Activity Report 2015

Team PLEIADE

patterns of diversity and networks of function

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Bordeaux - Sud-Ouest

THEME
Computational Biology

Table of contents

Team PLEIADE

Creation of the Team: 2015 January 01

Keywords:

Computer Science and Digital Science:

1.5. - Complex systems
3.1. - Data
3.2. - Knowledge
3.3. - Data and knowledge analysis
3.4.1. - Supervised learning
3.4.2. - Unsupervised learning
6.1. - Mathematical Modeling
6.2.8. - Computational geometry and meshes

Other Research Topics and Application Domains:

1.1. - Biology
1.1.11. - Systems biology
1.1.12. - Synthetic biology
1.1.5. - Genetics
1.1.6. - Genomics
1.1.8. - Evolutionnary biology
1.1.9. - Bioinformatics
1.2. - Ecology
1.2.1. - Biodiversity
3.1. - Sustainable development
3.4.1. - Natural risks
3.4.3. - Pollution
4.2.1. - Biofuels

1. Members

Research Scientists

David Sherman [Team leader, Inria, Senior Researcher, HdR]
Pascal Durrens [CNRS, Researcher, HdR]
Alain Franc [INRA]
Stephanie Mariette [INRA]

Engineers

Philippe Chaumeil [INRA]
Jean-Marc Frigerio [INRA]
Franck Salin [INRA]

Post-Doctoral Fellow

Razanne Issa [INRA]

Administrative Assistant

Anne-Laure Gautier [Inria]

Others

Ulysse Guyet [Inria, intern, from Mar 2015 until May 2015]
Leyla Mirvakhabova [Inria, intern, from Nov 2015]
Hugo Ignacio Campbell Sills [Univ. Bordeaux]
Anna Zhukova [Institut Pasteur]

2. Overall Objectives

2.1. Overall Objectives

Diversity, evolution, and inheritance form the heart of modern biological thought. Modeling the complexity of biological systems has been a challenge of theoretical biology for over a century [25] and flourished with the evolution of data for describing biological diversity, most recently with the transformative development of high-throughput sequencing. However, most concepts and tools in ecology and population genetics for exploiting diversity data are still not adapted to high throughput data production. A better connection is needed: *computational biodiversity*.

Paradoxically, diversity emphasizes differences between biological objects, while modeling aims at unifying them under a common framework. This means that there is a limit beyond which some components of diversity cannot be mastered by modeling. We need efficient methods for recognizing patterns in diversity, and linking them to patterns in function. It is important to realize that diversity in function is not the same as coupling observed diversity with function.

Diversity informs both the study of traits, and the study of biological functions. The double challenge is to measure these links quickly and precisely with pattern recognition, and to explore the relations between diversity in traits and diversity in function through modeling.

PLEIADE links pattern recognition with modeling in biodiversity studies and biotechnology. We develop distance methods for NGS datasets at different levels of organization: between genomes, between individual organisms, and between communities; and develop high-performance pattern recognition and statistical learning techniques for analyzing the resulting point clouds. We refine inferential methods for building hierarchical models of networks of cellular functions, exploiting the mathematical relations that are revealed by large-scale comparison of related genomes and their models. We combine these methods into integrated e-Science solutions to place these tools directly in the hands of biologists.

3. Research Program

3.1. Distances and pattern recognition

Diversity may be understood as a set of dissimilarities between objects. The underlying mathematical construction is the notion of distance. Knowing a set of objects, on the condition that pairwise distances can be measured, it is possible to build a Euclidan image of it as a point cloud in a space of relevant dimension. Then, diversity can be associated with the shape of the point cloud. It is still true that the reference for recognizing patterns or shapes is the human eye. One objective of our project is to narrow the gap between the story that a human eye can read, and the story that an algorithm can tell. Several directions will be explored. First, it is necessary to master dimension reduction, mainly classical algebraic tools (PCA, NGS, Isomap, eigenmaps, etc ...), and collaborate with experts in efficient methods in spectral methods. Second, a neighborhood in a point cloud naturally leads to graphs describing the neighborhood networks. There is a natural link between modular structures in distance arrays and communities on graphs. Third, points defined by DNA sequences (for example) are samples of diversity. Dimension reduction may show that they live on a given manifold. This leads to geometry (differential or Riemanian geometry). Knowing some properties of the manifold can inform us about the constraints on the space where the measured individuals live. The connection between Riemannian geometry and graphs, where weighted graphs are seen as meshes embedded in a manifold, is currently an active field of reasearch [23], [22].

