IN PARTNERSHIP WITH:
CNRS

Université de Bordeaux

Activity Report 2012

Project-Team MAGNOME

Models and Algorithms for the Genome

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

RESEARCH CENTER
Bordeaux - Sud-Ouest

THEME
Computational Biology and Bioinformatics

Table of contents

Project-Team MAGNOME

Keywords: Computational Biology, Genomics, Next Generation Sequencing, Machine Learning, High Performance Computing

Creation of the Project-Team: July 01, 2009 .

1. Members

Research Scientists
David James Sherman [Team leader; Inria, Senior Researcher (DR), HdR]
Pascal Durrens [CNRS, Junior Researcher (CR), HdR]

Faculty Member
Elisabeth Bon [U. Bordeaux Ségalen, Associate Professor (MCF)]

External Collaborator
Vsevolod Makeev [Russian Acad. Sci., since 2011-07-21, HdR]

Engineers
Tiphaine Martin [CNRS, Research engineer (IR), until 2012-08-31]
Aurélie Goulielmakis [U. Bordeaux 1, Inria, until 2012-09-14]
Florian Lajus [Inria, Contract engineer for Magus ADT]

PhD Students
Laetitia Bourgeade [U. Bordeaux]
Natalia Golenetskaya [Inria CORDI-S]
Razanne Issa [Exchange Fellowship Syria]
Nicolás Loira [Inria, until 2012-01-20]
Anna Zhukova [Inria CORDI-S, since 2011-10-15]

Post-Doctoral Fellow
Witold Dyrka [Inria, since 2012-12-14]

Visiting Scientist
Rodrigo Assar Cuevas [U. Chile]

Administrative Assistant
Anne-Laure Gautier [Inria]

2. Overall Objectives

2.1. Overall Objectives

One of the key challenges in the study of biological systems is understanding how the static information recorded in the genome is interpreted to become dynamic systems of cooperating and competing biomolecules. MAGNOME addresses this challenge through the development of informatic techniques for multi-scale modeling and large-scale comparative genomics:

- logical and object models for knowledge representation
- stochastic hierarchical models for behavior of complex systems, formal methods
- algorithms for sequence analysis, and
- data mining and classification.

We use genome-scale comparisons of eukaryotic organisms to build modular and hierarchical hybrid models of cell behavior that are studied using multi-scale stochastic simulation and formal methods. Our research program builds on our experience in comparative genomics, modeling of protein interaction networks, and formal methods for multi-scale modeling of complex systems.

New high-throughput technologies for DNA sequencing have radically reduced the cost of acquiring genome and transcriptome data, and introduced new strategies for whole genome sequencing. The result has been an increase in data volumes of several orders of magnitude, as well has a greatly increased density of genome sequences within phylogenetically constrained groups of species. MAGNOME develops efficient techniques for dealing with these increased data volumes, and the combinatorial challenges of dense multi-genome comparison.

3. Scientific Foundations

3.1. Overview

Fundamental questions in the life sciences can now be addressed at an unprecedented scale through the combination of high-throughput experimental techniques and advanced computational methods from the computer sciences. The new field of *computational biology* or *bioinformatics* has grown around intense collaboration between biologists and computer scientists working towards understanding living organisms as *systems*. One of the key challenges in this study of systems biology is understanding how the static information recorded in the genome is interpreted to become dynamic systems of cooperating and competing biomolecules.

MAGNOME addresses this challenge through the development of informatic techniques for understanding the structure and history of eukaryote genomes: algorithms for genome analysis, data models for knowledge representation, stochastic hierarchical models for behavior of complex systems, and data mining and classification. Our work is in methods and algorithms for:

- **Genome annotation** for complete genomes, performing *syntactic* analyses to identify genes, and *semantic* analyses to map biological meaning to groups of genes [20], [6], [10], [11], [56], [57].
- **Integration of heterogenous data**, to build complete knowledge bases for storing and mining information from various sources, and for unambiguously exchanging this information between knowledge bases [1], [4], [41], [44], [29].
- **Ancestor reconstruction** using optimization techniques, to provide plausible scenarios of the history of genome evolution [11], [8], [46], [62].
- **Classification and logical inference**, to reliably identify similarities between groups of genetic elements, and infer rules through deduction and induction [9], [7], [10].
- **Hierarchical and comparative modeling**, to build mathematical models of the behavior of complex biological systems, in particular through combination, reutilization, and specialization of existing continuous and discrete models [40], [27], [60], [34], [59].

