Activity Report 2017

Team PLEIADE

from patterns to models in computational biodiversity and biotechnology

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:

A3.1. - Data
A3.2. - Knowledge
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4. - Machine learning and statistics
A6.2.8. - Computational geometry and meshes

Other Research Topics and Application Domains:

B1.1.6. - Genomics
B1.1.9. - Bioinformatics
B1.1.11. - Systems biology
B1.1.12. - Synthetic biology
B3. - Environment and planet

1. Personnel

Research Scientists

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

Post-Doctoral Fellow

Pierre Blanchard [INRA, from Mar 2017 until Sep 2017]

Technical staff

Redouane Bouchouirbat [INRA, until Feb 2017]
Philippe Chaumeil [INRA]
Jean-Marc Frigerio [INRA]
Franck Salin [INRA]
Louise Amelie Schmitt [Inria]

Interns

Vassili Bernat [Inria, Nov 2017]
Charlotte Courtand [Inria, from Dec 2017]
Marta Iollo [Inria, from Dec 2017]
Mercia Ngoma Komb [Inria, from Jun 2017 until Aug 2017]
Dorian Rocuet [Inria, Nov 2017]

Administrative Assistants

Cecile Boutros [Inria]
Nathalie Robin [Inria]

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 [28] 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 capitalizing on this wealth of data are still not adapted to high throughput data production. A better connection between high-throughput data production and tool evolution is highly 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 (Figure 1). 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

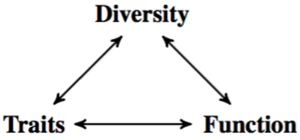

Figure 1. Diversity informs both the study of traits, and the study of biological functions

PLEIADE links recognition of patterns, classes, and interactions with applications 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.

Our methodology (Figure 2) is designed pragmatically to advance the state of the art in applications from biodiversity and biotechnology: molecular based systematics and community ecology, annotation and modeling for biotechnology.

3. Research Program

3.1. A Geometric View of Diversity

Diversity may be studied as a set of dissimilarities between objects. The underlying mathematical construction is the notion of distance. Knowing a set of objects, it is possible, after computation of pairwise distances, or sometimes dissimilarities, to build a Euclidean 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. The human eye is often far better than an algorithm at recognizing a pattern or shape. One objective of our project is to narrow the gap between the story

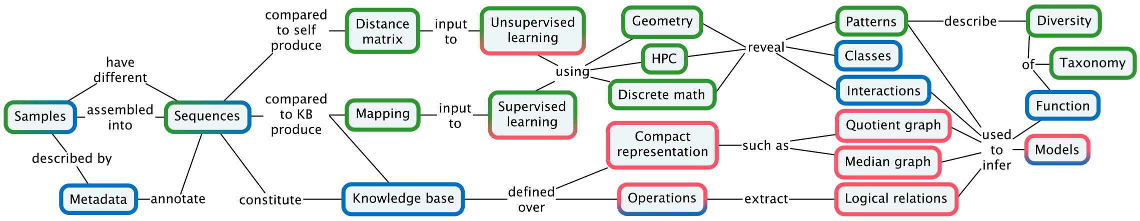

Figure 2. PLEIADE *is a pluridisciplinary team. Each application in biodiversity and biotechnology follows a path calling on methods from biology (blue), mathematics (green), and computer science (red).*

that a human eye can tell, and that an algorithm can tell. Several directions will be explored. First, this requires mastering classical tools in dimension reduction, mainly algebraic tools (PCA, NGS, Isomap, eigenmaps, etc ...). Second, neighborhoods in point clouds naturally lead to graphs describing the neighborhood networks. There is a natural link between modular structures in distance arrays and communities on graphs. Third, points (representing, say, DNA sequences) are samples of diversity. Dimension reduction may show that they live on a given manifold. This leads to geometry (differential or Riemannian geometry). It is expected that some properties of the manifold can tell something of the constraints on the space where measured individuals live. The connection between Riemannian geometry and graphs, where weighted graphs are seen as mesh embedded in a manifold, is currently an active field of research [27], [26]. See as well [29] for a link between geometric structure, linear and nonlinear dimensionality reduction.