To resolve these objectives computationally will require investment in research directions in computational geometry (such as convex hulls of high-dimension sets of points), on circumventing the curse of dimensionality, and on linking distance geometry with convex optimization procedures through matrix completion. None of these questions is trivial: most recent work has focused on two or three dimensions, for example for image analysis or for reconstruction of protein conformation from local distances between atoms. The methodological goal is to extend these approaches to higher dimension spaces.

3.2. Modeling by successive refinement

Describing the links between diversity in traits and diversity in function will require comprehensive models, assembled from and refining existing models. A recurring difficulty in building comprehensive models of biological systems is that accurate models for subsystems are built using different formalisms and simulation techniques, and hand-tuned models tend to be so focused in scope that it is difficult to repurpose them [15]. Our belief is that a sustainable effort in building efficient behavioral models must proceed incrementally, rather than by modeling individual processes *de novo*. *Hierarchical modeling* [11] is one way of combining specific models into networks. Effective use of hierarchical models requires both formal definition of the semantics of such composition, and efficient simulation tools for exploring the large space of complex behaviors. We have previously shown that this approach can be effective for certains kinds of systems in biotechnology [14], [16] and medicine [13]. Our challenge is to adapt incremental, hierarchical refinement to modeling organisms and communities in metagenomic and comparative genomic applications.

4. Application Domains

4.1. Genome and transcriptome annotation, to model function

Sequencing genomes and transcriptomes provides a picture of how a biological system can function, or does function under a given physiological condition. Simultaneous sequencing of a group of related organisms is now a routine procedure in biological laboratories for studying a behavior of interest, and provides a marvelous opportunity for building a comprehensive knowledge base of the relations between genomes. Key elements in mining these relations are: classifying the genes in related organisms and the reactions in their metabolic networks, recognizing the patterns that describe shared features, and highlighting specific differences.

PLEIADE will develops applications in comparative genomics of related organisms, using new mathematical tools for representing compactly, at different scales of difference, comparisons between related genomes. New methods based on computational geometry refine these comparisons. Compact representations can be stored, exchanged, and combined. They will form the basis of new simultaneous genome annotation methods, linked directly to abductive inference methods for building functional models of the organisms and their communities.

Our ambition in biotechnology is to permit the design of synthetic or genetically selected organisms at an abstract level, and guide the modification or assembly of a new genome. Our effort is focused on two main applications: genetic engineering and synthetic biology of oil-producing organisms (biofuels in CAER, palm oils), and improving and selecting starter microorganisms used in winemaking (collaboration with the ISVV and the BioLaffort company).

4.2. Molecular based systematics and taxonomy

Defining and recognizing myriads of species in biosphere has taken phenomenal energy over the past centuries and remains a major goal of Natural History. It is an iconic paradigm in pattern recognition (clustering has coevolved with numerical taxonomy many decades ago). Developments in evolution and molecular biology, as well as in data analysis, have over the past decades enabled a profound revolution, where species can be delimited and recognized by data analysis of sequences. We aim at proposing new tools, in the framework of E-science, which make possible (i) better exploration of the diversity in a given clade, and (ii) assignment of a place in these patterns for new, unknown organisms, using information provided by sets of sequences. This will require investment in data analysis, machine learning, and pattern recognition to deal with the volumes of data and their complexity.

One example of this project is about the diversity of trees in Amazonian forest, in collaboration with botanists in French Guiana. Protists (unicellular Eukaryots) are by far more diverse than plants, and far less known. Molecular exploration of Eukaryotes diversity is nowadays a standard in biodiversity studies. Data are available, through metagenomics, as an avalanche and make molecular diversity enter the domain of Big Data. Hence, an effort will be invested, in collaboration with other Inria teams (GenScale, HiePACS) for porting to HPC algorithms of pattern recognition and machine learning, or distance geometry, for these tools to be available as well in metagenomics. This will be developed first on diatoms (unicellular algae) in collaboration with INRA team at Thonon and University of Uppsala), on pathogens of tomato and grapewine, within an existing network, and on bacterial communities, in collaboration with University of Pau. For the latter, the studies will extend to correlations between molecular diversity and sets of traits and functions in the ecosystem.

