IN PARTNERSHIP WITH:
CNRS

INRA

Institut national de recherche pour l'agriculture, l'alimentation et l'environnement

Activity Report 2019

Project-Team PLEIADE

Patterns of diversity and networks of function

IN COLLABORATION WITH: Laboratoire Bordelais de Recherche en Informatique (LaBRI)

RESEARCH CENTER
Bordeaux - Sud-Ouest

THEME
Computational Biology

Table of contents

Project-Team PLEIADE

Creation of the Team: 2015 January 01, updated into Project-Team: 2019 March 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:

B1.1.7. - Bioinformatics
B1.1.10. - Systems and synthetic biology
B3. - Environment and planet

1. Team, Visitors, External Collaborators

Research Scientists

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

Post-Doctoral Fellow

Swarna Kanchan [Inria, Post-Doctoral Fellow]

PhD Student

Mohamed Anwar Abouabdallah [Inria, PhD Student, from Oct 2019]

Technical staff

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

Administrative Assistant

Catherine Cattaert Megrat [Inria, Administrative Assistant]

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 [29] 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

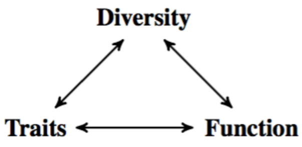

Figure 1. Diversity informs both the study of traits, and the study of biological functions

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.

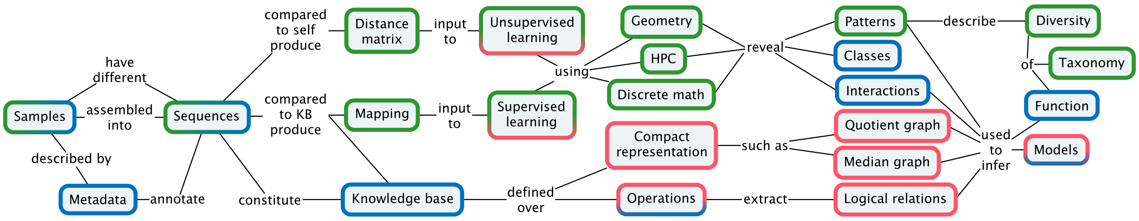

Figure 2. PLEIADE is a pluridisciplinary team. Each application in biodiversity and biotechnology follows a path calling on methods from biology (blue), mathematics (green), and computer science (red).

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 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 [28], [27]. See as well [30] 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 [22] with convex optimization procedures through matrix completion [15], [17].

