Activity Report 2016

Team PLEIADE

From patterns to models in computational biodiversity and biotechnology

RESEARCH CENTER
Bordeaux - Sud-Ouest

THEME
Computational Biology

Table of contents

Team PLEIADE

Creation of the Team: 2015 January 01

PLEIADE *is located in two sites: Inria Bordeaux Sud-Ouest on the Talence campus of the Université de Bordeaux and INRA Pierroton.*

Keywords:

Computer Science and Digital Science:

3.1. - Data
3.2. - Knowledge
3.3.2. - Data mining
3.3.3. - Big data analysis
3.4. - Machine learning and statistics
6.1.4. - Multiscale modeling
6.2.8. - Computational geometry and meshes

Other Research Topics and Application Domains:

1.1.6. - Genomics
1.1.9. - Bioinformatics
1.1.11. - Systems biology
1.1.12. - Synthetic biology
1.2. - Ecology
3. - Environment and planet

1. Members

Research Scientists

David Sherman [Team leader, Inria, Senior Researcher, HDR]
Pascal Durrens [CNRS, Researcher, HDR]
Alain Franc [INRA, Senior Researcher]
Stephanie Mariette [INRA, Researcher, until Aug 2016]

Engineers

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

Administrative Assistants

Cecile Boutros [Inria]
Anne-Laure Gautier [Inria, until Feb 2016]

Others

Arthur Demene [Univ. Bordeaux, from Mar 2016]
Christian Dutech [INRA, Researcher, from Feb 2016]
Razanne Issa [Univ. Bordeaux]
Adrien Lopez [Min. de l'Education Nationale, Feb 2016]
Remi Pellerin [Inria, from Jun 2016 until Jul 2016]
Anna Zhukova [Institut Pasteur, until Jun 2016]

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 [35] 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.

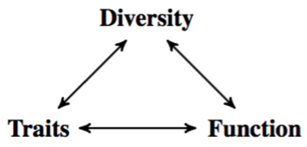

Figure 1. Diversity informs both the study of traits, and the study of biological functions. The double challenge is to measure these links quickly and precisely with pattern recognition, and to explore the relations between diversity in traits and diversity in function through modeling.

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 [33], [32].

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 [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 [22]. About one decade ago, some works [30] 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 [28]. 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 [25]. Several approaches were proposed to correct for paralogy, either by working directly on the sequences issued from mapped reads [25] 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. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Biotechnology

In collaboration with MIAT INRA and UMR 5234 CNRS/Université de Bordeaux, PLEIADE assembled and analyzed *Clavispora lusitaniae*, an ubiquist environmental ascomycetous yeast that can be pathogenic and is responsible for invasive candidiases in pediatric and onco-haematology patients [24].

In collaboration with UMR 5200 CNRS/Université de Bordeaux, PLEIADE assembled and analyzed transcriptomes from three tissues of the oil palm tree *Elaeis guineensis* Jacq., whose mesocarp contains oil up to 90% of its dry weight. Our goal is to increase, by synthetic biology approaches, the yield in oil for crops grown in Europe. The yield and the composition of oil measured from wild-type palm tree specimens varies dramatically, indicating a high level of bio-diversity.

5.1.2. Biodiversity

PLEIADE and the HIEPACS team developed connections between random projection methods and multidimensional scaling, in order to compute eigenvectors and eigenvelaues in space of reasonable dimension. The method for MDS developed by Pierre Blanchard has proved to be surprisingly efficient and precise. It was presented at PASC 2016 Lausanne. This work improves the analysis of microbial communities, where the shape of the point cloud built from pairwise distances between a large set of NGS reads is used to describe the diversity of the community.

6. New Software and Platforms

6.1. Magus

KEYWORDS: Bioinformatics - Genomic sequence - Knowledge database
FUNCTIONAL DESCRIPTION

Comparative genomics requires efficient and scalable tools for managing knowledge about genomes, genes, and the high-dimensional relations between them.