4.3. Community ecology and population genetics

Community assembly models how species can assemble or diassemble to build stable or metastable communities. It has grown out of inventories of countable organisms. Using *metagenomics* one can produce molecular based inventories at rates never reached before. Most communities can be understood as pathways of carbon exchange, mostly in the form of sugar, between species. Even a plant cannot exist without carbon exchange with its rhizosphere. Two main routes for carbon exchange have been recognized: predation and parasitism. In predation, interactions–even if sometimes dramatic–may be loose and infrequent, whereas parasitism requires what Claude Combes has called intimate and sustainable interactions [17]. About one decade ago, some works [21] have proposed a comprehensive framework to link the studies of biodiversity with community assembly. This is still incipient research, connecting community ecology and biogeography.

We aim at developping graph-based models of co-occurence between species from NGS inventories in metagenomics, i.e. recognition of patterns in community assembly, and as a further layer to study links, if any, between diversity at different scales and community assemblies, starting from current, but oversimplified theories, where species assemble from a regional pool either randomly, as in neutral models, or by environmental filtering, as in niche modeling. We propose to study community assembly as a multiscale process between nested pools, both in tree communities in Amazonia, and diatom communities in freshwaters. This will be a step towards community genomics, which adds an ecological flavour to metagenomics.

Convergence between the processes that shape genetic diversity and community diversity–drift, selection, mutation/speciation and migration–has been noted for decades and is now a paradigm, establishing a continuous scale between levels of diversity patterns, beyond classical approaches based on iconic levels like species and populations. We will aim at deciphering diversity pattern along these gradients, connecting population and community genetics. Therefore, some key points must be adressed on reliability of tools.

Next-generation sequencing technologies are now an essential tool in population and community genomics, either for making evolutionary inferences or for developing SNPs for population genotyping analyses. Two problems are highlighted in the literature related to the use of those technologies for population genomics: variable sequence coverage and higher sequencing error in comparison to the Sanger sequencing technology. Methods are developed to develop unbiased estimates of key parameters, especially integrating sequencing errors [20]. An additional problem can be created when sequences are mapped on a reference sequence, either the sequenced species or an heterologous one, since paralogous genes are then considered to be the same physical position, creating a false signal of diversity [18]. Several approaches were proposed to correct for paralogy, either by working directly on the sequences issued from mapped reads [18] or by filtering detected SNPs. Finally, an increasingly popular method (RADseq) is used to develop SNP markers, but it was shown that using RADseq data to estimate diversity directly biases estimates [12]. Workflows to implement statistical methods that correct for diversity biases estimates now need an implementation for biologists.

5. New Software and Platforms

5.1. Magus

KEYWORDS: Bioinformatics - Genomic sequence - Knowledge database

SCIENTIFIC DESCRIPTION

MAGUS can be used on small installations with a web server and a relational database on a single machine, or scaled out in clusters or elastic clouds using Apache Cassandra for NoSQL data storage and Apache Hadoop for Map-Reduce.

FUNCTIONAL DESCRIPTION

The MAGUS genome annotation system integrates genome sequences and sequences features, in silico analyses, and views of external data resources into a familiar user interface requiring only a Web navigator. MAGUS implements annotation workflows and enforces curation standards to guarantee consistency and integrity. As a novel feature the system provides a workflow for simultaneous annotation of related genomes through the use of protein families identified by in silico analyses this has resulted in a three-fold increase in curation speed, compared to one-at-a-time curation of individual genes. This allows us to maintain standards of high-quality manual annotation while efficiently using the time of volunteer curators.