The hundred- to thousand-fold decrease in sequencing costs seen in the past few years presents significant challenges for data management and large-scale data mining. MAGNOME's methods specifically address "scaling out," where resources are added by installing additional computation nodes, rather than by adding more resources to existing hardware. Scaling out adds capacity and redundancy to the resource, and thus fault tolerance, by enforcing data redundancy between nodes, and by reassigning computations to existing nodes as needed.

3.2. Comparative genomics

The central dogma of evolutionary biology postulates that contemporary genomes evolved from a common ancestral genome, but the large scale study of their evolutionary relationships is frustrated by the unavailability of these ancestral organisms that have long disappeared. However, this common inheritance allows us to discover these relationships through *comparison*, to identify those traits that are common and those that are novel inventions since the divergence of different lineages.

We develop efficient methodologies and software for associating biological information with complete genome sequences, in the particular case where several phylogenetically-related eukaryote genomes are studied simultaneously.

The methods designed by MAGNOME for comparative genome annotation, structured genome comparison, and construction of integrated models are applied on a large scale to:

- eukaryotes from the hemiascomycete class of yeasts [56], [57], [6], [10], [2], [11] and to
- prokaryotes from the lactic bacteria used in winemaking [20], [23], [21], [32], [36], [28].

3.3. Comparative modeling

A general goal of systems biology is to acquire a detailed quantitative understanding of the dynamics of living systems. Different formalisms and simulation techniques are currently used to construct numerical representations of biological systems, and a recurring challenge is that hand-tuned, accurate models tend to be so focused in scope that it is difficult to repurpose them. We claim that, instead of modeling individual processes *de novo*, a sustainable effort in building efficient behavioral models must proceed incrementally. *Hierarchical modeling* 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 combined uses theoretical results from formal methods and practical considerations from modeling applications to define BioRica [26], [40], [60], a framework in which discrete and continuous models can communicate with a clear semantics. Hierarchical models in BioRica can be assembled from existing models, and translated into their execution semantics and then simulated at multiple resolutions through multi-scale stochastic simulation. BioRica models are compiled into a discrete event formalism capable of capturing discrete, continuous, stochastic, non deterministic and timed behaviors in an integrated and non-ambiguous way. Our long-term goal to develop a methodology in which we can **assemble a model** for a species of interest using a library of reusable models and a organism-level "schematic" determined by comparative genomics.

Comparative modeling is also a matter of reconciling experimental data with models [5] [27] and inferring new models through a combination of comparative genomics and successive refinement [51], [52].

4. Application Domains

4.1. Function and history of yeast genomes

Yeasts provide an ideal subject matter for the study of eukaryotic microorganisms. From an experimental standpoint, the yeast *Saccharomyces cerevisiae* is a model organism amenable to laboratory use and very widely exploited, resulting in an astonishing array of experimental results. From a genomic standpoint, yeasts from the hemiascomycete class provide a unique tool for studying eukaryotic genome evolution on a large scale. With their relatively small and compact genomes, yeasts offer a unique opportunity to explore eukaryotic genome evolution by comparative analysis of several species.

- Yeasts are widely used as cell factories, for the production of beer, wine and bread and more recently of various metabolic products such as vitamins, ethanol, citric acid, lipids, etc.
- Yeasts can assimilate hydrocarbons (genera *Candida*, *Yarrowia* and *Debaryomyces*), depolymerise tannin extracts (*Zygosaccharomyces rouxii*) and produce hormones and vaccines in industrial quantities through heterologous gene expression.
- Several yeast species are pathogenic for humans, especially *Candida albicans*, *Candida glabrata*, *Candida tropicalis* and the Basidiomycete *Cryptococcus neoformans*.

The hemiascomycetous yeasts represent a homogeneous phylogenetic group of eukaryotes with a relatively large diversity at the physiological and ecological levels. Comparative genomic studies within this group have proved very informative [33], [48], [47], [35], [50], [30], [31], [6].

