plus précisement les bibliothèques tnt.tree, tnt.vis et tnt.tree.node qui sont utilisées pour construire un outil correspondant aux besoins de la plate-forme ATGC.

Remerciements

Cette participation à la 17e édition des Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM) a été financée grâce au département informatique de la Faculté des Sciences de l'Université de Montpellier (`http://deptinfofds.univ-montp2.fr/`) et au Labex Numev (`http://www.lirmm.fr/numev/`).

Références

Boc A., Diallo A.B, et Makarenkov V., 2012. "T-REX: A Web Server for Inferring, Validating and Visualizing Phylogenetic Trees and Networks". *Nucleic Acids Research* 40 (W1): W573-79. doi:10.1093/nar/gks485.

Han M.V., et Zmasek C.M., 2009. "phyloXML: XML for evolutionary biology and comparative genomics". *BMC Bioinformatics* 10 : 356. doi:10.1186/1471-2105-10-356.

Huerta-Cepas, J., Dopazo J., et Gabaldón T. 2010. "ETE: a python Environment for Tree Exploration". *BMC Bioinformatics* 11: 24. doi:10.1186/1471-2105-11-24.

Junier T., et Zdobnov E.M., 2010. "The Newick Utilities: High-Throughput Phylogenetic Tree Processing in the UNIX Shell". *Bioinformatics (Oxford, England)* 26 (13): 1669-70. doi:10.1093/bioinformatics/btq243.

Maddison D.R., et Maddison W.P., 1997. "NEXUS: an extensible file format for systematic information". *Systematic Biology* 46 590-621

Maddison, W. P., et Maddison D. R., 2015. "Mesquite: a modular system for evolutionary analysis. Version 2.75. 2011". [`http://mesquiteproject.org/`]

Olsen G., 1990. "Newick's 8:45, Tree Format Standard" [http://evolution.genetics.washington.edu/phylip/newick_doc.html]

Rambaut, A., 2014. "Figtree" [http://tree.bio.ed.ac.uk/software/figtree/]

Robinson O., Dylus D., et Christophe Dessimoz C., 2016. "Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web". *Molecular Biology and Evolution*, msw080. doi:10.1093/molbev/msw080.

Smits, S.A., et Ouverney C., 2010. "jsPhyloSVG: A Javascript Library for Visualizing Interactive and Vector-Based Phylogenetic Trees on the Web". *PLoS ONE* 5 (8):e12267. doi:10.1371/journal.pone.0012267.

Tilkov S., Vinoski S., 2010. "Node.js: Using JavaScript to Build High-Performance Network Programs". *IEEE Internet Computing*, v.14 n.6, p.80-83 doi:10.1109/MIC.2010.145

Yachdav G., Goldberg T., Wilzbach S., Dao D., Shih I., Choudhary S., Crouch S., et al. 2015. "Anatomy of BioJS, an Open Source Community for the Life Sciences". *eLife* 4:e07009. doi:10.7554/eLife.07009.

Mots clefs : arbre phylogénétique, visualisation, plate, forme ATGC, outil informatique

Whole genome duplications shaped the receptor tyrosine kinase repertoire of jawed vertebrates

Frédéric Brunet [*] [†1], Jean-Nicolas Volff[2], Manfred Schartl[3]

Poster 81

[1] Institut de Génomique Fonctionnelle de Lyon (IGFL) – CNRS : UMR5242, Institut national de la recherche agronomique (INRA) : UA1288, Université Claude Bernard - Lyon I (UCBL), École Normale Supérieure (ENS) - Lyon – École Normale Supérieure de Lyon, 46 allée d'Italie, F-69 364 Lyon Cedex 07, France
[2] Institut de Génomique Fonctionnelle de Lyon (IGFL) – École Normale Supérieure [ENS] - Lyon – 46, allée d'Italie, F-69 364 Lyon Cedex 07, France
[3] University of Würzburg – Physiologische Chemie, Biozentrum, University of Würzburg, Am Hubland, and Comprehensive Cancer Center, University Clinic Würzburg, Josef Schneider Straße 6, 97074 Würzburg, Allemagne

The receptor tyrosine kinase (RTK) gene family, involved primarily in cell growth and differentiation, comprises proteins with a common enzymatic tyrosine kinase intracellular domain adjacent to a transmembrane region. The amino-terminal portion of RTKs is extracellular and made of different domains, the combination of which characterizes each of the 20 RTK subfamilies among mammals. We analyzed a total of 7376 RTK sequences among 143 vertebrate species to provide here the first comprehensive census of the jawed vertebrate repertoire. We ascertained the 58 genes previously described in the human and mouse genomes and established their phylogenetic relationships. We also identified five additional RTKs amounting to a total of 63 genes in jawed vertebrates. We found that the vertebrate RTK gene family has been shaped by the two successive rounds of whole genome duplications (WGD) called 1R and 2R (1R/2R) that occurred at the base of the vertebrates. In addition, the Vegfr and Ephrin receptor subfamilies were expanded by single gene duplications. In teleost fish, 23 additional RTK genes have been retained after another expansion through the fish-specific third round (3R) of WGD. Several lineage-specific gene losses were observed. For instance, five RTKs were lost during the evolution of the tetrapods, carnivores lost another one, birds have lost three RTKs, and different genes are missing in several fish sublineages. The RTK gene family presents an unusual high gene retention rate from the vertebrate WGDs (58.75 % after 1R/2R, 64.4 % after 3R), resulting in an expansion that might be correlated with the evolution of complexity of vertebrate cellular communication and intracellular signaling.

Mots clefs : RTK

*. Intervenant
†. Corresponding author : frederic.brunet@ens-lyon.fr

CompPhy v2 : une plate-forme collaborative pour visualiser et comparer des phylogénies

Floréal Cabanettes [*1], Marc Chakiachvili[1], Vincent Lefort[1], Jean-François Dufayard[2], Frédéric De Lamotte[3], Nicolas Fiorini[4], Vincent Berry[1], Anne-Muriel Arigon Chifolleau [†1]

Poster 82

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – CNRS : UMR5506, Université de Montpellier – Campus St Priest, Bâtiment 5, 860 rue de St Priest, CC 05016, F-34 095 Montpellier Cedex 5, France
[2] CIRAD UMR AGAP (AGAP) – Montpellier SupAgro, Institut national de la recherche agronomique (INRA) : UMR1334, CIRAD-BIOS – TA A-108/03, Avenue Agropolis, F-34 398 Montpellier Cedex 5, France
[3] INRA UMR AGAP (AGAP) – Montpellier SupAgro, Institut national de la recherche agronomique (INRA) : UMR1334, CIRAD-BIOS – TA A-108/03, Avenue Agropolis, F-34 398 Montpellier Cedex 5, France
[4] Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P) – École Nationale Supérieure des Mines d'Alès – Site EERIE Parc Scientifique Georges Besse, F-30 035 Nîmes Cedex 1, France

Introduction

Les outils collaboratifs sont très utilisés dans le cas de projets impliquant des partenaires travaillant à distance. Depuis quelques années, les technologies du Web ont permis la construction de tels outils pour éditer conjointement des documents bureautiques et des données scientifiques, mais actuellement, aucun n'est disponible pour le traitement de phylogénies. Bien qu'un grand nombre d'études et de projets en biologie et en systématique évolutive impliquent des collaborations entre scientifiques de différents instituts, les logiciels et sites Web de visualisation et de comparaison d'arbres phylogénétiques existants (TreeView [1], TreeDyn [2], Dendroscope [3], Archaeopteryx [4], FigTree [5], iTOL [6], EvolView [7] and PhyloIO [8]) proposent un accès mono-utilisateur. En outre, les fonctionnalités liées à la comparaison des arbres sont dispersées entre différents logiciels qui, pour la plupart, se concentrent principalement sur des fonctionnalités de visualisation de haut niveau pour un arbre unique, au détriment des fonctions basiques de comparaison d'arbres. Pour répondre à ces besoins, nous avons proposé en 2014 la plate-forme Web CompPhy permettant le travail collaboratif sur des phylogénies et rassemblant des outils dédiés à la comparaison d'arbres.

CompPhy propose par une approche collaborative de visualiser et de comparer des arbres phylogénétiques. La première version de cette application Web [9] est actuellement disponible sur la plate-forme ATGC (`http://www.atgc-montpellier.fr/compphy/`). Dans sa version actuelle, elle offre des fonctionnalités pour l'édition et la comparaison d'arbres phylogénétiques, l'inférence de super-arbres et la gestion de ces données dans un environnement collaboratif. Cette plate-forme permet à plusieurs utilisateurs de gérer de manière synchrone ou asynchrone des arbres phylogénétiques. Cette première version permet à tous les collaborateurs d'un projet de voir les actions effectuées par un des partenaires sur les arbres partagés, et ce en temps réel, facilitant ainsi le travail à distance entre plusieurs personnes. En outre, CompPhy est un outil unique regroupant des opérations de comparaison d'arbres telles que la restriction de deux arbres à leurs taxons communs, la permutation automatique des branches autour des nœuds (« automated branch swaps ») afin de mettre en correspondance les feuilles de deux arbres, le calcul de consensus d'arbres et de super-arbres. L'application propose deux façons de manipuler les arbres : soit en

*. Intervenant
†. Corresponding author : `arigon@lirmm.fr`

choisissant deux arbres pour une visualisation détaillée en face à face, soit en traitement par lots par des opérations sur plusieurs centaines d'arbres. Enfin, cette première version offre une interface proposant les fonctionnalités habituelles d'édition d'arbres, telles que la coloration de feuilles ou de sous-arbres, la gestion des annotations, la permutation de sous-arbres et le ré-enracinnement.

La nouvelle version de CompPhy vise à améliorer les fonctionnalités existantes et à en proposer de nouvelles, notamment pour répondre aux besoins des utilisateurs biologistes. Cette nouvelle version sera mise en ligne pour JOBIM 2016.

Limitations de la version actuelle de CompPhy

Plusieurs aspects de la première version de CompPhy limitent son utilisation, notamment la fluidité des actions de l'utilisateur, la gestion des arbres et de leurs annotations. L'aspect collaboratif manque aussi de certaines fonctionnalités.

Un aspect important pour une application Web est la fluidité des actions pour l'utilisateur. Dans la première version de CompPhy, l'architecture impose que les images des arbres soient générées par le programme ScripTree [10], exécuté sur le serveur Web. Cette architecture limite considérablement la fluidité du site Web, notamment avec de grands arbres phylogénétiques car le temps de calcul de l'image dépend notamment du nombre de taxons. La visualisation de ces grands arbres pose également des problèmes. Ainsi une fonctionnalité de masquage de sous-arbres (« collapse ») manuel et automatique doit être proposé aux utilisateurs.

Par ailleurs, dans sa première version, CompPhy impose d'importer des arbres au format Newick et n'accepte que des annotations au format ScripTree. Il est donc nécessaire, d'une part, de permettre l'importation d'arbres de formats différents tels que NEXUS et NHX, et d'autre part d'améliorer la gestion de ces annotations afin de proposer une approche standard et automatisée à l'utilisateur.

L'aspect collaboratif est lui aussi primordial dans l'approche de CompPhy. Dans la première version, la gestion collaborative est faite au niveau d'un projet dans sa globalité : le contrôle total sur le projet est transféré d'un utilisateur à un autre. En effet, une gestion de demande de contrôle du projet permet à chacun de pouvoir manipuler à tour de rôle les arbres du projet. Cependant, les premiers retours montrent qu'il est indispensable d'affiner cette granularité et de gérer l'accès concurrentiel au niveau de chaque arbre plutôt qu'au niveau du projet.

Enfin, des fonctionnalités semblent manquer dans cette première version. Par exemple, la possibilité d'utiliser l'application Web en tant qu'utilisateur anonyme est importante. La création d'une API (Application Programming Interface) CompPhy constitue également une demande des utilisateurs, permettant ainsi de connecter automatiquement la plate-forme Web à une application existante.

Présentation de la nouvelle interface (CompPhy v2)

Lors de la création d'un projet, l'utilisateur peut fournir un ou plusieurs fichiers contenant des arbres. Ces fichiers peuvent maintenant être au format Newick, NEXUS et NHX. Les annotations peuvent également être fournies dans un fichier tabulé. Une fois les données importées et le projet créé, l'utilisateur peut visualiser ses arbres sur la page principale de CompPhy (voir Figure 1). Dans le cas de grands arbres phylogénétiques, une vision condensée masquant automatiquement certains sous-arbres est proposée. Par la suite, il peut manipuler les arbres du projet puis les exporter sous divers formats.

Deux modes de connexion sont possibles dans CompPhy : l'utilisateur peut utiliser l'interface directement en créant un projet en restant anonyme (ses données ne sont pas sauvegardées) ou en créant un compte (ses données sont conservées). En mode connecté, il peut partager ses projets avec d'autres utilisateurs et utiliser les fonctionalités de travail collaboratif de l'application.

La zone en haut de la page principale (nommée « Tree collections 1 / 2 » dans la Figure 1) présente la liste des arbres importés, répartis en deux collections. Depuis ces collections, l'utilisateur sélectionne les deux arbres qu'il souhaite afficher en face à face dans les espaces de travail de la zone centrale afin de les comparer.

Ces deux espaces de travail (nommée « Workbenches » dans la Figure 1) proposent un ensemble de fonctionnalités d'édition et de comparaison des deux arbres. Un utilisateur peut prendre le contrôle de certains arbres d'un projet, ces arbres ne sont alors plus éditables par les autres partenaires du projet. Au centre de ces deux espaces de travail, on retrouve un ensemble d'outils applicables aux deux arbres : un outil permettant de personnaliser les arbres affichés (couleur et épaisseur des branches, coloration de sous-arbres, masquage de sous-arbres (« collapse »), permutation de deux nœuds (« swap »)) ; un autre pour paramétrer le zoom applicable aux arbres affichés ; et un ensemble de fonctionnalités de comparaison d'arbres (calcul des arbres consensus, calcul de la distance entre les topologies des deux arbres, ou encore la permutation automatique des branches). Sous chaque espace de travail, les utilisateurs peuvent modifier les annotations de l'arbre (« Side annotations » dans la Figure 1) et modifier manuellement l'arbre au format Newick (« Manual newick editing » dans la Figure 1).

Dans la zone en bas de la page (nommée « Miscellaneous tools » dans la Figure 1), on retrouve des outils de gestion de projet et un ensemble d'outils applicables à plusieurs arbres du projet simultanément, notamment pour renommer des arbres ou des taxons, ré-enraciner des arbres en fonction d'un groupe externe précis et inférer des super-arbres. Pour ce dernier outil, nous utilisons les services de phylogénomiques présents sur la plate-forme ATGC. Enfin, nous retrouvons une barre d'outils latérale facilitant les aspects collaboratifs. Elle permet de modifier rapidement certains paramètres utilisateur applicables au projet et d'avoir accès aux différentes notifications et requêtes en attente de réponse (telles que les demandes de contrôle). Enfin, une messagerie instantanée est proposée pour faciliter la discussion entre les utilisateurs.

Implémentation de CompPhy v2

Fluidité de la navigation et interactivité avec l'utilisateur

Afin d'améliorer la fluidité des actions de l'utilisateur, les images des arbres sont directement générées par le navigateur de l'utilisateur. Nous utilisons pour cela la librairie javascript D3.js, une référence dans la création de graphiques pour le Web. Par ailleurs, D3.js nous permet d'améliorer considérablement l'interactivité des arbres avec l'utilisateur ainsi que la navigation sur le site Web.

Aspect collaboratif

Afin de faciliter la synchronisation des données entre différents utilisateurs d'un même projet connectés simultanément, nous utilisons la technologie Websocket par le biais de Ratchet, une librairie PHP. Cette technologie permet au serveur d'envoyer aux clients des informations spontanément sans que le client n'ait effectué de requête. Pour cela, lorsqu'un utilisateur se connecte à CompPhy, en plus de la gestion d'une session PHP, le client établit une connexion Websocket avec le serveur. Celle-ci est utilisée pour assurer la synchronisation des données entre les utilisateurs (voir Figure 2).

API

Nous avons développé une API REST qui permet de se connecter à CompPhy automatiquement depuis n'importe quelle application.

Conclusion et perspectives

CompPhy est une plate-forme Web qui permet de manipuler, visualiser et comparer des arbres phylogénétiques dans un environnement collaboratif. La version 2 apporte de nouvelles fonctionnalités et améliore considérablement celles existantes. Les principaux développements concernent l'interactivité et la fluidité de l'interface ainsi que les aspects collaboratifs. Pour ces derniers, la granularité est notamment affinée pour permettre une gestion au niveau de chaque arbre. Parmi les évolutions à venir, nous envisageons un service d'inférence d'arbres disponible dans CompPhy, un service de réconciliation d'arbres (comparaison d'arbres de gènes et d'espèces par exemple) et une gestion des versions des arbres.

Remerciements

Ce travail a bénéficié d'une subvention de l'état français gérée par l'Agence Nationale de la Recherche Française sous un Investissement pour le programme Future par le biais du LabEx NUMEV (référence n◦ANR-10-LABX-20).

Références

[1] Page, R. D. Visualizing phylogenetic trees using TreeView. *Curr Protoc Bioinformatics* Chapter 6, (2002).

[2] Chevenet, F., Brun, C., Banuls, A. L., Jacq, B. and Christen, R. TreeDyn : towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* 7, (2006).

[3] Huson ; D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M. and Rupp, R. Dendroscope : An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8, (2007).

[4] Han, M. V. and Zmasek, C. M. phyloXML : XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 10, (2009).

[5] Rambaut A. [`http://tree.bio.ed.ac.uk/software/figtree/`].

[6] Letunic, I. and Bork, P. Interactive Tree Of Life v2 : online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39, (2011).

[7] Zhang, H., Gao, S., Lercher, M. J., Hu, S. and Chen, W. H. EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Res* 40, (2012).

[8] Robinson, O., Dylus, D. and Dessimoz, C. Phylo.io : interactive viewing and comparison of large phylogenetic trees on the web. *arXiv* 1602.04258 [q-bio] (2016).

[9] Fiorini, N., Lefort, V., Chevenet, F., Berry, V. and Chifolleau, A.-M. A. CompPhy : a web-based collaborative platform for comparing phylogenies. *BMC Evolutionary Biology* 14, 253 (2014).

[10] Chevenet, F., Croce, O., Hebrard, M., Christen, R. and Berry, V. ScripTree : scripting phylogenetic graphics. *Bioinformatics* 26, (2010).

Mots clefs : Arbres phylogénétiques, Comparaison d'arbres, Ressource en ligne, Collaboration en temps réel

Emergence d'un clone de *Legionella pneumophila* subsp. *non-pneumophila* ST701

Amandine Campan-Fournier [*1,2], Christophe Ginevra[3,4], Christine Oger[1], Frédéric Jauffrit[2,5], Vincent Navratil[1], Anne-Gaëlle Ranc[3,4], Guy Perrière[1,2], Sophie Jarraud[3,4]

Poster 83

[1] Pôle Rhône-Alpes de Bioinformatique (PRABI) – Université Claude Bernard - Lyon I (UCBL) – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL) – 43 bd du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[3] Hospices Civils de Lyon, Centre National de Référence des Légionelles (HCL - CNR) – Hospices Civils de Lyon – 59 boulevard Pinel, F-69 500 Bron, France
[4] Centre International de Recherche en Infectiologie (CIRI) – École Normale Supérieure (ENS) - Lyon, Université Claude Bernard - Lyon I (UCBL), CNRS : UMR5308, Inserm : U1111 – 46 allée d'Italie, F-69 364 Lyon Cedex 07, France
[5] bioMérieux, Département de Recherche Technologique – BIOMÉRIEUX – 376 chemin de l'Orme, F-69 280 Marcy l'Étoile, France

La légionellose, ou maladie des légionnaires, est caractérisée par une pneumonie le plus souvent sévère causée par des bactéries du genre *Legionella*. Les *Legionella* sont des bactéries hydrotelluriques ubiquitaires qui ont la capacité de se multiplier chez un hôte intracellulaire (amibes, macrophages chez l'Homme). La voie de contamination la plus courante est l'inhalation d'aérosols contaminés, dont les sources peuvent être multiples : douches, tours aéroréfrigérantes, bassins et fontaines décoratives, eaux thermales, jacuzzi… [1, 2, 3] En France, environ 1 300 cas de légionellose sont recensés par an [3] et 98 % des patients doivent être hospitalisés [4]. Quand le traitement est bien conduit, l'évolution est favorable dans la majorité des cas [4]. Le décès survient cependant dans 5 à 10 % des cas [1].

Parmi la soixantaine d'espèces du genre *Legionella*, l'espèce *Legionella pneumophila* est responsable de plus de 90 % des cas de légionellose. Elle est subdivisée en trois sous-espèces : *L. pneumophila* subsp. *pneumophila*, *L. pneumophila* subsp. *fraseri* et *L. pneumophila* subsp. *pascullei*. Les infections impliquent principalement la sous-espèce *L. pneumophila* subsp. *pneumophila* [5]. Néanmoins, en l'absence d'outil simple permettant de distinguer les 3 sous-espèces dans les laboratoires réalisant le diagnostic, l'implication des deux autres sous-espèces dans les cas de légionellose est très peu connue.

Dans le cadre d'une analyse de génomique comparative préliminaire portant sur 11 génomes de référence et 32 génomes nouvellement séquencés, nous avons remarqué qu'une des souches séquencées, la souche HL 0641 3006 (ST701), semble être évolutivement plus proche des sous-espèces *fraseri* et *pascullei* que de la sous-espèce *pneumophila*. Les cas de légionellose dus à des souches appartenant à ce même ST701 semblent en émergence en France : un seul cas avait été recensé entre 2006 et 2008, puis 15 cas entre 2009 et 2011 et 25 cas entre 2012 et 2014 [6].

Nous avons entrepris une analyse phylogénétique afin de mieux caractériser les souches ST701. Les séquences des 45 protéines ribosomiques en unicopie de 80 génomes de *Legionella*, *Tatlockia*, *Fluoribacter* et *Coxiella* ont été extraites de la banque RiboDB [7] et alignées avec le programme MAFFT [8]. Les 45 alignements multiples ont été filtrés avec l'outil Gblocks [9] disponible dans SeaView [10] afin d'éliminer les régions mal alignées, puis concaténés. Une procédure de sélection de modèles a été utilisée pour déterminer le modèle le plus adapté aux données, en utilisant l'outil

*. Intervenant

ProtTest [11]. Le test AIC (*Akaike Information Criterion*) [12] utilisé au cours de cette procédure nous a permis de déterminer que le modèle le plus approprié était le LG (Le et Gascuel) [13] avec correction par la loi Gamma, prise en compte des invariants et utilisation des fréquences observées des acides aminés (LG + Γ4 + I + F). L'arbre phylogénétique a été reconstruit avec ce modèle d'évolution, une méthode de maximum de vraisemblance et en calculant le support des branches par aLRT (*approximate Likelihood Ratio Test*) [14], grâce au logiciel PhyML [15]. Enfin, il a été mis en forme avec TreeGraph2 [16].

L'arbre phylogénétique ainsi obtenu montre que les souches de *L. pneumophila* subsp. *fraseri* et *L. pneumophila* subsp. *pascullei* forment un clade fortement soutenu dans lequel se positionnent également les souches ST701. Au sein de ce clade, les souches de la sous-espèce *pascullei* forment un groupe monophylétique et les deux souches ST701 sont regroupées à proximité des souches de la sous-espèce *fraseri*.

Ces observations suggèrent pour la première fois que les souches ST701 pourraient être un clone émergent de *L. pneumophila* subsp. non-*pneumophila*.

Références

[1] Aide-mémoire Légionellose, Organisation Mondiale de la Santé (OMS), novembre 2014. `http://www.who.int/mediacentre/factsheets/fs285/fr/`.

[2] Légionellose, Agence Régionale de la Santé d'Île de France. `http://www.ars.iledefrance.sante.fr/Legionellose.93615.0.html`.

[3] Caractéristiques épidémiologiques des cas de légionellose en 2014 en France et en Europe. Campèse C. Communication orale à SympoLegio, Lyon, novembre 2015.

[4] Détection pulmonaire chronique de *Legionella* : échecs thérapeutiques, récidives ou réinfections ? Jarraud S. et Ginevra C. Communication orale à SympoLegio, Lyon, novembre 2015.

[5] *Legionella pneumophila* serogroup Lansing 3 isolated from a patient with fatal pneumonia, and descriptions of *L. pneumophila* subsp. *pneumophila* subsp. nov., *L. pneumophila* subsp. *fraseri* subsp. nov., and *L. pneumophila* subsp. *pascullei* subsp. nov. Brenner DJ, Steigerwalt AG, Epple P, Bibb WF, McKinney RM, Starnes RW, Colville JM, Selander RK, Edelstein PH et Moss CW. *Journal of Clinical Microbiology*, 1988, 26(9):1695-703.

[6] Données privées du Centre National de Référence des légionelles, Ginevra C, Lyon, 2015.

[7] RiboDB the Prokaryotes Ribosomal Proteins Database. Jauffrit F, Penel S, Delmotte S, Rey C, de Vienne DM, Gouy M, Charrier J-P, Flandrois J-P et Brochier-Armanet C. `https://ribodb.univ-lyon1.fr/ribodb/ribodb-in.cgi`.

[8] MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Katoh K et Standley DM. *Molecular Biology and Evolution*, 2013, 30(4):772–780.

[9] Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Castresana J. *Molecular Biology and Evolution*, 2000, 17(4):540-552.

[10] SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Gouy M, Guindon S et Gascuel O. *Molecular Biology and Evolution*, 2010, 27(2):221-224.

[11] ProtTest 3: fast selection of best-fit models of protein evolution. Darriba D, Taboada GL, Doallo R et Posada D. *Bioinformatics*, 2011, 27(8):1164-1165.

[12] A new look at the statistical model identification. Akaike H. *IEEE Transactions on Automatic Control*, 1974, 19(6):716-723

[13] An improved general amino acid replacement matrix. Le SQ et Gascuel O. *Molecular Biology and Evolution*, 2008, 25(7):1307-1320

[14] Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Anisimova M et Gascuel O. *Systematic Biology*, 2006, 55(4):539-552.

[15] A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Guindon S and Gascuel O. *Systematic Biology*, 2003, 52(5):696-704.

[16] TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. Stöver BC et Müller KF. *BMC Bioinformatics*, 2010, 11:7.

Mots clefs : Legionella pneumophila, émergence, ST701, phylogénie

Influence of transposable elements on the fate of duplicate genes in human

Margot Corréa [*] [†1], Cyril Dalmasso[1], Emmanuelle Lerat[2], Carène Rizzon[1]

Poster 84

[1] Laboratoire de Mathématiques et Modélisation d'Évry (LaMME) – Université d'Évry-Val d'Essonne, CNRS : UMR8071, ENSIIE, Institut national de la recherche agronomique (INRA) – 23 boulevard de France, F-91 037 Évry Cedex, France

[2] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – Université de Lyon, CNRS : UMR5558 – 43 boulevard du 11 novembre 1918, F-69 622 Villeurbanne Cedex, France

Since 1970 with Susumu Ohno's work [1], it is widely agreed that gene duplication is an engine for the apparition of new functions. Duplicate genes can primarily result from unequal crossing over, retroposition or polyploidy [2]. The fate of most duplicate genes are lost by pseudogeneisation or by deletion [3], however they can represent a significant part of genomes. For instance in mammals, about 65% of genes are duplicated in human and mouse genomes [4]. Three main mechanisms are proposed to explain the maintenance of duplicate genes in genomes. Neofunctionalization, when one of the copies evolve towards a novel function and the other retains the ancestral function, functional redundancy when the ancestral function is maintained in both copies and subfonctionalization when the two copies are necessary to provide the ancestral function. Even though today many studies try to decipher these mechanisms, they are still poorly understood.

TEs are repeated genomic sequences that have the intrinsic capacity to multiply and move within genomes. They are a major component of eukaryotic genomes, they represent approximately 45 % of the human genome [5,6]. TEs can be divided into two major classes based on their mechanism of transposition. DNA transposon move through DNA intermediate ("cut and paste" transposition mechanism) and retrotransposons transpose through RNA intermediates according to a "copy and paste" mechanism. Retrotransposons can be subdivided into two groups distinguished by the presence or absence of long terminal repeats (LTRs) [7]. In human, among the Non-LTR retrotransposons, we can distinguish the autonomous long interspersed nuclear elements (LINEs) and the non-autonomous short interspersed nuclear elements (SINEs). TEs are known to be implied in genome evolution [8, 9] and to have an influence on genome structure. Given the fact that they are repeated elements, they can induce chromosomal rearrangements between TE homologous regions. When TEs are inserted near genes (or regulatory regions) they can modify gene regulation. For example, the insertion of TEs into or near the promoter region can alter the normal pattern of expression [6]. Also, according to Lerat and Semon [10], the presence of TEs in the vicinity of genes can influence gene expression in different conditions, notably in cancer. TEs are not randomly distributed in genomes. Their distribution can rely on recombination rates, gene density and selective pressure. Epigenetics mechanisms are implied in TEs regulation. This can induce either TE silencing (defending genome against proliferation of TEs) or TE transcription (inducing the effects of transposition) [11]. It has been shown that TEs are associated with DNA methylation [12]. Indeed, Weisenberger et al. [12] have demonstrated by MethyLight assay technology, a technology to evaluate DNA methylation of repeatitive element, that methylation levels of TEs is signifi- cantly correlated with the global DNA methylation levels (in human). Most interestingly, in a recent study it has been shown that epigenetic modifications, such as DNA methylation, may contribute to duplicate gene evolution [13]. Keller et al. [13] examined human duplicate genes (young and old) associated with promoter DNA methylation

*. Intervenant

†. Corresponding author : margot.correa@genopole.cnrs.fr

levels, DNA methylation divergence and the gene expression divergence. They found that promoters of young duplicate genes are more methylated in both copies than old duplicate genes, and the differences in DNA methylation of duplicate genes are significantly correlated with functional differences in expression. In view of all these contributions one may ask whether the TE context on the vicinity of genes may influence the maintenance of duplicate genes. However, knowing that to our knowledge, no study has been conducted concerning the link between TE context and duplicate genes. Here we propose to decipher the relationship between the maintenance of duplicate genes and TEs in human. For that, we aim to answer the following questions : is TE context (in term of density and TE classes) is different between duplicate and singleton genes ? Is the observed pattern between duplicate and singleton genes can be explained by the selective pressure and GC content ?