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.

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.

- Participants: David Sherman, Pascal Durrens
- Partners: CNRS - INRA - Université de Bordeaux
- Contact: David James Sherman
- URL: http://magus.gforge.inria.fr

6.2. 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/

6.3. 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/

6.4. 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/

6.5. Declic

Metabarcoding relies on mapping large sets of reads on reliable databases, with taxonomically annotated sequences. Declic facilitates data analyses for metabarcoding.

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

6.6. Platforms

6.6.1. 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.

6.6.2. 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.

6.6.3. Team Platform

PLEIADE maintains a dedicated computing platform for software development and experimentation by the team, comprised of a private cloud, storage, and a Project Atomic cluster for hosting Docker containers

7. New Results

7.1. Clavispora lusitaniae genome

Clavispora lusitaniae is an ubiquist environmental 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 onco-haematology patients [24].

So far, two strains have had their genomes sequenced: ATCC 42720, isolated from the blood of a patient with myeloid leukemia [29], and MTCC 1001, a self-fertile strain isolated from citrus [27]. We performed the genome assembly of the *C. lusitaniae* type strain CBS 6936 [37], isolated from citrus peel juice.

Illumina sequences were obtained by our collaborator (T. Noel, UMR 5234 CNRS Université Bordeaux) and we ran the assemblies using several assemblers, e.g. MINIA [21], MIRA [20] and SPAdes [19]. Each assembly gave sequence scaffolds colinear with the already sequence genome of ATCC 42720. However the number of scaffolds varied dramatically in assemblies. SPAdes gave the best results by an order of magnitude (MINIA: 2913, MIRA: 930, SPAdes: 53). This last assembly will serve as a basis for further experiments and SNP detections in mutants strains derived from CBS 6936.

This work is a collaboration between Pleiade team, UMR 5234 CNRS/Université de Bordeaux, and MIAT INRA.

7.2. Elaeis guineensis transcriptome

The mesocarp of the oil palm tree (*Elaeis guineensis* Jacq.) contains oil up to 90% of its dry weight and is the oil richest known vegetal tissue [23]. Our goal is to understand how this tissue achieves this result, in order to increase, by synthetic biology approaches, the yield in oil for crops grown in Europe. In a first milestone, oil palm tree genes relevant for oil synthesis from fatty acids will be transiently expressed in tobacco leaves.

In order to select relevant gene candidates, we performed transcriptome assemblies on high quality Illumina sequences. As an annotated genome is available [34], we performed genome-guided assemblies with TRINITY assembler [26]. The sequence reads (total 300 millions) came from 5 RNA independent isolates from 3 tissues: leaf (2 isolates), kernel (1 isolate) and mesocarp (2 isolates). Transcriptomes coming from duplicate isolates show a good level of overlap as regards the predicted transcripts.

Using expression specificity, abundance and transcript annotation from the genome, we selected genes or transcrpt isoforms candidates, as well as transcription factors. A set of 19 sequences has been retained for expression in tobacco leaves and is under genetic engineering processing.

The 5 transcriptomes were fused into a pan-transcriptome which will be used in another angle of the project. The yield and the composition of oil measured from wild-type palm tree specimens varies dramatically, indicating a high level of bio-diversity. We are currently doing a sampling campaign in Africa, RNA extracts will be sequenced and compared to our pan-genome to serve in variant detection (SNPs) and association genetics studies.

This work is a collaboration between Pleiade team and UMR 5200 CNRS/Université de Bordeaux.