MAGNOME applies its methods for comparative genomics and knowledge engineering to the yeasts through the ten-year old *Génolevures* program (GDR 2354 CNRS), devoted to large-scale comparisons of yeast genomes with the aim of addressing basic questions of molecular evolution. We developed the software tools used by the CNRS's genolevures.org web site. For example, MAGNOME's MAGUS system for simultaneous genome annotation combines semi-supervised classification and rule-based inference in a collaborative web-based system that explicitly uses comparative genomics to simultaneously analyse groups of related genomes.

4.2. Alternative fuels and bioconversion

Oleaginous yeasts are capable of synthesizing lipids from different substrates other than glucose, and current research is attempting to understand this conversions with the goal of optimizing their throughput, production and quality. From a genomic standpoint the objective is to characterize genes involved in the biosynthesis of precursor molecules which will be transformed into fuels, which are thus not derived from petroleum. Biological experimentation by partner laboratories study lipid accumulation the oleaginous yeasts such as *Yarrowia lipolytica* starting from:

- pentoses, produced from lignin cellulose agricultural substrates following a biorefining strategy,
- glycerol, a secondary output of chemical production of biodiesel, and
- industrial residues.

Lipases from *Y. lipolytica* are of particular interest (see [38] for review). Experimental characterization of the lipid bodies produced from these substrates will aid in the identification of target genes which may serve for genetic engineering. This in turn requires the development of molecular tools for this class of yeasts with strong industrial potential. MAGNOME's focus is in acquiring genome sequences, predicting genes using models learned from genome comparison and sequencing of cDNA transcripts, and comparative annotation. Our overall goal is to define dynamic models that can be used to predict the behavior of modified strains and thus drive selection and genetic engineering.

4.3. Winemaking and improved strain selection

Yeasts and bacteria are essential for the winemaking process, and selection of strains based both on their efficiency and on the influence on the quality of wine is a subject of significant effort in the Aquitaine region. Unlike the species studied above, yeast and bacterial starters for winemaking cannot be genetically modified. In order to propose improved and more specialized starters, industrial producers use breeding and selection strategies.

Yeast starters from the *Saccharomyces* genus are used for primary, alcohol fermentation. Recent advances have made it possible to identify the genetic causes of the different technological differences between strains [55], [54], [53]. Manipulating the genetic causes rather than the industrial consequences is far more amenable to experimental development. An essential tool in identifying these genetic causes is comparative genomics.

Bacterial starters based on *Oenococcus oeni* are used in secondary, malolactic fermentation. Genetically, *O. oeni* presents a surprising level of intra-specific diversity, and clues that it may evolve more rapidly than expected. Studying the diversity of the emphO. oeni genomes has led to genetic tools that can be used to evaluate the predisposition of different strains to respond to oenological stresses. While identifying particular genes has been the leading strategy up to now, recently a new strategy based on comparative genomics has been undertaken to understand the impact and mechanisms of genetic diversity [20], [23], [21], [32], [36], [28] [3].

Starting from historical collaborations by Pascal Durrens and Elisabeth Bon with partners from the Institute for Wine and Vine Sciences in Bordeaux (ISVV), and local industry, and in the framework of an effective partnership, we apply our tools to large-scale comparative genomics of yeast and bacterial starters in winemaking.

4.4. Knowledge bases for molecular tools

Affinity binders are molecular tools for recognizing protein targets, that play a fundamental in proteomics and clinical diagnostics. Large catalogs of binders from competing technologies (antibodies, DNA/RNA aptamers, artificial scaffolds, etc.) and Europe has set itself the ambitious goal of establishing a comprehensive, characterized and standardized collection of specific binders directed against all individual human proteins, including variant forms and modifications. Despite the central importance of binders, they presently cover only a very small fraction of the proteome, and even though there are many antibodies against some targets (for example, >900 antibodies against p53), there are none against the vast majority of proteins. Moreover, widely accepted standards for binder characterization are virtually nonexistent.

Alongside the technical challenges in producing a comprehensive binder resource are significant logistical challenges, related to the variety of producers and the lack of reliable quality control mechanisms. As part of the ProteomeBinders and Affinomics projects, MAGNOME works to develop knowledge engineering techniques for storing, exploring, and exchanging experimental data used in affinity binder characterization. This work involves databases and tools for molecular interaction data [42] [45], standards for data exchange between peers [44], [49], [41] and reporting standards [4] [58].