To retrieve the duplicate genes, we used the different human and chimpanzee homologous gene families from the HOGENOM database [14]. This database (`http://pbil.univ-lyon1.fr/databases/hogenom/home.php`) provides users homologous genes families from fully sequenced genomes (bacteria, archeaea and eukarya). We identified duplicate and singleton genes, considering that duplicate gene is defined as more than one gene, of a same species, in a homologous gense families. Among the 19,582 human genes from HOGENOM, we identified 8,421 singletons genes and 11,161 duplicates genes. RepeatMasker (`http://www.repeatmasker.org/`) is a program which search the occurence of query sequences that matches in a library of known repeat families like RepBase. The human TE positions was obtained from RepeatMasker results and available at UCSC (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/`). In order to parse the RepeatMasker out- put files and assemble multiple hits corresponding to a tranposable element, a perl tool "One code to find them all" [15] was used. Genes positions were obtained from hg19, GRCh37.p10 using Ensembl MartView (`http://www.ensembl.org/`). To determine the TE context, for each gene, we considered the 2Kb flanking regions upstream and downstream of the gene in order to consider the promoter regions of genes. 772,837 TEs insertions were localized on the human genome, in the vicinity of 20,806 genes. The yn00 method from the PAML package [16], evaluates the ratio of non-synonymous to synonymous substitution rates (Ka/Ks ratios) which is a measure of the selective pressure. First, we used the Best Reciprocal hits (BRH) method to identify orthologous genes between human and chimpanzee. Then, we evaluated the Ka/Ks ratios between orthologous gene pairs. The TE density is defined as the number of TEs inserted per base pair into each gene and into 2 kb flanking regions. For each gene, the overall density in TEs was calculated as well as TE density of each of the four classes of TEs (LTR retrotransposon, LINEs, SINEs, and DNA transposon). GC content in the vicinity of each gene was calculated excluding the GC content of TEs. The relationship between TE, gene type, GC content and selective pressure context was studied using ANCOVA and logistic regression.

We first studied the relationship between the overall TE density and gene type (duplicate and singleton). As TE distribution can rely on selective pressure and as it is well-established that GC content is correlated with the distribution of repeated elements (Lander et al., 2001), we performed the analysis taking into account selective pressure and GC content effects. Results obtained from ANCOVA revealed that the mean density is significantly smaller for duplicate genes than for singletons genes. The relationship between TE density and GC content was also significant. No interaction was significant between gene type and Ka/Ks ratios, Ka/Ks ratios and GC and gene type and GC after Bonferroni correction. As it has been shown that the relationship between the TE context and the modification of the regulation of neighbor genes can depend on TE classes [10], we also performed an analysis for each TE class separately using logistic regressions. Our results indicate that the relationship between TE density and gene type is different for the four TE classes. For example, gene type was not correlated with LINEs and LTR retrotransposons. However, for DNA transposons and SINEs elements, density in the vicinity of genes is higher for singletons genes than duplicate genes. In addition, the logistic regression revealed that DNA transposon density was linked to selective pressure (P-value < 0.005) but no LINEs neither LTR retrotransposons and SINEs (P-value > 0.005). The relationship between

GC content and TE density for DNA transposons, LINEs and LTR retrotransposons was also significant.

Our result show that TE density in the vicinity of genes was found to be higher for singletons genes compared to duplicate genes. As we expected more TEs in regions with low selective pressure, it may be partly explained by weaker Ka/Ks ratios in duplicate genes compared to singleton genes [17] and sur-representation of essential genes in duplicate genes than singleton genes [17]. However, our results indicate that when adjusting for selective pressure and GC content, the gene type effect is significant, what suggests a specific relationship between TE density and gene type and a possible implication of the TE context in the maintenance of duplicated genes in genomes.

References

[1] Susumu Ohno. *Evolution by gene duplication*. Springer Science & Business Media, 2013.

[2] Jianzhi Zhang. Evolution by gene duplication : an update. *Trends in ecology & evolution*, 18(6):292–298, 2003.

[3] Michael Lynch and John S Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[4] Valia Shoja and Liqing Zhang. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Molecular Biology and Evolution*, 23(11):2134–2141, 2006.

[5] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[6] Richard Cordaux and Mark A Batzer. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10):691–703, 2009.

[7] Thomas Wicker, François Sabot, Aurelie Hua-Van, Jeffrey L Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, et al. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12):973–982, 2007.

[8] Cedric Feschotte and Ellen J Pritham. DNA transposons and the evolution of eukaryotic genomes. *Annual review of genetics*, 41:331, 2007.

[9] Margaret G Kidwell and Damon R Lisch. Transposable elements and host genome evolution. *Trends in Ecology & Evolution*, 15(3):95–99, 2000.

[10] Emmanuelle Lerat and Marie Semon. Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. *Gene*, 396(2):303–311, 2007.

[11] R Keith Slotkin and Robert Martienssen. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4):272–285, 2007.

[12] Daniel J Weisenberger, Mihaela Campan, Tiffany I Long, Myungjin Kim, Christian Woods, Emerich Fiala, Melanie Ehrlich, and Peter W Laird. Analysis of repetitive element DNA methylation by methylight. *Nucleic acids research*, 33(21):6823–6836, 2005.

[13] Thomas E Keller and V Yi Soojin. DNA methylation and evolution of duplicate genes. *Proceedings of the National Academy of Sciences*, 111(16):5932–5937, 2014.

[14] Simon Penel, Anne-Muriel Arigon, Jean-François Dufayard, Anne-Sophie Sertier, Vincent Daubin, Laurent Duret, Manolo Gouy, and Guy Perrière. Databases of homologous gene families for comparative genomics. *BMC bioinformatics*, 10(6):1, 2009.

[15] Marc Bailly-Bechet, Annabelle Haudry, and Emmanuelle Lerat. "one code to find them all" : a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA*, 5(1):1, 2014.

[16] Ziheng Yang and Rasmus Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular biology and evolution*, 17(1):32–43, 2000.

[17] Debarun Acharya, Dola Mukherjee, Soumita Podder, and Tapash C Ghosh. Investigating different duplication pattern of essential genes in mouse and human. *PloS one*, 10(3):e0120784, 2015.

Mots clefs : Gene duplication, Transposable Elements, Evolution

Exploring the dark side of phylogeny

Poster 85

Laura Do Souto [*1], Benjamin Noël[2], Pascal Bento[3], Corinne Da Silva[2], Jean-Marc Aury[2], Arnaud Couloux[2], Simone Duprat[2], Éric Pelletier[2], Jean-Luc Souciet[4], Teun Boekhout[5], Toni Gabaldón[6], Bernard Dujon[7], Betina Porcel[2,8], Patrick Wincker[2,8]

[1] Genoscope - Centre national de séquençage – CEA, Genoscope, universite rouen – 2 rue Gaston Crémieux CP5706, F-91 057 Évry Cedex, France
[2] Genoscope - Centre national de séquençage – CEA, Genoscope – 2 rue Gaston Crémieux CP5706, F-91 057 Évry Cedex, France
[3] Institut de Génétique humaine (IGH) – CNRS : UPR1142 – 141 rue de la Cardonille, F-34 396 Montpellier Cedex 5, France
[4] Université Louis Pasteur - Strasbourg 1 (ULP) – 7 rue René Descartes, F-67 084 Strasbourg, France
[5] CBS-KNAW Fungal Biodiversity Centre – Pays-Bas
[6] Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) – Espagne
[7] Génétique des génomes (GG) – Institut Pasteur de Paris – Département des biotechnologies – 25 rue du Docteur Roux, F-75 724 Paris Cedex 15, France
[8] Génomique métabolique (UMR 8030) – CEA, CNRS : UMR8030, Université d'Évry-Val d'Essonne – Genoscope, 2 rue Gaston Crémieux, F-91057 Évry Cedex, France

Advances in sequencing, assembly and syntactic annotation methods have been fast these past years, leading to the completion of the catalog of genes for many organisms. Howewer, still a large fraction of genes in annotated or even in re-annotated genomes do not have any homologs in the current databases. In fact, every genome contains 10 to 30 % of genes without similarity sequences even in the most exploited branches of the eukaryotic tree of life. These genes, named as 'taxonomically restricted genes' (TRGs), are still poorly investigated and their function are still unknown. Their absence of homology suggests their possible contribution to specific functions, probably more linked to the lifestyle of the organism bearing them, than to central functions for basic cellular processes.

How can we study such genes which do not have any associated counterparts in other organisms? Are they different from 'known' genes? How can they be identified and characterized by current methods?

We used the fungal kingdom as a model system. This Kingdom is a great model for such study, encompassing an enormous diversity of taxa with varied ecologies, life cycle strategies, and morphologies. Moreover, exploring little known areas of the fungal tree of life could give us a better understanding and a well-balanced image of the real diversity and the evolution of their genomes. So our objective was to characterize these "genetic signatures" of certain unknown branches of the fungal tree, by retracing their evolutionary history using a phylostratigraphic approach (Domazet-Lošo, Tomislav; Brajković, Josip; Tautz, Diethard (2007-01-11). After benchmark analysis in order to define the best parameters to be used in the homologs detection across the tree of life using BLAST, efforts were performed to date the emergence of genes in selected fungal genomes belonging to the *Debaryomyces* genus.

This approach uses the principle of last common ancestor and had allowed us to assign all the proteins from each species to the different phyla of the fungal taxonomic tree, recovering the evolutive history of each proteomes. Using these results we were able to identify and retrieve the proteins associated to the last phylum, the one that are taxonomically restricted to the species so the ones which have emerged more recently. Characterization of genes taxonomically-restricted

*. Intervenant

to newly *Debaryomyces* species such as *D. tyrocola* and *D. fabrii*, as well as *Priceomyces carsonii* was performed in order to compare these "young genes" data sets to "older" genes in these species. After manual inspection and validation of these datasets using a transcriptomic approach, secondary protein structure analysis were done by using *ab initio* protein structure modeling. Moreover, comparison of these selected *ab initio* structures to the structures from the PDB database allowed us to have some clues on the putative functions of these TRGs.

With this approach we found that the most of protein-coding genes in *Debaryomyces* species appeared before the divergence of the fungal kingdom and few are restricted to species. Although we highlighted in these species the presence of hundred 'genetic signatures' and comparing length of all genes through the time, we found that younger genes are significantly shorter than older genes. Furthermore, the *ab initio* modeling gave us some clues about the protein families these TRGs belong to and thus about their functions.

Mots clefs : 'taxonomically restricted genes', fungal kingdom, genetic signatures, phylostratigraphy, unknown, functions

Whole genome sequencing of the *Pteropus giganteus* genome and bioinformatic analysis of positively selected sites in bats relevant for their immuno-virologic peculiarity

Poster 86

Julien Fouret [*1,2], Magali Roche[1], Jeanne Blanchard[1,3], Noémie Aurine[2], Clément Delestre[1], François Enchéry[2], Marie Guinier[1], Kévin Dhondt[1,2], Marc Bailly-Bechet [†3], Branka Horvat [‡2], Catherine Legras-Lachuer [§1,4]

[1] Viroscan3D – Viroscan3D – France
[2] Centre International de Recherche en Infectiologie (CIRI) – École Normale Supérieure (ENS) - Lyon, Université Claude Bernard - Lyon I (UCBL), CNRS : UMR5308, Inserm : U1111 – 21 avenue Tony Garnier, 69 365 Lyon Cedex 07, France
[3] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[4] Écologie microbienne (ÉM) – Institut national de la recherche agronomique (INRA) : UR1193, CNRS : UMR5557, Université Claude Bernard - Lyon I (UCBL), École Nationale Vétérinaire de Lyon – Bâtiment Gregor Mendel (ex 741) - 4e ét., 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France

Although bats are reservoir for numerous viruses highly pathogenic for humans, including Ebola, SRAS and Nipah virus, very little is known about their immuno-virology [1]. Nevertheless, the understanding of mechanisms controlling the viral infection in bats is of major interest to explain the absence of pathogenicity in this species, in contrast to most of the others mammals, and may lead to identification of therapeutic targets in human.

Nipah virus is a *Henipavirus*, highly pathogenic and mortal for humans, which emerged in Malaysia in 1998 [2]. It is responsible for severe respiratory diseases and for encephalitis. Nipah is classified BSL4 (Biosafety Level) with a rate of mortality reaching 92 % [3]. This virus re-emerges every year in the South-East Asia [4]. The Indian flying fox *Pteropus giganteus* was described as main reservoir of numerous viruses, including the Nipah virus. However, the genome of *P. giganteus* is not sequenced yet, which prevents further advancement in this field of research [1].

Bats are the only mammals to have developed the capacity of flight during evolution. Recent studies focusing on their evolution hypothesize that the development of the flight could have led to the evolution of other unique biological features: a long life expectancy, a low ratio of tumorigenesis as well as their capacity to shelter numerous viruses without pathogenesis [5,6]. Interestingly, there are indeed very few lethal infections to bats [7]. Phylogenetic analyses demonstrated their efficiency to highlight relevant pathways for the immuno-virology of bats, based on the comparison of related fruit bat genomes. On one hand, PYHIN family has been shown absent in bats; in human, the PYHIN family is important for the regulation of the inflamasome and the interferons pathway, and PYHIN genes are acting as sensors of the intracellular DNA [6]. On the other hand, genes under positive selection in bats that have been detected are relevant in the immuno-virology. Among these genes there is NF-kB pathway, critical in regulating the inflammatory reaction [5] and its evolution in bats may be part of their immunovirologic pecuilarities.

*. Intervenant
†. Corresponding author : marc.bailly-bechet@univ-lyon1.fr
‡. Corresponding author : branka.horvat@inserm.fr
§. Corresponding author : catherine.lachuer@viroscan3d.com

To further advance in this analysis and understand better how bats control the viral infection, we aimed at identifying sites under positive selection in bat's genes, with a particular interest for genes under positive selection in *P. giganteus*. To conduct these studies, more data both on the species of interest and on the mammalian genetic background are needed. A lot of sequenced and annotated mammalian genomes are available for the background. Concerning the foreground, there is up to date ten genomes of bats available but all these genomes are not of a sufficient quality to realize phylogenetic analyses [6]. Nipah virus is mainly accommodated with *Pteropus* genus [1]. Two species of the *Pteropus* genus have a complete and well-annotated genome available on databases: *P. vampyrus* and *P. alecto* [8]. We think that having at least three species in our foreground branches is essential for the robustness of evolutionary analyses. Indeed, we cannot exclude that individuals used for genome sequencing may be subject to rare mutation making positive selection analysis biased with single nucleotide polymorphism. Then, before performing these analyses, we first sequenced and *de novo* assembled the genome of *P. giganteus*, the main reservoir of numerous viruses. This step is also essential for de novo annotation of new genes. In addition, it is of particular interest to have this genome sequenced in order to develop tools for functional studies using available *P. giganteus* biological material, which is supposed to complete this study.

In order to test our pipeline for phylogenetic analyses, we applied the branch-site method with two species of *Pteropus* (*vampyrus* and *alecto*) in the foreground branches and five other mammalian species in background branches. The functional annotation of sites identified under positive selection highlighted some changes in known motifs and domains relevant during the response to the infection. Furthermore, this method allows de novo putative annotation using known consensus of functional motif. Detection of bat specific motif involved in innate immunity would represent a very useful resource for functional characterization. To determine how the choice of the model of evolution (branch or branch-site) could be a source of variation, we analyzed all 19 genes highlighted by the team of Pr. Wang [5], with the tree from their study and alignments of the UCSC. While the majority of the results are model independent, some remain specific to a model. The analysis of parameters from different models reveals that genes detected only by the model branch-site are characterized by a lower proportion of sites under positive selection as well as higher ratio dN/dS (> 10). These two models may then raise distinct questions from an evolutionary point of view in terms of number of sites affected by the selection, as well as the intensity of this selection. The next step for us will be to complete the assembling and annotation of the *P. giganteus* genome before adding *P. giganteus* genes in our alignments.

We sequenced de novo the genome of *P. giganteus* with coverage of 51.7x. Reads have been de novo assembled in primary scaffold with SGA [9]. Our strategy for scaffolding is based on the use of one or several closely related species. Scaffold_builder is a method using the information of a reference genome for the scaffolding [10]. By using *P. alecto* as reference, only 2 long sequences (> 50 kb) of 1 Mb and 237 Mb are obtained. In comparison, 50 % of the total length of the genome of *P. alecto* is contained in 36 sequences between 15 Mb and 70 Mb; it is then surprising to obtain a long secondary scaffold of 237 Mb. Furthermore, 7 % of the long sequences (> 10 Mb) of the genome of *P. alecto* are not covered by the primary scaffolds produced by SGA [9], while 96.7 % of reads are aligned against this genome. Nevertheless, 97.3 % of reads align with the primary scaffolds with lower rates of error, mismatch and indel. This non-coverage of the genome of *P. alecto* by the primary scaffolds may be related to chromosomal rearrangements leading to structural variations between the genomes of *P. giganteus* and *P. alecto*. AlignGraph, RACA and SHEAR are methods of scaffolding based on the detection of syntenies and/or structural variations by using one or several reference genomes [11–13]. We are evaluating how far these methods are effective in our case for obtaining longer scaffolds suitable for annotation.

We built a method for detection and annotation of sites positively selected in bats. The aim is to facilitate the functional approach by reducing the number of sites to study. In the future, results of phylogenetic analysis will be compiled with the transcriptomic signatures of cellular infection in both *P. giganteus* and human cells, in order to bring a comprehensive view on how the efficient

response to Nipah virus infection works in fruit bat species.

References

[1] N. Gay, K.J. Olival, S. Bumrungsri, B. Siriaroonrat, M. Bourgarel, S. Morand, Parasite and viral species richness of Southeast Asian bats: Fragmentation of area distribution matters, *Int. J. Parasitol. Parasites Wildl.* 3 (2014) 161–170. doi:10.1016/j.ijppaw.2014.06.003.

[2] J.R.C. Pulliam, J.H. Epstein, J. Dushoff, S. a. Rahman, M. Bunning, a. a. Jamaluddin, et al., Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis, *J. R. Soc. Interface.* 9 (2012) 89–101. doi:10.1098/rsif.2011.0223.

[3] G.A. Marsh, L.F. Wang, Hendra and Nipah viruses: Why are they so deadly?, *Curr. Opin. Virol.* 2 (2012) 242–247. doi:10.1016/j.coviro.2012.03.006.

[4] B. Horvat, V. Guillaume, T.F. Wild, Les virus Nipah et Hendra?: des agents pathogènes zoonotiques émergents, *Virologie.* 11 (2007) 351–60.

[5] G. Zhang, C. Cowled, Z. Shi, Z. Huang, K.A. Bishop-Lilly, X. Fang, et al., Comparative Analysis of Bat Genomes Provides Insight into the Evolution of Flight and Immunity, *Science* (80-.). 339 (2013) 456–460. doi:10.1126/science.1230835.

[6] M. Ahn, J. Cui, A.T. Irving, L.-F. Wang, Unique Loss of the PYHIN Gene Family in Bats Amongst Mammals: Implications for Inflammasome Sensing, *Sci. Rep.* 6 (2016) 21722. doi:10.1038/srep21722.

[7] C.E. Brook, A.P. Dobson, Bats as "special" reservoirs for emerging zoonotic pathogens, *Trends Microbiol.* 23 (2015) 172–180. doi:10.1016/j.tim.2014.12.004.

[8] J. Fang, X. Wang, S. Mu, S. Zhang, D. Dong, BGD: A Database of Bat Genomes, *PLoS One.* 10 (2015) e0131296. doi:10.1371/journal.pone.0131296.

[9] J.T. Simpson, R. Durbin, Efficient de novo assembly of large genomes using compressed data structures, *Genome Res.* 22 (2012) 549–556. doi:10.1101/gr.126953.111.

[10] G.G. Silva, B.E. Dutilh, T.D. Matthews, K. Elkins, R. Schmieder, E. a Dinsdale, et al., Combining *de novo* and reference-guided assembly with scaffold_builder., *Source Code Biol. Med.* 8 (2013) 23. doi:10.1186/1751-0473-8-23.

[11] J. Kim, D.M. Larkin, Q. Cai, Asan, Y. Zhang, R.-L. Ge, et al., Reference-assisted chromosome assembly, *Proc. Natl. Acad. Sci. U. S. A.* 110 (2013) 1785–90. doi:10.1073/pnas.1220349110.

[12] S.R. Landman, T.H. Hwang, K.A.T. Silverstein, Y. Li, S.M. Dehm, M. Steinbach, et al., SHEAR: sample heterogeneity estimation and assembly by reference., *BMC Genomics.* 15 (2014) 84. doi:10.1186/1471-2164-15-84.

[13] E. Bao, T. Jiang, T. Girke, AlignGraph: Algorithm for secondary de novo genome assembly guided by closely related references, *Bioinformatics.* 30 (2014) 319–328. doi:10.1093/bioinformatics/btu291.

Mots clefs : phylogeny bats infectiology de novo

A resource on families of genes duplicated by whole-genome or small-scale duplication events in the human lineage : analysis on evolutionary, sequence and functional features

Poster 87

Solène Julien [*1,2], Margot Corréa[3], Carène Rizzon[3], François Artiguenave[2], Jean-François Deleuze[2], Vincent Frouin[1], Edith Le Floch [†2], Christophe Battail [‡2]

[1] IFR49 - Neurospin - CEA – CEA, Neurospin, IFR49 – Gif sur Yvette, France
[2] Centre National de Génotypage (CNG) – CEA – Centre National de Génotypage, 2 rue Gaston Crémieux CP5721, F-91 057 Évry Cedex, France
[3] Laboratoire de Mathématiques et Modélisation d'Évry (LaMME) – ENSIIE, CNRS : UMR8071, Université Paris-Est Créteil Val-de-Marne (UPEC), Université Paris V - Paris Descartes, Institut national de la recherche agronomique (INRA), Université d'Évry-Val d'Essonne – 23 boulevard de France, F-91 037 Évry, France

Introduction

In evolution history, duplicated genes originate from the duplication event of a common ancestor gene. In a species, a set of genes descending from one or more duplications of their ancestral gene are in the same gene family. Duplicates can come from a Small-Scale Duplication (SSD) event or from the two rounds of Whole Genome Duplication (2R-WGD) event. The WGD event occurred early in vertebrate evolution (around 450 million years ago) (Dehal & Boore, 2005) and is linked to extensive species diversity (Acharya & Ghosh, 2016). In a different manner, the SSD which is a mechanism that implies only one gene can happen at any evolutionary time. SSD events can be divided relatively to WGD period as older SSD (before the WGD period), younger SSD (after the WGD period) and WGD-old SSD (during the WGD period) (Chen et al., 2013, Singh et al., 2014 and Makino and McLysaght, 2010). After duplication events, most genes become pseudogenes or are lost because of disadvantageous changes to the organism (selection pressure). Maintained genes can share ancestral functions (subfunctionalization) or can adapt to new functions (neofunctionalization). These types of functionalization induce an increase in tissue complexity of the organism (Acharya & Ghosh, 2016). In the human genome, depending on the algorithm used, about 65 % of protein coding genes are considered as duplicated genes. According to Chen et al. duplicated genes (12346 genes) are gathered into 3692 gene families using TreeFam version 8.0 (Ruan et al., 2008) algorithm based on protein sequence similarity. Gene families have various sizes (from 2 to 57 genes) and many are small (47 % contain 2 genes). By taking in account Chen et al. and Singh et al., 6835 SSD genes and 7082 WGD genes are found.

These gene families can be characterized based on their expression profiles. Indeed expression can show the tissue specificity of an entire family or of a few duplicated genes in a particular family. In humans, the family SRGAP2 (Dennis et al., 2012) can be used as example because all its genes are specifically expressed in fetal brain and one of them is more expressed in the cerebellum. Moreover given that this family is only duplicated in the human lineage and that SRGAP2 genes seem involved in brain development (neoteny in spine maturation), increase of human brain

*. Intervenant
†. Corresponding author : lefloch@cng.fr
‡. Corresponding author : christophe.battail@cea.fr

complexity seems to be related to this SRGAP2 gene family. To study the tissue-specificity of duplicated genes, an accurate estimation of their expression value is required. Nevertheless the high sequence identity of some conserved duplicates could be challenging in RNA-Seq because of unspecific mappings of sequencing reads between conserved regions.

A last but not least particularity of duplicated genes is their implication into diseases. Chen et al. used monogenic disease associations from OMIM (Online Mendelian Inheritance in Man) and HGMD (Human Gene Mutation Database) databases to show that 55 % of monogenic disease genes were duplicated genes, which represents 23 % of duplicates that were also monogenic disease genes. A Fisher exact test was applied to indicate that duplicates were significantly enriched into genes implicated in monogenic diseases. They explained this implication by the mutations accumulated on one of the copies of a duplicated pair. Their hypothesis is that just after the duplication event, new mutations would be hidden by the functional redundancy of each copy but when the sequences become too divergent, mutations could create a disease gene on one copy.

The objective of our project is to characterize gene families by collecting and analyzing different evolutionary, functional and sequence features such as the types of duplication events, the association of duplicated genes with diseases and the identity of gene sequences within families.

Results

Firstly, we generated genomic and evolutionary characteristics of gene families in humans, originating from WGD and SSD events. Using gene families and duplication categories identified by Singh et al., 2014 and Chen et al., 2013, we compared sizes of families between both types of duplication and between different dating of SSD genes. We observed that the sizes of families of the WGD group (mean = 4.04) were significantly smaller than all SSD ones (mean = 4.47). Moreover, the sizes of families of the older SSD and WGD-old SSD groups (respective means = 4.35 and 4.75) were significantly smaller than the young SSD ones (mean = 6.03). Indeed the younger SSD group contained on average the largest families.

Secondly, we compared functions associated with each type of duplication. We computed GO enrichments on Molecular Function and Biological Process terms for SDD and WGD groups. The SSD group was significantly enriched in 32 terms (such as “receptor activity”) as Molecular Function and 104 terms (such as “sensorial perception”) as Biological Process. For WGD, the number of significant terms was 182 for Molecular Function (such as “kinase activity”) and 1024 for Biological Process (such as “anatomical structure” and “development”).

Thirdly, regarding implication of duplicated genes into diseases, we assessed enrichments with all genetic disease genes (monogenic and polygenic) collected from the ClinVar database. We found that only enrichments on WGD genes were significant compared to enrichments of all duplicates. Indeed, the proportion of WGD in genetic diseases (42 %) was significantly greater than their proportion in non-disease genes (35.5 %).

Finally, to assess each gene family we extracted its sequence identity between all its gene members. We used the Needleman and Wunsch (Needleman & Wunsch, 1970) algorithm to globally align all gene sequences belonging to the same family. Because of varying transcript lengths between families, we decided to study the relationship between the gene sequence identity and the variance of transcript lengths in each family. We observed that transcript length variance was anti-correlated (-0.33) with sequence identity. The highest the variance, the smaller the identity percentage is. This result indicates that transcripts with little divergent sequences are of similar lengths. We also calculated local alignments between transcripts of the same family using BLAST all-against-all algorithm. For example, we found that 66 % (8205 genes) of duplicated genes had at least a region of 80 bp with an identity greater than 75 % and 915 of these genes with regions greater than 98 % of identity. The high sequence identities should be problematic for the mapping of NGS data such as RNA-Seq and for a correct estimation of abundance value. That is why we complemented these alignment analyses with a mappability study to find the exact

number of regions (size =100 bp) repeated along the reference sequence of the duplicated gene transcriptome. Approximately 23 % (2191 genes) of duplicated genes have a non-unique sequence interval of 100/,bp.

Discussion & Conclusion

Results on family sizes of both types of duplication show that the WGD group contains small families whereas all SSD contain larger families and that the older SSD families are smaller than younger SSD ones. Older SSD events are ancient so the selection pressure has occurred on families. However, younger SSD events are more recent so they are not as altered yet as ancient duplication events by the selection pressure which probably explains their larger family size. Otherwise, young and old SSD groups have larger families than the WGD group, so another parameter, different from the dating, could impact the size of the family. Acharya and Ghosh highlights that WGD evolutionary rate is slower than SSD and that WGD functions are more essential (Acharya & Ghosh, 2016) than SSD ones so the sizes of WGD families remain stable in the genome evolution. This essentiality and stability of WGD families should be the other parameter that can explain the small size of WGD families.

Concerning specific ontological functions enriched in each duplication group, WGD gene terms are confirmed by Singh et al.. Regarding the number of significant terms in each type, we have found more terms in the WGD group than the SSD one for both Molecular Function and Biological Process terms. This result is also observed in Acharya and Ghosh which associates this higher number of terms for WGD genes with their multi-functional nature.

Previously we observed that WGD were implicated into genetic diseases. As explained in the introduction, genetic diseases can be induced by mutated duplicated genes. As WGD families are more implicated into essential functions, when a WGD gene is mutated the risk to induce a disease or lethality is increased.

The high proportion of duplicated genes with high sequence identity and our results on the duplicated transcript mappability show the potential problem of the unspecific expression measurement using RNA-Seq data because of inaccurate mappings of sequence reads. However, if only uniquely mapped reads are conserved, duplicated genes will probably be less covered than singletons. This coverage bias could lead to a drop in abundance estimation quality. As a perspective, an appropriate selection of read alignment parameters such as multiple read positions, mismatch numbers and *a priori* knowledge of SNPs should lead to better expression estimates of duplicated genes. Finally, more accurate estimates would allow us to study the expression profiles of all duplicated genes and to investigate their tissue-specificity.

In conclusion, we generated and aggregated with our project a large collection of evolutionary, functional and sequence features related to duplicated genes and their families. All information is gathered into organized personal files. We now want to organize this information into a public database to be able to merge all different characteristics on gene families and make it available to the research community.

References

Acharya, D., & Ghosh, T. C. (2016). Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. *BMC Genomics*, 17:71.

Chen, W. H., Zhao, X. M., van Noort, V., & Bork, P. (2013). Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. *PLoS Computational Biology*, 9(5).

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, 3(10). http://doi.org/10.1371/journal.pbio.0030314.

Dennis, M. Y., Nuttle, X., Sudmant, P. H., Antonacci, F., Tina, A., Nefedov, M., … Eichler, E. E. (2012). Human-specific evolution of novel SRGAP2 genes by incomplete segmental duplication. *Cell* 149(4):912–922.

Makino, T., & McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences of the United States of America*, 107(20), 9270–9274.

Needleman, & Wunsch. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453. http://doi.org/10.1016/0022-2836(70)90057-4.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J. M., Guo, Y., … Durbin, R. (2008). TreeFam: 2008 Update. *Nucleic Acids Research*, 36(December 2007):735–740. http://doi.org/10.1093/nar/gkm1005.

Singh, P. P., Affeldt, S., Malaguti, G., & Isambert, H. (2014). Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication. *PLoS Computational Biology*, 10(7).