7.3. A Geometric View of Diversity

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 which pairwise distances can be measured, it is possible to build a Euclidean image of it as a point cloud in a space of relevant dimension. The objects under study are microbial communities, given as a set of reads produced by NGS technologies. Distances between specimen are computed as genetic distances between associated reads (so called amplicon approach). Then, the diversity of a community can be associated with the shape of the point cloud built from such distances. Such an embedding is classically implemented by MDS (Multidimensional Scaling). Such an approach triggers two methodological questions, addressed in 2016:

- the numerical solution is through finding the eigenvectors and eigenvalues of a large, full, symmetric matrix. Current algorithm, parallelized or not, are in complexity $\mathcal{O}(n^3)$ if n is the number of specimen on which to study patterns of diversity, i.e. the size of the matrix. This is not feasible for dataset produced by NGS technologies, which can assemble 10^5 to 10^6 sequences. We have set up a collaboration with a PhD student in HIEPACS Team (Inria Bordeaux SO), Pierre Blanchard, to develop a connection between random projection methods and MDS. Random Projection Methods are methods relying on Johnson-Lindenstrauss Lemma, which states that the likelihood that the distances are very well conserved is very high when projecting a point cloud in a space of very large dimension (say n) to a random space of large dimension (say, proportional to $\text{Log}\, n$) . This permits to compute eigenvectors and eigenvelaues in space of reasonable dimension. The method for MDS has been studied by Pierre Blanchard, under supervision by Olivier Coulaud and Alain Franc, and proved to be surprisingly efficient and precise. This work builds one chapter of the PhD thesis of P. B. to be defended by early 2017. This collaboration has lead to a joint poster at Platform for Advanced Scientific Computing (PASC), June 2016, Lausanne, Switzerland.

- The eigenvalues of the matrix under study can be positive or negative. Positive eigenvalues lead to Euclidean structure behind MDS. Classically, negative eigenvalues are ignored. We have begun a study on the role of negative eigenvalues in the discrepancy between Euclidean distances computed between points in MDS, and genetic distances between reads produced by NGS, which adds to the well understood discrepancy in MDS due to dimension reduction. This has lead to three seminars or presentations:
  - a seminar at MIAT research Unit, Toulouse, on February 19
  - a seminar at LABRI on April, 28
  - a presentation at the days of mathematics and computing sciences organized by MIA INRA division (INRA global meeting), on October, 5

  These three events have permitted to “polish” the analysis of the problem through several exchanges, and to orientate its study towards quadratic embedding, or isometry into pseudo-euclidean spaces.

7.4. Topological Data Analysis

Leyla Mirvakhabova has defended and obtained her BSc memoir on "Distance geometry and biodiversity patterns" at the Department of Mathematics of the National Research University - Higher School of Economics, Moscow. Here is the abstract: In this work, we study the biodiversity of the tree species in French Guiana. We consider the matrix of the pairwise genetic distances between the 1501 species. The distances are measured by using the Smith-Waterman algorithm applied to the trnH regions - the chloroplasts of trees. The aim of the project is to analyze the shape of a point cloud in a Euclidean space built from the pairwise distances. To study the structure of the point cloud built from the given distances, we have considered the following methods: Hierarchical Clustering, Multidimensional Scaling (MDS), Nonlinear Mapping (NLM), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Topological Data Analysis (TDA). For the first four methods, we used the Python package scikit-learn 0.17.1 implementations and have written the program for the TDA algorithm. All of these methods were tested on the given dataset. This work has been performed as part of a collaborative research project of the PLEIADE team in the Inria Bordeaux – Sud-Ouest (supervisor Alain Franc) with the Laboratory of System Biology and Computational Genetics in the Vavilov Institute of General Genetics (supervisor Ivan Kulakovskiy).

7.5. Bespoke tools for comparative genomics

Large-scale comparison of strains of cell factory species is an indispensible tool for understanding the genetic origin of phenotypic variability, and can considerably optimize the selection and construction of high-performing industrial strains. For example, in oenological applications new strains may be selected based on their influence on aroma, their adaptation to grape musts, or their robustness during fermentation. In oil production applications, new strains may be selected based on their yield, or on the saturation degree of the lipids, or on their growth characteristics. Comparative genomics has proven quite effective in understanding cell factory diversity [1], [6], [5], [36], [31]. A typical project will involve 500 segregants and 50 genomes. Accurate and rapid analysis of the concomitant data volumes requires efficient tool sets that must be adapted to the real use cases of the industrial application.