Biodiversity and high-performance computing: Most methods and tools for characterizing diversity have been designed for datasets that can be analyzed on a laptop, but NGS datasets produced for metabarcoding are far too large. Data analysis algorithms and tools must be revisited and scaled up. We will mobilize both distributed algorithms like the Arnoldi method and new algorithms, like random projection or column selection methods, to build point clouds in Euclidean spaces from massive data sets, and thus to overcome the cubic complexity of computation of eigenvectors and eigenvalues of very large dense matrices. We will also link distance geometry [23] with convex optimization procedures through matrix completion [19], [20].

Intercalibration: There is a considerable difference between supervised and unsupervised clustering: in supervised clustering, the result for an item i is independent from the result for an item $j \neq i$, whereas in unsupervised clustering, the result for an item i (e.g. the cluster it belongs to, and its composition) depends on nearby items $j \neq i$. Which means that the result may change if some items are added to or subtracted from the sample. This raises the more global problem of how to merge two studies to yield a more comprehensive view of biodiversity?

3.2. Knowledge Management for Biology

The heterogenous data generated in computational molecular biology and ecology are distinguished not only by their volume, but by the richness of the many levels of interpretation that biologists create. The same nucleic acid sequence can be seen as a molecule with a structure, a sequence of base pairs, a collection of genes, an allele, or a molecular fingerprint. To extract the maximum benefit from this treasure trove we must organize the knowledge in ways that facilitate extraction, analysis, and inference. Our focus has been on the efficient representation of relations between biological objects and operations on those representations, in particular heuristic analyses and logical inference.

PLEIADE will develop 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 distance geometry will 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.

Since a goal of PLEIADE is to integrate diversity throughout the analysis process, it is necessary to incorporate **diversity as a form of knowledge** that can be stored in a knowledge base. Diversity can be represented using various compact representations, such as trees and quotient graphs storing nested sets of relations. Extracting structured representations and logical relations from integrated knowledge bases (Figure 2) will require domain-specific query methods that can express forms of diversity.

3.3. 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 [17]. 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* [14] 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 [2], [18] and medicine [16]. 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 [21]. About one decade ago, some works [25] 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 [24]. 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 [22]. Several approaches were proposed to correct for paralogy, either by working directly on the sequences issued from mapped reads [22] 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 [15]. 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.
NEWS OF THE YEAR: Magus is now available as a Docker image, and can be integrated with other containerized services using Pleiade's Alcyone system.

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

5.2. Pantograph

KEYWORDS: Systems Biology - Bioinformatics - Genomics - Gene regulatory networks
SCIENTIFIC DESCRIPTION: Pantograph requires a template model in SMBL format, where every reaction is annotated with a gene association that describes its gene-protein-reaction dependencies as a boolean formula over the genes of the organism.

Pantograph uses a consensus procedure to infer relationships between metabolic models, based on several sources of orthology between genomes. These inter-genome relations are used to rewrite the gene associations. Every successful rewrite is used as evidence that the corresponding reaction should be present in the inferred model.

The resulting models can be validated with respect to phenotypic information obtained from experimental results.
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 a reconstructions process in 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.

NEWS OF THE YEAR: Work is in progress to integrate Razanne Issa's Ab-Pantograph modules into Pantograph. Ab-Pantograph uses abductive logic to invert the inference process: a reaction explains the presence of the genes in its gene-protein-reaction formula, rather than the genes justify the reaction. Ab-Pantograph is driven by the goal of explaining all of the genes in the target organism.

- Participants: Anna Zhukova, David James Sherman, Nicolas Loira and Pascal Durrens
- Partner: University of Chile
- Contact: Nicolas Loira
- Publication: Pantograph: A template-based method for genome-scale metabolic model reconstruction
- URL: http://pathtastic.gforge.inria.fr/

5.3. 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. Mimoza generalizes genome-scale metabolic models, by factoring equivalent reactions and metabolites while preserving reaction consistency. 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 resulting map can be explored on-line, or downloaded in a COMBINE archive. 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.
NEWS OF THE YEAR: Mimoza is now available as a Docker image, and can be integrated with other containerized services using Pleiade's Alcyone system.