Mots clefs : Gene family, Duplicated gene, Whole Genome Duplication, Small Scale Duplication, Gene expression, Disease genes

Comparative genomics of gene families in relation with metabolic pathways for gene candidates highlighting

Poster 88

Delphine Larivière [*] [†1], David Couvin [*] [‡1], Gaëtan Droc[1], Jonathan Lorenzo[2], Valentin Guignon[3], Mathieu Rouard[3], Eric Lyons[4], Dominique This[5], Stéphanie Bocs[6], Jean-François Dufayard[6]

[1] Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) – Avenue Agropolis, Montpellier, France
[2] Institut Français de Bioinformatique (IFB) – Gif-sur-Yvette, Île-de-France, France
[3] Bioversity International – Consultative Group on International Agricultural Research [CGIAR] – Parc Scientifique Agropolis II, F-34 397 Montpellier Cedex 5, France
[4] University of Arizona – University of Arizona Tucson AZ 85721 USA/États-Unis
[5] Montpellier SupAgro – Montpellier SupAgro – 2 place Pierre Viala, F-34 060 Montpellier Cedex 02, France
[6] CIRAD UMR AGAP (AGAP) – Montpellier SupAgro, Institut national de la recherche agronomique (INRA) : UMR1334, CIRAD-BIOS – TA A-108/03, Avenue Agropolis, F-34 398 Montpellier Cedex 5, France

The study of gene families is an important field of comparative genomics, allowing, by the analysis of evolutionary history, to identify homology relationships, gene losses and to help in annotation transfers. The addition of metabolic information improves the identification of candidate genes by addition of functional and gene network data. We propose new web systems to facilitate and improve the analysis of gene family for the search of candidate genes in plants. GenFam (`http://genfam.southgreen.fr/`) is dedicated to the manual and precise analysis of gene families and includes specific workflows running under a Galaxy platform, allowing to gather several data sources, analysis and visualization tools, in order to (i) build custom families (ii) run analysis workflows dedicated to the analysis of gene families (iii) visualize analysis results and functional evidences through a dedicated visualization dashboard.

In complement to the integration of data sources, tools and visualizations, we also suggest a new way to find evidences for the identification of evolutionary events through syntenic analyses. The IDEVEN algorithm is based on the study of syntenic blocks linked to a gene family to identify speciation, Whole Genome Duplication (WGD) events, and other duplications in a family history. The identified events will be reported on the phylogeny and aim to bring complementary evidences to have a clearer view of the evolutionary history of a gene family. To extend this tool to the analysis of multiple gene families and integrate metabolic pathways data, this tool has been integrated in genesPath, which will allow a deep identification and highlighting of candidate genes of interest for a specific project called "Biomass For the Future (BFF)". This online tool will be soon available and could be notably used for searching candidate genes involved in biosynthesis of lignin and cellulose in various plant species (such as maize and sorghum).

Mots clefs : gene family, data integration, synteny, evolutive events, comparative genomics, metabolic pathways, phylogeny

*. Intervenant
†. Corresponding author : `lariviere.delphine@gmail.com`
‡. Corresponding author : `david.couvin@googlemail.com`

"AdaptSearch" : a galaxy pipeline for the search of adaptive mutations and positively selected genes from RNAseq orthologs

Poster 89

Misharl Monsoor [*] [†1], Éric Fontanillas[2], Julie Baffard[1], Pierre-Guillaume Brun[2], Erwan Corre[1], Christophe Caron[1], Didier Jollivet[2]

[1] Station biologique de Roscoff (SBR), Plateforme ABIMS, FR2424 – Université Pierre et Marie Curie (UPMC) – Paris VI, CNRS – Place Georges Teissier - BP 74, F-29 682 Roscoff Cedex, France
[2] Station biologique de Roscoff (SBR), Équipe ABICE, UMR 7144 – Université Pierre et Marie Curie (UPMC) – Paris VI, CNRS – Place Georges Teissier - BP 74, F-29 682 Roscoff Cedex, France

Dans le cadre d'une approche de génomique comparative à partir d'espèces proches issues d'habitats contrastés ou soumis à des conditions variables de l'environnement, notre volonté était de disposer d'une chaîne de traitement bioinformatique de données HTS (RNAseq par exemple) pour étudier l'effet de la sélection naturelle sur la partie codante du génome et sur les protéines traduites en comparant les biais de compositions en codons et en acides aminés entre espèces dans un cadre phylogénétique restreint et en respectant l'orthologie des gènes.

L'objectif est donc de proposer une suite de traitements bioinformatique de recherche de séquences codantes orthologues à partir d'assemblages issus des outils Velvet/Oases ou Trinity pour (1) obtenir un arbre phylogénomique des espèces, (2) quantifier les remplacements d'acides aminés dans les protéines des espèces et évaluer leurs taux d'évolution, et (3) mesurer la force de la sélection sur les différentes lignées évolutives à travers le ratio entre mutations non-synonymes et synonymes.

Ce package (AdaptSearch), reposant sur l'utilisation de scripts python et d'outils d'analyse phylogénétique de référence (RaxML, PaML), est déployé au travers de plusieurs briques développées sous Galaxy et disponibles sur l'instance `http://galaxy.sb-roscoff.fr/` de la plateforme ABiMS.

La première partie du pipeline se décompose en 5 briques successives incluant (1) le filtrage des transcriptomes assemblés soit par Trinity ou Velvet/Oases en ne prenant que la partie codante du transcrit, (2) une recherche d'orthologues entre paires d'espèces par TBLASTX réciproque, (3) la détection de groupes d'orthologues putatifs (POGs) sur l'ensemble des espèces à partir des résultats de BLAST en reprenant la totalité de la région codante du gène, (4) l'alignement des POGs incluant un nouveau filtrage des zones codantes et l'élimination des indels, et enfin (5) une analyse d'inférence phylogénétique sur les gènes concaténés.

1. Filtrage des transcriptomes (Filter_assemblies) : Le module utilise la fonction GetORF (inclus dans le package EMBOSS) pour récupérer les parties codantes des transcrits et sélectionne un seul variant par transcrit en fonction des critères de qualité et de longueur en modifiant le nom des séquences pour y faire apparaître l'abréviation des noms d'espèce. Une filtration préalable des transcrits est également effectuée sur la qualité des régions terminales de la séquence selon les recommandations de Tang et al. (2008). Les 30 premières bases sont systématiquement tronquées en région 5' alors que 10 % de la longueur de la séquence est supprimée à la fin de la région 3'. Les transcrits contenant des positions indéterminées (i.e. 5N et plus) sont supprimés du jeu de données.

*. Intervenant
†. Corresponding author : `misharl.monsoor@sb-roscoff.fr`

2. Blast réciproque pour recherche d'orthologues (Pairwise) : Une série de TBLASTX réciproques en prenant le meilleur hit de chaque recherche est effectuée sur l'ensemble des paires possibles entre transcriptomes selon les méthodes implémentées dans (Savard *et al.* 2006) avec l'option de « medium soft filtering for low similarity regions » recommandée par (Moreno-Hagelsieb and Latimer 2008) en choisissant une valeur seuil de e-value ($< 10^{-10}$).

3. Sélection des groupes d'orthologues (POGs) : Dans une approche conservative, chaque groupe de séquences orthologues avec plus de 2 espèces est archivé dans en répertoire allant de 3 à n espèces. Les groupes d'orthologues contenant plus d'une séquence pour une espèce donnée sont éliminés pour écarter l'insertion de gènes paralogues. Pour chaque gène (locus), les séquences ADN contenues dans les POGs sont localement alignées en utilisant l'algorithme de Blastalign qui prend en compte les longs indels de manière conservative dans l'alignement (Belshaw and Katzourakis 2005).

4. Filtrage des CDS et suppression des indels (Blastalign and CDS_search) : En raison des analyses suivantes basées sur les décomptes des codons, acides aminés traduits et le ratio des mutations non-synonymes et synonymes (dN/dS), ce module permet d'obtenir des groupes de séquences codantes sans codon stops dans le bon cadre de lecture en choisissant une valeur seuil de la longueur minimale de l'alignement. De plus, comme ce programme n'a pas vocation à utiliser l'information des indels et que ces derniers produisent souvent des erreurs d'alignement dans les régions flanquantes à même de conduire à une surestimation du taux de mutation (Fletcher and Yang 2010), le module supprime tout indel et sa région flanquante dans l'alignement et fournit un ensemble d'alignements en nucléotides ou en acides aminés.

5. Phylogénomique (ConcatPhyl) : Ce module permet d'effectuer une concaténation des différents locus présentant un alignement de séquences codantes dans le bon cadre de lecture sans indels en autorisant un nombre possible d'espèces manquantes. Selon le nombre d'espèces choisies (n), le script de concaténation utilise les POGs de n à N espèces pour reconstruire un arbre phylogénétique en Maximum de Vraisemblance. L'arbre en ML est construit sous RAxML en choisissant le modèle de substitution *adhoc* (selon que le jeu de séquences concaténées en nucléotides ou en acides aminés), et le nombre de bootstraps. Le module fournit plusieurs arbres (best tree, bi-partition tree with bootstraps) au format netwick et un jeu de séquences concaténées.

La seconde partie du pipeline, partiellement déployée sous Galaxy, se décompose en 3 étapes : (A) une recherche de branches ou de codons sous sélection positive en utilisant le module CodeML de PaML 5.0 (Yang 2010), l'arbre phylogénomique RaxML de référence (Best tree) et le jeu de séquences concaténées, (B) une identification de gènes sous sélection positive en comparant les modèles de branches M1 et M0 sur l'ensemble des locus (POGs) pour lesquels il n'y a pas d'espèces manquantes et enfin (C) effectue un ensemble de comptages (codons, acides aminés traduits) sur chaque espèce, calcule des indices (e.g. catégories d'a.a.) et effectue des tests de permutation sur ces comptes et indices. Ces modules constituent les étapes 6, 7 et 8 du pipeline.

6. Recherche de gènes sous sélection positive (CompCodeML : en cours d'intégration dans Galaxy)

Les locus sont filtrés sur la base de 2 critères : le nombre d'espèces qui partagent le même site pour un codon donné, et la longueur minimale de l'alignement pour mener l'analyse après suppression des sites (codons) qui ne satisfont pas le premier critère. Un premier module compare à l'aide d'un LRT les modèles de branches M0 (« one branch model ») et M1 (« free ratio model ») sur l'ensemble des locus pour lesquels les N espèces sont présentes. Le module parse ensuite les valeurs de likelihood, LRT et dN/dS des différentes branches en effectuant une filtration des données où les valeurs du ratio sont nulles ou infinies en appliquant un filtre sur la valeur S.dS (S.dS> 1).

7. Recherche de branches et codons sous sélection positive (GeneSearch : en cours d'intégration dans Galaxy)

Un deuxième module examine l'ajustement des modèles de branches (M0, M1, M2) et des

modèles de branche-sites (M1a et M2a) sur l'ensemble des locus concaténés. Dans ce cas, le module partitionne en répertoires les sorties des différents modèles en y copiant sur chacun l'arbre de référence (avec ou sans choix de branches à analyser spécifiquement) et l'alignement à ajuster au modèle. Dans le cas où un modèle de sélection positive serait significativement meilleur, les codons sous sélection positive présentant une probabilité bayésienne (BEB) supérieure à 0,90 sont filtrés et gardés dans un fichier `.csv`.

8. Composition des gènes et des protéines traduites (MutCount : intégré dans Galaxy)

Ce module permet le comptage des codons, acides aminés traduits et des types de résidus (polaire, polaire chargé, aliphatique et aromatique) et calcule certains indices tels que le GC content (GC12, GC3) ou l'excès de purines et les indices de thermophilie tels que IVYWREL, EK/QH, CvP index. Les différences entre espèces de ces comptes (proportions) et indices sont ensuite évaluées par un test de permutations. Les données et p-values associées au test de permutation sont ensuite fournies sous la forme d'un tableau (.csv) pouvant être utilisé par la fonction R aov.phylo du programme Geiger (ANOVA à composantes phylogénétiques) (Harmon et al. 2008) en utilisant la température comme variable explicative ou par une analyse factorielle des correspondances sur l'usage des codons en utilisant codonW 1.4.2 (`http://codonw.sourceforge.net/`). Ces modules R ne sont cependant pas implémentés dans le pipeline.

Ce travail a été initialement développé et utilisé dans le cadre d'une étude ayant pour but de caractériser les mutations à la base de la thermostabilité des protéines chez les polychètes thermophiles Alvinellidae vivant au niveau des sources hydrothermales profondes (Fontanillas et al. soumis). L'intégration de l'ensemble de la chaîne de traitement dans l'environnement Galaxy a permis d'assurer la portabilité et la diffusion d'un pipeline hétérogène initialement constitué de plus de 400 scripts bash et python, ainsi que d'outils de recherche de similarité, d'inférence phylogénétique et d'analyses de génomique évolutive au sein d'un environnement d'analyse homogène, documenté et extensible. L'outil en phase final de portage sera prochainement ouvert à la communauté sur l'instance `http://galaxy.sb-roscoff.fr/` de la plateforme ABiMS.

Références

Belshaw R, Katzourakis A. 2005. BlastAlign: a program that uses blast to align problematic nucleotide sequences. *Bioinformatics* 21:122-123.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol.* 27:2257-2267

Harmon Luke, Jason T. Weir, Chad D. Brock, Richard E. Glor and Wendell Challenger. GEIGER: investigating evolutionary radiations. *Bioinformatics* 24 (1):129-131. doi:10.1093/bioinformatics/btm538

Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24:319-324

Savard JI, Tautz D, Richards S, Weinstock GM, Gibbs RA, Werren JH, Tettelin H, Lercher MJ. 2006. Phylogenomic analysis reveals bees and wasps (*Hymenoptera*) at the base of the radiation of Holometabolous insects. *Genome Res.* 16:1334-1338.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18:1944-1954.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24(8):1586-1591.doi: 10.1093/molbev/msm088

Mots clefs : Phylogénie, Galaxy, Workflow, Pipeline, NGS, Evolution, Adaptative mutations, Positively selected genes, Orthologs

Origin and evolution of the haem-copper oxidases superfamily in Archaea

Anne Oudart [*] [†1], Simonetta Gribaldo [‡2], Céline Brochier [§1]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558 – France
[2] Institut Pasteur – Institut Pasteur de Paris – France

Poster 90

Understanding the origin and evolution of dioxygen reductases, the terminal electron acceptors of aerobic respiratory chains, can provide precious clues for the emergence of this energy process that is still debated. The heme-copper oxidases superfamily (HCO), belonging to the dioxygen reductases, contains three families of dioxygen reductases named A, B, and C according to a classification based on sequence similarity and phylogenetic analysis of the homologous catalytic subunits (Pereira et al., 2001). A phylogenomic study showed that these enzymes have very different evolutionary histories (Brochier-Armanet et al., 2009) with an ancient dioxygen reductase (A-HCO) presents prior to the divergence of major present-day bacterial and archaeal phyla, thus before the emergence of oxygenic photosynthesis. However, this result is in contradiction with those proposed by Ducluzeau et al. (2014) about the structure of the A-HCO suggesting that this dioxygen reductase would be the most recent. So the question about the origin of the haem-copper oxidases is still unresolved, and a new analysis is required.

The heme-copper oxidases superfamily of the ancestor of Archaea probably used dioxygen (Gribaldo et al., 2009). Nevertheless, the previous conclusions are old and the available data was very limited (73 archeal complete genomes in 2009). Today, with more available data in the public database (252 archeal complete genomes) it is interesting to reassess the questions about the origin and evolution of the heme-copper superfamily for Archaea. We will present our results about subunits of the heme-copper superfamily in Archaea.

Pereira, M. M., M. Santana, et M. Teixeira. "A Novel Scenario for the Evolution of Haem-Copper Oxygen Reductases". *Biochimica Et Biophysica Acta* 1505, no 2?3 (1 juin 2001): 185?208.

Brochier-Armanet, Celine, Emmanuel Talla, et Simonetta Gribaldo. "The Multiple Evolutionary Histories of Dioxygen Reductases: Implications for the Origin and Evolution of Aerobic Respiration". *Molecular Biology and Evolution* 26, no 2 (février 2009): 285?97. doi:10.1093/molbev/msn246.

Ducluzeau, Anne-Lise, Barbara Schoepp-Cothenet, Robert van Lis, Frauke Baymann, Michael J. Russell, et Wolfgang Nitschke. "The Evolution of Respiratory O2/NO Reductases: An out-of-the-Phylogenetic-Box Perspective". *Journal of The Royal Society Interface* 11, no 98 (9 juin 2014): 20140196. doi:10.1098/rsif.2014.0196.

Gribaldo, Simonetta, Emmanuel Talla, et Celine Brochier-Armanet. "Evolution of the Haem Copper Oxidases Superfamily: A Rooting Tale". *Trends in Biochemical Sciences* 34, no 8 (août 2009): 375?81. doi:10.1016/j.tibs.2009.04.002.

Mots clefs : aerobic respiration, Archaea, heme, copper oxidases

*. Intervenant
†. Corresponding author : anne.oudart@univ-lyon1.fr
‡. Corresponding author : simonetta.gribaldo@pasteur.fr
§. Corresponding author : celine.brochier-armanet@univ-lyon1.fr

The draft genome sequence of the rice weevil *Sitophilus oryzae* as a model to explore the host-symbiont interactions in a nascent stage of endosymbiosis

Poster 91

Carlos Vargas Chavez[1,2], Nicolas Parisot [*†1], Clément Goubert[3], Patrice Baa-Puyoulet[1], Séverine Balmand[1], Matthieu Boulesteix[3], Nelly Burlet[3], Hubert Charles[1], Stefano Colella[1], Gérard Febvay[1], Toni Gabaldón[4], Damian Loska[4], Justin Maire[1], Florent Masson[1], Andrés Moya[2], Rita Rebollo[1,3], Agnès Vallier[1], Aurélien Vigneron[1], Carole Vincent-Monégat[1], Anna Zaidman-Rémy[1], Federica Calevro[1], Cristina Vieira[3], Amparo Latorre[2], Abdelaziz Heddi[1]

[1] Biologie Fonctionnelle, Insectes et Interactions (BF2I) – Institut national de la recherche agronomique (INRA) : UMR0203, Institut National des Sciences Appliquées [INSA] - Lyon – France
[2] Institut Cavanilles de Biodiversitat i Biologia Evolutiva (ICBiBE) – Espagne
[3] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – France
[4] Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) – Espagne

Organisms across the tree of life are associated with diverse microbial partners that impact host adaptive traits and exhibit phenotypes ranging from parasitism to mutualism. For example, insects thriving on nutritionally unbalanced habitats are prone to house mutualistic intracellular bacteria (endosymbionts) that complement their diet, thus greatly improving their ecological performances. Within insects, endosymbiosis is very common in the Curculionoidea weevils superfamily, which constitutes a group with considerable worldwide biodiversity. Weevils include some of the most invasive insects and cause huge crop damages. Recent phylogenetic and molecular studies have shown that endosymbiosis history has been marked by several symbiotic displacements within this insect group. The most recent event may have occurred less than one million year ago within the cereal weevil *Sitophilus* clade resulting in the replacement of an ancestral symbiont Candidatus *Nardonella* by *Sodalis pierantonius* symbiont. *S. pierantonius* genome exhibits peculiar molecular features associated with a massive pseudogenization and the occurrence of a huge amount of repeated elements. Whether these phenomena are adaptive, and whether they impact host genome reshaping are puzzling questions that will be addressed thanks to the genome level investigation of the *Sitophilus-Sodalis* recent association.

Here, we present the draft genome sequence of the rice weevil *Sitophilus oryzae*. The full genome sequence has been obtained through a combination of short-read (Illumina HiSeq and Roche/454 GS FLX) and long-read (Pacific Biosciences PacBio RS) sequencing methods. After error correction, the data were assembled using the Platanus algorithm for an initial scaffolding and gap-filling. These scaffolds were then re-scaffolded several times using PacBio data. The final assembly consisted in 17,365 scaffolds of a total length of 652 Mbp (the *S. oryzae* genome size was estimated to be about 650 Mbp using flow cytometry), a N50 value of 110 kbp, a coverage of 101 X and a GC content of 38.4 %. Intriguingly, transposable elements (TE) analysis using both automated tools (dnaPipeTE, RepeatModeler and MITEhunter) and manual annotations

*. Intervenant
†. Corresponding author : nicolas.parisot@insa-lyon.fr

revealed an unexpected high amount of repeated DNA (> 50 %) in this weevil genome. Gene prediction was then performed using a combination of MAKER, GeneMark, Augustus and SNAP algorithms and taking advantage of the available transcriptomic data (EST and RNA-seq data) on *S. oryzae* to build more accurate gene models. Finally, the official gene set contained 17,026 protein-coding genes. Based on this gene set, the complete catalogue of gene phylogenies (phylome) was predicted through the PhylomeDB pipeline and will be publicly available in this database (`http://www.phylomedb.org/`). The weevil metabolic and signalling networks were also reconstructed using the CycADS pipeline in order to generate the SitorCyc database (a BioCyc interface of the *S. oryzae* metabolism). These metabolic pathways were integrated in the ArthropodaCyc database collection dedicated to comparative metabolic analyses among arthropods (`http://arthropodacyc.cycadsys.org/`). The interdependence of the metabolic networks of *S. oryzae* and its endosymbiont *S. pierantonius* will then be characterized thanks to their integration into the ArtSymbioCyc database that is being developed and will be dedicated to arthropod symbioses. All these annotations (TE, phylome and metabolic networks) will be integrated in a comprehensive genome database providing a genome browser with crosslinks to available resources.

Altogether, these results are expected to unravel basic molecular mechanisms and evolutionary features associated with the establishment and the maintenance of endosymbiosis in animals, and to permit identifying potential gene targets useful for the development of new ecologically-friendly strategies for pest insects control and management.

Mots clefs : Symbiosis, Evolution, Genome annotation, Data visualization, Database development

HOGENOM 666 : un réseau de 13 bases de données phylogénomiques

Poster 92

Simon Penel [*] [†1], Vincent Daubin[1], Manolo Gouy[1], Dominique Guyot[2], Vincent Miele[1], Laurent Duret[1], Guy Perrière[1,2], Guillaume Gence[1]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Pôle Rhône-Alpes de Bioinformatique (PRABI) – Université Claude Bernard - Lyon I (UCBL) – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France

Introduction

La génomique comparative est une approche centrale dans l'analyse des séquences biologiques, depuis l'annotation et l'identification de régions fonctionnelles jusqu'à l'étude des processus évolutifs tels que la perte ou la duplication de gènes, les transferts horizontaux voire la duplication complète de génomes.

HOGENOM

Dans ce contexte, la base de données HOGENOM permet d'accéder à un ensemble de familles de gènes homologues provenant d'organismes (eukaroytes, bactéries et archées) dont le génome est complètement séquencé. Ces familles sont associées à des alignements multiples ainsi que des arbres phylogénétiques. HOGENOM offre la possibilité d'utiliser des motifs d'arbres pour rechercher des ensembles de gènes paralogues et/ou orthologues dans les arbres des familles.

Nouvelles approches

L'explosion des données provenant du séquençage (des centaines de génomes complets d'eucaryotes sont accessibles dans la base Ensembl, et des dizaines de milliers de génomes complets bactériens sont proposés par le NCBI) a nécessité le développement de nouvelles approches pour la construction de HOGENOM, aussi bien en terme de temps de calcul pour le classement des séquences homologues en familles qu'en terme de méthodes et d'outils pour la visualisation des arbres associés à ces familles.

Classement en familles

La nouvelle stratégie plus rapide de calcul des familles repose sur un un pré-classement avec Kclust puis un post-traitement avec SiLiX sur les profils HMM calculés à partir des clusters obtenus. Elle a permis de classer les 50 millions de séquences de la version HOGENOM en cours, soit 593 eucaryotes 12 336 bactéries et 224 archées, en environ 1 million de familles.

Arbres phylogénétiques

Plusieurs bases ont été définies, une base « HOGENOM-CORE » contenant les génomes réprésentatifs et 12 bases « HOGENOM-PHYL[1-12] » contenant tous les génomes d'une sélection de 11 phylum principaux, la dernière base contenant les autres phylum non sélectionnés. Les

*. Intervenant
†. Corresponding author : simon.penel@univ-lyon1.fr

666 génomes représentatifs de HOGENOM-CORE sont choisis de manière semi-automatisée avec l'objectif de maximiser la représentation taxonomique, et les 11 phylum de manière à constituer des groupes de séquences de taille raisonnable et biologiquement pertinents. Ces phylums sont les suivants : eukaryota (462 espèces), archaea (169), betaproteobacteria (140), alphaproteobacteria (256), gammaproteobacteria (392), delta/epsilon subdivisions (106), firmicutes (483), bacteroidetes/chlorobi group (188), actinobacteria (294), spirochaetes (52), tenericutes (59).

Des alignements multiples sont calculés à partir des familles obtenues par le classement global de l'ensemble des séquences, alignements dont sont ensuite extraits des sous-alignements en sélectionnant d'une part les séquences provenant de la base HOGENOM-CORE et d'autre part celles de l'une des 12 bases HOGENOM-PHYL. On obtient ainsi 12 collections d'alignements contenant des séquences de génomes représentatifs ainsi que des séquences des génomes de l'un des 12 phylum. On extrait aussi une collection d'alignements contenant uniquement des séquences de génomes représentatifs. Ces derniers sont utilisés pour calculer les arbres phylogénétiques des familles de HOGENOM-CORE, qui sont utilisés ensuite comme contrainte pour calculer les arbres des alignements des familles HOGENOM-CORE+HOGENOM-PHYLA. Grâce à cette contrainte sur la partie HOGENOM-CORE des arbres, il sera possible de mettre en relation les arbres de familles HOGENOM-CORE+HOGENOM-PHYL de différents phylum afin d'obtenir une information phylogénétique étendue.

Référence

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M and Perrière G (2009). "Databases of homologous gene families for comparative genomics" *BMC Bioinformatics*, 10 (Suppl 6):S3

Mots clefs : génomique comparative, clustering, phylogénie, évolution, base de données

Analyse de l'évolution d'une épidémie bactérienne par séquençage haut débit

Marie Petitjean *[1,2], Xavier Bertrand[1,2], Benoît Valot[2], Didier Hocquet[1,2]

Poster 93

[1] Laboratoire d'hygiène hospitalière – CHRU Minjoz, Besançon – France
[2] UMR 6249 Chrono Environnement – Université de Franche-Comté, Besançon – France

Pseudomonas aeruginosa est une bactérie pathogène responsable d'épidémies au sein des hôpitaux et en particulier chez les patients dont le système immunitaire est défaillant. Entre 1997 et 2014, un clone épidémique de la bactérie a infecté ou colonisé plus de 300 patients au sein du CHRU de Besançon. Le nombre de patients atteints a rapidement augmenté avant qu'une diminution du nombre de nouveaux cas n'apparaissent, précédant la disparition complète du clone. Celle-ci ne peut être expliquée par la mise en place de mesures d'hygiènes. Les profils de résistance aux antibiotiques ainsi que l'aspect morphologique des colonies bactériennes ont évolués durant l'épidémie. Cinquante-cinq génomes séquencés en haut-débit vont être comparés afin de comprendre l'évolution génétique de la bactérie au cours des 17 ans de l'épidémie.

Le premier génome de l'épidémie est séquencé en utilisant la technique Pacific Bioscience qui pertmet d'obtenir un génome complet de bonne qualité. Les autres génomes, séquencés en Ion Torrent ou Illumina sont assemblés en utilisant MIRA et le core génome – partie conservée entre les différents isolats – va être utilisé afin de générer un arbre phylogénétique. Celui-ci est réalisé à partir des 3205 SNPs présents parmi les 4408996 nucléotides du core génome. Les clusters ainsi générés sont liés à la date et au service où à eu lieu le prélèvement.

Le plus proche ancêtre commun calculé à partir des mêmes données que précédemment est daté aux alentours de 11,8 $\pm$ 1,7 ans avant le début de l'épidémie. Nous pouvons supposer que le clone c'est installé à peu près au moment de l'ouverture de l'hôpital en 1983.

Tous les clones sont multi-résistant, ils sont tous porteurs d'un gène codant pour la résistance aux aminoglycosides. La présence de gènes de résistance à été montrée grâce à ResFinder, logiciel accessible en ligne et permettant d'identifier les gènes impliquant des résistances à certains types d'antibiotiques. De plus la présence de mutations dans certains gènes induisent d'autres résistances chez la grande majorité de ces clones. Ces mutations ont été recherchées en utilisant PAO1 – souche de référence pour Pseudomonas aeruginosa. L'existence phénotypique de ces résistances à été montrée par réalisation d'antibiogrammes.

La présence de gènes de virulence – réalisée par *clustering* des gènes en utilisant la base de données de Virulence Factor DataBase – a été observée. De plus les mutations existantes chez les différents gènes induisant de la virulence ont également été recherchées. Des mutations différentielles en fonction des isolats sont présentes dans les gènes impliqués dans la mise en place de systèmes de sécrétion, dans l'assemblage du pili, dans les méchanismes du quorum sensing ou encore dans la production de sidérophores appelés pyocheline et pyoverdine.

Le système CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeat) permet de protéger les bactéries contre l'intégration d'éléments géniques étrangers, tels que les bactériophages ou les plasmides. En effet ce système est composé de séquences répétées séparant des séquences issues d'éléments mobiles. Lorsque l'ADN de l'élément mobile est reconnu par le système, celui-ci va déclencher la fragmentation de l'ADN exogène empêchant ainsi son intégration dans le génome. Dans tous les isolats de l'épidémie, on retrouve la présence de deux systèmes CRISPR ainsi que les protéines Cas (CRISPR-associated). Le premier possède 12 spacers alors que

*. Intervenant

le deuxième en possède 8 permettant donc de reconnaitre 20 séquences exogéniques différentes. La relative stabilité du génome ainsi que l'absence de protéines agissant contre le système laissent penser que ce système est actif.

Pourtant, la présence d'éléments mobiles tels que les ICEs ou Integrative Conjugative Elements permet d'expliquer la taille relativement élevée de la souche de référence pour l'épidémie, DHS01. En effet, cinq potentiels ICEs sont présent dans le génome de la souche épidémique. Leur présence a été identifiée en recherchant la machinerie spécifique aux éléments mobiles ainsi que la présence des protéines de conjugaison. Ces ICEs sont stables tout au long de l'épidémie. L'ICE-1 contient essentiellement des gènes impliqués dans le métabolisme tel que la synthèse de purine et de pyrimidine, le métabolisme du méthane et du pyruvate ainsi que la synthèse de métabolites secondaires. Le deuxième – ICE-2 – présente une portion identique à un ICE connu, PAPI-I et cette séquence commune est impliquée dans le déplacement de l'ICE vers d'autres souches. L'ICE-3 est composé de gènes du métabolisme impliqués dans la synthèse de métabolites secondaires, d'un gène codant pour une toxine et de transporteurs. Un certains nombre de gènes de résistance incluant de la résistance aux aminoglycosides, sulfonamide et résistance au mercure est retrouvée dans l'ICE-4. Le dernier, l'ICE-5, porte un opéron codant pour une pompe RND qui est impliqué dans la résistance. Les formes circulaires de ces différents ICEs ont été recherchées en réalisant une *nested PCR*. Seul deux d'entre eux ont été validés, les ICEs 2 et 4.

Le clone épidémique de *Pseudomonas aeruginosa* ST395 a probablement commencé à diffuser au sein de l'hôpital une dizaine d'années avant de commencer à diffuser chez les patients. Il est porteur de nombreuses résistances aux antibiotiques et accumule les mutations au sein des gènes de virulence.