- Participants: Florian Lajus, David Sherman, Natalia Golenetskaya, Pascal Durrens and Xavier Calcas
- Partners: Université de Bordeaux - CNRS - INRA
- Contact: David James Sherman
- URL: http://magus.gforge.inria.fr

5.2. Mimoza

KEYWORDS: Systems Biology - Bioinformatics - Biotechnology

FUNCTIONAL DESCRIPTION

Mimoza uses metabolic model generalization and cartographic paradigms to allow human experts to explore a metabolic model in a hierarchical manner. The software creates an zoomable representation of a model submitted by the user in SBML format. The most general view represents the compartments of the model, the next view shows the visualization of generalized versions of reactions and metabolites in each compartment , and the most detailed view visualizes the initial model with the generalization-based layout (where similar metabolites and reactions are placed next to each other). The zoomable representation is implemented using the Leaflet JavaScript library for mobile-friendly interactive maps. Users can click on reactions and compounds to see the information about their annotations. The resulting map can be explored on-line, or downloaded in a COMBINE archive.

- Participants: Anna Zhukova and David James Sherman
- Contact: David James Sherman
- URL: http://mimoza.bordeaux.inria.fr/

5.3. Pantograph

KEYWORDS: Systems Biology - Bioinformatics - Genomics - Gene regulatory networks

FUNCTIONAL DESCRIPTION

Pantograph is a software toolbox to reconstruct, curate and validate genome-scale metabolic models. It uses existing metabolic models as templates, to start its reconstructions process, to which new, species-specific reactions are added. Pantograph uses an iterative approach to improve reconstructed models, facilitating manual curation and comparisons between reconstructed model's predictions and experimental evidence.

Pantograph uses a consensus procedure to infer relationships between metabolic models, based on several sources of orthology between genomes. This allows for a very detailed rewriting of reaction's genome associations between template models and the model you want to reconstruct.

- Participants: Nicolas Loira, Anna Zhukova, David James Sherman and Pascal Durrens
- Partner: University of Chile
- Contact: Nicolas Loira
- URL: http://pathtastic.gforge.inria.fr/

5.4. biorica

KEYWORDS: Systems Biology - Bioinformatics - Hierarchical models - Hybrid models - Stochastic models
FUNCTIONAL DESCRIPTION

BioRica is used to mathematically describe the behavior of complex biological systems.

It is a software platform that permits simulation of biological systems on the basis of their description. It allows one to reuse existing biological models and to combine them into more complex models.

- Partner: University of Chile
- Contact: David Sherman
- URL: http://biorica.gforge.inria.fr/

5.5. Platforms

5.5.1. Plafrim

Plafrim (http://plafrim.fr) is an essential instrument for PLEIADE. We use it for developing software data analysis methods and evaluating them at real world scale. The platform combines considerable computing power with excellent support, both in terms of the quality of the interactions with the local staff and of the ease of large-scale data transfer between Plafrim and PLEIADE's data storage infrastructure. Plafrim facilitates collaboration between team members who are not in the Bordeaux Sud-Ouest building, and furthermore allows us to share best practices and tools with other teams from the Center.

5.5.2. Inria forge and Inria continuous integration

The Inria forge (http://gforge.inria.fr) provides a secure collaboration platform for software project administration and source code management, and Inria's continuous integration platform (http://ci.inria.fr) provides a cloud-based service for automatic compilation and testing of software systems. PLEIADE uses these two services extensively for agile software development. The continuous integration platform allows us to verify the correct operation of our methods in different operating system and deployment environments.

6. New Results

6.1. Inference of metabolic networks

Participants: David Sherman [correspondant], Razanne Issa, Pascal Durrens.

We are particularly interested in incremental modeling of metabolic networks, where the target organism to be modeled is demonstrably similiar to other organisms for which whole or partial models are available. The other organisms are typically strains of the same species as the target, or species with a close phylogenetic relation to the target species. The similarity is measured genomically at different scales: sequence polymorphisms, expansions and contractions in conserved protein families, and genome rearrangements. We have defined and refined two complementary methods for inferring metabolic models for target species.

In the same way that comparative analysis of genomes and proteomes makes it possible to define protein families that summarize protein-coding genes into phyletic patterns [24], comparative analysis of related metabolic models makes it possible to define network generalizations [26] that factor families of reactions and metabolites into summary graphs that preserve stoichiometry. These summaries can be used for expert curation and visualization [5]. An online demonstration tool is made available at http://mimoza.bordeaux.inria.fr/.