5. Software

5.1. Inria Bioscience Resources

Participants: Olivier Collin [correspondant], Frédéric Cazals, Mireille Régnier, Marie-France Sagot, Hélène Touzet, Hidde de Jong, David James Sherman, Marie-Dominique Devignes, Dominique Lavenier.

Inria Bioscience Resources is a portal designed to improve the visibility of bioinformatics tools and resources developed by Inria teams. This portal will help the community of biologists and bioinformatians understand the variety of bioinformatics projects in Inria, test the different applications, and contact project-teams. Eight project-teams participate in the development of this portal. Inria Bioscience Resources is developed in an Inria Technology Development Action (ADT).

5.2. Magus: Collaborative Genome Annotation

Participants: David James Sherman [correspondant], Pascal Durrens, Natalia Golenetskaya, Florian Lajus, Tiphaine Martin.

As part of our contribution the Génolevures Consortium, we have developed over the past few years an efficient set of tools for web-based collaborative annotation of eukaryote genomes. 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 the 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 Génolevures standards of high-quality manual annotation while efficiently using the time of our volunteer curators.

MAGUS is built on: a standard sequence feature database, the Stein lab generic genome browser [61], various biomedical ontologies (http://obo.sf.net), and a web interface implementing a representational state transfer (REST) architecture [39].

For more information see magus.gforge.inria.fr, the MAGUS Gforge web site. MAGUS is developed in an Inria Technology Development Action (ADT).

5.3. YAGA: Yeast Genome Annotation

Participants: Tiphaine Martin, Pascal Durrens [correspondant], Elisabeth Bon, Aurélie Goulielmakis.

With the arrival of new generations of sequencers, laboratories, at a lower cost, can be sequenced groups of genomes. You can no longer manually annotate these genomes. The YAGA (Yeast Automatic Genome Annotation) software's objective is to annotate a raw sequence syntactically and functionally as well as generate EMBL files for publication. The annotation takes into account data from comparative genomics, such as protein family profiles.

After determining the constraints of the annotation, the YAGA software can automatically annotate *de novo* all genomes from their raw sequences.The predictors used by the YAGA software can also take into account the data RNAseq to reinforce the prediction of genes.The current settings of the software are intended for annotation of the genomes of yeast, but the software is adaptable for all types of species, and has been trained and used for the annotation of bacterial genomes.

5.4. BioRica: Multi-scale Stochastic Modeling

Participants: David James Sherman [correspondant], Rodrigo Assar Cuevas.

BioRica is a high-level modeling framework integrating discrete and continuous multi-scale dynamics within the same semantics field. A model in BioRica node is hierarchically composed of nodes, which may be existing models. Individual nodes can be of two types:

- Discrete nodes are composed of states, and transitions described by constrained events, which can be non deterministic. This captures a range of existing discrete formalisms (Petri nets, finite automata, etc.). Stochastic behavior can be added by associating the likelihood that an event fires when activated. Markov chains or Markov decision processes can be concisely described. Timed behavior is added by defining the delay between an event's activation and the moment that its transition occurs.
- Continuous nodes are described by ODE systems, potentially a hybrid system whose internal state flows continuously while having discrete jumps.

The system has been implemented as a distributable software package

The BioRica compiler reads a specification for hierarchical model and compiles it into an executable simulator. The modeling language is a stochastic extension to the AltaRica Dataflow language, inspired by work of Antoine Rauzy. Input parsers for SBML 2 version 4 are curently being validated. The compiled code uses the Python runtime environment and can be run stand-alone on most systems [40].

For more information see biorica.gforge.inria.fr, the BioRica Gforge web site. BioRica was developed as an Inria Technology Development Action (ADT).

5.5. Pathtastic: Inference of whole-genome metabolic models

Participants: David James Sherman [correspondant], Pascal Durrens, Nicolás Loira, Tiphaine Martin, Anna Zhukova.

Pathtastic is a software tool for inferring whole-genome metabolic models for eukaryote cell factories. It is based on *metabolic scaffolds*, abstract descriptions of reactions and pathways on which inferred reactions are hung are are eventually connected by an interative mapping and specialization process. Scaffold fragments can be repeatedly used to build specialized subnetworks of the complete model.

Pathtastic uses a consensus procedure to infer reactions from complementary genome comparisons, and an algebra for assisted manual editing of pathways.

For more information see pathtastic.gforge.inria.fr, the Pathtastic Gforge web site.