Mots clefs : Génomique, bactérie, évolution, hôpital

The Bgee database : gene expression data in animals

Poster 94

Frederic Bastian [*1,2], Julien Roux[1,2], Anne Niknejad[1,2], Valentine Rech De Laval [†1,2], Sébastien Moretti[1,2], Philippe Moret[1,2], Mathieu Seppey[1,2], Marc Robinson-Réchavi[1,2]

[1] Swiss Institute of Bioinformatics (SIB) – Suisse
[2] Department of Ecology and Evolution, Université de Lausanne (UNIL) – Suisse

Bgee is a unique resource which allows to retrieve and to accurately compare gene expression patterns in multiple animal species. Its database integrates all sources of expression data, from the anatomical detail of *in situ* hybridization, to the genome coverage of RNA-seq. It provides a reference dataset of wild-type and healthy, as well as high quality and comparable, gene expression data in animals. To this aim, we perform stringent quality controls, and annotate and re-analyze all RNA-seq, microarray and EST data integrated in Bgee. The database currently includes 17 animal species. Bgee is available at `http://bgee.org/`.

Bgee provides novel analytics tools, such as TopAnat, which allows to discover the organs where a set of genes is preferentially expressed. These analyses are quite similar to Gene Ontology (GO) enrichment tests, which determine the GO terms preferentially associated to a set of genes. In our case, however, the test is applied to terms from an anatomical ontology (Uberon ontology), mapped to genes by expression patterns. See `http://bgee.org/?page=top_anat`.

There is also a dedicated page for each gene present in the database. Each gene page displays information about a gene, summarizing all expression information. By using a new sorting algorithm, we rank the anatomical structures and life stages in which a gene is expressed, according to their relevance in term of gene function. For this, we take into account all sources of expression data together, which is unique to Bgee.

Bgee also allows to directly download data: i) calls of baseline presence/absence of expression, and of differential over-/under-expression, either in single species, or made comparable between multiple species; ii) annotations and experiment information (e.g., annotations to anatomy and development, quality scores used in quality controls, chip or library information); and iii) processed expression values (e.g., read counts, RPKM or TPM values, log values of Affymetrix probeset normalized signal intensities).

Bgee currently provides data for 17 species. We are currently in the process of integrating nine new species, plus a few very large datasets. We are also optimizing the use of anatomical homology information in the web interface, which will allow users to leverage expression data in an innovative way (e.g., conservation scores of expression of orthologous genes in homologous tissues).

We hope that Bgee will become a major hub for multi-species expression data integration, and a reference resource for the analysis of expression data in non-model organisms.

Mots clefs : gene expression patterns, multi species, homology, anatomical ontology, data integration, high throughput sequencing

*. Corresponding author : `Frederic.Bastian@unil.ch`
†. Intervenant

Beyond representing orthology relations with trees

Guillaume Scholz *[1]

[1] School of Computing Sciences (UEA) – School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK/Royaume-Uni

Poster 95

Reconstructing the evolutionary past of a family of genes is an important aspect of many genomic studies. To help with this, simple operations on a set of sequences called orthology relations may be employed. In addition to being interesting from a practical point of view they are also attractive from a theoretical perspective in that a characterization is known for when such a relation is representable by a certain type of phylogenetic tree.

For an orthology relation inferred from real biological data it is however generally too much to hope for that it satisfies that characterization. Rather than trying to correct the data in some way or another which has its own drawbacks, as an alternative, we propose to represent an orthology relation in terms of a structure more general than a phylogenetic tree called a phylogenetic network.

After a brief review of the work that has been done in the context of phylogenetic trees, we present the novel Network-Popping algorithm. This algorithm is aimed to build structurally simples phylogenetic networks, known as level-1 networks, representing a given orthology relation, should such a network exist.

Mots clefs : orthology relations, phylogenetic networks, trinets

*. Intervenant

Deciphering the biosynthetic pathways of ether lipids in Bacteria

Poster 96

Najwa Taib [*1], Béatrice Lauga[2], Cristiana Cravo-Laureau[3], Arnauld Vinçon-Laugier[4], Vincent Grossi[5], Céline Brochier-Armanet[6]

[1] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] IPREM – Université de Pau et des Pays de l'Adour – Pau, France
[3] Équipe Environnement et Microbiologie - IPREM UMR CNRS 5254 (EEM) – CNRS : UMR5254, Université de Pau et des Pays de l'Adour [UPPA] – IBEAS - UFR Sciences, BP 1155, F-64 013 Pau Cedex, France
[4] Laboratoire de géologie de Lyon (LGL-TPE) – Université Claude Bernard-Lyon I - UCBL (FRANCE), CNRS : UMR5276 – Bâtiment Géode, 2 rue Raphaël Dubois, F-69 622 Villeurbanne Cedex, France
[5] Université Claude Bernard Lyon 1, CNRS – Laboratoire de Géologie de Lyon (UMR 5276) – Campus de la Doua, Bâtiment GÉODE, 2 rue Raphaël Dubois, F-69 622 Villeurbanne Cedex, France
[6] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I, INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France

The "Woesian Revolution" in 1977 defined the three domains of life as the *Eucarya*, the *Bacteria* and the *Archaea*. The membrane lipid composition is one of the most remarkable feature distinguishing *Bacteria* and *Eucarya* from *Archaea*. In the *Archaea*, hydrocarbon chains consist in isoprenoid moieties linked with ether bonds to the enantiomeric glycerol backbone, glycerol-1-phosphate (G1P), while in bacteria and eukaryotes, fatty acid chains are linked to a glycerol-3-phosphate (G3P) molecule with ester bonds. These differences in lipid structures likely confer variations in the physical and physiological properties of the membranes of Bacteria and Archaea, with potential implications in terms of ecology and evolution. It should be stressed that ether-linked lipids are not unique to archaea per se, and have been reported in an increasing number of bacteria, including a few mesophilic species. Interestingly, while such lipids are widespread in the environment, little is known about their biosynthesis pathways, physiological role, and precise taxonomic distribution in *Bacteria*.

We conducted two complementary approaches in order to identify the genes and enzymes involved in the formation of bacterial ether lipids. To this end, we focused our investigations on *Desulfobacteraceae*, a family of sulfate-reducing bacteria (SRB) gathering species able to synthetize di-ester, mono-ether/mono-ester and/or di-ether membrane lipids. An in-depth genomic survey of 22 desulfobacteraceae proteomes allowed identifying a distant homologue of the di-0- geranyl-geranylglyceryl diphosphate (DGGGP) synthase – which catalyses the formation of the second ether bond in *Archaea*. We are now investigating experimentally the function of this candidate gene. In parallel we are carrying an accurate comparison of the 22 desulfobacteraceae proteomes in order to identify additional candidate genes.

Mots clefs : Phospholipids, Ether bonds, biosynthetic pathway of membrane lipids, Desulfobacteraceae

*. Intervenant

Organisation des Protéomes dans UniProtKB

Benoît Bely [*] [†1], Ramona Britto[1], Borisas Bursteinas[1], Andrea Auchincloss[2], Chuming Chen[3], Maria Martin [‡1], Uniprot Consortium[1,2,3]

[1] European Bioinformatics Institute [Cambridge] (EMBL-EBI) – EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK, Royaume-Uni
[2] Swiss Institute of Bioinformatics - Geneva (SIB) – Swiss Institute of Bioinformatics, Centre Médical Universitaire, 1 rue Michel Servet, Geneva, Switzerland/Suisse
[3] Protein Information Resource (PIR) – Protein Information Resource, University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA/États-Unis

Poster 97

La Ressource Universelle de Proteines (The Universal Protein Resource - UniProt) est une ressource centrale complète de séquences de protéines et d'annotations fonctionnelles avec de nombreux liens vers d'autres ressources complémentaires. UniProt est un consortium qui repose sur une large infrastructure bioinformatique et une vaste expertise scientifique. UniProt est le fruit de la collaboration entre l'Institut Européen de Bioinformatique (European Bioinformatics Institute – EBI), la Ressources d'Information des Proteins (Protein Information Resource - PIR) et l'Institut Suisse de Bioinformatique (Swiss Institute of Bioinformatics – SIB).

La base de connaissance d'UniProt (UniProt Knowledgebase – UniProtKB) fournit une large collection de protéomes à travers le portail de protéomes (1). Un protéome est défini par l'ensemble des protéines connu pour être exprimé par un organisme et il est généralement obtenu à partir de la traduction du génome entièrement séquencé et annoté. Ces dernières années ont vu une augmentation considérable des soumissions de plusieurs génomes pour le même organisme ou pour des organismes très proches dans la classification taxonomique. Les protéomes sont classés en trois types qui ne sont pas mutuellement exclusifs [1] : les Protéomes de Référence, les Pan Protéomes et les Protéomes Redondant.

Les Protéomes de Référence sont choisis pour fournir une large couverture de l'arbre des espèces et ils constituent un échantillon représentatif de la diversité taxonomique trouvée dans UniProtKB. Ces protéomes sont soit sélectionés par la communauté scientifique, notamment les organismes modèles et autres protéomes d'intérêt pour la recherche biomédicale et biotechnologique ; soit ils sont determinés de manière informatique [2]. Pour chaque Protéomes de Référence UniProtKB fournit des fichiers spécifiques disponibles sur le FTP d'UniProt (2). Dans les répértoires de dépot, le nom des fichiers sont préfixés avec l'identifiant du protéome (UPID) et l'identifiant taxonomique. Pour chaque Protéome de Référence, l'utilisateur peut récupérer : les séquences protéiques canoniques, une séquence par gène (`fasta.gz`) ; les séquences protéiques supplémentaires pour les gènes ayant des isoformes ou variants (`additional.fasta.gz`) ; les séquences d'ADN (CDS) des séquences protéiques canoniques (DNA.fasta.gz) ; la liste des noms de gène associés aux accessions UniProtKB des séquences protéiques canoniques (gene2acc.gz) ; toutes les bases de données de références croisées (xrefdb) et leurs identifiants associés aux accessions UniProtKB des séquences protéiques canoniques (`idmapping.gz`) [3].

Un Pan Protéome est l'ensemble complet des protéines considéré pour être exprimé par un groupe d'organismes très liés (par exemple plusieurs souches de la même espèce bactérienne). Les Pan Protéomes fournissent un ensemble représentatif de toutes les séquences au sein d'un groupe taxonomique et ils capturent les séquences uniques qui ne se trouvent pas dans le groupe des Protéomes de Référence. Les Pan Protéomes d'UniProtKB englobent tous les protéomes non

*. Intervenant
†. Corresponding author : benoit.bely@ebi.ac.uk
‡. Corresponding author : martin@ebi.ac.uk

redondants et ils sont destinés aux utilisateurs intéressés par des comparaisons phylogénétiques, par l'étude de l'évolution du génome et de la diversité génétique. Sur le portail des protéomes du site `http://www.uniprot.org/`, quand un protéome a des protéines qui font partie d'un plus grand Pan Protéome, la page décrivant ce protéome contient un lien 'Pan Proteome' vers le protéome de référence. Un lien pour télécharger les séquences fasta de l'ensemble complet du Protéome de Référence est aussi disponible. Les données des Pan Protéome sont disponibles sur le FTP pour téléchargement (3).

Ces dernières années, UniProtKB a connu une croissance exponentielle avec une augmentation de deux fois du nombre d'entrées en 2014. Cela fait suite à la soumission accrue de plusieurs génomes pour le même organisme ou pour des organismes très proches dans la classification taxonomique. Cela a conduit à un niveau élevé de redondance dans la section UniProtKB/TrEMBL. Cette section (UniProtKB/TrEMBL) représente les entrées générées de manière automatique en contraste avec la section UniProtKB/SwissProt qui correspond aux entrées annotées manualement par les curateurs. Par conséquent, de nombreuses séquences étaient sur-représentées dans la base de données. Cela était particulièrement vrai pour les espèces bactériennes où les gennomes de différentes souches de la même espèce ont été séquencés, annotés et soumis. Deux exemples extrêmes sont *Mycobacterium tuberculosis* et *Staphylococcus aureus* qui contenait respectivement 1.692 et 4.080 souches, ce qui correspond à 5,97 millions et 10,88 millions entrées UniProtKB/TrEMBL dans la version 2015_03. Pour réduire cette redondance, nous avons mis au point une procédure pour identifier les protéomes hautement redondantes au sein de groupes d'espèces en utilisant une combinaison de méthodes manuelles et automatiques [4]. Nous avons appliqué cette procédure aux protéomes bactériens à compter de la version 2015_04. Dans la version 2015_03 les entrées UniProtKB/TrEMBL bactérienes constituaient 82 % de l'ensemble des entrées UniProtKB/TrEMBL. Les entrées UniProtKB/TrEMBL correspondant aux potéomes redondants ont été retirés de UniProtKB qui représente une baisse de 51 % ; 47,0 millions d'entrées supprimées sur 92,6 millions d'entrées au total dans UniProtKB. Depuis la version 2015_04, nous ne générons plus de nouvelles entrées UniProtKB/TrEMBL pour protéomes identifiés comme redondants. Entre les version 2016_04 et 2015_04, cela représente 81,2 millions de séquences protéique supplémentaires maintenus en dehors UniProtKB. Les séquences des protéome redondants sont disponibles dans la base d'archive de séquences UniParc. Tous les protéomes redondants et non redondants restent consultables via le portail de protéomes (1). Récemment, nous avons observé même tendance des protéomes redondants pour les champignons/levures et nous prévoyons d'appliquer la suppression des protéomes redondants sur les champignons/levures avant la fin de 2016.

En conclusion, les différentes façons d'organiser les protéomes en différents types dans UniProtKB aident les utilisateurs à trouver les données qu'ils recherchent de manière très efficace et cohérente. Le retrait ds protéomes redondants d'UniProtKB permet également aux utilisateurs d'effectuer leur recherche de manière plus efficace. Les utilisateurs qui veulent récupérer des séquences protéiques de protéomes redondants peuvent aussi le faire à partir de Uniparc.

Références :

(1) `http://www.uniprot.org/proteomes/`

(2) `ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/`

(3) `ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/pan_proteomes/`

[1] UniProt Consortium. "Reorganizing the protein space at the Universal Protein Resource (UniProt)." *Nucleic acids research* (2011): gkr981.

[2] Chen, Chuming, et al. "Representative proteomes: a stable, scalable and unbiased proteome set for sequence analysis and functional annotation." *PloS one* 6.4 (2011): e18910.

[3] Altenhoff, Adrian M., et al. "Standardized benchmarking in the quest for orthologs." *Nature Methods* (2016).

[4] Bursteinas, Borisas, et al. "Proteome Redundancy in UniProt" *(in press)*

Mots clefs : Protéome de référence, Protéome Redundant, Pan Protéome, Base de données de protéines

Appliances « clé-en-main » pour des applications bioinformatiques avec CYCLONE

Poster 98

Bryan Brancotte [*1], Mohamed Bedri[1], Jonathan Lorenzo[1], Sandrine Perrin[1], Awa Sepou Ngaïlo[1], Christophe Blanchet[1,2], Jean-François Gibrat[1,3]

[1] IFB-CORE – Inserm : US21, CNRS : UMS3601, Université Paris XI - Paris Sud – 1 avenue de la Terrasse, Bâtiment 21, F-91 190 Gif sur Yvette, France
[2] Institut de Biologie et Chimie des protéines (IBCP) – CNRS : FR3302, Université de Lyon – France
[3] INRA, UR1404 Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE) – Institut national de la recherche agronomique (INRA) – Bâtiment 210-233, Domaine de Vilvert, F-78 350 Jouy en Josas Cedex, France

Faire face au déluge de données en bioinformatique est un challenge tant scientifique que technique. Répondre à ce challenge technique passe par la proposition de nouvelles technologies supportant les traitements des dites données. Le projet CYCLONE est un projet Horizon 2020 « Actions d'innovation » financé par la Commission Européenne. CYCLONE vise à intégrer et améliorer des solutions open-source pour la gestion de clouds telles que StratusLab, OpenStack, OpenNaaS, SlipStream, et TCTP. Le but est de permettre une gestion unifiée de différents clouds, le déploiement et le maintien de plateformes de traitement de données complexes dans une architecture multi-clouds afin d'assurer une fiabilité de service, une réactivité et une élasticité dans l'utilisation des plates-formes proposées. Dans le cadre de CYCLONE, l'Institut Français de Bioinformatique (IFB) est en charge de la définition des besoins et cas d'utilisation en bioinformatique, la proposition d'images de machines virtuelles prédéfinies (appliances) répondant à ses besoins, la formations des utilisateurs à l'utilisation des appliances, mais aussi à la création de nouvelles.

Deux domaines applicatifs, et leurs cas d'utilisation associés, guident les développements des outils de CYCLONE : l'un d'eux est le cloud IFB et certaines de ses applications bioinformatiques encapsulées sous forme d'appliance.

Le premier cas d'utilisation concerne la sécurisation des données biomédicales humaines dans le cadre d'un traitement sur le cloud. Dans ce cas, les politiques de confidentialité des données sont strictes. La proposition de CYCLONE pour répondre à ce besoin est une authentification unifiée basée sur la fédération européenne d'identités numériques eduGAIN (utilisée par ailleurs pour eduroam) à la fois pour un accès web ou en ligne de commande (ssh). Le transport sécurisé des données est assuré par un chiffrage des données de bout-en-bout avec TCTP. Finalement les machines virtuelles hébergeant les données et les plateformes les traitant sont placées dans un réseau isolé en utilisant OpenNaaS. Ce dernier permet d'orchestrer le déploiement d'un VPN pour un cluster de machines virtuelles en cours d'exécution dans un environnement multi-clouds.

Un second cas d'utilisation concerne le traitement de données -omics toujours plus volumineuses. Dans l'ère du post-NGS l'obtention de génomes est devenue peu coûteuse (quelques centaines d'euros) pour une grande partie des fournisseurs de séquençage. Dans le cadre de CYCLONE, nous avons pour objectif l'intégration multi-clouds d'un pipeline pour l'annotation de génomes microbiens (AGMIAL, [1]). Dans un premier temps, nous proposons l'outil Insyght pour visualiser les syntonies (conservations locales de gènes dans les génomes). Ces applications demandent le déploiement en un clic d'un cluster de machines virtuelles configurées à la volée pour traiter les données, le tout dans un réseau informatique isolé avec une authentification basée sur eduGAIN. Le déploiement de ce cluster inclurera aussi à terme une élasticité, c'est-à-dire le dimensionnement de la puissance de calcul en fonction des besoins même après la phase

*. Intervenant

d'initialisation.

De nombreuses appliances sont proposées par le cloud IFB. Afin de permettre aux utilisateurs de trouver rapidement celles répondant à leurs besoins, nous avons mis en place RAINBio (`http://cloud.france-bioinformatique.fr/rainbio/`). RAINBio intègre les informations extraites depuis bio.tools [2], un registre d'outil bioinformatiques où les services et leurs entrées-sorties sont étiquetés avec l'ontologie bioinformatique EDAM [3]. Cette ontologie propose des termes sémantiques pour les données, les formats, les opérations de traitement et les domaines thématiques.

Pour répondre à la volumétrie de données toujours plus importante, CYCLONE permet d'intégrer des solutions logicielles matures afin de répondre 'a certaines des problématiques bioinformatiques. CYCLONE permet ainsi une facilité d'utilisation de pipelines complexes d'analyse de données biologiques avec un déploiement effectif de ces pipelines d'analyse en environnement multi-clouds, un haut niveau de sécurité réalisé par une authentification reposant sur la fédération d'identité et un placement des appliances au sein de réseaux isolés et sécurisés. CYCLONE a également comme perspective la mise en place la flexibilité dynamique de l'allocation et de la répartition des ressources de calcul, dite 'élasticité' du cloud, pour les application considérées.

Références

[1] Bryson, K., Loux, V., Bossy, R., Nicolas, P., Chaillou, S., Van De Guchte, M., ... & Gibrat, J. F. (2006). AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. *Nucleic Acids Research*, 34(12):3533-3545.

[2] Ison, J., Rapacki, K., Ménager, H., Kalaš, M., Rydza, E., Chmura, P., ... & Booth, T. (2016). Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic acids research*, 44(D1):D38-D47.

[3] Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., ... & Rice, P. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, 29(10):1325-1332.

Mots clefs : cloud, turnkey appliance, traitement de données biomédicale, appliance clé en main

Using Docker for automatic Galaxy deployment

Jocelyn Brayet [*1], Vivien Deshaies[1], Nicolas Servant[1], Alban Lermine[1]

[1] Institut Curie, PSL Research University, Mines Paris Tech, Bioinformatics and Computational Systems Biology of Cancer, INSERM U900, F-75005, Paris, France – Institut Curie – 26 rue d'Ulm, F-75 248 Paris Cedex 05, France

Poster 99

In this abstract, we present a project of an automatic one-clic deployable fully operational Galaxy [1] instance. Despite the Docker's Galaxy distribution [2], that helps in setting up and configuring a Galaxy server, there are still difficulties in setting up tools from the Toolshed. Indeed, getting a tool from the Toolshed means setting up all dependencies required, such as packages (Python, R), binaries, system librairies, etc, which are all OS dependent. This can lead to error during installation, or even differences in the tool results.

We propose here a tool allowing an easy Galaxy and ToolShed tools install, by reducing tools dependencies in one Docker [3] container. With this scheme, each component of a Galaxy instance runs its own Docker image, which ensure the deployment whatever the infrastructure.

As a proof of concept, we developed and shared 20 Nebula [4] tools (ChIP-seq analysis) and their associated Docker's container in the main ToolShed. We also created a web form which allows thes user to design its own Galaxy instance by selecting each Galaxy component, such as the required packages. When this form is submitted, shell scripts are automatically created, to install Galaxy Core (Docker version) and add required tools (Docker version) from the main ToolShed. Biologists can deploy an instance by project with useful tools. That leads to make evolve the classic scheme, one Galaxy server centralising data, to an on demand service deployable anywhere at anytime.

References

[1] Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biol.* 2010 Aug 25;11(8):R86.

[2] Web site: `https://github.com/bgruening/docker-galaxy-stable`

[3] Web site: `https://www.docker.com/`

[4] Boeva, V, Lermine, A, Barette, C, Guillouf, C, and Barillot, E. "Nebula – a web-server for advanced ChIP-seq data analysis." *Bioinformatics*, 28 (19), 2517-2519 (2012).

Mots clefs : Galaxy, ToolShed, Docker, automatic deployment, Nebula

*. Intervenant

WAVES : a web application for versatile evolutionary bioinformatic services

Marc Chakiachvili [*1], Floréal Cabanettes[1], Vincent Berry[1], Anne-Muriel Arigon Chifolleau[1], Vincent Lefort [†1]

Poster 100

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – CNRS : UMR5506, Université de Montpellier – CC 477, 161 rue Ada, F-34 095 Montpellier Cedex 5, France

Abstract

Background: Over the last decade, the number of available online bioinformatic services has dramatically increased. The BioCatalogue (Bhagat et al. 2010) references almost 1,200 services, from which nearly 250 propose REST APIs. These services rely on different software platforms, web-oriented frameworks and content management systems. They provide a large number of disparate ways to interact and a huge variety of user interfaces. Moreover, most of the services dedicated to evolutionary studies require access to powerful computing resources to answer the users' queries in a reasonable time. The abundance of software layers leads to an increasing complexity which is becoming hard to deal with. Thus, nowadays the development of new services requires a large range of different skills.

Results: We propose a versatile service-oriented web application, named WAVES, designed to gather a comprehensive selection of evolutionary bioinformatic software under a unique application programming interface (API), independent of the underlying computing system. The use of WAVES through its API allows bioinformatician to easily integrate predefined tools or workflows and let them focus on designing high-level user interfaces. WAVES is currently being integrated into the ATGC platform [`http://www.atgc-montpellier.fr/`].

Conclusion: The WAVES application serves as a basis for further web interface developments to provide user-friendly interfaces for running analysis tools and workflows. WAVES may be used through its RESTful API. It also automatically generates web forms to be integrated into any web interface, easing its integration in bioinformatic service platforms.

Introduction

On-line web portals such as Galaxy (Goecks et al. 2010) or Mobyle (Néron et al. 2009) help the integration of bioinformatic tools and make them accessible through a general web user interface. As these interfaces are not customizable, it is difficult to provide a high-level web user experience. Furthermore, these interfaces were not designed to be integrated in a global front-end which would present the underlying tools as Web services, hence specific code has to be written every time a new tool is added to the interface. WAVES was developed with the aim to overcome some of these difficulties, especially to prevent the web administrators from having to configure each service separately.

*. Intervenant

†. Corresponding author : `Vincent.Lefort@lirmm.fr`

System overview

The main goal of WAVES is to provide an integrated web-oriented interface for evolutionary bioinformatic tools. To achieve this task, WAVES implements the Facade design pattern (Gamma et al. 1994): it provides an abstraction layer between the computing infrastructure and the user interface (see Figure 1). Thus, it gives access to job submission via simple web forms or direct RESTful API calls. WAVES was designed to communicate with several computing environments. It is able to run analyses on Galaxy instances or through direct cluster connexions. It may also redirect service calls to other remote system, such as CIPRES (Miller et al. 2010).

Tools as services

Within WAVES, each available tool is seen as a black-box service defined by its inputs, its parameters and its outputs. For each service, WAVES dynamically generates a web form based on its inputs and parameters. This web form is intended to be integrated into any web interface, letting web developers focus on their web site design. WAVES also provides a JSON/XML RESTful API to gain access to analysis submission configuration.

Authentication / authorization

Each WAVES API service call is subject to the authentication and authorization policies defined by the administrator. WAVES allows fine-grained configurations to define access rules. Depending on the user, WAVES may show different services or hide those which are not intended for public access. For instance, a new tool which is still under review can be made available only to the reviewers for whom an authenticated access has been set up. Administrators can grant or restrict access to services for users or groups.

WAVES installation

WAVES is a Django web application. As a stand-alone application, WAVES is able to deliver web pages to end users and it provides comprehensive back-end functions to set up services. Therefore, it can be deployed as easily as any other Django project. Moreover, WAVES has been conceived as a Django standard application, meaning that it may be integrated into any existing Django project as an application dependency.

Implementation

WAVES is developed in Python 2.7 and is based on Django (Wiles et al. n.d.) and the Django-REST framework. Front-end pages are designed with Bootstrap 3 standard components for responsiveness, and some jQuery elements in a single page application. Django and its components are open-source software and developed by a large community (more than 10k people) and benefit from a large set of reusable packages. Rendering WAVES pages in responsive design enables the application to be displayed on many platforms, from desktop computers to tablets and mobile devices. WAVES provides simple job spool management and while not meant to be a complex job scheduler, it takes advantage of existing and popular tools such as Galaxy or the DRMAA API (Rajic et al. 2001) to manage jobs submission, load balancing and jobs monitoring.

Back-end:

– *User management*

WAVES users can be either humans, local or remote platforms (web sites or stand-alone programs). The user management provides configuration tools for services and users. It may be used to define user access rights. Each service may be accessed by registered or anonymous users, depending

on service policy. Once the user accessrights are defined, each non-human user is automatically provided an API key. This API key is used to identify the API calls. Then, each subsequent API call must include the user's API key, avoiding the need for the users to provide their credentials at each API call.

– *Service management*
WAVES service definition includes several information such as name, description, inputs, parameters, related outputs, etc. The service definition also links services to their related papers, URLs for download, dedicated websites, etc. Each service has to be associated with a job runner adapter. A runner adapter is responsible for launching a job on a specific computation platform and retrieving its outputs. Current implemented runner adapters are: Galaxy tools and workflows execution, DRMAA compliant computing resources, we plan to develop specific remote API calls adapter to access dedicated RESTful services.

Database

The objects relational model (ORM) of WAVES is currently hosted on a sqlite3 database. Django allows ORM integration of many standard DBMS such as MySQL or PostgreSQL and non relational models like MongoDB. WAVES can therefore be deployed on any of these persistence layers.

Job spool

The back-end of WAVES displays the current jobs submitted by the users. Once submitted, jobs enter a FIFO (First In First Out) queue to be treated and relayed to the relevant adapter at regular time intervals. Each job can be cancelled by its owner as long as it is not handled by a runner. After completion, jobs can be deleted either by administrators or by allowed users.

Service Import Adapter

WAVES is able to read the Galaxy tools (or workflow) descriptors in order to automatically integrate new services. Through the Bioblend API (Sloggett et al. 2013), WAVES can set up new services directly from Galaxy tool and workflow definitions, hence automatically synchronize related forms and API entry point to final users. This capability may be extended to other dedicated runners.

Front-end

WAVES comes with some already made front-end pages. These pages are designed for service list displays, service details, job submissions and result retrievals. All publicly available registered WAVES services can be listed. For each service within the list, WAVES provides a link to the service details, including a job submission form. Other pages are intended to list jobs outputs, with a link to job details, and download links for output data.

REST API end-points

WAVES exposes its available services through a RESTful API. It gives access to service lists, service details and job submissions in the well-known JSON or XML data exchange formats. Another available output format is plain HTML for specific end-pointsdedicated to web form access, which allows the integration of job submission web forms on remote web sites.

Perspectives and future works

Websites using WAVES

The ATGC platform will soon migrate its evolutionary analysis tools on WAVES, allowing ATGC to provide an integrated on-line environment dedicated to phylogenomic studies. New services will also be deployed.
Two websites dedicated to evolutionary analyses will use WAVES. The on going next version of phylogeny.fr (Dereeper et al. 2008) will use WAVES capabilities to launch dedicated workflows on a Galaxy remote platform. The upcoming version of CompPhy (Fiorini et al. 2014) will use WAVES REST services for supertree computation.

Implementing runners

Other implementations of runner adapters can be set up in WAVES. These implementations could be integrated in the main source upon request. The CIPRES science gateway provides an API (Miller et al. 2015) which could be integrated within WAVES as a dedicated job runner. As long as they comply with the job submission interface contract, runners can be extended to send queries to remote phylogenomic databases, in this context "jobs" are in fact data retrieval services.

Django CMS standard application

Nowadays, most modern web sites rely on Content Management Systems. By separating the web content from its formatting, such systems allow fast information updates. Strong CMS solutions such as Django-CMS [`http://www.django-cms.org/`] or Mezzanine [`http://mezzanine.jupo.org/`] are largely used in production servers. Currently, WAVES cannot be easily installed within this kind of system. But the development of plugins dedicated to CMS integration should be straightforward. WAVES has been designed with this final idea in mind.

Conclusion

WAVES is a service-oriented web application meant to gather many computation architectures within a single unified API, mainly dedicated to run evolutionary bioinformatic tools. Presenting tools as RESTful services and enabling the use of current web standards will help web developers to focus on designing user friendly interfaces. Thus, WAVES will ease the development of new bioinformatic services, independent of the underlying computing infrastructure.

WAVES is delivered as a stand-alone web application. It can easily be downloaded and installed on any web server running Python. Initially conceived to provide evolutionary bioinformatic services, WAVES is an open-source project and as such is intended to evolve and upgrade in the near future. We hope that a community will grow around the project, and help it become a standard tool.