- Participants: Anna Zhukova and David James Sherman
- Contact: David James Sherman
- Publications: Knowledge-based generalization of metabolic models - Knowledge-based zooming for metabolic models - Knowledge-based generalization of metabolic networks: a practical study
- URL: http://mimoza.bordeaux.inria.fr/

5.4. Declic

FUNCTIONAL DESCRIPTION: Declic is a Python library that provides several tools for data analysis in the domains of multivariate data analysis, machine learning, and graph based methods. It can be used to study in-depth the accuracy of the dictionary between molecular based and morphological based taxonomy.

Declic includes an interpreter for a Domain Specific Language (DSL) to make its Python library easy to use for scientists familiar with environments such as R.

- Partner: INRA
- Contact: Alain Franc

5.5. Diagno-Syst

diagno-syst: a tool for accurate inventories in metabarcoding
KEYWORDS: Biodiversity - Clustering - Ecology

FUNCTIONAL DESCRIPTION: Diagno-syst builds accurate inventories for biodiversity. It performs supervised clustering of reads obtained from a next-generation sequencing experiment, mapping onto an existing reference database, and assignment of taxonomic annotations.

- Participants: Alain Franc, Jean-Marc Frigerio, Philippe Chaumeil and Franck Salin
- Partner: INRA
- Contact: Alain Franc
- Publication: diagno-syst: a tool for accurate inventories in metabarcoding

5.6. Alcyone

Alcyone instantiates bioinformatics environments from specifications committed to a Git repository
KEYWORDS: Docker - Orchestration - Bioinformatics - Microservices - Versioning
SCIENTIFIC DESCRIPTION: Alcyone conceives the user's computing environment as a microservices architecture, where each bioinformatics tool in the specification is a separate containerized Docker service. Alcyone builds a master container for the specified environment that is responsible for building, updating, deploying and stopping these containers, as well as recording and sharing the environment in a Git repository. The master container can be manipulated using a command-line interface.
FUNCTIONAL DESCRIPTION: Alcyone defines a file structure for the specifying bioinformatics analysis environments, including tool choice, interoperability, and sources of raw data. These specifications are recorded in a Git repository. Alcyone compiles a specification into a master Docker container that deploys and orchestrates containers for each of the component tools. Alcyone can restore any version of an environment recorded in the Git repository.
NEWS OF THE YEAR: Alcyone was designed and implemented this year.

- Participants: Louise-Amelie Schmitt and David Sherman
- Contact: David Sherman
- URL: https://team.inria.fr/pleiade/alcyone/

6. New Results

6.1. Alcyone system for repeatable e-science

One of PLEIADE's goals is to assist scientific users in deploying analysis software in their desktop environments. Increasingly, this is not a question of installing software packages locally, but of building bespoke environments that comprise many cooperating software tools. A typical example is a local Galaxy instance, communicating with a project-specific database that is shared with visualization and analysis tools, and cooperating with an electronic notebook such as Jupyter. In order to foster repeatable science, the configuration of each such environment should be reliably recorded, in a way that allows it to be redeployed in the future or shared with a colleague.

PLEIADE's **Alcyone** system provides a mechanism for specifying and deploying software environments for scientific users in bioinformatics and biodiversity. Alcyone offers three facilities:

1. A *specification* using configuration-by-convention style, combining specification files in YAML format and raw data files.
2. A *collection of Dockerized services* that can be chosen in the specification.
3. A *deployment system* that compiles the specification into a master container image, which orchestrates the deployment and management of the service containers.

The user's environment is fully specified in files that can be archived and shared, allowing future reuse. The use of Docker containers guarantees that future deployments run exactly as before, since the precise versions of the service containers are recorded.

Furthermore, Alcyone specifications are files, that can be managed by the Git source code control system. Different versions of the environment, including different analysis pipelines and intermediate results, are stored in the Git history and any version can be resurrected and deployed. Git branches can also be used to share configurations between users in the same lab.