5.6. Génolevures On Line: Comparative Genomics of Yeasts

Participants: Pascal Durrens [correspondant], Natalia Golenetskaya, Tiphaine Martin, David James Sherman.

The Génolevures online database provides tools and data for exploring the annotated genome sequences of more than 20 genomes, determined and manually annotated by the Génolevures Consortium to facilitate comparative genomic studies of hemiascomycetous yeasts. Data are presented with a focus on relations between genes and genomes: conservation of genes and gene families, speciation, chromosomal reorganization and synteny. The Génolevures site includes an area for specific studies by members of its international community.

Génolevures online uses the MAGUS system for genome navigation, with project-specific extensions developed by David Sherman, Pascal Durrens, and Tiphaine Martin. An advanced query system for data mining in Génolevures is being developed by Natalia Golenetskaya. The contents of the knowledge base are expanded and maintained by the CNRS through GDR 2354 Génolevures. Technical support for Génolevures On Line is provided the CNRS through UMR 5800 LaBRI.

For more information see genolevures.org, the Génolevures web site.

6. New Results

6.1. Yeast comparative genomics

Participants: Pascal Durrens [correspondant], Tiphaine Martin, David James Sherman.

By using MAGNOME's MAGUS system and YAGA software, we have successfully realized a full annotation and analysis of seven new genomes, provided to the Génolevures Consortium by the CEA–Génoscope (Évry)[15]. Two distant genomes from the *Debaryomycetaceae* and *mitosporic Saccharomycetales* clades of the *Saccharomycetales* were annotated using previously published Génolevures genomes [6], [10], [11] as references (in prep.). A further group of five species, comprised of pathogenic and nonpathogenic species, was analyzed with the goal of identifying virulence determinants [37]. By choosing species that are highly related but which differ in the particular traits that are targeted, in this case pathogenicity, we are able to focus of the few hundred genes related to the trait (in rev.). The approximately 40,000 new genes from these studies were classified into existing Génolevures families as well as branch-specific families.

In collaboration with partners in the ISVV, Bordeaux, we have assembled and analyzed 12 wine starter yeasts, with the goal of understanding genetic determinants of performance (in prep.).

6.2. Assembly, annotation and comparison of bacterial Omics data

Participants: Elisabeth Bon [correspondant], Laetitia Bourgeade, Pascal Durrens, Aurélie Goulielmakis, Tiphaine Martin, David James Sherman.

Oenococcus oeni is part of the natural microflora of wine and related environments, and is the main agent of the malolactic fermentation (MLF), a step of wine making that generally follows alcoholic fermentation (AF) and contributes to wine deacidification, improvement of sensorial properties and microbial stability. The start, duration and achievement of MLF are unpredictable since they depend both on the wine characteristics and on the properties of the *O. oeni* strains. In collaboration with Patrick Lucas's lab of the ISVV Bordeaux that is currently proceeding with genome sequencing, explorative and, and comparative genomics, Elisabeth Bon coordinates our efforts into the OENIKITA project (since 2009), a scale switching challenge including highthrouput exploratory and comparative genomics for oenological bacterial starters, and the development of an online web-collaborative multigenomic comparative platform based on the the Génolevures database architecture and MAGUS / YAGA systems.

OENI-Genomics: In comparative genomics, we investigated gene repertoire and genomic organization conservation through intra- and inter-species genomic comparisons, which clearly show that the *O. oeni* genome is highly plastic and fast-evolving. Results reveal that the optimal adaptation to wine of a strain mostly depends on the presence of key adaptive loops and polymorphic genes. They also point up the role of horizontal gene transfer and mobile genetic elements in O. oeni genome plasticity, and give the first clues of the genetic origin of its oenological aptitudes[3], [14], [29], [33], [35], [36]. As a result of the scaling out challenge, we participated to the assembly and annotation of 19 fully sequenced *O. oeni* genome variants.

KITA-Genomics (E. Bon, D. Sherman): This project that is focused on the sequencing, assembly, exploration and comparison of the *O. kitaharae* genome, has benefited to an international collaboration involving Dr V. Makeev. MAGNOME was involved into the pilot assembly, exploration and comparison of the *O. kitaharae* genome.