Project home page: `https://sourcesup.renater.fr/projects/waves/`

Acknowledgements

This research was supported by the "Institut Français de Bioinformatique" (RENABI-IFB, Investissements d'Avenir).

Mots clefs : Web based platform, Bioinformatics, Workflow, RESTful API

Functional proteomics analysis using Galaxy workflows

Cathy Charlier [*1], Mathilde Laine[1], Delphine Feron [*1], Julien Gras[2], Stéphane Téletchéa [*3], Damien Eveillard[2], Pierre Weigel[1]

Poster 101

[1] Plateforme IMPACT (Interactions Moléculaires Puces ACTivités) – Universté de Nantes, CNRS : UMR6286 – Plateforme IMPACT « Interactions Moléculaires Puces Activités » UMR CNRS 6286 - UFIP Université de Nantes - Faculté Sciences & Techniques, 2 rue de la Houssinière, BP 92208, F-44 322 Nantes Cedex 3, France

[2] Laboratoire d'Informatique de Nantes Atlantique (LINA) – CNRS : UMR6241, Université de Nantes, École Nationale Supérieure des Mines - Nantes – LINA - Faculté des Sciences, 2 rue de la Houssinière, BP 92208, F-44 322 Nantes Cedex 3, France

[3] Unité de Fonctionnalité et Ingénierie des Protéines (UFIP) – Université de Nantes, CNRS : UMR6286 – 2 rue de la Houssinière, Bâtiment 25, F-44 322 Nantes Cedex 3, France

La plateforme IMPACT (Interactions Moléculaires Puces ACTivités) propose différentes approches de criblages et d'analyses multi-paramétrées en protéomique fonctionnelle. Dans le cadre de ses activités dédiées au « Bio-profiling », la plateforme conçoit et développe des puces à protéines.

Ces systèmes miniaturisés et multiplexés génèrent un grand nombre de données nécessitant une expertise importante en terme d'analyse et d'interprétation. Pour faciliter ces étapes d'analyse pour les utilisateurs de la plateforme, nous avons développé des outils bioinformatiques accessibles sur un portail Galaxy. Ces applications permettent de réaliser le traitement des données et leurs analyses statistiques et proposent des méthodes complémentaires d'exploration et de représentation des données. Ces méthodes d'analyses sont et seront intégrées dans le portail galaxy mis en place au sein de la plateforme BiRD. L'objectif est de pouvoir utiliser ces outils individuellement ou de les incorporer dans des workflows spécifiques dédiés à chaque jeu de données.

L'analyse expérimentale complémentée par le traitement bioinformatique des données a déjà été appliquée pour étudier les sérums d'une cohorte de patients sur puces à protéines dédiées. L'automatisation de ce protocole d'analyse globale a permis d'améliorer la qualité du traitement des données et des profils d'expression caractéristiques de certains groupes de patients ont pu être mis en évidence.

La plateforme IMPACT continue cette stratégie de développement en couplant les analyses de protéomique fonctionnelle avec une analyse experte des données sur des projets en cours, ou sur des projets spécifiques en fonction des besoins des collaborateurs. Ces protocoles d'analyse automatisée sont proposés à la communauté scientifique et industrielle. Pour tous renseignements vous pouvez nous joindre sur le site de la plateforme : `http://www.impact-plateforme.com/`.

Mots clefs : galaxy workflow, protein array analysis

*. Intervenant

Recent updates on Norine, the nonribosomal peptide knowledge-base

Poster 102

Yoann Dufresne [*] [†1,2], Areski Flissi[1,2], Laurent Noé[1,2], Valérie Leclère[2,3], Maude Pupin [‡1,2]

[1] Équipe Bonsai, Laboratoire CRIStAL, UMR 9189 (CRIStAL) – CNRS : UMR9189 – Université Lille 1, Bâtiment M3, France

[2] INRIA Lille - Nord Europe (INRIA Lille - Nord Europe) – INRIA – Parc Scientifique de la Haute Borne, 40 avenue Halley, Bâtiment A, Park Plaza, F-59 650 Villeneuve d'Ascq, France

[3] Institut Charles Violette, Équipe ProBioGEM (probiogem) – Université Lille 1, Sciences et Technologies – Université Lille 1, France

Non-Ribosomal Peptides (NRP) [1,2] are a valuable source of biologically active molecules. These molecules, produced by several bacteria and fungi have various activities and pharmacological properties like antibiotics, siderophores, anti-tumor agents… For example, the very famous penicillin compound discovered a century ago, is an antibiotic derived from a NRP precursor.

These molecules are small polymers produced by enzymatic complexes called Non-Ribosomal Peptide Synthetases (NRPS). This alternative synthesis pathway allows the incorporation of a huge diversity of monomers. Some of them are the proteinogenic amino-acids and their variants but those polymers also include other kinds of monomers like sugars or fatty acids. In addition, the NRPSs produce sometimes linear polymers and often more complex 2D structures with cycles and branches. From this diversity of compositions and structures comes the diversity of activities that exceed the one of classical ribosomally synthesized peptides.

Since 2006, Norine [3] is the unique knowledge-base dedicated to nonribosomal peptides. Although other databases contain NRPs, Norine is the only one focused on NRPs and their annotations. The database now contains 1,179 peptides composed of 533 different monomers and grouped in 11 biological activities. The current annotations were manually extracted from research articles making it a valuable resource.

Since the very beginning, the database is paired with an online user friendly interface. Each NRP and each monomer have a dedicated web page showing detailed annotations. For example, experimentally validated biological activities are given, associated with the articles from which the information was extracted. But the most precious annotations are the monomeric structures, that are the monomers composing a peptide and the chemical bonds between them. These allow users to perform searches by structures over the database, with dedicated smart query tools. Recent developments give the possibility to directly draw a monomeric structure on the web browser and run a structural search.

As the quantity of articles describing NRPs is growing exponentially, we designed two complentary strategies. First, we opened Norine database to crowdsourcing [4]. Since 2015, each user can create an account on myNorine tool and start to submit new NRPs or suggest modifications on already added ones. All these submissions must contain a minimal set of annotations like the monomeric structure and a bibliographical reference to confirm the existence of the compound. After a verification process, the new peptides are available online with the name of the submitters on the peptide pages. Second, we are developping scripts to automatically extract peptides and their annotations from other databases. As in the UniProt knowledge-base, we will provide both manually annotated and automatically annotated entries.

*. Intervenant

†. Corresponding author : yoann.dufresne@ed.univ-lille1.fr

‡. Corresponding author : maude.pupin@lifl.fr

We also give the possibility to download the data through a REST API. If a user need the data for an external tool, he/she can now create a script for the downloading of the annotations that he/she need. This API also provides a direct way to integrate the last version of the database in any software. All the methods of this API are well documented on the Norine website.

Recently, we published a tool named Smiles2Monomers (s2m) [5] to infer monomeric structure of polymers from their atomic structure. We use this tool for several purposes in Norine. s2m is included in myNorine to help users determining the monomeric structure of their peptides. If they add a SMILES (textual format for 2D atomic structures), the tool provides automatically a predicted monomeric structure that can be manually corrected. Moreover, we use s2m to verify the correctness of the peptide structures stored in Norine. We detected few petides with differences between their manually designed monomeric structure and the automatically predicted one with s2m. We are currently curating these peptides. Finally, we plan to run s2m on all compounds of huge databases like PubChem to automatically detect NRP candidates that are not in Norine. s2m is a very helpful tool for increase the quality and quantity of data in Norine.

In conclusion, Norine is an essential software plateform that helps scientists analysing nonribosomal peptides by providing both valuable data and powerful tools.

References

[1] Marahiel M.A. A structural model for multimodular NRPS assembly lines. *Nat. Prod. Rep.* 2016;33:136-140. doi: 10.1039/C5NP00082C

[2] Walsh C.T. Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Nat. Prod. Rep.* 2016;33:127-135. doi: 10.1039/C5NP00035A

[3] Caboche S, Pupin M, Leclere V, Fontaine A, Jacques P, Kucherov G. NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.* 2008;36(Database issue):D326-31. doi: 10.1093/nar/gkm792

[4] Flissi A, Dufresne Y, Michalik J, Tonon L, Janot S, Noé L, Jacques P, Leclère V, Pupin M. Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowdsourcing. *Nucleic Acids Res.* 2016;44(D1):D1113-8. doi: 10.1093/nar/gkv1143

[5] Dufresne Y, Noé L, Leclère V, Pupin M. Smiles2Monomers: a link between chemical and biological structures for polymers. *Journal of Cheminformatics* 2015 7:62. doi: 10.1186/s13321-015-0111-5

Mots clefs : Non, Ribosomal Peptides, NRP, Norine, database, crowdsourcing

Integration and query of biological datasets with Semantic Web technologies : AskOmics

Poster 103

Aurélie Évrard [*] [†1], Charles Bettembourg [‡2], Melanie Jubault [§1], Olivier Dameron [¶2], Olivier Filangi [1,3], Anthony Bretaudeau [**1,3], Fabrice Legeai [††1,4]

[1] Institut de Génétique, Environnement et Protection des Plantes (IGEPP) – Institut national de la recherche agronomique (INRA) : UMR1349, Agrocampus Ouest – Agrocampus Ouest, UMR1349 IGEPP, F-35 042 Rennes, France, France

[2] DYLISS (INRIA - IRISA) – INRIA, Université de Rennes 1, CNRS : UMR6074 – Campus de Beaulieu, F-35 042 Rennes Cedex, France

[3] Plateforme bioinformatique GenOuest – Université de Rennes 1, Biogenouest – France

[4] GENSCALE (INRIA - IRISA) – Université de Rennes 1, CNRS : UMR6074, INRIA – Campus de Beaulieu, F-35 042 Rennes Cedex, France

Over the past few years, research programs involving genetic, genomic and post-genomic sequencing of various living organisms have become fast growing areas of biology. Once the computational challenges of processing datasets have been dealt with; large and complex biological data still remain in the hands of biologists for interpretation. Projects such as Biomart and Intermine have been developed for the international community to facilitate exchange and comparison of complex biological data. However for non-model organisms, large heterogeneous biological datasets can be difficult to associate in order to obtain a comprehensive view. Overall, access and interrogation remain time consuming for biologists and integrating publicly available data is still an open challenge.

Linked data and Semantic Web technologies benefit to biologists. Using RDF (Reference Description Framework); biological data can be described using triples that associate an entity (called subject), a relation (called property) and a value for the relation (called object). Data from different datasets can be integrated and the SPARQL query language support their analysis. Nevertheless; understanding and acquiring the query language can be a daunting task for biologists.

Here we present AskOmics, a tool supporting both intuitive data integration and querying while shielding the user from most of the technical difficulties underlying RDF and SPARQL. The virtualization-based deployment of AskOmics makes the tool easy to manage, reliable and simple to install. For data integration, the user loads his data as tabulation-separated files structured according to simple principles. This structure allows AskOmics to generate automatically the corresponding RDF triples, and to store them into a triplestore such as Fuseki or Virtuoso. At this point, the user's data are available just like in any SPARQL endpoint. AskOmics automatically generates an abstract representation of the dataset based on the types of the subject and object of its triples. For data querying, AskOmics provides a visually intuitive interface compatible with any SPARQL endpoint (that is one generated by AskOmics data generation function, or any regular triplestore). The user can then select a sequence of nodes in this simplified view, and AskOmics generates the corresponding SPARQL query that can be executed on the original dataset.

*. Intervenant
†. Corresponding author : aurelie.evrard@rennes.inra.fr
‡. Corresponding author : charles.bettembourg@irisa.fr
§. Corresponding author : melanie.jubault@agrocampus-ouest.fr
¶. Corresponding author : olivier.dameron@univ-rennes1.fr
Corresponding author : olivier.filangi@irisa.fr
**. Corresponding author : anthony.bretaudeau@rennes.inra.fr
††. Corresponding author : fabrice.legeai@rennes.inra.fr

For example, it could be difficult for biologists to identify features such as genes underlying localised genomic regions limited by genetic markers as it requires the users to combine different files. Tabulation-separated files containing genes and genetic markers could be uploaded in AskOmics with the following criteria: genetic markers and genes identified as entity, each entity is related to a chromosome and a position start and end with numerical values. AskOmics interface allows the user, without knowledge in SPARQL language, to either select genomic regions with distinct markers or simply provide numerical values as the lower and upper position. The intersection with additional features could be computed for producing lists of features such as genes underlying specific genomic regions. The result can then be downloaded as a tabulation-separated file. Currently under development, AskOmics will also support the integration of external databases to compare or complete new findings.

AskOmics' principle is generic. It has been applied successfully to the analysis of large scale datasets including genetic, epigenomic, transcriptomic profiles and orthologous relationships to identify genomic regions that are involved in the variability of Brassicaceae (Arabidopsis, cabbage, turnip and oilseed rape) in response to clubroot disease. About 2.6 millions of triples were stored from 370,000 uploaded entities corresponding to genomic positions of genes amongst the four species of the Brassicaeae family as well as relationship data (orthology and transcriptomics). The fast queries allowed to identify lists of genes with specific expression profiles and their corresponding orthologs in the three others species.

Mots clefs : Semantic Web technology, linked data, RDF, SPARQL, comprehensive view

Association des jeunes bioinformaticiens de france (JeBiF)

Poster 104

Léopold Carron[1], Sylvain Léonard[1], Gwenaëlle Lemoine[1], Emmanuelle Lastrucci[1], Nolwenn Lavielle[1], Julien Fouret[1], Bérénice Batut[1], Romy Chen-Min-Tao[1], Cédric Midoux[1], Lambert Moyon[1], Hugo Pereira[1], Julien Fumey [*] [†1]

[1] Association des Jeunes Bioinformaticiens de France (RSG France - JeBiF) – ISCB SC – France

L'association des Jeunes Bioinformaticiens de France (RSG France - JeBiF) a pour mission de structurer la jeune communauté bioinformatique au niveau local, national et international. Nous vous présenterons les activités de l'assocation.

Mots clefs : JeBiF, communauté, jeunes, bière, réseau, structuration, vulgarisation

*. Intervenant

†. Corresponding author : contact@jebif.fr

Bioinfo-fr.net : présentation du blog communautaire scientifique francophone par les *Geekus biologicus*

Poster 105

Nicolas Allias[1], Jérôme Audoux[1], Bérénice Batut[1], Marouen Ben Guebila[1], Pierre Bertin[1], Adem Bilican[1], Emmanuel Bouilhol[1], Lucas Bourneuf[1], Julien Buratti[1], Léopold Carron[1], Aurélien Chateigner[1], Guillaume Collet[1], Olivier Dameron[1], Julien Delafontaine[1], Clément Delestre[1], Gregory Farrant[1], Arnaud Ferre[1], Julien Fumey[1], Maxime Garcia[1], Nils Giordano[1], Laura Grégoire[1], Vincent Henry[1], William Jarassier[1], Fabrice Jossinet[1], Mathilde Leboudic-Jamin[1], Gwenaëlle Lemoine [*1], Sylvain Léonard[1], Gildas Lepennetier[1], Jean-Emmanuel Longueville[1], Nicolas Maillet[1], Christophe Malabat[1], Pierre Marijon[1], Clémentine Mercé[1], Alexis Michon[1], Ismaël Padioleau[1], Anaïs Painset[1], Thibaut Payen[1], David Picard Druet[1], Sylvain Prigent[1], Estelle Proux-Wéra[1], Raoul Raffel[1], Louise-Amélie Schmitt[1], Jonathan Sobel[1], Maria Sorokina[1], Rayna Strombiskova[1], Axel Thieffry[1], Aurélien Tylski[1], Yohan Jarosz[1], Nolwenn Lavielle[1], Yoann Mouscaz[* †1], Isabelle Stévant[1]

[1] Bioinfo-fr.net – Bioinfo-fr.net – France

Pourquoi ?

« Bioinformatique ? C'est quoi ça ? De l'informatique respectueuse des terres ??? ». Qui n'a jamais eu ce genre de remarque sur notre profession ? Qui n'a jamais peiné pour expliquer à son entourage son travail de tous les jours ?

Depuis quelques années, une petite communauté de bioinformaticien·ne·s francophones s'est formée suite à un simple constat : il y avait bel et bien un trou dans l'Internet francophone par rapport à notre science ! Nous ne pouvions pas laisser cela tel quel, c'était devenu notre mission. `www.bioinfo-fr.net` était né !

Qui ?

Au commencement, nous n'étions qu'une petite dizaine avec tout plein d'idées par dessus la tête et une envie commune de faire avancer les choses. Aujourd'hui la recette a séduit, les fondations ont été posées, le mécanisme est bien huilé et nous sommes plus d'une soixantaine de collaborateurs bénévoles dispersés de partout à travers le monde, toujours avec la même motivation. Pas de hiérarchie stricte clairement mise en place, si ce n'est un petit groupe d'administrateurs dont les rôles principaux sont de maintenir le site à jour, veiller à la bonne cohérence des articles, fournir un planning de publication, accueillir les nouveaux venus, relancer les auteurs et les relecteurs de

*. Intervenant
†. Corresponding author : admin@bioinfo-fr.net

temps en temps, communiquer avec l'extérieur et essayer de mettre en place les nouvelles idées venant de tout un chacun.

Comment ?

D'abord via un canal IRC (#bioinfo-fr, réseau FREENODE). Mais très vite nous avons constaté que de nombreuses questions similaires ressortaient et qu'il serait plus efficient de garder les réponses sur un support écrit et de façon pérenne.

Le blog nous a semblé être le support le mieux adapté cela. Chaque semaine nous nous efforçons de fournir un article qui a subit un processus de relecture scientifique robuste et qui doit avoir un rapport de près ou de loin avec la bioinformatique. Un article doit entre-autre chose permettre au lecteur, averti ou non, de comprendre une méthode, de reproduire une expérience, de plonger directement dans un code solutionnant un problème biologique ou encore de l'informer sur une toute nouvelle découverte. Au fil des années et des articles nous avons essayé de décomposer ces articles en catégories afin de faciliter la visite et la recherche du lecteur.

Bilan

En quelques années nous avons réussi à tisser un large réseau de professionnels/étudiants/passionnés capable de produire des articles d'intérêt public et de répondre à des problématiques axées autour de la bioinformatique. Cela nous a permis également de mieux connaitre nos spécialités, nos manières de vivre, notre métier, et de représenter à notre niveau la force de la bioinformatique francophone. Nous ne comptons pas en rester là et vous encourageons fortement à venir discuter avec nous les *Geekus biologicus*. Peut-être, qui sait, franchirez-vous le cap et nous rejoindrez-vous dans l'aventure !

Mots clefs : blog, communauté, scientifique, bénévole, article, entre, aide, bioinformatique, web, tutoriels, astuces, explications, geek

Présentation d'Infogene

Gwenaëlle Lemoine *[1]

[1] INFOGENE – 19 rue d'Orléans, F-92 200 Neuilly sur Seine, France

Poster 106

INFOGENE est une Entreprise de Services du Numérique (ESN) spécialisée dans la conception, gestion et réalisation de projets (MOE et AMOA) dans les domaines suivants :

- DIGITAL : nous intervenons dans toutes les phases d'un projet (Architecture, conception, réalisation, intégration, maintenance) en assistance technique ou au forfait (nous avons un plateau de développement à Neuilly/Seine). INFOGENE met à votre disposition son expertise technique dans les technologies .NET/ORACLE/SQL SERVER, PHP/SYMFONY, et autres technologies plus spécifiques associées à de fortes compétences métier.
- BI & BIG DATA : Maîtrise technologique sur QlikView, SAPBI4, Microsoft BI (SSIS,SSAS, SSRS), SAS BI, COGNOS. Notre savoir-faire : Schéma directeur, migration, architecture, conception, développement, recette, maintenance.
- INGENIERIE R&D : Logiciels et Progiciels spécifiques à l'industrie pharmaceutique (ELN, LIMS, CTMS, CDMS, PV…) et plus largement à l'informatique scientifique. Notre savoir-faire : Accompagnement à la conception, développement, déploiement et support.

Nous recherchons à ce sujet des bioinformaticiens qui pourraient apporter leurs connaissances et compétences à nos clients. Nous avons à cœur que nos collaborateurs expriment tout leur potentiel et leur talent au travers des projets qui leurs sont confiés.

Nos domaines d'intervention en terme de périmètre fonctionnel sont multiples mais s'orientent principalement en R&D, Industrielle avec comme clients actuels SERVIER, SANOFI, LFB, IPSEN, NESTLÉ, ABBVIE, STALLERGÈNES, L'ORÉAL, TOTAL…

Mots clefs : Bioinformatique, Informatique scientifique, SSII, Big data, Emploi

*. Intervenant

Cluster de calcul de machines virtuelles en bioinformatique

Poster 107

Jonathan Lorenzo [*,†,1], Sandrine Perrin [‡,1], Awa Sepou Ngaïlo [§,1], Bryan Brancotte [¶,1], Mohamed Bedri [1], Blanchet Christophe [**,1], Jean-François Gibrat [††,1]

[1] IFB-CORE – Inserm, CNRS : UMS3601, Université Paris XI - Paris Sud – 1 avenue de la Terrasse, Bâtiment 21, F-91 190 Gif sur Yvette, France

Le projet BioDataCloud concerne l'utilisation d'un système de type cloud computing adapté à une production efficace de données à haute valeur ajoutée dans le domaine de la biotechnologie végétale. Ici, les besoins sont principalement basés sur l'assemblage de novo, la visualisation dans une même interface (type 'genome browser') de plusieurs individus pour une espèce donnée, le traitement des données génétiques ou encore l'analyse de données RNAseq sur des génomes non séquencés précédemment.

De façon à favoriser l'utilisation des outils bio-informatiques appropriés sur une infrastructure de cloud computing, nous avons encapsulé certains de ces logiciels dans des machines virtuelles (VM) prédéfinies, des appliances, pouvant être exécutées sur le Cloud IFB. A l'aide du cloud, nous pouvons adapter nos besoins en termes de calcul en spécifiant le nombre de processeurs et la taille de mémoire associés, mais aussi en termes de stockage en spécifiant la taille voulue pour les disques virtuels associés. Le cloud permet la mutualisation des ressources, un accès en self-service, un accès standard par le réseau, une « élasticité » des ressources informatiques utilisées et aussi une mesure fine de l'utilisation de ces ressources.

Dans le cadre du projet BioDataCloud, l'intégration d'outils sur des VMs a été effectuée dans un premier temps avec un système de scripts 'shell' appelé 'approver' mais ce dernier est dépendant de la version du système d'exploitation (OS) utilisé (CentOS 6 ou 7) ainsi que des paquets logiciels des dépendances. Une autre façon d'automatiser l'installation est l'utilisation de l'environnement de virtualisation légère Docker [1]. Ce dernier permet une installation facile d'un outil sans se soucier des dépendances qui peuvent poser problème selon les mises à jour et les versions différentes de l'OS au contraire du cas précédent. Un certain nombre d'outils ont déjà été intégrés avec Docker et importés dans le dépôt BioShaDock [2]. BioShaDock est un dépôt Docker, créé et maintenu par la plateforme GenOuest de l'IFB, et dédié à la bioinformatique et la biologie computationnelle. Une dernière façon d'automatiser l'installation est l'utilisation de l'outil Puppet (`https://puppet.com/`) qui permet le déploiement de logiciels et leurs configurations sur un ensemble de serveurs en quelques minutes. Une recette Puppet a été développée pour déployer DRAP (`http://www.sigenae.org/drap/index.html`) qui est le pipeline RNAseq utilisé pour l'analyse de données RNAseq sur des génomes non séquencés dans le cadre du projet BioDataCloud. Cette recette peut également être utilisée pour installer ce pipeline via une interface graphique avec le portail Foreman (`http://theforeman.org/`) pour simplifier d'avantage l'installation d'outils sans passer par une ligne de commande.

*. Intervenant
†. Corresponding author : jonathan.lorenzo@france-bioinformatique.fr
‡. Corresponding author : Sandrine.PERRIN@france-bioinformatique.fr
§. Corresponding author : Awa.SEPOUNGAILO@france-bioinformatique.fr
¶. Corresponding author : brancotte@lri.fr
Corresponding author : Mohamed.BEDRI@france-bioinformatique.fr
**. Corresponding author : Christophe.BLANCHET@france-bioinformatique.fr
††. Corresponding author : Jean-Francois.GIBRAT@france-bioinformatique.fr

Après avoir installé les outils nécessaires sur une VM, il est possible d'effectuer les calculs pour récupérer les résultats. Mais actuellement la taille maximale d'une VM sur le cloud IFB est de 16 cœurs et 250 Go de mémoire. Lorsque les calculs nécessiteront plus de ressources pour améliorer la rapidité de l'analyse, il sera utile d'utiliser plusieurs VM et de les configurer comme un cluster de calcul virtuel. De plus cela peut permettre à l'utilisateur d'avoir le même environnement que celui existant dans les plateformes et laboratoires. La configuration d'un cluster virtuel passe tout d'abord par l'installation d'un gestionnaire de calculs comme SGE [3], Torque [4] ou, plus récemment, SPARK (`http://spark.apache.org/`). Les différents gestionnaires ont été adaptés chacun dans une appliance permettant ensuite la création d'un cluster. L'installation est réalisée via approver. La configuration se fait ensuite à la volée via un script permettant de configurer les paramètres du cluster selon les VM créées (une VM maitre et plusieurs VM de calcul). Le script va aussi gérer l'installation du système de partage de données utilisé (NFS, SCP, HDFS).

Pour tous les types de gestionnaires de calculs proposés, le cluster de calcul est prêt pour des analyses après les configurations effectuées par le script adapté au gestionnaire. Le développement de ce script permet d'adapter et de simplifier la configuration du cluster de calcul sur des machines virtuelles. Les objectifs à terme sont de simplifier la création d'un cluster de calcul à l'aide d'une interface graphique, de réaliser l'intégration continue des nouvelles versions du pipeline et l'utilisation de conteneurs Docker n'incluant qu'un seul outil, combinés par un outil de workflow.

Références :

[1] Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, *2014*(239), 2.

[2] Moreews, F., Sallou, O., Ménager, H., Monjeaud, C., Blanchet, C., & Collin, O. (2015). BioShaDock: a community driven bioinformatics shared Docker-based tools registry. *F1000Research*, *4*.

[3] Gentzsch, W. (2001). Sun grid engine: Towards creating a compute power grid. In *Cluster Computing and the Grid, 2001. Proceedings. First IEEE/ACM International Symposium on* (pp. 35-36). IEEE.

[4] Staples, G. (2006, November). TORQUE resource manager. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing* (p. 8). ACM.

Mots clefs : Cloud computing, BigData, BigMem, Cluster, intégration d'outils, RNAseq

Rencontre autour de l'enseignement en bioinformatique en france (REBIF)

Poster 108

Alban Mancheron [*] [†1], Morgane Thomas-Chollier [*2], Céline Brochier-Armanet[3], Jacques Van Helden[4], Claudie Fabry[5], Sophie Schbath[6]

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – Université de Montpellier, CNRS : UMR5506 – CC 05016, 860 rue de St Priest, F-34 095 Montpellier Cedex 5, France
[2] Institut de biologie de l'école normale supérieure (IBENS) – École normale supérieure [ENS] - Paris, Inserm : U1024, CNRS : UMR8197 – 46 rue d'Ulm, F-75 005 Paris, France
[3] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL) – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[4] Technologies avancées pour le génome et la clinique (TAGC) – Inserm : U1090, Université de la Méditerranée - Aix-Marseille II – Parc scientifique de Luminy, 163 avenue de Luminy, F-13 288 Marseille Cedex 9, France
[5] Master Sciences & Numérique pour la Santé (SNS) – Université de Montpellier - Faculté des Sciences – Bâtiment 2, 860 rue Saint Priest, F-34 090 Montpellier, France
[6] Institut National de Recherche Agronomique - Centre de Jouy-en-Josas (MaIAGE) – Institut national de la recherche agronomique (INRA) – France

Les formations diplômantes en Bioinformatique (IUT, Masters) sont à l'heure actuelle assez cloisonnées, alors qu'elles rencontrent les mêmes difficultés liées, entre autre, à l'interdisciplinarité inhérente à ce domaine. Pour les futurs étudiants, le paysage des formations reste assez flou, et il n'est pas évident de s'orienter vers la formation la plus adaptée à leurs profils. Fin 2015, l'association des Jeunes Bioinformaticiens de France (JeBiF) a mis à jour la liste des formations en Bioinformatique (`https://jebif.fr/fr/bioinformatique/les-formations/`). La Société Française de Bioinformatique (SFBI) propose de réunir les acteurs de la formation en Bioinformatique au niveau national, afin de leur permettre d'échanger et surtout d'établir un véritable réseau de formations.

Les objectifs de cette manifestation sont multiples :

- partager les expériences et les stratégies mises en œuvre pour l'élaboration des offres de formation et assurer leur développement et leur pérennité ;
- permettre la mise en place de collaborations / échanges entre les formations ;
- développer des projets pédagogiques participatifs innovants entre les formations ;
- permettre un meilleur examen des candidatures des étudiants et leur proposer des formations en adéquation avec leur projet professionnel ;
- permettre une meilleure reconnaissance des diplômés, notamment pour les étudiants titulaires d'une Licence vis-à-vis de leur poursuite en Master (Bioinformatique ou autre) et pour les étudiants titulaires d'un Master vis-à-vis de leur poursuite en thèse (sélection des écoles doctorales) ;
- mettre en place une stratégie de reconnaissance de la transdisciplinarité comme une(des) science(s) à part(s) entière(s) (sections CNU, établissements) et des difficultés inhérentes au cloisonnement scientifique très largement instauré dans la structuration de la recherche mais davantage encore de l'enseignement en France.

Ces premières Journées REBIF auront lieu les 30 et 31 mai 2016 à Marne-la-Vallée (`http://www.sfbi.fr/rebif2016`) et accueilleront entre 30 et 35 personnes (plus de 20 formations seront représentées).

*. Intervenant
†. Corresponding author : `alban.mancheron@lirmm.fr`

Les deux posters présenteront le paysage actuel des formations et de leur spécificités, ainsi que les points discutés et les grandes conclusions de cette rencontre.

REBIF est soutenu par la Société Française de Bioinformatique (SFBI), l'Institut Français de Bioinformatique (IFB) et la Faculté des Sciences de l'Université de Montpellier.

Mots clefs : Enseignement, bioinformatique, LMD, étudiants, Licence, Master, formations, diplômes

The bioinformatics timeline of the data integration in BIOASTER

Poster 109

Yoann Mouscaz [*1], Audrey Cauchard[*1], Amila Malinovic[1], Laurène Picandet[1], Nicolas Sapay [†1], Pierre Veyre[1], Guillaume Boissy[1]

[1] Institut de Recherche Technologique BIOASTER – 40 avenue Tony Garnier, F-69 007 Lyon, France

Background

BIOASTER is the French Technology Research Institute in the fields of infectious diseases and microbiology. The Institute aims at fostering the transition between the proof of concept and the mature technology in the fields of diagnostics, vaccines, antimicrobials, or microbiota-based applications.