Alcyone is being tested internally by PLEIADE and is undergoing intense development. Existing service containers are PLEIADE's Magus knowledge base, Magecal gene prediction pipeline, and Mimoza metabolic network explorer ; as well as third-party tools Galaxy, Gbrowse, and Jbrowse.

6.2. Clavispora lusitaniae

Clavispora lusitaniae, an environmental saprophytic yeast belonging to the CTG clade of *Candida* and a teleomorph of *Candida lusitaniae*, is an environmentally ubiquitous ascomycetous yeast with no known specific ecological niche. It can be isolated from different substrates, such as soils, waters, plants, and gastrointestinal tracts of many animals including birds, mammals, and humans. In immunocompromised hosts, *C. lusitaniae* can be pathogenic and is responsible for about 1% of invasive candidiasis, particularly in pediatric and hematology-oncology patients.

The Laboratoire de Microbiologie Fondamentale et Pathogénicité UMR-CNRS 5234 and PLEIADE sequenced and annotated the genome of *C. lusitaniae* type strain CBS 6936, and analyzed it in comparison with the strains ATCC 42720, isolated from the blood of a patient with myeloid leukemia, and MTCC 1001, a self-fertile strain isolated from citrus. In spite of a conserved genome structure, the genomes have undergone significant divergence. In particular the SNP density of 1 SNP per 90 bp is twice the level observed between strains SC5314 and WO-1 of *Candida albicans*, which are members of different subgroups within the species and qualified as having diverged relatively recently.

This work contributes to PLEIADE's long-term goal of developing understanding how diversity measured at the genome level can be made to correspond with observed functional diversity.

6.3. Introgressions as a source of diversity

Several prominent mechanisms of genomic evolution have been described for the yeasts, among them inter-specific hybridization, reticulated evolution, aneuploidization, recent or ancient poly-ploidization events, large chromosomal duplication or more limited gene duplication, and horizontal transfer. These mechanisms are usually so closely intertwined that it is difficult to determine which ones are causes or consequences. Regardless of mechanisms the result has been a drastic reshaping of yeasts genome along evolution. Understanding these mechanisms is important, not only for strain construction in biotechnology, but also more fundamentally for insight into the causes and effects of genome reshaping on much shorter time scales.

Introgression, the transfer of large or more limited genetic information from one species to another, is an evolutionary mechanism of particular interest in industrial applications such as wine making where large vat cultures are used. Introgression results in mosaic genomes, and can be the result of interspecific hybridization fol- lowed by the extensive loss of one parental genome, either through repeated backcross with one parental species or through missegregation of the hybrid at meiosis.

In collaboration with the Institut des Science de la Vigne et du Vin and Bordeaux Sciences Agro, PLEIADE developed tools to rapidly assess the presence of introgressed regions in a large population of *Saccharomyces uvarum* isolates (104 strains), focusing on Holarctic isolates from natural, cider and wine environments since introgressed regions are absent in Southern hemisphere isolates. The overall number of introgressed regions is significantly higher in cider-associated strains compared to wild strains, and is higher in wine isolates. However, only a subset of the introgressed regions were found to be overrepresented in anthropic activities and their number and quality varied between cider- and wine-making processes.

Paradoxically, the low Holarctic genetic diversity observed in [1] contrasts with the relative high phenotypic diversity found for technological traits. This contradiction suggests that interspecific introgressions found among Holarctic *S. uvarum* strains could be the most important source of genetic, and by extention of phenotypic, diversity.