PLEIADE addresses this problem through the definition of bespoke software systems that associate integrated sets of tools, including its Magus software platform (section 6.1). A key challenge in defining this kind of integrated system is the need to connect the components. We develop configuration formalisms whose solutions are orchestrations of weakly-coupled microservices running in independant containers. These services may be data banks, genome browser and visualization software, workflow tools like Galaxy, machine learning algorithms for classification, or shared workbooks like Jupyter or Zeppelin. By formalizing the connections between services, we can simplify deployment, and also create an opportunity for *continuous deployment*.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANTICOR – 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.

8.1.2. 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.

8.2. International Initiatives

8.2.1. Supervised clustering

One way to build an inventory in a community on a molecular basis is to map unknown reads onto a taxonomically annotated reference database. We (AF, PC, JMF, FS) have developed a cooperation with UMR Carrtel (A. Bouchez, F. Rimet) and SLU at Uppsala (Sweden, M. Kahlert) for industrializing molecular based inventories from data production (NGS facilities, PGTB, Pierroton) to data analysis. Molecular based inventories of about 200 samples have been done, for diatoms Mayotte rivers, and the same number for diatoms in Fennoscandian rivers. The method has been published in [13]. As far as those tools and metagenomics are concerned, a complementary partnership has been established with UMR BioGER (V. Laval) on metabarcoding of fungal communities.

8.2.2. Metagenomics for zoonoses

In the framework of CEBA Cluster of Excellence (Centre d'Etude de la Biodiversité Amazonienne), Pleiade team has been successful in an application for being part of a so called long term strategic project (2017-2019) called microbiome, chaired by Institut Pasteur in Cayenne and UMR MIVEGEC (CNRS-IRD) at Montpellier. The role of the team is twofold: (i) develop methods for metabarcoding of viral and bacterial communities in some hosts (bats, birds, ...) and (ii) run some data analysis for scaling up from microbiomes to landscape ecology, having in mind the dilution effect, i.e. pristine forest offer a better protection against disease spread than disturbed ones. The project starts on January 1, 2017.

8.2.3. *Historical biogeography of plant families*

In the framework of CEBA too, AF and David Sherman have worked in providing some tools for mapping paleoclimatic conditions on the Earth over geological times, elaborating on datasets of paleoclimates produced by running General Circulation Models (work done by UMR LSCE, Orsay, in a previous ANR project lead by AF). These maps will be part of a collaboration established with The Royal Botanical Gardens at Kew (UK) and several Brasilian Universities in a join project on historical biogeography of Myrtaceae, a large family of trees and shrubs, well developed in the Neotropoics. A. Franc has been visiting E. Lucas, at Kew Botanical gardens, in March 2016 for setting up a cooperation. A first workshop has been organized by F. Salgeiro and AF at Rio in May 2016. The next one will be held in August 2017, organized by E. Lucas and coll. An an open access paper on historical biogeography of the genus Quercus, in collaboration with University of Padova and Museum of Natural History of Stockholm, is [11].

8.2.4. *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). Our focus is inference of metabolic and regulatory models by comparative genomics, and their description using stochastic transition systems.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. *Journal*

9.1.1.1. Member of the Editorial Boards

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

9.1.1.2. Reviewer - Reviewing Activities

Alain Franc has been reviewing in 2016 manuscripts for BMC Evultionary Biology, Nature reports, Methods in Ecology and Evolution, Research in Microbiology, Molecular Ecology.

9.2. Teaching - Supervision - Juries

9.2.1. *Juries*

A. Franc has been supervisor of PhD Thesis of François Keck, UMR Carrtel, Thonon, Grenoble University, with Agnès Bouchez and Frédéric Rimet as co-supervisors. The PhD has been defended on April, 26, 2016. Reference: https://www.theses.fr/196768691. The topic is the use of phylogenetic signal for improving the assessment of water quality from inventories of diatoms. Three papers have been published by François Keck [7], [8], [9].