One of the main objectives of BIOASTER is to develop the capacity to analyze a same biological sample with several lab technologies in parallel: next generation sequencing, mass spectrometry, mass cytometry, nuclear magnetic resonance, imaging, etc. Hence, a single biological sample is the origin of a large set of molecular data.

However, the generation of a large data set is not enough to create a useful knowledge that can be directly interpreted by a clinician or a biologist. The data need to be contextualized and integrated by bioinformatics workflows. At first, the experimental data has to be cross-referenced with phenotypes, preclinical or clinical data in order to allow the interpretation of the biological results. Then, all those data have to be contextualized by metadata, such as the sampling conditions, the association to a particular metabolic pathway, etc. Finally, all the data have to be cured, packaged and shipped to the bioinformatician who will analyze them.

The data workflow of BIOASTER has also to take into account the industrial environment in which the institute takes root. As all the projects are issued from partnerships with academics, SMEs or industrial groups, the access and the availability of the data have to be secured.

Results

Addressing all these issues requires the implementation of a well-managed data life cycle within the institute. Initially, BIOASTER and the CNRS/IN2P3 computing center (CCIN2P3) have initiated in 2013 a partnership in order to deploy the capacity to host the data storage and computing resources for the BIOASTER's projects.

We are now implementing a set of platforms and services to collect, store, tag, cross-link, analyze and visualize our data. At first, we have defined a set of core services, defined as mandatory for all the projects and platforms:

- GitLab and Redmine to manage our developments.
- Shibboleth, a universal identity provider. The Shibboleth server provides all the information related to the users, in particular identification/authentication tokens.
- OpenStack, the cloud platform of the CCIN2P3. It can host the web services and computing resources required by the bioinformaticians.
- Ansible, a machine deployment manager.

*. Intervenant

†. Corresponding author : nicolas.sapay@bioaster.org

In parallel, we have also developed or deployed several web interfaces in order to collect the data generated by the projects:

— OwnCloud, a data sharing service opened to all projects and partners. The OwnCloud instance is interfaced with one of the CCIN2P3 file systems, in order to provide a scalable storage space accessible by all our projects. Its role is to centralize the raw data generated by the labs.
— ERDC, an electronic case report form (eCRF). ERDC is a web serviced developed and edited by Clinfile to collect the clinical data generated by clinical studies. ERDC is fully compliant with the legal aspects associated to those data.
— Biospecimens (biospecimens.bioaster.org), a collaborative platform which brings together the project leaders and the biological sample holder in the field of infectious diseases and microbiota. Biospecimens is developed internally.
— NoE, a lab information management system (LIMS). The project is currently under development. NoE will collect both the lab data and metadata at the source, as well as manage the biological sample within the institute. Its main goal is to allow a perfect reproducibility of the experiments described in its database.

Conclusions

The core services and the data collection services constitute the first links of the data supply chain in BIOASTER. The chain has to be continued until the final user, i.e. the biologist or the clinician that will interpret the data.

At present, we are focusing our efforts on data analysis and management platforms:

— The Symbiosis and Selfy projects will provide a web interface to the bioinformatics analysis workflow developed internally. They will be interfaced with OpenStack and the file systems of the CCIN2P3.
— TranSMART as a Web service to browse clinical studies associated to high dimensional data (for example gene expression data). The tranSMART service is provided by the eTRIKS platform (`http://www.etriks.org/`). It needs to be fed by an internal curation workflow that will cross reference the clinical data from the eCRF with high dimensional data from our omics technologies.

As a conclusion, the data time line in BIOASTER is a unique approach where the bioinformatics workflows are built upon open-source solutions, high throughput computing, a rigorous software engineering, legal constraints and regulatory requirements. All those aspects strongly promote the deployment of reusable bricks, allowing a better scalability and adaptability to the projects.

Hence, each brick is a step toward an integrative scientific information system dedicated to the conversion of raw massive data from heterogeneous origins into actionable knowledges understandable by the clinicians and the biologists.

Mots clefs : analyze, sample, next generation sequencing, bioinformatics workflow, storage, data, experiment, LIMS, eCRF, biocollection, database, visualisation, metadata, scalability, scientific information, clinical data, security

La plateforme PRABI-AMSB – analyse et modélisation des systèmes biologiques

Poster 110

Dominique Guyot[1], Vincent Navratil[1], Philippe Veber[1,2], Christine Oger[1], Guy Perrière [*] [†1,2]

[1] Pôle Rhône-Alpes de Bioinformatique (PRABI) – Université Claude Bernard - Lyon I (UCBL) – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[2] Laboratoire de Biométrie - Biologie Évolutive (LBBE) – CNRS : UMR5558 – France

Le PRABI (Pôle Rhône-Alpes de Bioinformatique, est l'un des six centres régionaux membre de l'Institut Français de Bioinformatique (IFB). Il est labellisé IBiSA depuis la mise en place de ce label. En tant que centre IFB, le PRABI fédère les activités de recherche, de service et de formation en bioinformatique de la Région Auvergne Rhône-Alpes. Cette fédération regroupe sept composantes comprenant aussi bien des équipes de recherches que des plateformes de services (`http://www.prabi.fr/`).

La composante PRABI-AMSB (Analyse et Modélisation des Systèmes Biologiques) est une plateforme de l'Université Claude Bernard – Lyon 1 (UCBL) rattaché à la Fédération de Recherche Bio-Environnement et Santé (BioEnviS). L'activité de cette plateforme est orientée vers :

- La recherche et les développements méthodologiques en bioinformatique (modélisation des interactions hôtes-pathogènes, calcul haute performance, bases de données et interfaces utilisateur).
- L'analyse experte des données massives en biologie (séquençage massif, interactomique, métabolomique, génomique comparative, phylogénie, métagénomique).
- L'accompagnement de projets et la mise à disposition de services web.
- La formation aux méthodes d'analyse de données (RNA-seq, Galaxy, R).

L'accès à la plateforme se fait dans le cadre de collaborations scientifiques et dans l'optique de réponses à des appels d'offre. Dans le cadre de ses « ateliers du jeudi », le PRABI-AMSB propose un conseil expert en bioinformatique. La plateforme est également ouverte aux prestations de services pour les entreprises privées. Les demandes d'accès aux ressources du PRABI-AMSB se font directement par mail à l'adresse `contact@prabi.fr`.

Mots clefs : Plateforme de service, formation, IFB

*. Intervenant
†. Corresponding author : `guy.perriere@univ-lyon1.fr`

Tackling the issues of reproducibility in NGS analyses with snakemake workflows deployed on virtual environments

Claire Rioualen [*1], Jocelyn Brayet[2], Lucie Khamvongsa[1], Jacques Van Helden [†1]

Poster 113

[1] Technologies avancées pour le génôme et la clinique (TAGC) – Inserm : U1090, Université de la Méditerranée - Aix-Marseille II – Parc scientifique de Luminy, 163 avenue de Luminy, F-13 288 Marseille Cedex 9, France
[2] Institut Curie, PSL Research University, Mines Paris Tech, Bioinformatics and Computational Systems Biology of Cancer, INSERM U900 – 26 rue d'Ulm, F-75 248 Paris Cedex 05, France

Introduction

It is widely known that next-generation sequencing (NGS) has become a mainstream technology in genomics, and it has gotten increasingly cheaper and faster to obtain genomic data. However, the development of reliable tools for the analysis of the huge amount of data generated is still lagging behind. Indeed, the full analysis of a dataset can require a high number of operations, and for each of those, there is no obvious way to choose the most appropriate tool and parameters.

High-throughput biology also comes with an increased concern about reproducibility of all the processing from the raw data to the published results. This challenge is however far from being met in current publications: journals impose to deposit the raw reads in some public database (ENA, SRA) but the readers have no formal description of the path from raw data to the final results.

Snakemake

We have decided to use the Snakemake workflow engine (Köster & Rahmann, 2012)[1] in order to develop comprehensive analysis workflows allowing the parallelization of tasks, and the automatic chaining of analysis steps. Just like GNU Make, it relies on the specification of "targets", corresponding to the list of desired output files, and "rules", defining a generic way to obtain a given output from one or several input files.

A typical Snakefile enables to define all the steps of a workflow, and includes a set of targets and all the rules to produce them. For example, a minimal ChIP-seq workflow could define as targets a bed-formatted peak file, and provide rules for read mapping and peak-calling.

We developed a public library of re-usable rules [2] which can be combined into different workflows, and parametrized in order to better handle the data-dependent specificities, by using human-editable configuration files in the YAML format.

To illustrate the use of these building bricks, we distribute two ready-to-use workflows handling respectively ChIP-seq and RNA-seq data. These workflows can be executed as simply as just typing one command line, and produce files and reports such as:

— mapped reads (SAM files, BAM files…),
— ChIP-seq peaks (BED files),

*. Intervenant
†. Corresponding author : Jacques.van-Helden@univ-amu.fr

— FastQC reports (HTML files),
— motif search reports (HTML reports from the RSAT suite),
— IGV session files (XML files containing coverage profiles, ChIP-seq peaks, genome annotations…).

This method was already applied to a number of study cases, including a variety of model organisms (Bacteria, yeast, drosophila, plants, vertebrates…), and diverse data types (ChIP-seq, RNA-seq, nascent strands of replication origins).

Virtualization

It can be cumbersome to install all the tools that are required for an NGS analysis, especially when these tools come in different versions, and with a number of dependencies. Moreover, some of these tools are designed to work with a certain operating system, some others might corrupt one's own computing set up, or even produce different results depending on the version.

There's a crucial need for tools with a perfect reproducibility, and there's a need for the simplification of the use of these tools.

For all of these reasons, in order to facilitate the distribution of our library of rules and workflows, we have decided to rely on virtualization tools. We have created our own virtual solutions, and we have also written tutorials [3] that can help users setting up their own virtual machines:

— creation/use of a virtual machine under the VirtualBox software [4];
— creation/use of a ready-to-use appliance on the IFB cloud [5];
— creation/use of a Docker image available in the Docker Hub repository [6].

Conclusion

Developing Snakemake workflows and virtual machines allowed us to solve a number of issues that can be restricting, when it comes to analyzing NGS data:

— modularity: one can pick up their own rules in order to build a custom pipeline;
— recycling: a given rule can be used in several workflows, by several people;
— collaboration: several developers can share their rules instead of "reinventing the wheel";
— portability: pipelines can be run on several servers, personal computers, different OS;
— flexibility: the configuration files allow the easy tuning of parameters, so we can cater workflows to different datasets;
— benchmarking: several tools can be run simultaneously in order to compare the results (mapping, peak-calling…).

These can be powerful tools for bioinformaticians and bioanalysts, but also for biologists with a basic training in using Unix commands on the terminal: when performing data analyses, we can now provide them with a ready-to-use system that allows them to reproduce the results rigorously on any computer, at any time.

This project is funded by France Génomique.

References

[1] Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biol.* 2010 Aug 25;11(8):R86.

[2] Web site: `https://github.com/bgruening/docker-galaxy-stable`

[3] Web site: `https://www.docker.com/`

[4] Boeva, V, Lermine, A, Barette, C, Guillouf, C, and Barillot, E. "Nebula – a web-server for advanced ChIP-seq data analysis." *Bioinformatics*, 28 (19), 2517-2519 (2012).

Mots clefs : Galaxy, ToolShed, Docker, automatic deployment, Nebula

Utilisation de Docker en bioinformatique dans le cloud de IFB

Poster 111

Sandrine Perrin [*] [†1], Bryan Brancotte[1], Jonathan Lorenzo[1], Mohammed Bedri[1], Awa Sepou Ngaïlo[1], Fréderic Sene[1], Olivier Sallou[2], François Morrews[2], Olivier Collin[2], Christophe Blanchet [‡1,3], Jean-François Gibrat [§1]

[1] IFB-CORE – Inserm : US21, CNRS : UMS3601, Université Paris XI - Paris Sud – 1 avenue de la Terrasse, Bâtiment 21, F-91 190 Gif sur Yvette, France

[2] Plateforme GenOuest INRIA/Irisa – INRIA – INRIA/Irisa, Campus de Beaulieu, F-35 042 Rennes Cedex, France

[3] Institut de Biologie et Chimie des protéines (IBCP) – CNRS : FR3302, Université de Lyon – France

L'Institut Français de Bioinformatique (IFB) est une infrastructure nationale de service en bio-informatique qui a pour mission principale de fournir des services de base en bioinformatique à la communauté des sciences du vivant. À cette fin, l'IFB met à disposition une infrastructure informatique, sur laquelle est déployé un Cloud, fournissant des moyens matériels (CPU, stockage), mais également un accès à des collections de données publiques de référence et à un large ensemble d'outils d'analyse de données biologiques. Ce cloud académique propose des machines virtuelles préconfigurées avec des applications prêtes à l'emploi.

L'IFB s'investit dans l'utilisation de nouvelles technologies pour améliorer les outils mis à disposition des utilisateurs. Dans cette optique, la technologie Docker s'est imposée comme une solution adaptée à l'intégration de logiciels dans des machines virtuelles répondant aux besoins en traitement de données biologiques.

Docker se présente comme une solution de virtualisation légère permettant de construire, diffuser et d'exécuter des applications sur Linux. Il repose sur la technologie du conteneur logiciel qui permet d'isoler une application et ses dépendances du système d'exploitation. Ainsi, sur une seule machine, il est possible d'installer et d'exécuter très rapidement un grand nombre d'applications sans installation préalable autre que Docker, cela sans avoir à gérer les problèmes de conflits entre logiciels, dépendances et versions de langages, qui est une problématique fréquente en bioinformatique.

Un des principaux inconvénients de cette solution est la connexion en tant que superutilisateur root sur les serveurs, d'où les recommandations de recourir à des environnements sécurisés tels que les machines virtuelles. Cette contrainte est en effet un frein important à l'utilisation de Docker par les services informatiques. De plus cette technologie est en constante évolution, de nouvelles versions sont publiées régulièrement, et toutes les questions de sécurité ne sont pas encore résolues.

Une autre contrainte à son utilisation est le coût en termes d'espace disque nécessaire pour stocker ces images, plus volumineuses que les binaires. En effet, pour permettre des déploiements rapides, les machines sont calibrées pour être de petites tailles. Diverses solutions émergent, elles passent par la compression des images obtenues.

Actuellement, les sciences du vivant font face à l'accroissement du volume des données biologiques produites et à l'augmentation de la complexité des pipelines de traitement de ces données.

*. Intervenant

†. Corresponding author : sandrine.perrin@france-bioinformatique.fr

‡. Corresponding author : christophe.blanchet@ibcp.fr

§. Corresponding author : Jean-Francois.GIBRAT@france-bioinformatique.fr

Le cloud IFB se présente comme un moyen simple et rapide de disposer de ressources pour utiliser, tester ou valider des applications portées sous Docker, voire à terme les mettre à disposition de la communauté dans de nouvelles machines virtuelles. Il fournit une infrastructure adaptée pour manipuler des données volumineuses et déployer des clusters.

Le recours à des dépôts d'images, soit le dépôt général Docker soit des dépôts thématiques, représente un moyen efficace de diffusion des outils dans un format commun et portable dans différents environnements. Effectivement avec une simple ligne de commande, un utilisateur peut installer rapidement une version spécifique d'un logiciel qui est immédiatement disponible.

L'utilisation des applications portées sous Docker par des utilisateurs finaux doit passer par des gestionnaires d'outils et de workflow. En effet, la méthode d'intégration des outils doit être transparent pour lui, il doit uniquement manipuler les outils. Galaxy, à partir de la version 16, offre la possibilité d'exécuter des conteneurs Docker, mais il existe d'autres solutions telles que Nextflow, un autre gestionnaire de workflow.

L'IFB contribue à la création d'images Docker d'applications bioinformatiques. À ce jour, 35 images ont été déposées sur le dépôt BioShaDock. BioShaDock [1] est un dépôt, créé et maintenu par la plateforme GenOuest de l'IFB, dédié à la bioinformatique et la biologie computationnelle. Les images sont indexées grâce à l'ontologie EDAM [2], promue par ELIXIR [3], qui est également employée par biotools [4], le catalogue des outils bioinformatiques.

Des machines virtuelles dédiées à l'utilisation de Docker sont disponibles sur le cloud IFB pour les développeurs, notamment une instance de Galaxy version 16.01 prête à l'emploi et configurée pour exécuter des conteneurs Docker.

Parallèlement le portage des outils sous Docker permet d'envisager de nouvelles solutions simples et efficaces pour créer des machines virtuelles préconfigurées grâce à des outils de configuration des systèmes tels que Puppet.

L'IFB fournit avec son cloud un environnement (sous la forme d'une machine virtuelle) où un utilisateur peut simplement et rapidement exécuter un pipeline de traitements sur ses données. De plus, l'IFB favorise l'utilisation de la technologie des conteneurs grâce aux moyens de développement et d'analyse mis à disposition sur son cloud. En outre, son cursus de formation à l'utilisation du cloud IFB inclut la présentation de la technologie Docker pour exploiter les images existantes et pour créer de nouvelles images afin d'enrichir le registre BioShaDock avec les outils communément utilisés par la communauté.

La reproductibilité et la traçabilité des analyses sont des questions récurrentes en biologie. Pour y répondre, la technologie des conteneurs offre des perspectives intéressantes.

Références :

[1] Moreews, François et al. "BioShaDock: A Community Driven Bioinformatics Shared Docker-Based Tools Registry." *F1000Research* 4 (2015):1443. *PMC*. Web. 29 Apr. 2016. `http://docker-ui.genouest.org/app/#/`

[2] Ison, Jon et al. "EDAM: An Ontology of Bioinformatics Operations, Types of Data and Identifiers, Topics and Formats." *Bioinformatics* 29.10 (2013): 1325–1332. *PMC*. Web. 29 Apr. 2016. `http://edamontology.org/page`

[3] ELIXIR `https://www.elixir-europe.org/`

[4] Ison, Jon et al. "Tools and Data Services Registry: A Community Effort to Document Bioinformatics Resources." *Nucleic Acids Research* 44.Database issue (2016): D38–D47. *PMC*. Web. 29 Apr. 2016. `https://bio.tools/`

Mots clefs : Cloud computing, intégration d'outils, docker, virtualisation

neXtProt : a knowledgebase on human proteins

Poster 112

Pascale Gaudet [*1], Pierre-André Michel[1], Monique Zahn[1], Isabelle Cusin[1], Paula Duek[1], Alain Gateau[1], Anne Gleizes[1], Daniel Teixeira[1], Frédéric Nikitin[1], Valentine Rech De Laval [†1,2], Mathieu Schaeffer[1,2], Lydie Lane[1,2], Amos Bairoch[1,2]

[1] CALIPHO group, Swiss Institute of Bioinformatics (SIB) – Suisse
[2] Department of Human Protein Sciences, Université de Genève (UNIGE) – Suisse

neXtProt (`http://www.nextprot.org/`) is a comprehensive human protein-centric knowledgebase. Its web-based platform offers to its users a seamless integration of and navigation through protein-related data. Focused solely on human proteins, neXtProt aims to provide a state of the art resource for the representation of human biology by capturing a wide range of data, precise annotations, fully traceable data provenance and a web interface, which enables researchers to find and view information in a comprehensive manner.

neXtProt is both a new and an old resource: new, because we try to create an innovative integrative resource around human proteins and old, because we are building it on top of the high-quality solid work that has been the hallmark of UniProtKB/Swiss-Prot. The extensive efforts made by Swiss-Prot to functionally annotate human proteins and curate their sequences and many other features is the foundation on which neXtProt relies. However, this is not enough to populate a resource that needs to address the complexity of the universe of human proteins. In order to remedy this, neXtProt continuously adds new content to the database. The major data sources include Bgee, Human Protein Atlas (HPA), Peptide Atlas, SRMAtlas, UniProtKB, GOA, COSMIC, and IntAct.

neXtProt allows querying data to access entries, publications and terminology. There are several ways to query neXtProt content.

The simple search system (`http://search.nextprot.org/`) is a Google-like full text search which allows you to search the data in neXtProt using Solr technology.

The so called advanced search system uses the SPARQL Protocol and RDF Query Language (SPARQL) to access all the neXtProt protein entry data. This search was designed to support the retrieval of proteins based on highly precise criteria taking into account the richness of the annotations and evidences. Tools were developed and integrated in our user interface to help users to learn how and work out SPARQL queries (`http://snorql.nextprot.org/`, `http://sparql-playground.nextprot.org/`).

It is also possible querying data via a REST API (`http://api.nextprot.org/`). This decouples the database from all our services; in particular, the search and the export services. The API services make it possible for our users to develop their own applications / tools that will benefit from neXtProt's clean and high-quality data.

neXtProt data are downloadable in two ways. After querying data with simple or advanced search, you can download one or all entries of the displayed result of the search. Entries can be exported in their entirety, or the users can customize which content they wish to export, for instance the sequence or a subset of annotation types like PTMs or expression profiles. Several formats are available: FASTA for sequences, and XLS, JSON and XML for entries. The second possibility is to download current and past releases of neXtProt, controlled vocabularies and ontologies developed

*. Corresponding author : `Pascale.Gaudet@isb-sib.ch`
†. Intervenant

specifically for neXtProt, as well as other files from the FTP site (`ftp://ftp.nextprot.org/`) in different formats (FASTA, PEFF, XML, RDF, etc.).

It is possible to explore data via our web interface. We are developing viewers for the display of our data as modular generic components to make them reusable by the life sciences community independently of neXtProt. One of them is a viewer displaying detailed information about the mass spectrometry-derived peptides observed in a protein sequence. Another one displays the different features of a specific protein. These new tools are pretty generic and could be applied to many other contexts. The source code of our components is available on Github (`http://github.com/calipho-sib`).

The data in neXtProt can be analyzed using tools: BLAST and a list manager. We provides access to a simple BLAST implementation to find other entries in neXtProt containing the same or a similar user sequence. It is possible to create a collection of protein entries. neXtProt allows you to create and manage private lists.

The big challenge is to be able to understand how a given set of proteins share or differ in their various features. Indeed data is useful, but we want neXtProt to be much more than a well-organized comprehensive data repository. In addition to enhanced search capabilities, we offer tools that help to make sense of the contents.

Mots clefs : knowledgebase, human, protein

Prolégomènes à la classification des protéines basée sur une représentation vectorielle des acides aminés

Poster 114

Ariane Bassignani [*] [†1], Jean-Michel Batto[1], Nicolas Pons[1], Stephan Fischer[2], Karel Zimmermann[2], Dusko Ehrlich[1], Magali Berland [‡1]

[1] INRA US MetaGenoPolis 1367 – Institut national de la recherche agronomique (INRA) : US1367 – INRA Domaine de Vilvert, Unité MGP - Bâtiment 325, F-78 352 Jouy-en-Josas Cedex, France
[2] INRA UR MaIAGE 1404 – Institut national de la recherche agronomique (INRA) : UR1404 – France

Résumé en français :

La classification fonctionnelle des protéines à l'échelle d'un métagénome est un élément clé de la compréhension des écosystèmes microbiens et de leur modélisation. Actuellement, l'annotation fonctionnelle systématique repose d'une part sur un nombre restreint de fonctions qui ont été déterminées expérimentalement, mais aussi sur un ensemble d'annotations prédites et non curées manuellement. Les limitations des approches actuelles sont liées à la fiabilité du système de transfert de l'information entre les séquences homologues et à la propagation et à l'accumulation d'erreurs dans les bases de données. Dans le microbiote intestinal humain, environ 80 % des espèces bactériennes ne sont pas cultivables [1, 2] et une vaste majorité reste donc inconnue [3, 4]. L'annotation des nouvelles protéines identifiées dans les catalogues de gènes du microbiote intestinal humain [5] est une tâche complexe liée au manque d'annotations disponibles ainsi qu'à l'absence de génomes de référence.

Nous présentons ici le travail sur une nouvelle méthode de classification des protéines qui utilise pour la première fois la représentation vectorielle des acides aminés dans l'espace euclidien proposée par Zimmermann et Gibrat [6]. Dans ce cadre, chaque acide aminé est remplacé par un vecteur issu de la décomposition en valeurs singulières d'une matrice BLOSUM ou PAM qui prend en compte sa fréquence de substitution avec les autres acides aminés. Cette numérisation de l'enchainement des acides aminés permet de représenter une protéine sous forme de trajectoire dans un espace euclidien et donc de calculer une similarité entre protéines. Elle permet également de déterminer la séquence consensus d'un alignement multiple en calculant l'acide aminé le plus proche de la moyenne des mutations possibles à une position. Notre hypothèse de travail est que les protéines partageant une fonction commune ont une trajectoire type similaire qui ne serait pas nécessairement directement dépendante de l'homologie entre les séquences. Nous proposons de comparer différentes méthodes pour définir une trajectoire type, appliqué à plusieurs ensembles de protéines simulées avec des pourcentages d'identité fixés à l'avance. L'enveloppe de ces trajectoires nous donne un faisceau dans lequel la trajectoire doit s'inscrire. La similarité entre la trajectoire d'une nouvelle protéine et le faisceau est calculée, ainsi que la trajectoire d'une protéine contenant les mêmes acides aminés, mais dans un ordre aléatoire. Nous montrerons que la pertinence du faisceau pour effectuer la discrimination entre ces deux protéines dépend du pourcentage d'identité entre les séquences de protéines utilisées pour construire le faisceau.

Ce travail propose une méthode originale pour tenter de classifier les protéines en s'affranchissant des limitations des méthodes d'annotations classiques. Un des perspectives de ce travail sera

*. Intervenant
†. Corresponding author : ariane.bassignani@jouy.inra.fr
‡. Corresponding author : magali.berland@jouy.inra.fr

de tester les avantages et les limitations de cette méthode par rapport aux méthodes existantes. Un package R est en cours d'implémentation et sera mis à la disposition de la communauté scientifique.

Summary in english:

Functional classification of proteins at a metagenomic level is a key tool for the microbial ecosystems understanding and modeling. Currently, systemic functional annotation is based, on one hand, on a restricted number of functions which was determined experimentally, and on the other hand, on numerous predicted annotations and not manually cured. Limitations of existing approaches are linked to the reliability of the information transfer between homologous sequences and to the propagation and accumulation of mistakes in databases. In the human gut microbiota, about 80 % of bacterial species are not cultivable [1, 2] and a vast majority is still unknown [3, 4]. Annotation of newly identified proteins from human gut microbiota gene catalogs [5] is a complex task, linked to the lack of available annotations and to the absence of any reference catalog.

We are presenting here the work about a new method of proteins classification, which uses for the first time the vector representation of amino acids in Euclidian space proposed by Zimmermann and Gibrat [6]. In this representation, each amino acid is replaced by a vector coming from singular value decomposition of BLOSUM or PAM matrices, which takes into account the substitution frequency with other amino acids. This numeric view of amino acids sequence allows to represent a protein like a trajectory in a Euclidian space, and thus to calculate a similarity between proteins. It also allows determining the consensus sequence of a multiple alignment by calculating the amino acid closest to the mean of possible mutations for a given position. Our hypothesis is that proteins sharing a common function have a similar trajectory which would not be necessary directly dependent of sequences homology. We propose to compare various methods to define a trajectory, applied to many simulated proteins sets with fixed identity percentages. The borders of these trajectories give us a beam in which the trajectory must fit. The similarity between the trajectory of a new sequence and the beam is calculated, as well as the trajectory of a protein containing same amino acids, but in a random order. We will show that the relevance of the beam to discriminate these two proteins depends on the identity percentage between amino acids sequences used to construct the beam.

This work proposes a novel framework to classify proteins that tries to overcome the limits of the classical annotations methods. One of the perspectives will be to test the advantages and limitations of this method compared with existing methods. An R package is being implemented and will be made available to the scientific community.

References

[1] Ericsson AC,Franklin CL (2015). Manipulating the gut microbiota: Methods and challenges. *ILAR journal*, 56(2):205-17.

[2] Stewart EJ (2012). Growing uncultivable bacteria. *Journal of bacteriology*, 194(16):4151-60.

[3] Eckburg PB et al. (2005). Diversity of the human intestinal microbial flora. *Science*, 308(5728):1635-8.

[4] Scanlan PD and Marchesi JR (2008). Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. *The ISME journal*, 2(12):1183-93.

[5] Qin et al. (2010). A human gut microbial gene catalog established by metagenomic sequencing. *Nature*, 464(7285):59-65.

[6] Zimmermann K, Gibrat JF (2010). Amino acid "little Big Bang": representing amino acid substitution matrices as dot products of Euclidian vectors. *BMC Bioinformatics*, 11:4

Mots clefs : Functional classification, amino acids vectors, gut microbiota, prediction of function, proteins

Global approach for assessing the link between DNA features and gene expression

Chloé Bessière [*] [†1], May Taha [*] [‡1,2], Charles Lecellier[1], Laurent Bréhélin[3], Sophie Lebre[2], Jimmy Vandel[4], Florent Petitprez[1]

Poster 115

[1] Institut de génétique moléculaire de Montpellier (IGMM) – CNRS : UMR5535 – 1919 route de Mende, F-34 293 Montpellier Cedex 5, France
[2] Institut de Mathématiques et de Modélisation de Montpellier (IMAG) – CNRS : UMR5149 – Case Courrier 051, Place Eugène Bataillon, F-34 095 Montpellier Cedex 5, France
[3] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – CNRS : UMR5506, Université de Montpellier – Bâtiment 5, 860 rue de St Priest, F-34 095 Montpellier Cedex 5, France
[4] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) – CNRS : UMR5506 – CC 477, 161 rue Ada, F-34 095 Montpellier Cedex 5, France

Gene expression is tightly controlled to ensure a wide variety of cell types and functions. These controls take place at the DNA and/or RNA levels and are associated with different regulatory regions: close and distal promoters, enhancers, untranslated regions, etc. While these regulatory regions remain the same for the different cells and cell types, the diversity of their responses is ensured by the combination of hundreds of features that can come into play: specific nucleotidic composition, presence/absence of binding sites of specific TFs or RNA-binding proteins… One today challenge is then to decipher the different DNA features that are responsible for a specific gene expression profile in a specific cell type.

We address this issue using a global regression approach. In this framework, the expression of the different genes and the DNA/RNA features of their regulatory regions are the predicted and predictive variables, respectively. The advantage of this approach is that it allows to unveil the most important regulatory mechanisms used by the studied cells into a single model. The combinatorial nature of the mechanisms and the relative contribution of each feature can thus be easily assessed.

Our computational approach is based on two steps. First, a linear model with LASSO penalty is trained to build a gene expression predictor on the basis of hundreds of sequence features. We then evaluated the performances of our model by computing the correlation between the predicted and the observed expression. Depending on the data, we show that the inferred model is not equally efficient for all genes but only fits certain classes of genes with specific genomic features. In a second step, we thus run a classification tree on the results of the linear model to identify these genes (and their genomic features) that are well or badly fitted by the model.