Transcriptomic axis (E. Bon, A. Goulielmakis, P. Durrens): Under the supervision of E. Bon, Aurélie Goulielmakis has completed for the ANR DIVOENI a detailed manual annotation of a new reference strain of *O. oeni* and performed comparative transcriptome analysis to identify genes differentially expressed under different culture conditions. We explored and compared how the expression system is solicited when *O. oeni* strains adapted to grow in some niches are placed under stress-exposure conditions. The monitoring of gene expression status between strains, through the definition of a global expression pattern proper to each gene, partially lift the veil on how *O. oeni* genome adapts function to its environment. The weight of genetic background and ecological niche pressure on gene expression flexibility was evaluated, and the *O. oeni* pan-transcriptome architecture characterized. The first guidelines revealed a supra-spatial organization of stress response into activated and repressed larger macro-domains defining functional landmarks and intra-chromosomal territories. Decryption of stress-sensitive gene repertoires promises to be an efficient tool in the conquest of *O. oeni* "domestication" through the identification of molecular markers responsible for different physiological capabilities, and the selection of the best adapted strains [21], [43].

Gene plasticity modelisation (E. Bon, L. Bourgeade): A novel axis of research recently emerged under the initiative of E. Bon (pseudOE project) around the detection, characterization and conservation of pseudogenes populations in *Oenococcus* bacteria. Such topic presents a double interest: phylogenetic at first because it should allow to better estimate the degree of genic/genomic plasticity of these bacteria, and algorithmic then because the pseudogenes are a source of confusion for the automatic prediction of genes. Through a transversal collaboration and a cooperative supervision with the Algorithms for Analysis of Biological Structures Group (P. Ferraro, J. Allali) at LaBRI, Laetitia Bourgeade (PhD, Univ. Bordeaux1) was recruited to develop dedicated methods to improve pseudogenes automatic detection, and therefore gene predictions, and to reconstruct fossil and modern genes evolutionary history [20], [23].

6.3. Big Data in comparative genomics

Participants: David James Sherman [correspondant], Pascal Durrens, Natalia Golenetskaya, Florian Lajus, Tiphaine Martin.

Data growth in comparative genomics presents a significant scaling challenge that requires novel informatic methods. Increase in sequence data is already a challenge, but in addition, the *relations* between the biological objects increase supralinearly (geometrically in the worst case) for every linear increase in sequence data.

MAGNOME's Tsvetok system proposes a highly-scalable distributed approach for data and computation in comparative genomics, targeting projects of the "comparative genomics of related species" type, where a set of genomes is sequenced and analyzed as part of the same process. Tsvetok combines a novel NoSQL storage schema with domain-specific MapReduce algorithms, to efficiently handle the fundamentally data-parallel analyses encountered in comparative genomics. Natalia Golenetskaya with Florian Lajus derived use cases from web site log analyses to identify standard queries, define an appropriate query-oriented storage schema, and map structured values to this schema. This was tested in MAGNOME's dedicated computing cluster.

Natalia Golenetskaya furthermore defined new distributed algorithms for two important large-scale analyses in MAGNOME's pipeline: systematic identification of gene fusion and fission events in eukaryote genomes (following [7]), and large-scale consensus clustering for protein families (following [9]). For fusions and fissions, she defined a new MapReduce algorithm that avoids graph-based analysis (which is notoriously slow in MapReduce), to achieve both significant speed ups and excellent scaling to much larger data sets. For protein family clustering, she defined a novel iterative sampling strategy that combines parallel clustering of submatrices of pairwise relations, to succesively approximate the result of a complete clustering, without the need to store the entire matrix of relations in memory.

6.4. Inferring metabolic models

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

In collaboration with Prof Jean-Marc Nicaud's lab at the INRA Grignon, we developed the first functional genome-scale metabolic model of an oleaginous yeast. Most work in producing genome-scale metabolic models has focused on model organisms, in part due to the cost of obtaining well-annotated genome sequences and sufficiently complete experimental data for refining and verifying the models. However, for many fungal genomes of biotechnological interest, the combination of large-scale sequencing projects and in-depth experimental studies has made it feasible to undertake metabolic network reconstruction for a wider range of organisms.