Starting from an existing reference metabolic network and measures of similarity between the reference and the target organisms' genomes, we can use knowledge-based inference to rewrite the reference network based on these differences, and thus obtain a draft network for the metabolisms of the target organism [2]. This rewriting, formalized in the Pantograph system, can be extended to an abductive logic framework as described in Razanne Issa's thesis [19]. Current work aims at extending the Pantograph and ab-Pantograph frameworks to leverage reaction classifications obtained by network generalization.

6.2. Bio-medicine and biotechnology

Participants: Pascal Durrens [correspondant], David Sherman.

6.2.1. Genome assembly for bio-medicine

We performed the assembly of the *Clavispora lusitaniae* (aka *Candida lusitaniae*) genome. Yeasts from the genus *Candida* are opportunistic human pathogens in immunocompromised patients, linked to a high mortality rate. Although *Candida albicans* is the major pathogen, related species are more and more isolated, such as *Clavispora lusitaniae* which is responsible for candidaemia in newborn babies and in onco-hematology patients.

Even though the genome of a *Clavispora lusitaniae* strain (ATCC 42720), isolated from a patient, has already been sequenced by the Broad Institute, we achieved the genome assembly of the wild type reference strain (CBS 6936) as patient isolates tend to harbor genome modifications. The assembly was computed from Illumina reads with a coverage of 30X, using the MINIA assembler from Inria GENSCALE team. We also looked for single nucleotide polymorphismw (SNPs) in the reads coming from 3 hypovirulent mutants impaired in the beta oxidation metabolic pathway. Some detected SNPs are now under experimental validation and we are going to make a Genome Announcement for the CBS 6936 genome.

6.2.2. Transcriptome assembly for bio-technology

We carried out the assembly of transcriptomes from different tissues of the African oil palm tree *Elaeis guineensis*. The goal of this project is twofold: (i) Select the most relevant genes involved in oil synthesis in order to implement heterologous expression of some of these genes in a cultivated plant recipient such as tobacco. Preliminary results on heterologous expression of 2-3 key genes/factors ended in 15% of dry weight of oil synthesis. New expression technology allows for simultaneous expression of 15-20 genes. Identifiying the best candidates for co-expression will permit efficient heterologous oil synthesis. (ii) Identify the polymorphism of genes in a panel of 25 wild type isolates and of 5 production lineages of *Elaeis guineensis* in relation to the oil yield in different environmental conditions. In addition to a high variability of oil quantity (1-12 tons/ha/year), the relative amount of unsaturated fatty acids spans widely (15-55% dry weight) among the 30 *Elaeis guineensis* strains. Identification of polymorphisms will pave the way to genome-wide association genetics (GWAS) for the improvement of the oil resource.

In a first step, we produced assembled transcriptomes of ca. 300 million reads from 3 tissues (leaf, mesocarp, kernel) ocoming form a single tree, using state-of-the-art assembler TRINITY. Tuning of the software parameters was performed on the Inria PLAFRIM computation platform. About 20% of the assembled sequences revealed to be tissue-specific. Computation of the protein sequences deduced from the assembled transcripts gave a protein repertoire which was annotated using related sequences available in public databases. These transcript and protein sets will be used as a framwork in the polymorphism studies.

6.3. Biodiversity and ecology

Participants: Alain Franc [correspondant], Jean-Marc Frigerio, Philippe Chaumeil, Razanne Issa, Leyla Mirvakhabova.

Our main activity has been code development in the framework of a research project with ONEMA, and preparing future development. Code development has been fostered with the work of Razanne Issa (CDD ONEMA) in the last three months of 2015, and preparation of further development has been fostered by welcoming Leyla Mirvakhabova (L3, National Research University Higher School of Economics (NRU HSE), Moscow, mathematics) . Declic is a python library providing tools for analysing molecular data for biodiversity stydies. The main object is a distance matrix, from which one can either build a point cloud in a Euclidean space with distances between points as close as possible from distances between reads (Multidimensional scaling), or to build a graph with edges between reads when their distance is smaller than a given threshold. Meanwhile, the team has developed the network around the Galaxy server where an early verison of tools has been installed and made available, especially with SLU at Uppsala (Maria Khalert). Alain Franc has developed a collaboration with Olivier Coulaud and Pierre Blanchard (Hiepacs) for efficient computation of eigenvectors and eigenvalues of large, dense and symmetric matrices, needed for scaling in Multidimensional scaling.