This approach was run on numerous gene expression data of the TCGA database and allowed us to highlight several important features of gene expression control. First, basic information like nucleotide and dinucleotide frequency of promoter regions have very high predictivepower. By distinguishing distal and proximal promoters, we show that both sequences carry different and complementary information. On the contrary, adding predictions about presence/absence of TF binding site motifs derived from PWMs do not lead to high improvement of the model. This might be due to limited accuracy of some TF binding predictions and indicates that more efforts are still required to improve TF binding modeling.

*. Intervenant
†. Corresponding author : chloe.bessiere@gmail.com
‡. Corresponding author : may-taha@hotmail.com

Mots clefs : regulatory network inference, cis, regulatory motifs, regulatory sequence, cancer

Une nouvelle approche de comparaison de séquences ADN à l'aide d'une fonction de hachage perceptuel

Poster 116

Jocelyn De Goër De Herve [*1,2,3], Hayfa Azibi[1,2,4], Myoung-Ah Kang[1,2], Xavier Bailly[3], Engelbert Mephu Nguifo[1,2]

[1] Université Blaise Pascal - Clermont-Ferrand 2 (UBP) – Université Blaise Pascal - Clermont-Ferrand II, Clermont Université – 34, avenue Carnot, BP 185, F-63 006 Clermont-Ferrand Cedex, France
[2] Laboratoire d'Informatique, de Modélisation et d'optimisation des Systèmes (LIMOS) – Institut Français de Mécanique Avancée, Université Blaise Pascal - Clermont-Ferrand II, Université d'Auvergne - Clermont-Ferrand I, CNRS : UMR6158 – Bâtiment ISIMA, Campus des Cézeaux, BP 10025, F-63 173 Aubière Cedex, France
[3] Unité de recherche d'Épidémiologie Animale (UEA) – Institut national de la recherche agronomique (INRA) : UR0346 – France
[4] University FSJEG (FSJEG) – Jendouba, Tunisie

DNA sequence similarity searching, to identify homologous sequences, is a fundamental task in genomics studies. In this context, we present a novel DNA sequence comparison method, based on concepts, from digital image processing and more particularly from perceptual hashing. Perceptual Hashing Function for Sequence (*PHS*) proposed here, has been adapted to the characteristics of DNA sequences, as well as the comparison method. Hashing process uses Discrete Cosine Transform Sign Only (DCT-SO), for sequence indexing, and we exploit DCT-SO correlation abilities, to determine the similarity between two sequences. In order to use *PHS* function, sequences have to be converted as grayscale images and hash key are calculated from significant frequencies characteristics of images. Hash keys are generated in binary format and are eight time smaller than original sequences. Validation process by theoretical simulations shows the stability of *PHS* function, and evaluation with real dataset, demonstrates that it is possible to identify the reference genome of a query sequence.

Mots clefs : Comparaison de séquences ADN, identification de séquences, hachage perceptuel, méthode de corrélation, transformée en cosinus discrète à coefficients signés

*. Intervenant

Wengan : a versatile scaffolder

Poster 117

Alex Di Genova [*1,2,3], Hongphong Pham[3], Arnaud Mary[3,4], Laurent Bulteau[3], Gonzalo Ruz[1], Alejandro Maass[2,5,6], Marie-France Sagot [†3,4]

[1] Universidad Adolfo Ibañez (UAI) – Chili
[2] Centre de Modélisation Mathématique / Centro de Modelamiento Matemático (CMM) – Chili
[3] Université Claude Bernard Lyon 1 (UCBL) – Université Lyon 1 – 43 boulevard du 11 novembre 1918, 69 622 Villeurbanne Cedex, France
[4] Laboratoire de Biométrie et Biologie Évolutive (LBBE) – CNRS : UMR5558, Université Claude Bernard - Lyon I (UCBL), INRIA – 43 boulevard du 11 Novembre 1918, F-69 622 Villeurbanne Cedex, France
[5] Departamento de Ingeniería Matemática [Santiago] (DIM) – Blanco Encalada 2120 (5to piso), Santiago, Chili
[6] Universidad de Chile [Santiago] – v. Libertador Bernardo O'Higgins 1058, Santiago, Chili

Scaffolding, which consists in the process of ordering and orienting a set of contigs along a chromosome sequence, is an old problem in bioinformatics. The information commonly used to produce the scaffolds comes from libraries of multiple mate-pairs with different insert-size. It could also come from reference genomes (not necessarily complete) or long reads (PacBio, Oxford nanopore), however these last two sources of information have been poorly exploited by the currently available scaffolding methods.

Our strategy consists in building a general undirected weighted graph (called a scaffold graph) using information from reference genomes, mate-pair libraries or long reads, and in solving an optimization problem to produce the scaffolds. The vertices of the graph represent the contig sequences and the edges define the relation in terms of distance, orientation and strength (weight) for a pair of contigs. Independently of the information used as input, we can always define these three variables for each edge in the graph. For instance, if the information comes from mate-pair libraries, the edge orientation is defined by how the read pairs align to the contig, the distance corresponds to the average insert size observed in the library, and the weight is the number of different pairs that link the two contigs. In a similar way, if we use long reads or reference genomes as sources, the same variables are defined. Given a scaffold graph $G=(V, E)$, we compute a maximal weighted path cover of the graph, which represents the scaffold solution and consists in a set of vertex disjoint paths $P1, \ldots, Ps$ of maximal weight that cover all the vertices of G. To solve this problem, we developed two algorithms, one greedy and other using a matching technique. The main difference between the two algorithms consists in how the orientation of the contig is solved. The greedy approach solves first the order and in a separate step the orientation, while instead the matching solves both simultaneously. Both algorithms are implemented in C++.

The performance of Wengan is being evaluated against other available scaffolders using the datasets and evaluation tools provided in the paper: "A comprehensive evaluation of assembly scaffolding tools" [1].

References

[1] Hunt, Martin, et al. "A comprehensive evaluation of assembly scaffolding tools." *Genome biology* 15.3(2014):1-15.

*. Intervenant
†. Corresponding author : marie-France.sagot@inria.fr

Mots clefs : Wegan, Scaffolder, graph algorithms, optimization, genome

Gmove : eukaryotic gene predictions using various evidences

Poster 118

Marion Dubarry [*1], Benjamin Noël[1], Tsinda Rukwavu[1], Sarah Farhat[1], Corinne Da Silva[1], Manuel Lebeurrier[1], Jean-Marc Aury[1]

[1] Genoscope - Centre national de séquençage – CEA, Genoscope – 2 rue Gaston Crémieux CP5706, F-91 057 Évry Cedex, France

The NGS make the sequencing faster and cheaper, so affordable for more and more laboratories. Consequently, the number of sequenced genome explodes, and it becomes feasible to sequence more complex genomes (like large genomes, highly repeated genome) as well as non-model organisms. Pipelines of gene predictions have to get used to these technological improvements to keep going to improve the quality of their predictions, make easier the calibration step and handle the large amount of data available. We present the Eukaryotic genome annotation pipeline used at Genoscope (the French national sequencing center), particularly the tool we use to predict coding genes, named Gmove (Gene Modelling using Various Evidence). It can use several source of data, like RNAseq, conserved proteic alignment and ab initio gene predictions. Gmove combines these data and finds a consensus without any prerequisite calibration. A graph is built where a node represents an exon and a vertex represents an intron, extracting paths are potential genes models. In these models, Gmove searches for an open reading frame based on existing protein alignments. We can select one or several isoform of a given gene. On the de novo annotation context, we already run this tool on a variety of current projects such as plant, fungus, insect and dinoflagellate (genomes not published yet). Moreover, Gmove could also improve existing gene prediction by combining the former annotation with new data.

Mots clefs : genome, annotation, rna, seq

*. Intervenant

Statistical methods for gene-based gene-gene interaction detection in R

Mathieu Emily *[1,2], Nicolas Sounac[1], Florian Kroell[1], Magalie Houée-Bigot †[1]

Poster 119

[1] Agrocampus Ouest – Ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt, Institut supérieur des sciences agronomiques, agroalimentaires, horticoles et du paysage – Centre de Rennes, 65, rue de St Brieuc, CS 84215, F-35 042 Rennes Cedex, France
[2] Institut de Recherche Mathématique de Rennes (IRMAR) – CNRS : UMR6625 – France

Case-control genome-wide association studies (GWAS) aim at investigating the genetic components of binary traits like major diseases. Single-locus approaches, whereby a large number of Single Nucleotide Polymorphisms (SNPs) are tested independently for association, have first been developed to analyse GWAS. Although such single-locus approaches have successfully identified regions of disease susceptibility, findings were of modest effect and a large proportion of the genetic heritability is still not covered for common complex diseases (Manolio, 2009). Epistasis is often cited as one of the main biological mechanism carrying the "missing heritability" in GWAS (Phillips, 2008). Since human complex diseases are generally caused by the combined effect of multiple genes, the detection of genetic interactions is thus essential to improve our knowledge of the etiology of complex diseases (Cordell, 2009).

Genetic interactions have first been investigated at the SNP level with the development of many statistical methods to detect SNP-SNP interactions. In contrast to SNP-level approach, gene-level testing can help characterizing functional, compositional and statistical interactions (Phillips, 2008). Such tests allow for all the SNPs within the region of a gene to be jointly modeled as a set and can take into account the Linkage Disequilibrium (LD) structure within a gene. Thus, by aggregating signals across variants in a gene, statistical power is likely to be increased in situations when multiple causal interactions influence the phenotype of interest. Furthermore, if the interacting variants are only tagged, rather than directly observed, gene-based tests can aggregate signals from different tag SNPs in partial LD. Finally, the use of the gene as the statistical unit can greatly facilitate the biological interpretation of findings. For these reasons, gene-based gene-gene interactions methods are considered as a promising and an attractive alternative to single-locus and SNP-SNP methods.

Although several methods have been proposed to test for interaction between two genes at the gene level, no software package is available to compute all these methods from large-scale data sets. To facilitate the search for gene-based gene-gene interaction in GWAS, we propose in this work a novel R package GeneGeneInteR that implements a collection of 10 methods: 6 methods that aim at modeling the joint distribution of SNPs within and between two genes (PCA, CLD, CCA, KCCA, PLSPM, GBIGM) and 4 methods that aggregate interaction tests performed at the SNP level (minP, GATES, *t*TS and tProd).

Principal component analysis (PCA) has first been used to test the association between synthetic variables (i.e. principal components) from each gene (Li *et al.* 2009). The PCA based method has been implemented in the PCA.Std and PCA.GenFreq functions. In another approach, Peng *et al.* proposed a U-statistic, called CCU, to measure the difference of correlation between two genes in cases and controls (Peng *et al.* 2010). In CCU, correlations in cases and controls are based on canonical correlation analysis in order to detect gene-gene co-association. CCU

*. Corresponding author : mathieu.emily@agrocampus-ouest.fr
†. Intervenant

test has been implemented in the CCA.test function of our package GeneGeneInteR. CCU has further been extended to KCCU, where correlation is estimated by kernel canonical correlation (Yuan *et al.* 2012, Larson *et al.* 2013). In our packge, KCCU can used with KCCA.test function. Partial Least Squares Path Modeling (PLSPM) has also been proposed as an alternative measure of correlation between two genes (Zhang, 2013). The implementation of the plspm based method is available through the use of the plspm.test function of our package.

Rather than focusing on a single measure of correlation between genes, Rajapakse *et al.* proposed a test to compare the whole covariance structure between two genes in cases and controls (Rajapakse *et al.* 2012). Such a method, based on the composite linkage disequilibrium, has been implemented in the CLD.test function. Our GeneGeneInteR also includes the GBIGM.test function that computes a non-parametric statistic based on information theory recently introduced as an attractive option to detect non-linear relationship between two genes (Li *et al.* 2015).

Rather than considering multiple markers in both gene as part of a joint model, an alternative strategy has recently been developed in order to aggregate p-values obtained at the SNP level into a test at the gene level (Emily, 2016). Several procedures can be used to combine p-values such as the minimum p-value (minP), the Gene Association Test using Extended Simes (GATES) procedure, the tail truncated strength (tTS) and the truncated product (tProd) (Ma *et al.* 2013). (1) MinP test is based on the minimum SNP-SNP interaction p-value. MinP p-value is calculated by integrating the observed test statistics multinormal distribution. (2) GATES test, as with minP, is based on the strongest signal, where the strongest signal is not defined as the minimum p-value but as the minimum value Simes procedure using the effective number of independant tests. Multiple methods exist to estimate the number of effective tests. (3) tTS test does not consider only the strongest signal but all signals that are inferior to a threshold. For these p-values, a weighted sum is computed and represents the test statistic. (4) This is the same as tTS but the test statistic is a product of the p-values, not a sum. For (3) and (4), the empirical p-value is calculated using a multivariate normal distribution with the estimated covariance matrix. These 4 methods have been implemented in the minP.test, GATES.test, tTS.test and tProd.test functions respectively.

In order to perform a complete statistical analysis of gene-based gene-gene interaction at the genome level, our GeneGeneInteR package provides a series of functionalities that allow to (1) download data in various standardized format (PED, PLINK, VCF), (2) impute missing genotypes, (3) perform a gene-based gene-gene interaction analysis and (4) visualize the results. Importation and imputation of data can be performed with the import.file and imputeSnpMatrix functions that are based on the snpStats package available on Bioconductor. Step (3) can be performed by using one of the gene-gene method, implemented in the functions PCA.test, CLD.test, CCA.test, KCCA.test, PLSPM.test, GBIBM.test, minP.test, GATES.test, tTS.test or tProd.test, for all pairs of gene in the dataset. Finally, the visualization of the results can be done with a matrix-like representation (function GGI.plot in the GeneGeneInteR package) or the design of gene-gene interaction network (function draw.network in the GeneGeneInteR package).

To test for the performances of our package, we conducted an extensive simulation study with respect to various disease models and different gene correlation structures. Although our results show a large heterogeneity among the different methods, the aggregation of p-values is the most powerfull method in many situations. Furthermore, the analysis of the true phenotype in the dataset GSE39428 gives also new insight in the understanding of the etiology of Rheumatoid Arthritis, thus paving the way for further investigation of gene-gene interaction a the gene level.

The GeneGeneInteR package is available on GitHub: `https://github.com/MathieuEmily/GeneGeneInteR`.

References

[1] Manolio, T. A. et al. (2009) Finding the missing heritability of complex diseases, *Nature*,

461:747–753.

[2] Phillips, P. (2008): Epistasis, the essential role of gene interactions in the structure and evolution of genetic systems, *Nature Review Genetics*, 9:855–867.

[3] Cordell, H. J. (2009): Detecting gene-gene interactions that underlie human diseases, *Nature Review Genetics*, 10:392–404.

[4] Li, J. et al. (2009): Identification of gene-gene interaction using principal components, *BMC Proceedings*, 3:S78.

[5] Peng, Q. et al. (2010): A gene-based method for detecting gene-gene co-association in a case-control association study, *European Journal of Human Genetics*, 18:582–587.

[6] Yuan, Z. et al. (2012): Detection for gene-gene co-association via kernel canonical correlation analysis, *BMC Genetics*, 13:83.

[7] Larson, N. B. et al. (2013): A kernel regression approach to gene-gene interaction detection for case-control studies, *Genetic Epidemiology*, 37:695–703.

[8] Zhang, X. et al. (2013): A plspm-based test statistic for detecting gene-gene coassociation in genome-wide association study with case-control design, *PLoS ONE*, 8:e62129.

[9] Rajapakse, I. et al. (2012): Multivariate detection of gene-gene interactions, *Genetic Epidemiology*, 36:622–630.

[10] Li, J. et al. (2015): A gene-based information gain method for detecting gene-gene interactions in case-control studies, *European Journal of Human Genetics*, 23:1566-1572.

[11] Emily, M. (2016): AGGrEGATOr: A Gene-based GEne-Gene interActTiOn test for case-control association studies, *Statistical Application in Genetics and Molecular Biology*, 15:151-171.

[12] Ma, L. et al. (2013): Gene-based testing of interactions in association studies of quantitative traits, *PLoS Genetics*, 9:e1003321.

Mots clefs : Genome Wide Association Studies, Gene based methods, R package

Évaluation de données HLA à partir d'informations de SNPs

Marc Jeanmougin [*] [†1], Cédric Coulonges[1], Josselin Noirel[1]

Poster 120

[1] Conservatoire National des Arts et Métiers (CNAM Paris) – EA4627 – 292 rue Saint-Martin, F-75 141 Paris Cedex 03, France

Introduction

Chez l'Homme, les gènes du système HLA (*Human Leukocyte Antigen*) codent pour des protéines de surface des cellules responsables de la régulation du système immunitaire. Ces gènes sont situés sur le bras court du chromosome 6, dans une région appelée le complexe majeur d'histocompatibilité (CMH). Ces gènes de reconnaissance du soi ont une importance majeure dans tous les domaines impliquant le système immunitaire : les maladies auto-immunes (comme la sclérose en plaque), les maladies infectieuses (comme le SIDA), l'inflammation et les transplantations d'organes (notamment afin d'assurer la compatibilité donneur/receveur).

La région CMH possède la particularité de présenter une grande variabilité génétique (elle est très polymorphe). Aujourd'hui, près de 4000 allèles sont répertoriés pour chacun des gènes de classe I (HLA -A, HLA-B, HLA-C) et jusqu'à 2000 allèles sont connus pour les gènes de classe II (HLA-DP, HLA-DQ, HLA-DR).

Durant les dernières décennies, le typage allélique de la région HLA a constitué un défi technologique majeur. A d'abord été utilisée une approche sérologique de microlymphocytotoxicité (LCT), supplantée par des techniques de biologie moléculaire basées sur de la PCR : les techniques SSP et SSOP (*Sequence-Specific Primers* et *Sequence-Specific Oligonucleotide Probes*), qualifiées de gold standard, qui ont plus récemment laissé place à des techniques de séquençage massivement parallèle (NGS). En dépit de leur fiabilité et même si leur coût décroît, ces méthodes restent chères (de l'ordre de milliers d'euros pour les NGS en 2015), Elles sont donc peu adaptées aux applications bio-informatiques comme les études génome entier où l'on recherche des associations statistiques entre un phénotype et un génotype dans des cohortes rassemblant usuellement des milliers d'individus.

En raison du fort polymorphisme de la région HLA, les puces de génotypage existant depuis une dizaine d'années et couvrant le génome entier, menant aux très nombreuses études d'associations (« GWAS ») sont, elles, inadaptées au typage direct des gènes de la région HLA.

Pour pallier ce problème et grâce aux déséquilibres de liaison existants entre allèles, il est possible d'utiliser une population de référence pour imputer les données manquantes issues d'une puce de génotypage.

Des outils d'imputation ont donc été développés ces dernières années (Impute2, Beagle, Mach) et pour certains adaptés à l'imputation d'allèle de gènes HLA (SNP2HLA, HIBAG, HLA*IMP).

Cependant ces méthodes ont une importante limitation : la taille, la diversité génétique et la qualité du typage du panel de référence sont cruciales pour la fiabilité du résultat.

Pour limiter l'impact de la qualité du typage de la référence, et effectuer une validation croisée du résultat d'imputation, nous avons développé un outil destiné à évaluer la plausibilité d'une information de typage HLA, à partir du génotype de l'individu et de la définition officielle des haplotypes HLA (IMGT).

*. Intervenant

†. Corresponding author : marc.jeanmougin@cnam.fr

Matériels et Méthodes

Nous avons utilisé trois cohortes :

- Une cohorte de référence de 5225 individus issue du « Type 1 Diabetes Genetics Consortium (T1DGC) » dont les SNP de la région HLA ont été génotypés ainsi que les allèles HLA par des méthodes directes « gold standard à 4 digits ».

- Les 5000 haplotypes de référence du projet « 1000 Genomes » pour l'imputation des SNPs de la région.

- La cohorte « 1958 British Cohort (1958BC) » constituée de 2434 individus partiellement typés HLA et genotypés sur les SNPs du CMH.

Pour imputer le HLA, nous avons représenté les allèles HLA comme des marqueurs binaires de SNP afin de pouvoir les traiter avec les outils de phasage et d'imputation.

Nous avons utilisé les logiciels ShapeIT (phasage) et Impute2 (imputation), adoptés pour le projet EMBL-EBI/IGSR « 1000 Genomes ».

Pour notre outil de vérification, nous avons privilégié une approche simple et générale : nous commençons par imputer les génotypes à tester avec les haplotypes du projet « 1000 Genomes », afin d'avoir le maximum de SNP connus dans les gènes HLA.

Les génotypes attendus par paire d'allèles HLA sont ensuite reconstitués à partir de l'appariement des haplotypes HLA donnés par la référence IMGT.

Nous réalisons ensuite une étude fondée sur les probabilités postérieures d'imputation des SNP des exons des gènes HLA, comparant, pour chaque gène HLA et pour chaque paire d'allèles de ce gène HLA connus, le résultat obtenu de l'imputation à la paire de SNPs prédits par les haplotypes IMGT.

La discordance statistique obtenue est ensuite incorporée au score, pour ce génotype, de la paire de HLA étudiée, et ceci pour chacun des SNP exoniques trouvés.

Par exemple, si un SNP est dans HLA-A*01 :01 une guanine et dans HLA-A*02 :01 une adénine et si l'imputation du SNP prédit 10% AA, 25%GG et 65%AG, nous considérons qu'il y a un désaccord de 0.35 avec la paire de HLA (01 :01 ; 02 :01). On somme ensuite les scores pour les différents SNP pour obtenir le score de la paire de HLA.

Nous cherchons ensuite le génotype HLA le plus proche du génotype observé, et nous mesurons la discordance entre ce type HLA, et le type HLA dont on cherche à évaluer la plausibilité dans notre test. Notre hypothèse de travail est que cette différence est directement reliée à la probabilité que le HLA soit incorrect.

Résultats

En testant notre outil sur la cohorte T1DGC, nous avons été en mesure d'évaluer la pertinence de notre fonction de score utilisant la base de données HLA de l'IMGT.

Par une méthode de permutation aléatoire, nous avons pu établir pour chaque gène HLA, un seuil de score permettant d'éliminer les paires d'allèles d'un type HLA probablement faux.

Nous avons ensuite filtré les individus imputés sur allèles HLA à deux niveaux : d'une part pour éliminer de notre objet d'étude les individus dont le génome présentait trop d'incohérences avec les HLA déclarés, et d'autre part, après l'imputation, pour identifier les individus ayant été mal imputés

Ainsi, même si des individus sont par là exclus de l'étude, l'important reste ici de privilégier la qualité de l'imputation (que nous pouvons comparer à la spécificité : la capacité à prédire un HLA correct lorsque nous le prédisons) plutôt que le nombre brut d'individus (comparable à la

sensibilité : la capacité à admettre lorsque nous ne parvenons pas à identifier un HLA plausible pour un individu).

Par ce protocole, nous avons été en mesure d'améliorer significativement la qualité de l'imputation, en particulier sur les HLA de classe I : pour le HLA-A en précision 4-digit, nous passons de 97.4 % de réussite sur l'imputation à 99.6 % en éliminant 3.5 % de notre cohorte (le pourcentage de réussite est calculé en terme de succès ou d'erreur à correctement prédire totalement la paire d'allèles).
Similairement, pour HLA-B, on passe de 96 % à 98.5 % en éliminant 7 % des individus, et pour le HLA-C de 95.8 % à 99.5 % avec 9 % de pertes (autrement dit, en partant d'une population à 95.8 % correctement imputée, le score permet d'éliminer une sous-population dont plus de la moitié des individus sont mal imputés).

Discussion

Un avantage de cette approche est que celle-ci est complètement indépendante de la population étudiée. En effet, elle se base sur 1000 genomes avec des individus d'origines variées, et sur les définitions du HLA de la base de données de l'IMGT/HLA, exhaustive (par définition). Cette base étant régulièrement mise à jour pour préciser des définitions ou en inclure de nouvelles, la précision de notre outil évoluera en conséquence.

Nous envisageons également plusieurs pistes pour améliorer encore nos résultats.

Mots clefs : génomique, hla, imputation, SNP

Analysis of microRNA sequences identifies conserved families of microRNAs

Christophe Le Priol [*] [†1], Laurent Guyon [‡1], Chloé-Agathe Azencott[2], Xavier Gidrol [§1]

Poster 121

[1] Laboratoire de Biologie à Grande Échelle, Biomics (BGE - UMR_S 1038) – Université Grenoble Alpes, Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) - Grenoble, Inserm – CEA-Grenoble/BIG Laboratoire Biologie à Grande Échelle, 17 rue des Martyrs, F-38 054 Grenoble Cedex 9, France
[2] Centre de Bioinformatique (CBIO) – MINES ParisTech - École nationale supérieure des Mines de Paris – 35 rue Saint-Honoré, F-77 300 Fontainebleau, France

microRNAs are small non-coding RNAs of typically 20-24 nucleotides. Plant microRNAs have been shown to bind to their target mRNA(s) with near perfect match in order to regulate the corresponding gene(s) expression [1]. By contrast, for animal microRNAs, only the seed sequence has to bind near perfectly [2]. The seed sequence is defined as a 6 to 8 nucleotide domain at the 5' end of the mature microRNA that determines target mRNA binding [3].

We downloaded mature microRNAs sequences from the last release of miRBase (release 21, `ftp://mirbase.org/pub/mirbase/CURRENT/mature.fa.gz`). In this release, the database is composed of 35,858 mature microRNAs from 223 animal and plant species. The most represented species are *Homo sapiens* and *Mus musculus* with respectively 2,588 and 1,915 microRNAs, followed by other model organisms and *Mammalia* species. Such a high number of microRNA sequences enables statistical analysis of the microRNAs and comparisons between species. With the present analysis, we propose to extend the current definition of families, *i.e.* microRNAs having the exact same seed sequences, and to identify new conserved families of microRNAs.

To investigate the presence of particular motifs, we performed a sequence analysis by counting the number of occurrences of all the k-mers, *i.e.* subsequences of length k, together with their positions for all microRNAs. We analyzed the content of k-mers of various lengths, from 5 to 8, close to the typical length of 6 nucleotides of the seed sequences.

Under the null hypothesis that the k-mers are randomly positioned in the sequences of the miRNAs, the number of occurrences for a given k-mer is at a given position is modeled by the binomial distribution. Using this distribution, the p probability (p-value) to obtain an observed number of occurrences for a given k-mer (one sided test) at a given position is estimated for each species. In particular, the probability of k-mers being in seed position has been calculated. The obtained p-values emphasize motifs frequently observed at particular positions in the sequence of microRNAs of a species.

As the median microRNA length in miRBase is 22 nucleotides, we calculated the p probability for hexamers (*i.e.* k-mers of length 6) at each position between 1 and 17 for all the species. The distribution of $-\log_{10}(p)$ values obtained at all the positions is heterogeneous with higher values at the 5' end of the sequences of the mature microRNAs in *Metazoa*, confirming that groups of microRNAs share similar sequences at the 5' end of mature microRNAs. On the contrary, no enrichment was observed at any position for vegetal species, also in agreement with the near perfect sequence matching required for microRNAs to bind their target(s) in plants.

*. Intervenant
†. Corresponding author : `christophe.lepriol@cea.fr`
‡. Corresponding author : `laurent.guyon@cea.fr`
§. Corresponding author : `xavier.gidrol@cea.fr`

More precisely, we observed the highest values of $-\log_{10}(p)$ for hexamers at position 2 in *Metazoa*, which corresponds to the definition of the seed sequence as previously described in the literature. However, significantly high values of –log10(p) are also observed at positions 1 and 3, and to a lesser extent at positions 4 and 5, which suggests a more flexible definition of the seed in microRNA sequences. Finally, we also observed an increased number of conserved hexamers at positions 13 and 14, which corresponds to the positions of some of the microRNA recognition sites for AGO1 proteins, a protein used to bind the microRNA to its mRNA target [4].

We then identified seed sequences specific to species or sets of species, thus corresponding to a gain during evolution. To do so, we retrieved taxonomic information of the miRBase species having at least 100 microRNAs using the NCBI taxonomy. We then built a phylogenetic tree of these species thanks to phyloT (`http://phylot.biobyte.de/index.html`). The p probability was finally visualized on the phylogenetic tree, which allowed to detect the emergence of seed sequences in clades. For instance, we bring a clear evidence of the appearance of the hexamer 'AAGUGC' as a seed sequence in *Craniata* species. We also detected other motifs specific to other clades, like plants or worms, at seed positions or not.

We also analyzed the 3,707 human mature microRNAs newly discovered by Londin et al [5], with the same methodology. We will show the similarities and the specificities brought by these more tissue specific microRNAs compared to the miRBase microRNAs regarding their k-mer content.

To conclude, we propose here an overview of the k-mer content of miRBase microRNA sequences. Our results confirm the importance of the seed sequence, consisting in hexamers at position 2, but also show a significant enrichment of clade-specific hexamers at positions 1 and 3 and in a smaller extent at positions 4 and 5, which tends to give a more flexible definition of the seed sequence. The method presented here also gives evidence for the emergence of a few seed sequence families of microRNAs during the evolution.

References

[1] M. W. Rhoades, B. J. Reinhart, L. P. Lim, C. B. Burge, B. Bartel, and D. P. Bartel, "Prediction of Plant MicroRNA Targets," pp. 513–520, 2002.

[2] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen, "Principles of MicroRNA – Target Recognition," no. March, 2005.

[3] S. L. Ameres and P. D. Zamore, "Diversifying microRNA sequence and function," *Nat. Publ. Gr.*, vol. 14, no. 8, pp. 475–488, 2013.

[4] V. Huang, J. Zheng, Z. Qi, J. Wang, R. F. Place, J. Yu, H. Li, and L. Li, "Ago1 Interacts with RNA Polymerase II and Binds to the Promoters of Actively Transcribed Genes in Human Cancer Cells," no. September, 2013.

[5] E. Londin, P. Loher, A. G. Telonis, K. Quann, P. Clark, Y. Jing, E. Hatzimichael, Y. Kirino, S. Honda, M. Lally, B. Ramratnam, C. E. S. Comstock, K. E. Knudsen, L. Gomella, G. L. Spaeth, L. Hark, L. J. Katz, A. Witkiewicz, A. Rostami, S. a Jimenez, M. a Hollingsworth, J. J. Yeh, C. a Shaw, S. E. McKenzie, P. Bray, P. T. Nelson, S. Zupo, K. Van Roosbroeck, M. J. Keating, G. a Calin, C. Yeo, M. Jimbo, J. Cozzitorto, J. R. Brody, K. Delgrosso, J. S. Mattick, P. Fortina, and I. Rigoutsos, "Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 10, pp. E1106–15, Mar. 2015.