A. Franc has been

- member of the committee of the PhD of Cyril Noël at IPREM, University of Pau and Pays de l'Adour (PhD advisor: Cristiana Cravo-Laureau)
- reviewer of the HdR of Jean-Daniel Bontemps on large scale forest growth models, University of Nancy
- reviewer of the HdR of Benoit de Thoisy, University of Cayenne, on “from Pleistocene to likely to the dawn of the sixth extinction crisis: the tormented history of Amazonian mammals”.
- member of the jury for PhD defense of Arielle Salmier, University of Cayenne, on the response of chiroptera to changing environment: viral diversity and adaptation, at Cayenne on December, 13, 2016.

D. Sherman was president of the jury for Claire Capdevieille in the Unversity of Bordeaux on Novembre 3, 2016.

D. Sherman was president of a first-year jury for the Mathematics and Computer Science Doctoral School at the University of Bordeaux.

9.2.2. Internships

Rémi Pellerin of the ENS Lyon was an intern in PLEIADE during June–July 2016, and contributed to Declic, a software package written in Python by A. Franc that provides several tools for data analysis, in the domains of multivariate data analysis, machine learning, and graph based methods. It permits users to study in depth the accuracy of the dictionary between molecular based and morphological based taxonomy.

Adrien Lopez of the Collège Henri Brisson in Talence spent a week in PLEIADE for his "stage du troisième".

9.3. Popularization

David Sherman participated in popularization activities based on Thymio-II mobile robots for education, coordinated by the Mobsya association and EPFL (Switzerland). He helped organize a team in the R2T2 event (http://r2t2.org) on November 2, 2016. He contributed code to the Aseba project for piloting Thymio-IIs from the Scratch programming language, and with Thibault Lainé of Inria Bordeaux Sud-Ouest helped improve a photo-realistic simulator for multiple robots.

10. 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

Publications of the year

Articles in International Peer-Reviewed Journals

[7] F. Keck, A. Bouchez, A. Franc, F. Rimet. *Linking phylogenetic similarity and pollution sensitivity to develop ecological assessment methods: a test with river diatoms*, in "Journal of Applied Ecology", March 2016, vol. 53, n^o^ 3, pp. 856 - 864 [*DOI :* 10.1111/1365-2664.12624], https://hal.inria.fr/hal-01426853

[8] F. Keck, F. Rimet, A. Bouchez, A. Franc. *phylosignal: an R package to measure, test, and explore the phylogenetic signal*, in "Ecology and Evolution", March 2016, vol. 6, n^o^ 9, pp. 2774 - 2780 [*DOI :* 10.1002/ECE3.2051], https://hal.inria.fr/hal-01426773

[9] F. Keck, F. Rimet, A. Franc, A. Bouchez. *Phylogenetic signal in diatom ecology: perspectives for aquatic ecosystems biomonitoring*, in "Ecological Applications", April 2016, vol. 26, n^o^ 3, pp. 861 - 872 [*DOI :* 10.1890/14-1966], https://hal.inria.fr/hal-01426854

[10] F. Rimet, P. Chaumeil, F. Keck, L. Kermarrec, V. Vasselon, M. Kahlert, A. Franc, A. Bouchez. *R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring*, in "Database - The journal of Biological Databases and Curation", February 2016, vol. 2016 [*DOI :* 10.1093/DATABASE/BAW016], https://hal.inria.fr/hal-01426772

[11] M. C. Simeone, G. W. Grimm, A. Papini, F. Vessella, S. Cardoni, E. Tordoni, R. Piredda, A. Franc, T. Denk. *Plastome data reveal multiple geographic origins of Quercus Group Ilex*, in "PeerJ", April 2016, vol. 4, e1897 [*DOI :* 10.7717/PEERJ.1897], https://hal.inria.fr/hal-01426766