6.4. New results Biodiversity

The activity of PLEIADE in computational biodiversity has consisted mainly in reinforcing a cooperation with actors in High Performance Computing, namely Inria team Hiepacs, for method developments in metabarcoding. Metabarcoding is a supervised or unsupervised statistical learning method, to build taxonomic inventories from so called environmental samples, i.e. sets of short reads of a same marker for a whole community or guild. Most of tools used therefore still rely on some classical ones shaped in Multivariate Data Analysis. Those tools are indeed well known, but still are often behind the scene in current developments in Machine Learning (like kernel PCA, Support Vector Machines, etc. ...). Most of them, if not all, are based on Singular Value Decomposition of a matrix. If p features are observed on n items, the size of the matrix is $n \times p$. The complexity of such algorithms is in $O(p^3)$. The recent development of NGS data has had as a consequence to multiply by a factor $10^2/10^3$ the size of data sets. This leads to a factor $10^6/10^9$ of required computation time. Reaching such a goal is beyond resources currently offered by parallelization. Hence, a new approach has been selected, by using other methods. Indeed, it has been known for some years now that concentration of measure phenomena (a sort of extension of law of large numbers) leads to a blessing of dimensionality, i.e. some randomized methods are available as heuristics to make some matrix computations efficiently and accurately. This is the case for running SVD. Therefore, a cooperation has been set up between HiePacs and PLEIADE through Pierre Blanchard (a former Hiepacs PhD student who has held a post-doc position during 7 months in PLEIADE) to implement those methods in the framework of metabarcoding. Former work in PLEIADE had led (with a DARI project 2014-2016) to the production of many high-dimensional pairwise distance matrices of DNA environmental samples (amplicon based metabarcoding). Classical Multidimensional Scaling of some of those matrices has been programmed in C++, with dedicated libraries in domain of so called random projection, or column selection (`fmr` library). This has permitted to build a point cloud of an environmental sample of 1.2×10^5 reads, and see its "shape", with eyes, from projections on first axis, and build a low dimensional approximation of it. The outcome is twofolds: (i) build a point cloud attached to an environmental sample, for further ecological studies and (ii) delivery of a scientific library in High Performance Computing for randomized matrix computations. These research lines will be carried on in 2018, and the cooperation extended to mésocentre GRICAD in Grenoble for HPC and C++ code development.

PLEIADE has carried on statistical learning methods, both supervised and unsupervised in metabarcoding. A cooperation with IMBE at Marseille has permitted to associate MDS as developed above with graph based methods (building connected components of a graph built from pairwise distance matrices after thresholding), and test these methods for unsupervised statistical learning (OTU building) of data sets from an ongoing PhD in Marseille Bay. Cooperation with Institut Pasteur at Cayenne has lead to a joint publication [12] for a proof of concept of an inventory by metagenomics of viromes of bats in French Guiana, with two objectives: (i) detect as soon as possible some strains which could potentially be transmitted to man and (ii) develop a viral ecology by studying further how environmental factors and nature of the host drive the virome composition.

Meanwhile, PLEIADE has carried on cooperation with SLU Universty at Uppsala especially on metabarcoding of diatom communities in rivers and lakes in Sweden (co-direction of a PhD student located at Uppsala in SLU) , and first steps in biogeography of diatoms in Fennoscandia (cooperation with a PostDoc in SLU).

7. Partnerships and Cooperations

7.1. Regional Initiatives

7.1.1. COTE – Continental to Coastal Ecosystems

The Labex cluster of excellence COTE (Continental To coastal Ecosystems: evolution, adaptability and governance) develops tools to understand and predict ecosystem responses to human-induced changes as well as methods of adaptative management and governance to ensure their sustainability. The LabEx includes nine laboratories of the University of Bordeaux and major national research institutes involved in research on terrestrial and aquatic ecosystems (INRA, CNRS, IFREMER and IRSTEA). PLEIADE is a partner in one project funded by COTE:

- *Aerobarcoding: détection de pollens allergenisants*. 2017-18.

7.2. National Initiatives

7.2.1. Biocontamination in aircraft reservoirs

ANTICOR is an industrial-academic research and development working group coordinated by Dassault Aviation, investigating the causes of microbial contamination in aircraft reservoirs and aimed at developing mitigating procedures and equipment. Previous results have shown that this contamination forms biofilms at the fuel-water interface and is comprised of complex communities of hundreds of bacterial and fungal species. PLEIADE is particularly interested in measuring and modeling these communities, especially as concerns understanding how they change based on environmental conditions and on reservoir geometry.

This working group continues work started in CAER – Alternative Fuels for Aeronautics, 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.