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

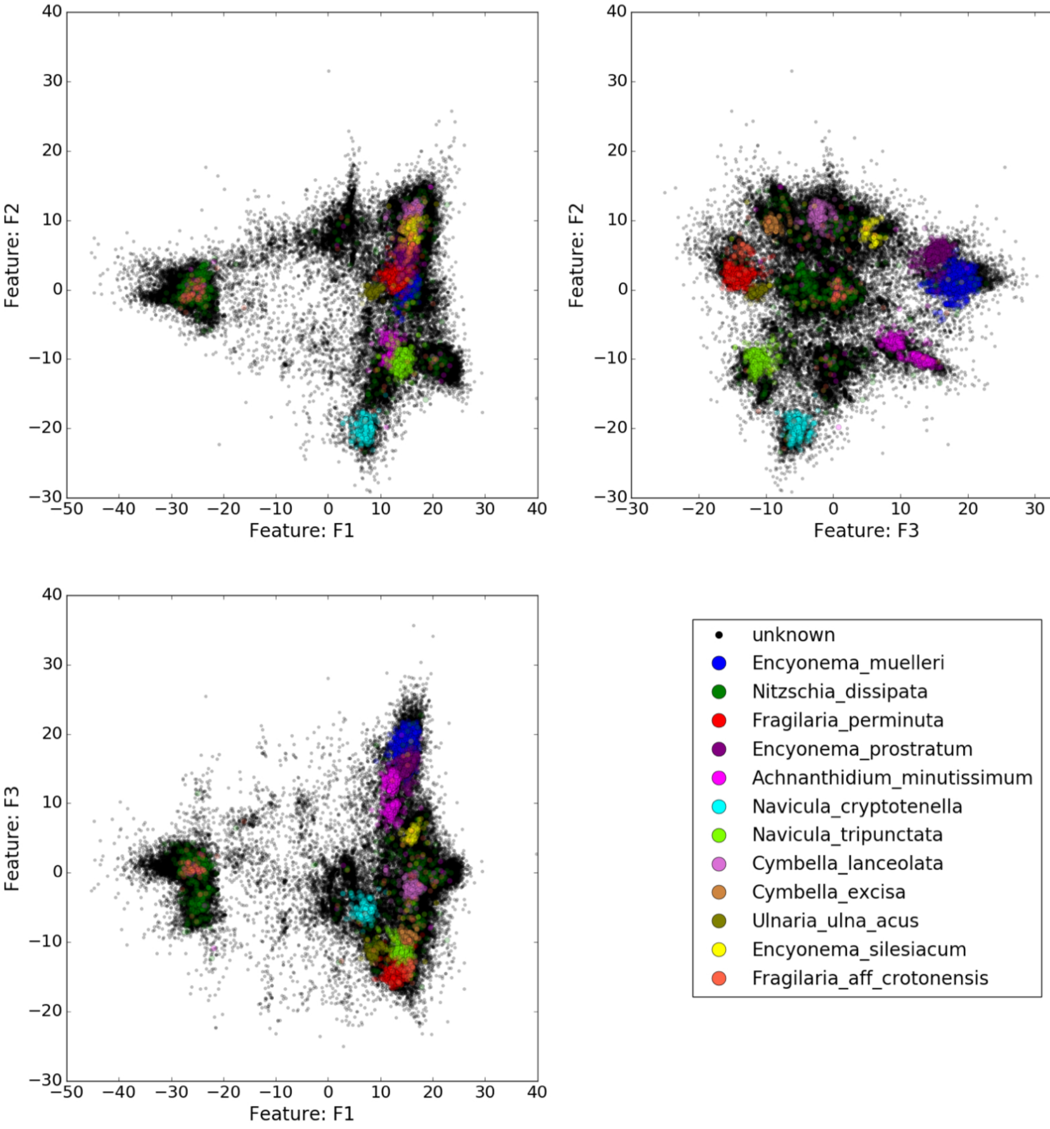

Figure 3. Validation of high density islands using supervised classification. Metagenomic reads from diatoms in Lake Geneva [26] were analyzed by the method from [16] and colored by species according to a reference database.

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 [13]. 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* [10] 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], [14] and medicine [12]. 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).

4.1.1. Oil Palm lipid synthesis

The largest source of vegetable oil [1] is the fruit mesocarp of the oil palm *Elaeis guineensis*, a remarkable tissue that can accumulate up to 90% oil, the highest level observed in the plant kingdom. The market share of oil palm is expected to increase in order to meet increased demand for vegetable oil, predicted to double by 2030 [18], be it as food or as a source of biofuels in Africa. A significant proportion of palm oil is produced on small estates that do not have access to efficient milling facilities, and run a great risk of spoilage through oil acidification. Improving palm oil quality through genetics and selection will result in economic gains [24] by addressing several targets such as improvement of oil yield, tuning of oil quality through the rate of unsaturated fatty acids or impairment of degradation processes. Furthermore, as genome biodiversity resides mostly in Africa, oil from African oil palms can vary greatly in fatty acid composition according to cultivar genetic differences and to weather conditions, and the precise mechanisms regulating this variability are not yet understood.

[1] 32% of the world market share [24]

A growing body of molecular resources for studying oil palm fruit are making it possible to study and improve the quality and quantity of oil produced by oil palms. In particular, these oils can vary greatly in fatty acid composition, and while the precise mechanisms regulating this variability are not completely understood, establishing a link between oil palm genotype and phenotype appears increasingly feasible. PLEIADE will work with the CNRS/UB UMR 5200 (LBM), a laboratory with an established reputation in studying fatty acid metabolism in *E. guineensis*, to improve understanding of the links between genetic diversity and oil production, and participate in developing applications.