The work of Razanne Issa has made it possible to extend the declic library in the direction of machine learning, by incorporating tools from support vector machines through library sklearn. This development will be pursued in 2016. The work of Leyla Mirvakhabova has permitted a first incursion into topological data analysis as a possible approach for studying the shape of point clouds produced by multidimensional scaling. The collaboration with NRU HSE on this topic will be pursued in 2016.

7. Partnerships and Cooperations

7.1. National Initiatives

7.1.1. CAER – Alternative Fuels for Aeronautics

CAER is a 6 M-Euro contract with the Civil Aviation Directorate (Direction Générale de l'Aviation Civile, DGAC), coordinated by the French Petroleum Institute (Institut français de pétrole-énergies nouvelles, IFPEN) on behalf of a large consortium of industrial (EADS, Dassault, Snecma, Turbomeca, Airbus, Air France, Total) and academic (CNRS, INRA, Inria) partners to explore different technologies for alternative fuels for aviation. PLEIADE's role concerns the genomics of highly-performant oleaginous microorganisms.

7.2. International Initiatives

7.2.1. Inria International Partners

7.2.1.1. Informal International Partners

PLEIADE collaborates with Rodrigo Assar of the Universidad Andrès Bello, and Nicolás Loira and Alessandro Maass of the Center for Genomic Regulation, in Santiago de Chile (Chile).

7.3. International Research Visitors

7.3.1. Visits of International Scientists

Rodrigo Assar, assistant professor in the ICBM Human Genetics Program of the School of Medicine of the University of Chile, was invited by PLEIADE in the context of an ongoing collaboration on hybrid, stochastic modeling of complex biological systems.

7.3.1.1. Internships

Leyla Mirvakhabova, student at the National research University Higher School of Economics, Moscow, was invited by PLEIADE for an internship to work on faster mathematical methods for nonlinear mapping, to be applied to very large distance matrices.

Ulysse Guyet, Masters student in Bioinformatique-Biostatistique at the University of Nantes, was invited by PLEIADE for an intership to work on software components for transferring DNA sequence annotations from reference genomes to newly sequenced strains.

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Journal

8.1.1.1. Member of the editorial boards

Pascal Durrens is a member of the editorial board of the journal ISRN Computational Biology. David Sherman is a member of the editorial board of the journal Computational and Mathematical Methods in Medicine.

8.1.1.2. Reviewer - Reviewing activities

Pascal Durrens was reviewer for the journal BMC Genomics.

8.1.2. Scientific expertise

Pascal Durrens is an expert in Genomics for the Fonds de la Recherche Scientifique-FNRS (FRS-FNRS), Belgium. David Sherman is an expert for INRA's "Microbial Ecosystems and Metaomics" program.

8.2. Teaching - Supervision - Juries

8.2.1. Supervision

PhD: Razanne Issa, *Analyse symbolique et inférence de modèles métaboliques*, Université de Bordeaux, July 10, 2015. Thesis director: David Sherman.

8.2.2. Juries

Mid-term PhD review: Julie Laniau, Université de Rennes, October 1, 2015. Thesis director: Anne Siegel. Examinator: David Sherman

8.3. Popularization

David Sherman participated in popularization activities based on Thymio-II mobile robots for education, coordinated by the Mobsya association and EPFL (Switzerland). He contributed code to the Aseba project for piloting Thymio-IIs from the Scratch programming language, assisted in teaching at the Flornoy elementary school, and organized a team in the R2T2 event (http://r2t2.org) on November 4, 2015.

9. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] H. CAMPBELL-SILLS, M. EL KHOURY, M. FAVIER, A. ROMANO, F. BIASIOLI, G. SPANO, D. J. SHERMAN, O. BOUCHEZ, E. COTON, M. COTON, S. OKADA, N. TANAKA, M. DOLS-LAFARGUE, P. M. LUCAS. *Phylogenomic Analysis of Oenococcus oeni Reveals Specific Domestication of Strains to Cider and Wines*, in "Genome Biology and Evolution", May 2015, vol. 7, n^{o} 6, pp. 1506-18 [*DOI :* 10.1093/GBE/EVV084], https://hal.inria.fr/hal-01202801

[2] N. LOIRA, A. ZHUKOVA, D. J. SHERMAN. *Pantograph: A template-based method for genome-scale metabolic model reconstruction*, in "Journal of Bioinformatics and Computational Biology", January 2015, vol. 10, 1550006 p. [*DOI :* 10.1142/S0219720015500067], https://hal.inria.fr/hal-01123733

[3] S. MARIETTE, F. WONG JUN TAI, G. ROCH, A. BARRÉ, A. CHAGUE, S. DECROOCQ, A. GROPPI, Y. LAIZET, P. LAMBERT, D. TRICON, M. NIKOLSKI, J.-M. AUDERGON, A. G. ABBOTT, V. DECROOCQ. *Genome-wide association links candidate genes to resistance to Plum Pox Virus in apricot (Prunus armeniaca)*, in "New Phytologist", September 2015, DOI: 10.1111/nph.13627 p. [*DOI :* 10.1111/NPH.13627], https://hal.archives-ouvertes.fr/hal-01198840

[4] S. POQUE, G. PAGNY, L. OUIBRAHIM, A. CHAGUE, J.-P. EYQUARD, M. CABALLERO, T. CANDRESSE, C. CARANTA, S. MARIETTE, V. DECROOCQ. *Allelic variation at the rpv1 locus controls partial resistance to Plum pox virus infection in Arabidopsis thaliana*, in "BMC Plant Biology", June 2015, vol. 15, 159 p. , https://hal.inria.fr/hal-01262384

[5] A. ZHUKOVA, D. J. SHERMAN. *Mimoza: web-based semantic zooming and navigation in metabolic networks*, in "BMC Systems Biology", February 2015, vol. 9, 10 p. [*DOI :* 10.1186/s12918-015-0151-5], https://hal.inria.fr/hal-01123715

Invited Conferences

[6] D. J. SHERMAN. *Diversity and Domestication in Oenological Yeasts*, in "Moscow Conference on Computational and Molecular Biology (MCCMB'15)", Moscow, Russia, July 2015, https://hal.inria.fr/hal-01212044

Conferences without Proceedings

[7] M. DAYDÉ, B. DEPARDON, A. FRANC, J.-F. GIBRAT, R. GUILLIER, Y. KARAMI, C. PÉREZ, F. SUTER, M. CHABBERT, B. TADDESE, S. THÉROND. *E-Biothon: an experimental platform for BioInformatics*, in "International Conference on Computer Science and Information Technologies", Yerevan, Armenia, September 2015, https://hal.inria.fr/hal-01207320

[8] W. DYRKA, P. DURRENS, M. PAOLETTI, S. J. SAUPE, D. J. SHERMAN. *Deciphering the language of fungal pathogen recognition receptors*, in "8th Symposium of the Polish Bioinformatics Society", Lublin, Poland, Polish Bioinformatics Society, September 2015, https://hal.inria.fr/hal-01203357

Other Publications

[9] N. DUBOIS PEYRARD, S. DE GIVRY, A. FRANC, S. ROBIN, R. SABBADIN, T. SCHIEX, M. VIGNES. *Exact and approximate inference in graphical models: variable elimination and beyond*, June 2015, working paper or preprint, http://arxiv.org/abs/1506.08544, https://hal.archives-ouvertes.fr/hal-01197655

[10] W. DYRKA, P. DURRENS, S. J. SAUPE, M. PAOLETTI, D. J. SHERMAN. *Deciphering the language of fungal pathogen recognition receptors*, July 2015, EMBO Young Scientists Forum 2015, Poster, https://hal.inria.fr/hal-01171745