7.2.2. Agence Française pour la Biodiversité

The AFB is a public law agency of the French Ministry of Ecology that supports public policy in the domains of knowledge, preservation, management, and restoration of biodiversity in terrestrial, aquatic, and marine environments. PLEIADE is a partner in two AFB projects developed with the former ONEMA:

- *Methods for metabarcoding*. 2017-18.
- *Molecular diagnosis of freshwater quality*. 2014-present.

7.2.3. Inria Projet Lab in silico Algae

In 2017 PLEIADE joined the IPL "In silico Algae" coordinated by Olivier Bernard. The IPL addresses challenges in modeling and optimizing microalgae growth for industrial applications. PLEIADE worked this year on comparative genomic analysis of genes implicated in lipid production by the picoalgae *Ostreococcus tauri*, in collaboration with Florence Corellou of the CNRS UMR 5200 (Laboratoire de Biogénèse Membranaire). The goal of this work is the production of long-chain polyunsaturated fatty acids, developed as nutritional additives. Mercia Ngoma Komb's two-month internship in PLEIADE contributed to this work.

7.3. European Initiatives

7.3.1. Collaborations in European Programs, Except FP7 & H2020

Alain Franc has been appointed co-chair of Working Group 4 (Data Analysis and Storage) of COST DNAqua.net [1], at the Sarajevo meeting in Fall 2017, with the main task of developing contact with HPC and metabarcoding for serving the whole community. The goal of DNAqua-Net is to nucleate a group of researchers across disciplines with the task to identify gold-standard genomic tools and novel eco-genomic indices and metrics for routine application for biodiversity assessments and biomonitoring of European water bodies.

[1] http://dnaqua.net/

7.4. International Initiatives

7.4.1. CEBA – Center for the study of biodiversity in Amazonia

The Laboratoire of excellence CEBA promotes innovation in research on tropical biodiversity. It brings together a network of internationally-recognized French research teams, contributes to university education, and encourages scientific collaboration with South American countries. PLEIADE participates in three current international projects funded by CEBA:

- *MicroBIOMES: Microbial Biodiversities*. 2017-19.
- *Neutrophyl: Inferring the drivers of Neotropical diversification*. 2017-19.
- *Phyloguianas: Biogeography and pace of diversification in the Guiana Shield*. 2015-present

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events Organisation

8.1.1.1. General Chair, Scientific Chair

Alain Franc organized on September 25-29, 2017 an ANF (Action Nationale de Formation) of CNRS on “Data analysis for massive data”. There were about 20 participants, from Astronomy to Bioinformatics, over Fluid Mechanics.

8.1.2. Journal

8.1.2.1. Member of Editorial Boards

Alain Franc is member of the editorial board of BMC Evolutionary Biology.

Pascal Durrens is a member of the editorial board of the journal ISRN Computational Biology.

8.2. Teaching - Supervision - Juries

8.2.1. Juries

David Sherman was a thesis reviewer for Julie Laniau (University of Nantes) and member of her defense jury, October 23, 2017. The title of the dissertation was “Structure de re´seaux biologiques : roˆle des nœuds internes *vis-a`-vis* de la production de compose´s” and concerned the methodological analysis of metabolite essentiality in metabolic modeling, applied to algae.

Alain Franc was a thesis reviewer for Cyril Noël (University of Pau and the Pays d’Adour) and member of his defense jury, 2017. The title of the dissertation was “Réseaux microbiens de dégradation des hydrocarbures aux interfaces oxie/anoxie des sédiments marins côtiers” and concerned metabarcoding, metagenomics and functional metagenomics of some Bactera and Archea.

Alain Franc was president of the jury for PhD defense of Pierre Blanchard (University of Bordeaux and Inria project-team HiePACS) on February 16, 2017. The title of the dissertation was “Fast hierarchical algorithms for the low-rank approximation of matrices with applications to materials physics, geostatistics and data analysis”

8.3. Popularization

David Sherman of PLEIADE coached two teams in Thymio R2T2 Challenges [2], organized by the Mobsya association and the EPFL in Spring and in Summer 2017. An R2T2 challenge brings together 16 teams (for the Mars mission, 4 teams for the Lunar mission) of children who must cooperate to remotely program Thymio robots. The Lunar mission in July was a public demonstration during the Scratch 2017 conference in Bordeaux.