4.1.2. Engineering pico-algae

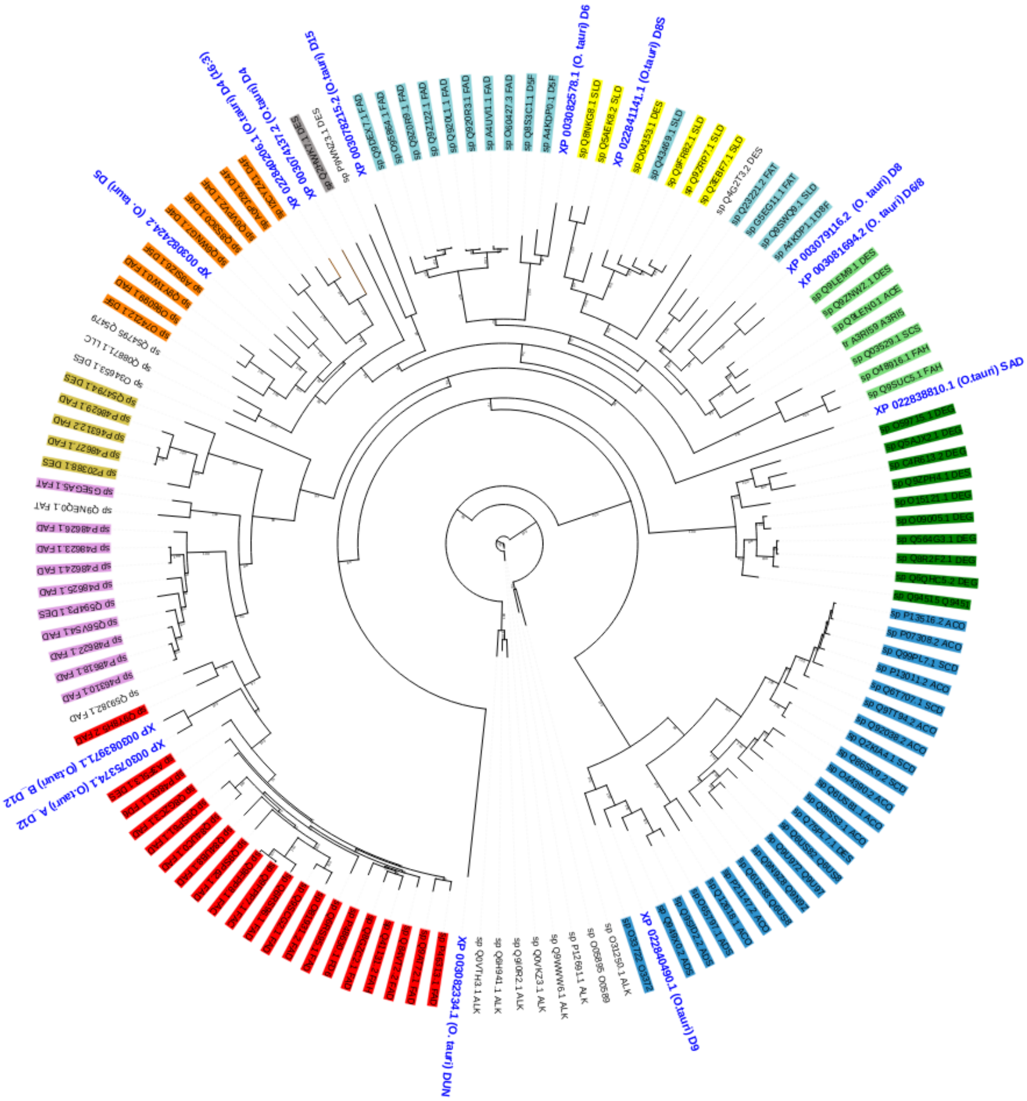
XP 003078215.2 (O.tauri) D15
XP 003082578.1 (O. tauri) D6
XP 022841141.1 (O.tauri) D8S
XP 003079116.2 (O. tauri) D8
XP 003081694.2 (O. tauri) D6/8
XP 022838810.1 (O.tauri) SAD
XP 022840490.1 (O.tauri) D9
XP 003082334.1 (O. tauri) DUN
XP 003075374.1 (O.tauri) A_D12
XP 003083971.1 (O.tauri) B_D12
XP 003082424.2 (O. tauri) D5
XP 022840206.1 (O.tauri) D4 (16:3)
XP 003074137.2 (O. tauri) D4

Figure 4. Phylogenetic structure of long-chain polyunsaturated fatty acid desaturase specificity. Highlighted are thirteen desaturases from Ostreococcus tauri

Docosahexaenoic acid (DHA) is an essential nutriment for human brain tissue and can only be obtained from marine or riverine fish that live on phytoplankton and zooplankton, since human neurons lack the delta desaturase required for *de novo* synthesis of DHA. [21] Unfortunately, fishing is become less and less a sustainable resource. Since phytoplankton and zooplankton are the ultimate source of DHA consumed, there is considerable interest in obtaining DHA directly rather than through the intermediary of fish. A very promising approach is through the bio-engineering of pico-algae.

In order to produce the long-chain polyunsaturated fatty acids (LC-PUFA) needed for human nutrition, it is necessary to precisely engineer the desaturases that produce them. Desaturases are enzymes responsible for the introduction of double bonds into fatty acids. Desaturases are specific in recognizing their substrates and in placing the double bond in the proper place. The desaturases that produce the LC-PUFA necessary for human nutrition are present only in some species.

Our goal is to design methods to predict the substrate and region specificities for desaturases in algal species, particularly *Ostreococcus tauri*, the smallest photosynthetic eukaryote that can be cultivated. Thirteen desaturases are known in *O. tauri* and can be placed in the phylogeny of the desaturase family (figure 4). The biochemical and structural characterization of these enzymes is as yet very incomplete. This work is ongoing (see section 8.2.2) and requires close collaboration between biologists and computer scientists.