References in notes

[11] R. ALUR. *SIGPLAN Notices*, in "Generating Embedded Software from Hierarchical Hybrid Models", 2003, vol. 38, n^{o} 7, pp. 171–82

[12] B. ARNOLD, R. CORBETT-DETIG, D. HARTL, K. BOMBLIES. *RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling*, in "Mol. Ecol.", 2013, vol. 22, n^{o} 11, pp. 3179–90

[13] R. ASSAR, A. V. LEISEWITZ, A. GARCIA, N. C. INESTROSA, M. A. MONTECINO, D. J. SHERMAN. *Reusing and composing models of cell fate regulation of human bone precursor cells.*, in "BioSystems", April 2012, vol. 108, n^{o} 1-3, pp. 63-72 [*DOI :* 10.1016/J.BIOSYSTEMS.2012.01.008], https://hal.inria.fr/hal-00681022

[14] R. ASSAR, M. A. MONTECINO, A. MAASS, D. J. SHERMAN. *Modeling acclimatization by hybrid systems: Condition changes alter biological system behavior models*, in "BioSystems", June 2014, vol. 121, pp. 43-53 [*DOI :* 10.1016/J.BIOSYSTEMS.2014.05.007], https://hal.inria.fr/hal-01002987

[15] R. ASSAR, D. J. SHERMAN. *Implementing biological hybrid systems: Allowing composition and avoiding stiffness*, in "Applied Mathematics and Computation", August 2013, vol. 223, pp. 167–79, https://hal.inria.fr/hal-00853997

[16] R. ASSAR, F. VARGAS, D. J. SHERMAN. *Reconciling competing models: a case study of wine fermentation kinetics*, in "Algebraic and Numeric Biology 2010", Hagenberg, Austria, K. HORIMOTO, M. NAKATSUI, N. POPOV (editors), Springer, July 2010, vol. 6479, pp. 68–83 [*DOI :* 10.1007/978-3-642-28067-2_6], https://hal.inria.fr/inria-00541215

[17] C. COMBES. *Parasitism: The Ecology and Evolution of Intimate Interactions*, University of Chicago Press, 2001

[18] P. GAYRAL, J. MELO-FERREIRA, S. GLEMIN. *Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap*, in "PLoS Genetic", 2013, vol. 9, n^{o} 4, e1003457

[19] R. ISSA. *symbolic analysis and inference of metabolic models*, Université de Bordeaux, July 2015, https://tel.archives-ouvertes.fr/tel-01216599

[20] M. LYNCH. *Estimation of Nucleotide Diversity, Disequilibrium Coefficients, and Mutation Rates from High-Coverage Genome-Sequencing Projects*, in "Mol. Biol. Evol.", 2008, vol. 25, n^{o} 11, pp. 2409–19

[21] R. E. RICKLEFS. *A comprehensive framework for global patterns in biodiversity*, in "Ecology Letters", 2004, vol. 7, n^{o} 1, pp. 1–15, http://dx.doi.org/10.1046/j.1461-0248.2003.00554.x

[22] S. T. ROWEIS, Z. GHAHRAMANI. *A unifying review of linear Gaussian Models*, in "Neural Computation", 1999, vol. 11, n^{o} 2, pp. 305–45

[23] L. K. SAUL, S. T. ROWEIS. *Think globally, fit locally: unsupervised learning of low dimensional manifolds*, in "Journal of Machine Learning Research", 2003, vol. 4, pp. 119–55

[24] D. J. SHERMAN, T. MARTIN, M. NIKOLSKI, C. CAYLA, J.-L. SOUCIET, P. DURRENS. *Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes*, in "Nucleic Acids Research (NAR)", 2009, pp. D550-4 [*DOI :* 10.1093/NAR/GKN859], http://hal.inria.fr/inria-00341578/en/

[25] D. W. THOMPSON. *On Growth and Form*, Cambridge University Press, 1917

[26] A. ZHUKOVA, D. J. SHERMAN. *Knowledge-based generalization of metabolic models*, in "Journal of Computational Biology", 2014, vol. 21, n^{o} 7, pp. 534-47 [*DOI :* 10.1089/CMB.2013.0143], https://hal.inria.fr/hal-00925881