Mots clefs : microRNAs, statistics, phylogenetics, kmers, sequence, seed

Sparse regression models to predict the antibiotic resistance profile of a bacteria from its genome

Pierre Mahé [*1], Antoine Bonnefoy[2], Ismael Ouamlil[1], Philippine Barlas[1], Stéphane Schicklin[1], Jean-Baptiste Veyrieras[1]

Poster 122

[1] Bioinformatics Research Department – BIOMÉRIEUX – France
[2] Laboratoire d'Informatique Fondamentale – Université Aix-Marseille – France

Recent advances in sequencing technologies provide new instruments to sequence large amount of DNA for a reasonable cost and in a limited time. This technological breakthrough is expected to significantly modify the landscape and practices in the field of clinical microbiology. Microorganisms, either isolated or within their environment, can now be characterized with an unprecedented level of resolution, which can have significant impact for both research and diagnostics purposes. In terms of diagnostics, next-generation sequencing indeed holds the promise of addressing with a single experiment the main questions of clinical interest, namely identifying an isolate and determining its antibiotic resistance profile (Didelot et al. 2012). Identifying a microorganism from its complete genome is indeed straightforward: the concept of bacterial species is usually defined in terms of conservation or divergence of well-established markers (like 16s rRNA gene for bacteria), that can be readily detected once the genome is available. This is also the case of further "typing" makers (e.g., MLST) that can be used to identify a microorganism at sub-species levels, which can be interesting for epidemiology purposes. The genetic bases of antimicrobial resistance, on the other hand, are still largely unknown, and it is still an open question whether the resistance of a microorganism can be inferred from its genome only. Several recent studies have demonstrated the feasibility of such a so-called genotypic approach, where a good concordance has been observed between resistance phenotypes predicted from microorganisms genomes, and their actual phenotypes, determined experimentally by assessing their ability to develop in the presence of antibiotics. This was for instance the case for *Staphylococcus aureus* (Gordon et al., 2014) and *Pseudomonas aeruginosa* (Kos et al., 2015), the resulting prediction rules involving the detection of genes, or specific mutations within these genes, known to cause – or at least to be associated to – antibiotic resistance.

Motivated by these latter proofs of concept, we propose in this work to address the issue from the supervised "machine" learning perspective using penalized regression models. Starting from a panel of strains that have been sequenced, assembled, and phenotypically characterized for their antibiotic resistance, we first propose a systematic procedure to extract a large list of candidate resistance-related genetic determinants, collectively referred to as their "resistome". This procedure relies on the prior knowledge of a list of genes known to be involved in antibiotic resistance, which is available in public databases like CARD and ARDB, among others. It consists in detecting the presence of each of these genes among the strains of the training panel, carrying out a multiple alignment of the hits found for a given gene, and encoding any possible mutation (SNP, insertion or deletion) encountered within each strain. Starting from a list of 100 to 200 genes, this procedure allows to define a genotypes matrix that may contain up to 10000 to 20000 features, encoding either the presence of a given gene within a strain, or the presence of a mutation at a given position within one of these genes. Having defined such a high-dimensional genotypes matrix, we propose to rely on standard penalized regression models to build predictive models of microorganisms resistance phenotypes. Since we expect the underlying biological mechanisms to involve a limited number of genetic determinants, and we want to retain some level of interpretability in the models obtained, we target sparse models, and naturally rely on the L1 (lasso) penalty.

*. Intervenant

We demonstrate the relevance of this approach using two datasets from the public domain, respectively focusing on the antibiotic resistance of *Staphylococcus aureus* and *Pseudomonas aeruginosa*, two important human pathogens involved, in particular, in hospital-acquired infections. We demonstrate indeed performance comparable to that obtained by alternative methods reported in the literature (Gordon et al., 2014, Kos et al., 2015, Drouin et al., 2016). Most of these works also base their predictions on the presence of genes or of specific mutations identified by means of a manual and iterative process. Our approach can be seen as an attempt to automate this process. The resulting models are in some cases in accordance with the models that are expected when the underlying resistance mechanisms are well known. This is for instance the case of the model predicting the resistance of *Staphylococcus aureus* to methicillin, were the model identified by our model selection process solely relied on the presence of the mecA gene. In other cases, the models obtained are not as sparse as what can be expected from a biological perspective. This may be a model selection issue, or an intrinsic limitation of the L1 regression in terms of support recovery in the presence of correlated predictors. Interestingly, we demonstrate significantly higher performances of the L1-penalized regression models over their ridge counterparts.

Finally, we discuss some on-going work aiming to improve the models obtained by this approach by addressing several specificities of the problem. Regarding strain genotyping, in particular, we note that our approach leads to strongly correlated features, which is due both to linkage disequilibrium and to the fact that our simple genotyping process proceeds position per position along the gene sequences, hence breaks haplotypes. We note moreover that the features we consider, defined as presence or absence of genes and mutations within these genes, exhibit a hierarchical structure. We describe attempts to take into account the structured nature of these predictors within the framework of structured sparsity. We also discuss the relevance of multi-task learning approaches, since one is usually interested in practice in determining the "antibiogram" of a microorganism, that is, in predicting its resistance to several drugs.

References

X. Didelot, R. Bowden, D. J. Wilson, T. E. A. Peto, and D. W. Crook. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, 13(9):601–612, 2012.

A. Drouin et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. bioRxiv, 2016.

N. C. Gordon et al. Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *Journal of Clinical Microbiology*, 52(4):1182–1191, 2014.

V. N. Kos et al. The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrobial Agents and Chemotherapy*, 59(1): 427–436, 2015.

Mots clefs : résistance aux antibiotiques, bactériologie, diagnostic, apprentissage statistique, machine learning

ALFA : A generic tool to compute and display reads distribution by genomic categories and biotypes

Poster 123

Benoît Noël [*] [†1], Mathieu Bahin [*] [‡1], Leila Bastianelli[1], Hervé Le Hir[1], Alice Lebreton[1], Auguste Genovesio [§1]

[1] Institut de Biologie de l'ENS (Paris - Ulm) (IBENS) – École Normale Supérieure de Paris - ENS Paris, CNRS : UMR8197, Inserm : U1024 – 46 rue d'Ulm, F-75 005 PARIS, France

The last ten years have witnessed the rise of a myriad of applications that take advantage of Next-Generation Sequencing (NGS) technologies. In the vast majority of cases, whatever the species, whatever the sequencing technique, the first analysis step of this type of data consists of a quality control of the reads while the second step consists of a mapping of those reads to a reference genome. However, the subsequent steps are often very specific to the type of NGS experiment.

With this work, we aim at introducing a third systematic step after mapping which would be common to any NGS experiment. This step consists in producing a global overview of the distributions of the mapped reads across genomic categories (stop codon, 5'-UTR, CDS, intergenic, etc.) and biotypes (protein coding, miRNA, ncRNA, etc.) at nucleotide resolution. Our approach turns out to be very useful for a broad range of NGS applications we are dealing with, as it brings a sort of post-mapping quality control and a first global functional insight. In any case, it adds information to the usual mapped/unmapped read count and other post-mapping statistics.

A few tools providing this type of information have been proposed in the literature for specific NGS applications. For instance, Homer or CEAS, dedicated to ChIP-seq data, count detected peaks found in each of a predefined set of categories. However, as those tools cannot conveniently deal with mapped reads, their application to other sequencing techniques is precluded. In fact, to the best of our knowledge, there is no available ready-made tool that proposes such a quantitative overview at a nucleotide precision. Furthermore, using directly the mapped reads allows us to propose a framework working for any species and whatever the sequencing technique.

The tool we propose works in two steps. First, a provided genome annotation file (GTF format) is processed to generate an index. Each nucleotide of the genome is annotated according to either a standard or a custom priority definition between features. Then the program computes the nucleotide fraction mapped to each predefined feature in one or more BAM files. By default, the program outputs a raw count and a normalized count plots for the categories and respectively for the biotypes. The normalization is achieved according to the relative importance of a given category or biotype in the genome in order to provide a view in term of enrichment.

We will show results obtained by the proposed tool on various types of NGS experiments such as: 1) RIP-Seq data on *Saccharomyces cerevisiae* samples to quantify whether the IP and Input reads are equally represented in the 3'-UTR region of the genes, 2) MeRIP-Seq data on *Arabidopsis thaliana* samples to identify the type of RNA preferentially methylated and 3) Ribosome Profiling on human and mouse data to discover at an early step of the analysis that some low quality samples should be discarded.

*. Intervenant
†. Corresponding author : bnoel@biologie.ens.fr
‡. Corresponding author : mathieu.bahin@biologie.ens.fr
§. Corresponding author : auguste.genovesio@ens.fr

Overall, we present a versatile, open source and freely available tool that is of a potential wide interest for the bioinformatics community.

Mots clefs : mapped reads, genomics, tool, quality control, NGS, visualization

Analysis of nanoparticle mixtures from the environment

Nina Paffoni [*1], Marc Bailly-Bechet[2], Yasmina Fedala[3], Claude Boccara[3], Chris Bowler[1], Martine Boccara [2,1]

Poster 124

[1] Institut de biologie de l'école normale supérieure (IBENS) – École normale supérieure [ENS] - Paris, Inserm : U1024, CNRS : UMR8197 – 46 rue d'Ulm, F-75 005 Paris, France
[2] Atelier de BioInformatique (ABI) – Université Pierre et Marie Curie (UPMC) - Paris VI – 12 rue cuvier, F-75 005 Paris, France
[3] Institut Langevin - Ondes et Images - ESPCI ParisTech, CNRS UMR 7587, INSERM U979 (Institut Langevin - Ondes et Images) – ESPCI ParisTech, Inserm, Université Paris Diderot - Paris 7, CNRS : UMR7587 – 1 rue Jussieu, F-75 238 Paris Cedex 05, France

Viruses are the most abundant entities on earth. In addition of being human pathogens they play important roles in regulating bacterial communities and in biogeochemical cycles. Other very common entities in environmanetal samples are membrane vesicles. There are more and more reports of the 'vesiculome' of various biotopes. Membrane vesicles are secreted by all organisms, they are made of lipids and proteins and may contain nucleic acids. Their role is only starting to be revealed. It is crucial to quantify and distinguish these biotic nanoparticles (viruses and membrane vesicles) in any environment to estimate their respective abundance, in order to understand the environment ecology. However unlike bacteria or eukaryotic micro-organisms, there is no universal conserved genetic element shared between viral genomes, such as the 16S ribosomal RNA (rRNA) gene to determine their abundance. It is thus necessary to count them, which is usually done by quantitative electronic microscopy or fluorescent staining wich is unprecise and biased, a slow and costly process.

We developed an optical microscope which uses interferometry to detect these small objects and follow them by amplification of their scattering signal. In addition to the measurement of the intensity of their scattering signal, the method allows the study of the Brownian motion of the particles. The individual localization of each particle allow their counting. To characterize further the particles we determine their diameters with two different methodologies: first from the average jump between two successive frames, which is a signal inversely related to the volume of the particle, and second from the maximum intensity of their scattering signal, which is proportional to their refractive index (i.e their density). We validated our detection method with calibrated beads and known viruses from various families with different genetic material (double or single stranded DNA or RNA). We tentatively established a reliable method to analyze the composition of samples by clustering analysis using the two measurements of diameter. The R package Mclust which is based on Gaussian mixture model appears to be adapted in our context. We can distinguish different types of particles both by size and refractive index and estimate their proportion.

We applied our method to analyze marine samples from TARA Oceans. Indeed, sea water is rich in various biologic particles of nanometer size which are mostly viruses and vesicles. Because viruses constitute 90 % of the biomass in the seas, distributions of marine viruses and vesicles are indicatives of the richness of the local environment. For example, our analysis revealed that coastal samples are more abundant in vesicles and diversified in viruses than oligotrophic environment samples.We will present the results and the limits of the Mclust analysis. We plan to confirm our analysis later with metagenomic data.

*. Intervenant

Mots clefs : interferometry, viruses, vesicles, TARA oceans

CORGI : un outil web versatile pour l'identification de groupes de co-régulation

Sandra Pelletier [*] [†1], Sylvain Gaillard [*] [‡1], Hugo Pereira[1], Herman Höfte[2], Marie-Laure Martin-Magniette [3,4], Jean-Pierre Renou [§ 1], Sébastien Aubourg [¶ 1]

Poster 125

[1] Institut de recherche en Horticulture et Semences (IRHS) – Institut national de la recherche agronomique (INRA) : UMR1345, Agrocampus Ouest, Université d'Angers – IRHS 42 rue Georges Morel, F-49 071 Beaucouzé Cedex, France
[2] Institut Jean-Pierre Bourgin, INRA-AgroParisTech, Saclay Plant Science – Institut national de la recherche agronomique (INRA) : UMR1318 – Centre de Versailles-Grignon, Route de St-Cyr (RD10), F-78 026 Versailles Cedex, France
[3] Mathématiques et Informatique Appliquées (MIA) – Institut national de la recherche agronomique (INRA) : UMR0518 – F-75 231 Paris Cedex 05, France
[4] Institute of Plant Sciences Paris Saclay – Institut national de la recherche agronomique (INRA) : UMRIPS2 – Bâtiment 630, Rue Noetzlin, F-91 405 Orsay, France

L'integration des données de transcriptomique fournie une voie d'accès à l'inférence de fonction des gènes inconnus, pour peu que l'on puisse établir des relations entre les profils d'expression des gènes et des conditions particulières de developpement ou d'environnement. Toutefois les outils de clustering couramment utilisés montrent vite leurs limites pour repondre à ce type de question avec une demarche non-supervisee. De plus les groupes obtenus sont souvent biaisés par le poids trop important donné à la valeur quantitative des ratios. Nous avons donc développé un logiciel de recherche de groupes de gènes co-régulés combinant une discrétisation des données pour s'affranchir de la valeur de différence d'expression, et un test statistique basé sur la convergence de la loi binomiale vers la loi normale, nommé CORGI : CO-Regulated Genes Identification.

D'un point de vue général cet outil permet d'identifier les membres d'un groupe ayant la plus forte probabilité de se retrouver ensemble dans un grand nombre de circonstances, et en corollaire d'identifier les circonstances les plus explicatives pour la constitution de ce « noyau » au sein du groupe. Appliqué à la transcription des gènes, il permet d'identifier au sein d'une liste de gènes dérégulés dans le même sens dans une condition donnée, quels sont ceux qui sont le plus souvent co-régulés par interrogation d'un ensemble de conditions. Les conditions les plus explicatives de la co-régulation pour ce groupe permettront ainsi d'identifier la réponse biologique commune aux membres de ce groupe.

L'outil pourra être appliqué ensuite à tous types de données biologiques, si les mesures peuvent être représentées en ratio entre différentes conditions. Elle pourront alors être discrétisées en trois valeurs (+1 : sur-expression, 0 : pas de différence, -1 : sous-expression).

Ainsi la première étape consiste à produire une matrice de données discrétisées pour l'ensemble des expériences et des mesures (transcrits ou autres) qui constitue la ressource à interroger. À partir d'une liste de choix, l'outil sélectionne un sous-ensemble de la matrice, et applique le test de convergence de la loi binomiale vers la loi normale, de façon successive sur les deux dimensions, avec itérations jusqu'à stabilisation des dimensions du bi-cluster.

*. Intervenant
†. Corresponding author : sandra.pelletier@angers.inra.fr
‡. Corresponding author : Sylvain.Gaillard@angers.inra.fr
§. Corresponding author : jean-pierre.renou@angers.inra.fr
¶. Corresponding author : sebastien.aubourg@angers.inra.fr

Test de convergence de la loi binomiale vers la loi normale :

$$(X\text{-}n\pi)/\mathrm{sqrt}(n\pi(1\text{-}\pi)) < \mu(1\text{-}\alpha/2) \text{ et } n\pi > 10$$

avec, pour un élément étant soit un gène soit une expérience :

- X : nombre total d'élément dérégulé (+1 et -1),
- n : nombre d'élément dérégulé dans le même sens (+1 ou -1),
- π : proportion d'élément dérégulé dans le même sens (+1 ou -1),
- $\mu(1\text{-}\alpha/2)$: seuil à partir duquel le test est accepté (seuil de confiance).

La proportion π est choisie par l'utilisateur. Pour permettre de sélectionner seulement des biclusters conséquents, l'outil propose un seuil supplémentaire de n minimum.

CORGI est développé en JavaScript en utilisant le framework Qooxdoo [1]. C'est une application web dynamique s'appuyant sur les dernières spécifications du HTML5 tirant notamment profit de l'API File [2] pour charger et enregistrer des fichiers sur le système de fichier de l'utilisateur. Les représentations graphiques utilisent la librairie D3js [3] pour générer des figures SVG interactives. CORGI est un logiciel libre publié sous licence CeCiLL dont le code est hébergé par la forge SourceSup de RENATER : `https://sourcesup.renater.fr/projects/corgi/`.

Références

[1] `http://qooxdoo.org/`

[2] `https://www.w3.org/TR/FileAPI/`

[3] `https://d3js.org/`

Mots clefs : biclustering, genes expression, transcriptomic, coregulation

HiC-Pro : an optimized and flexible pipeline for Hi-C data processing

Nicolas Servant [*] [†1], Nelle Varoquaux[1], Bryan Lajoie[2], Éric Viara[1], Chong-Jian Chen[1,3], Jean-Philippe Vert[1], Edith Heard[3], Job Dekker[2], Emmanuel Barillot[1]

Poster 126

[1] Institut Curie, PSL Research University, Mines Paris Tech, Bioinformatics and Computational Systems Biology of Cancer, INSERM U900, F-75 005 PARIS, France
[2] Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School (UMASS) – Worcester, Massachusetts 01605-0103, États-Unis
[3] Institut Curie, PSL Research University, Genetics and Developmental Biology Unit, INSERM U934/CNRS UMR3215, F-75 005 PARIS, France

High-throughput chromosome conformation capture methods are now widely used to map chromatin interactions within regions of interest and across the genome. The Hi-C technique involves sequencing pairs of interacting DNA fragments, where each mate is associated with one interacting locus. In practice, Hi-C usually requires several millions to billions of paired-end sequencing reads, depending on genome size and on the desired resolution. Managing these data thus requires optimized bioinformatics workflows able to extract the contact frequencies in reasonable computational time and with reasonable resource and storage requirements.

Here, we present HiC-Pro [1], an easy-to-use and complete pipeline to process Hi-C data from raw sequencing reads to normalized contact maps. HiC-Pro allows the processing of data from Hi-C protocols based on restriction enzyme or nuclease digestion such as DNase Hi-C or Micro-C. HiC-Pro is organized into four distinct modules following the main steps of Hi-C data processing: i) read alignment, ii) detection and filtering of valid interaction products, iii) binning and iv) contact map normalization. To assess the quality of a Hi-C experiment, HiC-Pro performs a variety of quality controls at different steps of the pipeline, such as mapping statistics, details about ligation products, or fractions of intra-/inter-chromosomal interactions. When phased genotypes are available, HiC-pro is able to distinguish allele-specific interactions and to build both maternal and paternal contact maps.

It is optimized and offers a parallel mode for very high-resolution data as well as a fast implementation of the iterative correction method. The main steps of the pipeline are implemented in Python and C++ programming languages and are based on efficient data structures, such as compressed sparse raw matrices for contact counts data. We applied HiC-Pro to different Hi-C datasets, demonstrating its ability to easily process large data in a reasonable time. For exemple, the IMR90 sample from the Rao et al dataset [2] (1.5 billions read pairs) was processed in parallel using 320 CPUs to generate up to 5 kb contact maps in 12 hours. We also provide an implementation of the iterative correction procedure which emphasizes ease of use, performance, memory-efficiency and maintainability. We obtain higher or similar performance on a single core compared with the original ICE implementation from the hiclib library [3] and from the HiCorrector package [4]. In addition, we also compared the intra and inter-chromosomal contact maps generated by HiC-Pro with the results of the hiclib python library. Both pipelines generate concordant results across processing steps.

HiC-Pro is a flexible and efficient pipeline for Hi-C data processing. It is freely available under the BSD licence as a collaborative project at `https://github.com/nservant/HiC-Pro`. The raw

*. Intervenant
†. Corresponding author : `nicolas.servant@curie.fr`

and normalized contacts maps are compatible with visualization software such as HiCPlotter or Juicebox, and with the HiTC BioConductor package for further analysis.

References

[1] Servant N, Varoquaux N, Lajoie BR, Viara É, Chen CC, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* 2015, 16:259 doi:10.1186/s13059-015-0831-x

[2] Rao SSP, Huntley MH, Durand NC, Bochkov SID, Robinson JT, Sanborn AL, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159(7):1665–80. doi:10.1016/j.cell.2014.11.021.

[3] Imakaev M, Fudenberg G, Patton McCord R, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*. 2012;9(10):999–1003. doi:10.1038/nmeth.2148.

[4] Li W, Gong K, Li Q, Alber Fand Zhou XJ. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics*. 2015;31(6):960–2. doi:10.1093/bioinformatics/btu747.

Mots clefs : HiC, chromatin, pipeline, normalization

Analysis of transposable elements within heterogeneous NGS data

Aurélie Teissandier [*1,2], Nicolas Servant[1], Emmanuel Barillot[1], Déborah Bourc'his[2]

Poster 127

[1] U900 – Inserm : U900, Institut Curie, MINES ParisTech - École nationale supérieure des Mines de Paris – France
[2] U934 – Inserm : U934, CNRS : UMR3215, Institut Curie – France

Retrotransposable elements are present in millions of copies and represent around 40% of mammalian genomes. Several classes of retrotransposons have co-evolved, and they are extremely diverse in sequence, length, structure, number and functional properties. Over the course of evolution, retrotransposons have provided beneficial innovations and have participated to genome shaping and speciation. However, in the short term, they can contribute to genome instability by altering gene organization and expression. Retrotransposons represent a significant challenge for next generation sequencing (NGS) analyses. Notably, the majority of sequence reads derived from these elements map to multiple positions in the genome, preventing unambiguous conclusions about their origin. For this reason, these multi-mapped reads are usually discarded from classical NGS analysis. Moreover, flawed bioinformatics analyses can induce erroneous conclusions.

Our goal is to develop a systematic pipeline adapted to the analysis of all retrotransposons classes in various mammalian genomes and from various NGS datasets (RNA-seq, small RNA-seq, ChIP-seq and WGBS). Using a simplified genome, we carried out a comprehensive analysis of the different parameters which can influence the processing of transposon sequences, such as sequencing library approach (single or paired-end reads, read length) and alignments tools. Our comparison proves that Novoalign has the best specificity and sensitivity to map reads over repetitive sequences. Moreover, we were able to determine the quality and specificity of the information that can be gained, depending on the the classes of retrotransposons, and the evolutionary age of individual elements.

Mots clefs : retrotranposable elements, analysis, Next Generation Sequencing, mapping, multi, mapped reads

*. Intervenant

Integration and visualization of epigenome and mobilome data in crops

Robakowska Hyzorek Dagmara [*1], Marie Mirouze [†2], Pierre Larmande [‡2,3,4]

Poster 128

[1] Parcours « Bioinformatique, Connaissances, Données » du Master Sciences & Numérique pour la Santé de l'Université de Montpellier (UM) – Université de Montpellier, 163 rue Auguste Broussonnet, F-34 090 Montpellier, France
[2] Institut de Recherche pour le Développement (IRD), UMR232 DIADE, Laboratoire Génome et Développement des Plantes, Perpignan (IRD) – Centre IRD de Montpellier, 911 avenue Agropolis, BP 64501, F-34 394 Montpellier Cedex 5, France
[3] Institut de Biologie Computationelle (IBC), Montpellier (IBC) – Université de Montpellier, 860 rue St Priest, Bâtiment 5, CC05019, F-34 095 Montpellier Cedex 5, France
[4] Équipe Zenith, INRIA et LIRMM, Montpellier (LIRMM) – Université Montpellier 2, LIRMM UMR 5506, CC477, 161 rue Ada, F-34 095 Montpellier Cedex 5, France

Epigenetics is a field of increasing interest in Biology because epigenetic mechanisms strongly influence developmental and cellular processes, and determine gene expression, DNA repair, and recombination. Therefore epigenetics data integration is especially relevant, given the complex nature and interactions among the mechanisms responsible for the deposition of various epigenetic marks. Over the last decade, the number of publications having 'epigenetics' in their title increased around 10 times (Deans et al., 2015). In addition the amount of high-throughput epigenomic data generated has increased exponentially. Archiving, curating, analysing, and interpreting all of these datasets represent a major challenge. Therefore, the development of methods that allow the proper storage, searching, and retrieval of information becomes critical (Aguiar-Pulido et al., 2015).

In parallel, recently a growing interest for the genomic compartment containing active transposable elements, or mobilome, has also gain a considerable interest, notably in crops with large genomes (Springer, 2013). In the coming years, the study of the interaction between the epigenome and the mobilome is likely to give insights on the role of TEs on genome stability and evolution (Zhao et Zhou., 2012).

The present project aims at studying existing approaches that overcome the challenges of epigenetic data analysis. Integration of heterogeneous measurements of epigenetics variation is non-trivial due to the diversity and variety of output data formats. To address these issues, we have created tools to collect epigenetic datasets generated in different laboratories as well as from different databases and translate them to a standard format to be integrated, analysed and finally visualized. I will present our online epigenome and mobilome database for the rice data and will highlight the tools we have used for linear and circular visualization of the data (Bules et al., 2016; Krzywinski et al.,2009).

I would like to thank to Department of Informatics of the "Faculté des Sciences de l'Université de Montpellier" (`http://deptinfofds.univ-montp2.fr/`) and the Labex Numev (`http://www.lirmm.fr/numev/`) for funding my participation at the "Journées Ouvertes en Biologie, Informatique et Mathématiques" (JOBIM).

Aguiar-Pulido V., Suarez-Ulloa J. M., Eirin-Lopez J., Pereira J., Narasimhan G. (2015). "Computational Methods in Epigenetics, In Personalized Epigenetics" Editor: T. Tollefsbol, Springer,

*. Intervenant
†. Corresponding author : `marie.mirouze@ird.fr`
‡. Corresponding author : `plarmande@gmail.com`

Book Chapter 6:153-180.

Buels R., Yao E., Diesh C.M., Hayes R.D, Munoz-Torres M., Helt G., Goodstein D.M., Elsik C.G., Lewis S.E., Stein L., Holmes I.H. (2016). "JBrowse: a dynamic web platform for genome visualization and analysis." *Genome Biol.* 17(1):66.

Deans C., Maggert K. A. (2015). "What Do You Mean, 'Epigenetic'?" *Genetics* 199(4):887-896.

Krzywinski M., Schein J., Birol, I., Connors J., Gascoyne R., Horsman D., Jones S. J.,Marra M. A.(2009). "Circos: An information aesthetic for comparative genomics." *Genome Research* 19(9):1639-1645.

Springer N.M. (2013). "Epigenetics and crop improvement." *Trends Genet.* 29(4):241-7.

Zhao Y., Zhou D.-X. (2012). "Epigenomic Modification and Epigenetic Regulation in Rice." *Journal of Genetics and Genomics* 39(7):307-315.

Mots clefs : epigenetics, databases, data integration, data visualisation, bioinformatics

SPADEVizR : an R package for visualization, analysis and integration of SPADE results

Poster 129

Guillaume Gautreau [*] [†1], David Pejoski[1], Ludovic Platon[1], Brice Targat[1], Anne-Sophie Beignon[1], Nicolas Tchitchek [‡1]

[1] Commissariat à l'Énergie Atomique et aux Énergies Alternatives, Département Immunologie des Infections Virales et des Maladies Auto-immunes (CEA) – CEA, Inserm : U1184, Université Paris Sud - Paris XI. CEA Fontenay-aux-Roses – 18 route du Panorama, F-92 260 Fontenay-aux-Roses, France

Flow and mass cytometry are experimental techniques used for the characterization of cell phenotypes. With the increase of usable cell markers (up to 50 markers), the identification of cell populations through manual gating is impossible. These high-dimensional data require new computational algorithms to automatically identify cell clusters. The SPADE algorithm has been proposed as a new way to analysis and explore mass-cytometry data. This algorithm performs a density-based down-sampling combined with an agglomerative hierarchical clustering.

While SPADE offers new opportunities for identifying cell populations, complementary approaches are needed to improve the characterization of identified cell populations. We present here SPADEVizR, an R package to better visualize and analyze SPADE clustering results. We extended the original SPADE outputs with techniques such as parallel coordinates, heatmaps, multidimensional scaling, volcano plots or streamgraph representations. Moreover, the proposed statistical methods allow the identification of SPADE clusters with relevant biological behaviors. For instance, significantly abundant clusters or differentially enriched clusters can be identified using SPADEVizR. In addition, the integration of cell cluster behaviors with additional phenotypical variables can be performed.

We illustrate the capabilities of our R package using a dataset of 15 cytometry profiles with 25 markers each, obtained in the context of a MVA macaque vaccine study. SPADEVizR has been designed in a way that it can be easily used by non bioinformatician experts, but can also be easily customizable by users with more expertise in bioinformatics. Through the multiple visualization and analysis features, SPADEVizR is also a powerful analysis pipeline for high-dimensional cytometry data.

Mots clefs : High dimensional cytometry data, Automatic gating, Visualization, Statistical analyses, SPADE

*. Intervenant

†. Corresponding author : guillaume.gautreau@free.fr

‡. Corresponding author : nicolas.tchitchek@gmail.com

Cassandre : A web-application for large-scale data management and visualisation

Simon Malesys [*1], Hervé Ménager[2], Cosmin Saveanu[1], Christophe Malabat[*1,3]

Poster 130

[1] Unité de Génétique des Interactions Macromoléculaires (GIM) – Institut Pasteur de Paris, Centre National de la Recherche Scientifique - CNRS – Unité de Génétique des Interactions Macromoléculaires, Institut Pasteur, 25-28 rue du Docteur Roux, F-75 724 Paris Cedex 15, France
[2] Centre d'informatique pour la biologie – Institut Pasteur de Paris – 25 rue du Dr Roux, F-75 015 Paris, France
[3] Hub de bioinformatique & biostatistique, Centre de Bioinformatique, biostatistique et Biologie Intégrative (C3BI) – Institut Pasteur de Paris, Centre National de la Recherche Scientifique - CNRS – Institut Pasteur, 25-28 rue du Docteur Roux, F-75 724 Paris Cedex 15, France

Large-scale experiments that generate quantitative data, such as changes in the levels of thousands of RNA or protein molecules under various conditions, are a hallmark of modern biology. Many other experiments, like those measuring functional interactions between mutant alleles or the variation in translation status of mRNAs are also characterized by the same basic information unit: an experimental condition, a genomic feature (usually a gene) and an associated numerical value. We have developed of a tool for the exploration of this type of large-scale data. This new "window" to quantitative experimental results is based on an existing prototype used to handle hundreds of genome-wide genetic interaction screen results. The architecture of this open-source tool (GPL v3.0) is based on the MEAN stack (MongoDB, Express, Angular, Node). Visualisations use the Plotly library. The interface provides users with multiple ways to visualize and explore data with a high degree of interaction with the generated figures and with minimal constraints on the data format.

Mots clefs : visualisation, data management

*. Intervenant

Liste des auteurs

C

N

O

P