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 [19]. 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 [23]. 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 [20]. Several approaches were proposed to correct for paralogy, either by working directly on the sequences issued from mapped reads [20] 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 [11]. 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. Genomic determinants:

The study [7] appearing in *BMC Genomics* resolves a longstanding enigma in winemaking. While the *fact* of resistant strains of wine yeasts has been known for many years, the actual mechanism underlying the phenomena have only now been elucidated by a combined genetic and bioinformatic analysis.

5.1.2. INRAE-Inria PhD:

PLEIADE was succesful in applying for a PhD in 2019 INRA-Inria call for PhD. The submitted topic is: “Statistical Learning for OTU identification and Biodiversty characterizaton.” This PhD is a collaboration between PLEIADE (INRAE-Inria, supervision), HiePACS (Inria) and MIAT Toulouse (INRAE).

6. New Software and Platforms

6.1. magecal

KEYWORD: Genomics

SCIENTIFIC DESCRIPTION: Magecal independently runs training and prediction steps for Augustus, Conrad, GeneID, GeneMark, and Snap. The results are cleaned and integrated into a common format. Jigsaw is trained and used for model reconciliation. Consistency constraints are applied to ensure that phase and intron structure are biologically plausible.

FUNCTIONAL DESCRIPTION: Magecal predicts a set of protein coding genes in fungal genomic sequences, using different de novo prediction algorithms, and reconciling the predictions with the aid of comparative data. Magecal applies consistency constraints to guarantee that the predicted genes are biologically valid.

RELEASE FUNCTIONAL DESCRIPTION: Dockerization and compatibility with Alcyone

- Participants: Pascal Durrens and David James Sherman
- Contact: David James Sherman
- URL: https://gitlab.inria.fr/magecal/magecal

6.2. Declic

KEYWORDS: Data analysis - Machine learning - Taxonomies

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
- URL: https://gitlab.inria.fr/pleiade/declic

6.3. 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 James 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

6.4. 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/

6.5. 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

6.6. 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 James Sherman
- Contact: David James Sherman
- URL: https://team.inria.fr/pleiade/alcyone/

6.7. family-3d

KEYWORDS: Biodiversity - Point cloud - 3D modeling

SCIENTIFIC DESCRIPTION: The method statistically selects a subset of pairwise distances between proteins in the family, constructs a weighted graph, and lays it out using an adaptation of the three-dimensional extension of the Kamada-Kawai force-directed layout.

FUNCTIONAL DESCRIPTION: Family-3D lays out high-dimension protein family point clouds in 3D space. The resulting lower-dimension forms can be printed, so that they can be explored and compared manually. They can also be explored interactively or stereographically.

Comparison of the 3D forms reveals classes of structurally similar families, whose characteristic shapes correspond to different evolutionary scenarios. Some of these scenarios are: neofunctionalization, subfunctionalization, founder gene effect, ancestral family.

To facilitate curator training, Family-3D includes an interactive terminal containing a microcontroller, an RFID reader, and an LED ring. A set of shapes that fall in predetermined classes is printed, with a unique RFID tag in each shape. Trainees classify family shapes by manual inspection and submit their classes to the terminal, which evaluates the proposed class and provides visual feedback.

- Participant: David James Sherman
- Contact: David James Sherman
- URL: https://gitlab.inria.fr/pleiade/family-3d

6.8. Yapotu

Yet Another Pipeline for OTU building

KEYWORDS: Taxonomies - Distance matrices - Clustering - Metagenomics

FUNCTIONAL DESCRIPTION: OTU building is one of the key operation in metabarcoding: how to delimit Operational Taxonomic Units in the "soup" where all amplicons of thousands of unicellular organisms have been lumped together (from visible plankton to nano- and picoplankton, for example). Yapotu is a software tool that enables approaches to unsupervised clustering on very large matrices of distances: each element is a distance between two reads produced by a sequencer. It permits one to select one method among several, like building MultiDimensional Scaling to produce an Euclidean image, and clustering within the point cloud, or building a graph where a node is a sequence and there is an edge if both sequences are at a distance smaller than a given threshold, and then after build connected components or communities on this graph. Other functions visualize the OTUs as clusters or subgraphs.

- Contact: Alain Franc
- URL: https://gitlab.inria.fr/afranc/diodon/tree/master/yapotu

6.9. Diodon

KEYWORDS: Dimensionality reduction - Data analysis

FUNCTIONAL DESCRIPTION: Most of dimension reduction methods inherited from Multivariate Data Analysis, and currently implemented as element in statistical learning for handling very large datasets (the dimension of spaces is the number of features) rely on a chain of pretreatments, a core with a SVD for low rank approximation of a given matrix, and a post-treatment for interpreting results. The costly part in computations is the SVD, which is in cubic complexity. Diodon is a list of functions and drivers which implement (i) pre-treatments, SVD and post-treatments on a large diversity of methods, (ii) random projection methods for running the SVD which permits to bypass the time limit in computing the SVD, and (iii) an implementation in C++ of the SVD with random projection at prescribed rank or precision, connected to MDS.