[2] Remote Rescue Thymio II https://www.thymio.org/en:thymio-r2t2

David Sherman contributes open-source software development to the Aseba platform for educational robotics [3], deployed in Thymio II robots used by children as well as in the simulator used by Class'Code [4] to train teachers.

9. Bibliography

Major publications by the team in recent years

[1] P. Almeida, C. Gonçalves, S. Teixeira, D. Libkind, M. Bontrager, I. Masneu-Pomarède, W. Albertin, P. Durrens, D. J. Sherman, P. Marullo, C. Todd Hittinger, P. Gonçalves, J. P. Sampaio. *A Gondwanan imprint on global diversity and domestication of wine and cider yeast Saccharomyces uvarum*, in "Nature Communications", 2014, vol. 5, 4044 p. [*DOI :* 10.1038/ncomms5044], https://hal.inria.fr/hal-01002466

[2] 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

[3] A. B. Canelas, N. Harrison, A. Fazio, J. Zhang, J.-P. Pitkänen, J. Van Den Brink, B. M. Bakker, L. Bogner, J. Bouwman, J. I. Castrillo, A. Cankorur, P. Chumnanpuen, P. Daran-Lapujade, D. Dikicioglu, K. Van Eunen, J. C. Ewald, J. J. Heijnen, B. Kirdar, I. Mattila, F. I. C. Mensonides, A. Niebel, M. Penttilä, J. T. Pronk, M. Reuss, L. Salusjärvi, U. Sauer, D. J. Sherman, M. Siemann-Herzberg, H. Westerhoff, J. De Winde, D. Petranovic, S. G. Oliver, C. T. Workman, N. Zamboni, J. Nielsen. *Integrated multilaboratory systems biology reveals differences in protein metabolism between two reference y east strains*, in "Nature Communications", December 2010, vol. 1, n^{o} 9, 145 p. [*DOI :* 10.1038/ncomms1150], https://hal.inria.fr/inria-00562005

[4] L. Kermarrec, A. Franc, F. Rimet, P. Chaumeil, J.-M. Frigerio, J.-F. Humbert, A. Bouchez. *A next-generation sequencing approach to river biomonitoring using benthic diatoms*, in "Freshwater Science", 2014, vol. 33, n^{o} 1, pp. 349-363, http://www.jstor.org/stable/10.1086/675079

[5] 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", 2009, vol. 37, pp. D550-D554 [*DOI :* 10.1093/nar/gkn859], https://hal.inria.fr/inria-00341578

[6] J.-L. Souciet, B. Dujon, C. Gaillardin, M. Johnston, P. V. Baret, P. Cliften, D. J. Sherman, J. Weissenbach, E. Westhof, P. Wincker, C. Jubin, J. Poulain, V. Barbe, B. Ségurens, F. Artiguenave, V. Anthouard, B. Vacherie, M.-E. Val, R. S. Fulton, P. Minx, R. Wilson, P. Durrens, G. Jean, C. Marck, T. Martin, M. Nikolski, T. Rolland, M.-L. Seret, S. Casaregola, L. Despons, C. Fairhead, G. Fischer, I. Lafontaine, V. Leh, M. Lemaire, J. De Montigny, C. Neuveglise, A. Thierry, I. Blanc-Lenfle, C. Bleykasten, J. Diffels, E. Fritsch, L. Frangeul, A. Goeffon, N. Jauniaux, R. Kachouri-Lafond, C. Payen, S. Potier, L. Pribylova, C. Ozanne, G.-F. Richard, C. Sacerdot, M.-L. Straub, E. Talla. *Comparative genomics of protoploid Saccharomycetaceae*, in "Genome Research", 2009, vol. 19, pp. 1696-1709 [*DOI :* 10.1101/gr.091546.109], https://hal.inria.fr/inria-00407511