Other Publications

[12] P. BLANCHARD, O. Coulaud, E. Darve, A. Franc. *FMR: Fast randomized algorithms for covariance matrix computations*, June 2016, Platform for Advanced Scientific Computing (PASC), Poster, https://hal.archives-ouvertes.fr/hal-01334747

[13] J.-M. Frigerio, F. Rimet, A. Bouchez, E. Chancerel, P. Chaumeil, F. Salin, S. Thérond, M. Kahlert, A. Franc. *diagno-syst: a tool for accurate inventories in metabarcoding*, November 2016, working paper or preprint, https://hal.inria.fr/hal-01426764

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] A. Bankevich. *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing*, in "Journal of Computational Biology", 2012, vol. 19, pp. 455-477

[20] B. Chevreux. *Genome sequence assembly using trace signals and additional sequence information*, in "Proceedings of the German Conference on Bioinformatics (GCB)", 1999

[21] R. Chikhi, G. Rizk. *Space-Efficient and Exact de Bruijn Graph Representation Based on a Bloom Filter*, in "Proceedings of the 12th Workshop on Algorithms in Bioinformatics (WABI)", 2012

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

[23] R. Corley, P. Tinker. *The Oil Palm, 4th ed*, Blackwell, 2003, 562 p.

[24] A. Favel. *Colony Morphology Switching of Candida Lusitaniae and Acquisition of Multidrug Resistance During Treatment of a Renal Infection in a Newborn: Case Report and Review of the Literature*, in "Diagn. Microbiol. Infect. Dis.", 2003, vol. 47, pp. 331-339

[25] 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

[26] M. Grabherr. *Full-length transcriptome assembly from RNA-Seq data without a reference genome*, in "Nat Biotechnol.", 2011, vol. 29, pp. 644-652

[27] M. Lachance. *The D1/D2 domain of the large-subunit rDNA of the yeast species Clavispora lusitaniae is unusually polymorphic*, in "FEMS Yeast Res.", 2003, vol. 4, pp. 253–258

[28] 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

[29] D. Pappagianis. *Development of resistance to amphotericin B in Candida lusitaniae infecting a human*, in "Antimicrob Agents Chemother.", 1979, vol. 16, pp. 123-126

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

[31] A. Romano, H. Trip, H. Campbell-Sills, O. Bouchez, D. Sherman, J. S. Lolkema, P. M. Lucas. *Genome Sequence of Lactobacillus saerimneri 30a (Formerly Lactobacillus sp. Strain 30a), a Reference Lactic Acid Bacterium Strain Producing Biogenic Amines*, in "Genome Announcements", January 2013, vol. 1, n° 1, pp. e00097-12 [*DOI :* 10.1128/genomeA.00097-12], https://hal.inria.fr/hal-00863284

[32] S. T. Roweis, Z. Ghahramani. *A unifying review of linear Gaussian Models*, in "Neural Computation", 1999, vol. 11, n° 2, pp. 305–45

[33] 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

[34] H. Teh. *Differential metabolite profiles during fruit development in high-yielding oil palm mesocarp*, in "PLoS One.", 2013, vol. 8, n° 4, e61344 p.

[35] D. W. Thompson. *On Growth and Form*, Cambridge University Press, 1917

[36] A. Zimmer, C. Durand, N. Loira, P. Durrens, D. J. Sherman, P. Marullo. *QTL dissection of Lag phase in wine fermentation reveals a new translocation responsible for Saccharomyces cerevisiae adaptation to sulfite*, in "PLoS ONE", 2014, vol. 9, n° 1, e86298 p. [*DOI :* 10.1371/journal.pone.0086298], https://hal.inria.fr/hal-00986680

[37] R. de Miranda. *Clavispora, a new yeast genus of the Saccharomycetales*, in "Antonie van Leeuwenhoek.", 1979, vol. 45, pp. 479-483