- Contact: Alain Franc
- URL: https://gitlab.inria.fr/afranc/diodon

7. New Results

7.1. Genetic Determinisms of Aborted Fermentation in Winemaking

This study contributes to the understanding of the mechanisms leading to stuck fermentation in winemaking, an economic prejudice [7]. A number of factors can trigger stuck or aborted fermentation such as high temperature, high ethanol concentration, low pH. The biodiversity of natural yeast strains used in winemaking starters has as a consequence that some of them are more prone to abort fermentation than others, indicating a genetic determinism. Crosses between strains called “sensitive” or “resistant” to stuck fermentation occurrence, followed by back-crosses with the “sensitive” parent while selecting for the “resistant” phenotype, allowed us to reduce the amount of genetic material inherited from the “resistant” parent in the progeny, ending to 3 small introgression areas after 4 generations. Quantitative Trait Locus (QTL) detection in this progeny (77 strains) involved characterization of SNP inheritence (circa 1200 validated SNPs) from either parent, through micro-array hybridization, mapping of the SNP on the reference genome and phenotypic measurements on the progeny. This analysis made it possible to detect two genes which, when inactivated by naturally occurring mutations, act as major perturbators of several fermentation parameters in winemaking physiological conditions. Consequently, our industrial partner incorporated into its catalogue of winemaking starters, strains carrying the functional forms of these genes.

7.2. Characterization of Molecular Biodiversity

In 2019 PLEIADE developed new methods for characterizing molecular biodiversity (see [4], [5] for applications). This point itself has been developed with two approaches in 2019, each with the beginning of a PhD.

- Building OTUs from a pairwise distance matrix typically is an unsupervised clustering issue. In this new development, the PhD student (PLEIADE) tests whether SBM (Stochastic Block Model) approach yields relevant results for a global characterization of biodiversity beyond a summary by a scalar index. This is done in collaboration with MIAT INRAE research unit in Toulouse and EPC HiePACS. It represents a connection between metabarcoding and statistical modeling, a topic which deserves investigation and is expanding. (Figure 5 from [8])
- A major goal of PLEIADE is to develop a geometric view on biodiversity. The tool selected up to now is to associate a point cloud to a dataset (pairwise distances between sequences) and study its shape. In 2019, PLEIADE has been associated in a collaboration with HiePACS to begin a new topic: comparison between point clouds, each cloud being associated to a data set. Indeed, the development of metabarcoding leads to the new issue of comparison between OTUs built from different dataset. This approach is part of the issues raised in a PhD supervized by HiePACS, in collaboration with PLEIADE.

7.3. Scaling Metabarcoding Programs

Metabarcoding is a series of technical procedures to build molecular based inventories from large datasets of amplicons. We derived new methods and tools to scale metabarcoding programs in collaboration with EPC HiePACS. This has been realized through following participation in research projects:

- Contribution to ADT Gordon project in Inria BSO. The objective of this project (partners: Tadaam (coordinator), STORM, HiePACS, PLEIADE) is to integrate SVD as an available tool in Chameleon, starPU and new Madeleine. The contribution of Pleiade is to bring metabarcoding as a use case, and random projection ([6]) as a method for scaling Multidimensional Scaling (which requires an SVD) in collaboration with HiePACS with a template implemented in Diodon. In 2019, PLEIADE has bought to the project a series of 55 matrices with size about 10^5 rows/column which, assembled, yield a full pairwise distance matrix between one million of sequences. The objective is to reach the million in 2020.
- Contribution to Region Nouvelle Aquitaine project “HPC Scalable Ecosystem.” This project is chaired by HiePACS. In collaboration with this EPC, PLEIADE is involved in developing a new approach for comparing OTUs built from different datasets.

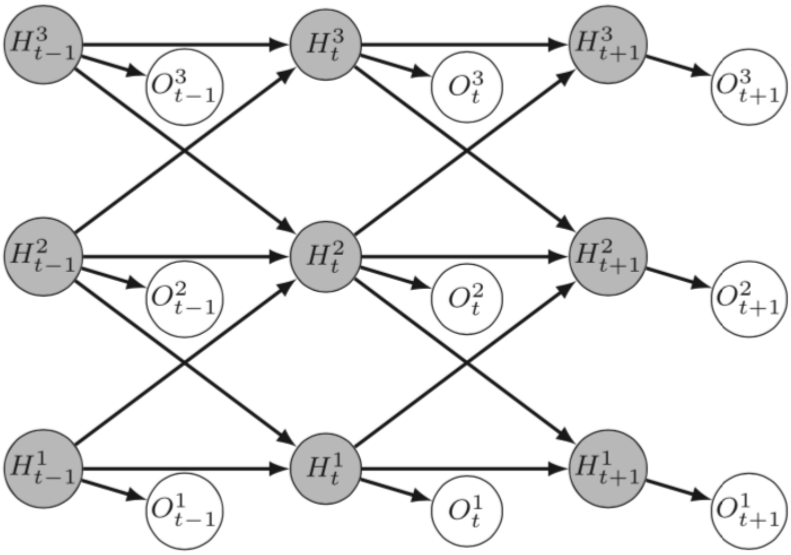

Figure 5. Graphical representation of a coupled HMM with three hidden chains (from [8])