[3] http://aseba.io/

[4] https://pixees.fr/classcode-la-formation-associee-a-pixees/

Publications of the year

Articles in International Peer-Reviewed Journals

[7] W. Albertin, M. Chernova, P. Durrens, E. Guichoux, D. J. Sherman, I. Masneuf-Pomarede, P. Marullo. *Many interspecific chromosomal introgressions are highly prevalent in Holarctic Saccharomyces uvarum strains found in human-related fermentations*, in "Yeast", August 2017, pp. 1-20 [*DOI :* 10.1002/yea.3248], https://hal.inria.fr/hal-01585808

[8] A. Bretagnolle, A. Franc. *Emergence of an integrated city-system in France (XVIIth–XIXth centuries): Evidence from toolset in graph theory*, in "Historical Methods: A Journal of Quantitative and Interdisciplinary History", 2017, vol. 50, n^o 1, pp. 49-65 [*DOI :* 10.1080/01615440.2016.1237915], https://hal.archives-ouvertes.fr/hal-01603847

[9] J. Brusini, M. L. Wayne, A. Franc, C. Robin. *The impact of parasitism on resource allocation in a fungal host: the case of Cryphonectria parasitica and its mycovirus, Cryphonectria Hypovirus 1*, in "Ecology and Evolution", 2017, vol. 7, n^o 15, pp. 5967–5976 [*DOI :* 10.1002/ece3.3143], https://hal.archives-ouvertes.fr/hal-01608310

[10] H. Campbell-Sills, M. El Khoury, M. Gammacurta, C. Miot-Sertier, L. Dutilh, J. Vestner, V. Capozzi, D. J. Sherman, C. Hubert, O. Claisse, G. Spano, G. De Revel, P. Lucas. *Two different Oenococcus oeni lineages are associated to either red or white wines in Burgundy: genomics and metabolomics insights*, in "OENO One", September 2017, vol. 51, n^o 3, pp. 309 - 322 [*DOI :* 10.20870/oeno-one.2017.51.4.1861], https://hal.inria.fr/hal-01666981

[11] P. Durrens, C. Klopp, N. Biteau, V. Fitton-Ouhabi, K. Dementhon, I. Accoceberry, D. J. Sherman, T. Noël. *Genome Sequence of the Yeast Clavispora lusitaniae Type Strain CBS 6936*, in "Genome Announcements", August 2017, vol. 5, n^o 31 [*DOI :* 10.1128/genomeA.00724-17], https://hal.inria.fr/hal-01583349

[12] V. Lacoste, A. Salmier, S. Tirera, A. Franc, E. Darcissac, D. Donato, C. Bouchier, A. Lavergne, B. De Thoisy, N. Forrester. *Virome analysis of two sympatric bat species (Desmodus rotundus and Molossus molossus) in French Guiana*, in "PLoS ONE", November 2017, vol. 12, n^o 11 [*DOI :* 10.1371/journal.pone.0186943], https://hal-riip.archives-ouvertes.fr/pasteur-01633803

Research Reports

[13] P. Blanchard, P. P. Chaumeil, J.-M. Frigerio, F. Rimet, F. Salin, S. Thérond, O. Coulaud, A. Franc. *A geometric view of Biodiversity: scaling to metagenomics*, Inria ; INRA, January 2018, n^o RR-9144, pp. 1-16, https://hal.inria.fr/hal-01685711

References in notes

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

[15] 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

[16] 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

[17] 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

[18] 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

[19] M. BAKONYI, C. R. JOHNSON. *The Euclidean Distance Matrix Completion Problem*, in "SIAM J. Matrix Anal. App.", 1995, vol. 16, n^{o} 2, pp. 646-654

[20] E. J. CANDÈS, B. RECHT. *Exact Matrix Completion via Convex Optimization*, in "Found. Comput. Math.", 2009, vol. 9, pp. 717-772

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

[22] 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

[23] L. LIBERTI, C. LAVOR, N. MACULAN, A. MUCHERINO. *Euclidean Distance Geometry and Applications*, in "SIAM review", 2014, vol. 56(1), pp. 3-69

[24] 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

[25] 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

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

[27] 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

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

[29] J. WANG. *Geometric structure of high-dimensional data and dimensionality reduction*, Springer & Higher Education Press, 2012