7.4. Linking Homology and Function for Algal Desaturases

Polyunsaturated fatty acids (PUFA) such as Omega-3 that are essential for human health cannot be synthesized by the human body and must be acquired through the consumption of certains foods, such as oily fish, that are becoming increasing difficult to produce sustainably. However, fish do not produce them either: their role is to concentrate PUFA through the food chain. There is consequently considerable interest in producing these essential nutriments directly, through the cultivation of domesticated strains of naturally occurring or of engineered strains of green algae.

Ultimately, polyunsaturated fatty acids are produced by molecular machines called **desaturases**. While desaturases are abundant in all branches of life, the link between gene sequence and the precise activity of the corresponding enzyme is poorly understood. The particular challenge is that, while the catalytic active site is well conserved, the features that recognize the substrate and that determine the regiospecificity of the enzyme are not. In order to produce specific PUFA at industrial scale, it is necessary to develop efficient tools for high-throughput identification of candidate genes in algal species, and precise models for designing desaturases through synthetic biology.

In collaboration with the LBM (UMR 5200 CNRS) and with the support of the Inria Project Lab *In silico algae*, we used a core collection of thirteen desaturases from *Osteococcus tauri* to explore the link between homology and function in 23 species ([9]). The study reinforced our understanding of the evolutionary conservation of desaturases and confirmed the identification of substrate and regio- specificity through graph neighborhoods. We were further able to extend the identification of PPR motifs correlated with specificity. This work is ongoing. Since most of the pertinent desaturases are membrane bound, the prediction of protein structure has proved perilous, but we are hopeful that future work will allow us to use structure-inspired prediction to narrow in on the sites responsible for specificity despite their poor sequence conservation.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Malabar

This is a project funded by labex COTE (University of Bordeaux) as a collaboration with IFREMER at Arcachon, EPOC (Talence), and ETI chair of the labex. The guideline of the project is to build models in statistical ecology on a series of molecular based invetories (300 samples) from occurence matrices of OTUs in samples, with environmental variables. The samples have been collected in 2018-2019, the sequences produced by BioGeCo in 2019, and data analysis will begin in 2020.

8.1.2. High-performance computing and metabarcoding

PLEIADE is member of two projects, one funded by the Région Nouvelle Aquitaine and one funded as Inria ADT Gordon, connecting Chameleon, StartPU and NewMadeleine, where the use case of metabarcoding (questions, data sets) hase been selected to link these layers together. This will permit us to address unsupervised clustering of one million reads next year. These projects are in collaboration with the HiePACS, TADAAM, and STORM project-teams.

8.1.3. 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).

8.2. National Initiatives

8.2.1. 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: one funded by ONEMA, the second by labex COTE, where BioGeCo/Pleiade is responsible for data analysis, with implementaton of the tools recently developed for scaling MDS. Calculations have been made on CURTA at MCIA and PlaFRIM at Inria.

8.2.2. 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.

8.3. European Initiatives

8.3.1. Collaborations in European Programs, Except FP7 & H2020

Program: COST

Project title: COST Action DNAqua.net

Abstract: PLEIADE is responsible for the WG "Data Analysis and storage" in this action. As such, we have organized with CNR Verbana (Italy) two Europeanwide workshops: one in Lyon in February 2019, and one in Limassol (Cyprus) in October 2019. As a follow up of these workshops, Pleiade and BioGeCo will be responsible for taking in charge data analysis of OTU picking in two European wide projects:

- a benchmark for different tools for OTU picking, with datasets from different European teams
- a comparison between different organisms (metabarcoding inventories) for assessing the quality of the water of Danube river, in collaboration with raparian countries

Program: EOSC

Project title: EOSC-Pillar

Abstract: This is a follow up of our former participation in EOSC-Pilot. In collaboration with HiePACS, PLEIADE is involved in task 7.4, for bringing use cases in metabarcoding as testbeds for circulation of codes between different infrastructures, including PlaFRIM.

8.4. International Initiatives

8.4.1. Vitapalm – Food and nutrition security and sustainable agriculture in Africa

PLEIADE participates in the Vitapalm program financed by LEAP-Agri [2], the joint Europe Africa Research and Innovation (R&I) initiative related to Food and Nutrition Security and Sustainable Agriculture. Vitapalm uses genomics and selection to improve the nutritional quality and the stability of palm oil produced by Africa smallholdings for local consumption. Project partners are from Cameroon, France, Germany, and Ghana.

8.4.2. Simulation of metacommunities

In collaboration with the Pasteur Institute in Cayenne and the INRA MIA Research Team in Toulouse, PLEIADE is developing a stochastic model for simulation of metacommunities, in the framework of patch occupancy models. The objective is a better understanding of zoonose propagation, namely rabies through bat hosts in connection with disturbances of pristine forests in French Guiana, which have an impact on the exposure of human populations to wildlife that act as reservoirs of zoonoses.

8.4.3. 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

PLEIADE is involved with BioGeCo as partner of Institut Pasteur de Guyane at Cayenne for developing the domain of so-called Ecoviromics for some zoonoses in French Guiana. The spine of this collaboration is co-supervizing of a PhD student at IPG in cayenne, in bioinformatics and statistical ecology to decipher the respective roles of host phylogeny and environmetal variables in the virome of different hosts (bats, rodents, birds).

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.

[2] http://www.leap-agri.com/

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

9.1.2. Scientific Expertise

David Sherman is a member of the Scientific Advisory Board of Enlightware GmbH, Zürich.

9.1.3. Research Administration

Alain Franc has been appointed “chargé de mission calcul” at INRA by INRA Delegate for Digital Transition. As such, his mission is to propose animations and solutions for the development of scientific computing at INRA, whatever the Research Department.

David Sherman is president of the Commission for Technology Development (CDT) of the Inria Bordeaux Sud-Ouest research center. The CDT has two roles. First, it evaluates funding requests for Technology Development and Technology Transfer projects, which typically involve hiring technical staff. Second, the CDT is responsible for validating and overseeing contract engineers hired by Inria project-teams.

David Sherman represents Inria in the steering committee of the Region Nouvelle Aquitaine’s regional research network “Biodiversity and Ecosystemic Services” (BIOSENA).

9.2. Popularization

9.2.1. Education

David Sherman contributed to the *Journée APEIA (Activités Pour l’Enseignement de l’IA)*
David Sherman help train teachers in programming mobile robots, in preparation for R2T2 events (http://r2t2.org) coordinated by the Mobsya association and EPFL (Switzerland).

9.2.2. Interventions

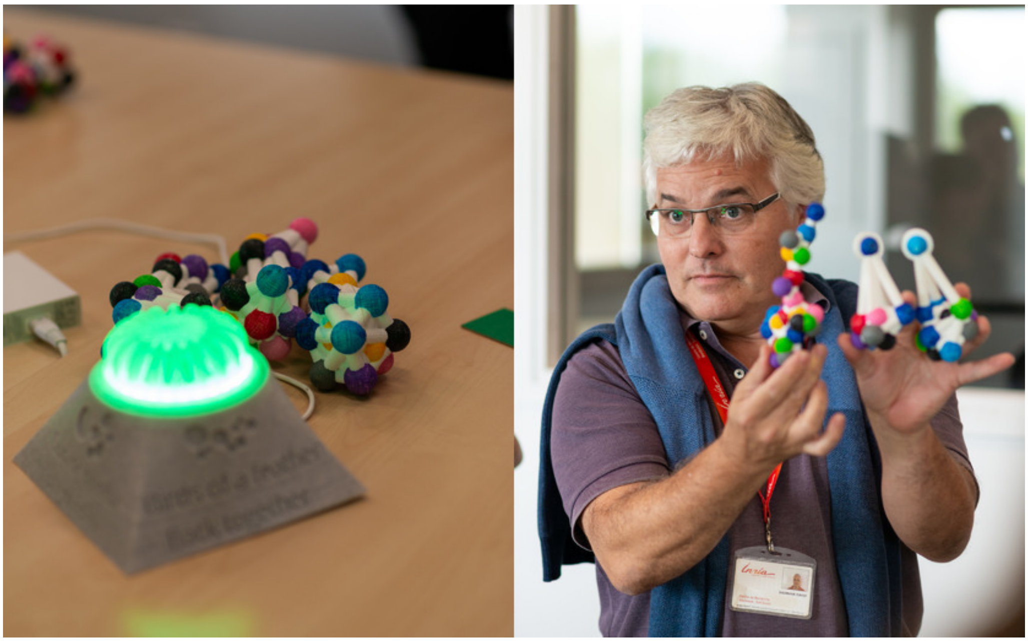

Figure 6. ***Qui se ressemble s’assemble*** *during the Circuit Scientifique in 2019: (left) classification oracle (right) comparing similar shapes (Photo copyright Gautier DUFAU)*

In collaboration with a local PME, David Sherman helped design an interactive exhibit on tree classification for a museum in the Region Nouvelle Aquitaine, based on a specialization of **Qui se ressemble s'assemble**.

David Sherman presented the **Qui se ressemble s'assemble** activity to 100 middle school students during the Circuit Scientifique in 2019 (Figure 6), and to 40 undergraduate students from the ENS Lyon.

9.2.3. Creation of media or tools for science outreach

David Sherman refined his design for the electronics of the **Qui se ressemble s'assemble** activity, to explain the methods and uses of pattern classification of protein families. The activity uses 20 3D-printed point clouds and 4 oracles, containing a microcontroller, an RFID reader with a custom-designed inductive coil as input, and an LED ring as output. Participants propose groups of 3D shapes that they believe belong to the same class, and the oracle evaluate the group.

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] 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

Publications of the year

Articles in International Peer-Reviewed Journals

[4] B. Bailet, A. Bouchez, A. A. Franc, J.-M. Frigerio, F. Keck, S.-M. Karjalainen, F. Rimet, S. Schneider, M. Kahlert. *Molecular versus morphological data for benthic diatoms biomonitoring in Northern Europe freshwater and consequences for ecological status*, in "Metabarcoding and Metagenomics", June 2019, vol. 3 [*DOI :* 10.3897/mbmg.3.34002], https://hal.inria.fr/hal-02432989

[5] H. Caron, J. Molino, D. Sabatier, P. P. Léger, P. Chaumeil, C. Scotti-Saintagne, J.-M. Frigerio, I. Scotti, A. A. Franc, R. J. Petit. *Chloroplast DNA variation in a hyperdiverse tropical tree community*, in "Ecology and Evolution", March 2019, vol. 9, n^{o} 8, pp. 4897-4905 [*DOI :* 10.1002/ece3.5096], https://hal.umontpellier.fr/hal-02108230

[6] A. A. Franc, P. Blanchard, O. Coulaud. *Nonlinear mapping and distance geometry*, in "Optimization Letters", May 2019 [*DOI :* 10.1007/s11590-019-01431-y], https://hal.inria.fr/hal-02124882

[7] P. MARULLO, P. DURRENS, E. PELTIER, M. BERNARD, C. MANSOUR, D. DUBOURDIEU. *Natural allelic variations of Saccharomyces cerevisiae impact stuck fermentation due to the combined effect of ethanol and temperature; a QTL-mapping study*, in "BMC Genomics", December 2019, vol. 20, n^o 1 [*DOI :* 10.1186/S12864-019-5959-8], https://hal.archives-ouvertes.fr/hal-02378470

[8] N. D. P. PEYRARD, S. DE GIVRY, A. A. FRANC, S. ROBIN, R. R. SABBADIN, T. SCHIEX, M. VIGNES. *Exact or approximate inference in graphical models: why the choice is dictated by the treewidth, and how variable elimination can be exploited*, in "Australian and New Zealand Journal of Statistics", June 2019, vol. 61, n^o 2, pp. 89-133 [*DOI :* 10.1111/ANZS.12257], https://hal.inria.fr/hal-02433018

Research Reports

[9] S. KANCHAN. *In silico comparative function prediction of enzymes, applied to fatty acid metabolism in microalgae : Final Report*, Inria Bordeaux Sud-Ouest, December 2019, https://hal.inria.fr/hal-02431250

References in notes

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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] P. BLANCHARD, 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://arxiv.org/abs/1803.02272 , https://hal.inria.fr/hal-01685711

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

[18] A. CARLSSON, J. YILMAZ, A. GREEN, S. STYMNE, P. HOFVANDER. *Replacing fossil oil with fresh oil - with what and for what?*, in "Eur J Lipid Sci Technol", 2011, vol. 113, n^{o} 7, pp. 812-831

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

[20] 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

[21] M. HASIMOTO, S. HOSSAIN, A. AL MAMUN, K. MATSUZAKI, H. ARAI. *Docosahexaenoic acid: one molecule diverse functions*, in "Crit Rev Biotechnol.", Aug 2017, vol. 37, n^{o} 5, pp. 579-597, http://dx.doi.org/10.1080/07388551.2016.1207153

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

[23] 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

[24] F. MORCILLO, D. CROS, N. BILLOTTE, G. NGANDO-EBONGUE, H. DOMONHÉDO, M. PIZOT, T. CUÉLLAR, S. ESPÉOUT, R. DHOUIB, F. BOURGIS, S. CLAVEROL, T. TRANBARGER, B. NOUY, V. ARONDEL. *Improving palm oil quality through identification and mapping of the lipase gene causing oil deterioration*, in "Nat Commun", 2013, vol. 4, 2160 p. , http://dx.doi.org/10.1038/ncomms3160

[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] 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

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

[28] 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

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

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