An excellent representative of this new class of organisms is *Yarrowia lipolytica*, an oleaginous yeast studied experimentally for its role as a food contaminant and its use in bioremediation and cell factory applications. As one of the hemiascomycetous yeasts completely sequenced in the Génolevures program it enjoys a high quality manual annotation by a network of experts. It is also an ideal subject for studying the role of species-specific expansion of paralogous families, a considerable challenge for eukaryotes in genome-scale metabolic construction. To these ends, we undertook a complete reconstruction of the *Y. lipolytica* metabolic network.

Methods: A draft model was extrapolated from the *S. cerevisiae* model iIN800, using *in silico* methods including enzyme conservation predicted using Génolevures and reaction mapping maintaining compartments. This draft was curated by a group of experts in *Y. lipolytica* metabolism, and iteratively improved and validated through comparison with experimental data by flux balance analysis. Gap filling, species-specific reactions, and the addition of compartments with the corresponding transport reactions were among the improvements that most affected accuracy. These methods, initially implemented in an *ad hoc* way in the *Pathtastic* software tool, have been redefined and formalized by Razanne Issa using a novel logical framework.

Results: We produced an accurate functional model for *Y. lipolytica*, MODEL1111190000 in Biomodels.net, that has been qualitatively validated against gene knockouts. This model has been enriched by Anna Zhukova with ontology terms from ChEBI and GO.

6.5. Summarized visualization of metabolic models

Participants: David James Sherman [correspondant], Anna Zhukova.

In collaboration with Romain Bourqui and Antoine Lambert of the LaBRI, we defined new strategies for exploring whole genome metabolic models. There is an inherent tension between detail and understandability in these large networks: on the one hand, detailed description of individual reactions is needed for accurate simulation, but on the other hand, high-level views of reactions are needed for describing partways in human terms. We are defining knowledge-based simplification rules, that permit the user to factor similar reactions into one "generic" reaction in order to visualize a whole pathway or compartment, while maintaining the underlying model so that the user can later "drill down" to the specific reactions if need be. New layout rules implemented in the Tulip platform are used to draw the resulting networks in a familiar way.

In collaboration with Bruno Pinaud of the LaBRI, rule-based rewriting of metabolic models was used to define these simplifications using his PORGY software tool.

6.6. Hierarchical modeling with BioRica

Participants: David James Sherman [correspondant], Rodrigo Assar Cuevas, Nicolás Loira.

A recurring challenge for *in silico* modeling of cell behavior is that experimentally validated models are so focused in scope that it is difficult to repurpose them. Hierarchical modeling 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.

BioRica is a high-level hierarchical modeling framework for models combining continuous and discrete components. By providing a reliable and functional software tool backed by a rigorous semantics, we hope to advance real adoption of hierarchical modeling by the systems biology community. By providing an understandable and mathematically rigorous semantics, this will make is easier for practicing scientists to build practical and functional models of the systems they are studying, and concentrate their efforts on the system rather than on the tool.

Building on previous work that formalized strategies for integrating discrete control with continuous models, Rodrigo Assar defined a new framework for BioRica models using Kaufman's Quantized State Systems (in prep.).

7. Bilateral Contracts and Grants with Industry

7.1. Contracts with Industry

SARCO, the research subsidiary of the Laffort group, has entered into a contract with MAGNOME to develop comparative genomics tools for selecting wine starters. This contract will permit SARCO to take a decisive step in the understanding of oenological microorganisms by obtaining and exploiting the sequences of their genomes. Comparison of the genomes of these strains has become absolutely necessary for learning the genetic origin of the phenotypic variations of oenological yeasts and bacteria. This knowledge will permit SARCO to optimize and accelerate the process of selection of the highest-performing natural strains. With the help of MAGNOME members and their rich experience in comparative analysis of related genomes, SARCO will acquire competence in biological analysis of genomic sequences. At the same time, MAGNOME members will acquire further experience with the genomes of winemaking microorganisms, which will help us define new tools and methods better adapted to this kind of industrial cell factory.

7.2. Grants with Industry

The French Petroleum Institute (*Institut français de pétrole-énergies nouvelles*) is coordinating a 6 M-Euro contract with the Civil Aviation Directorate (*Direction Générale de l'Aviation Civile*) 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. The CAER project studies both biofuel products and production, improved jet engine design, and the impact of aircraft. Within CAER MAGNOME via CNRS, works with partners from Grignon and Toulouse on the genomics of highly-performant oleaginous yeasts.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Aquitaine Region "SAGÉSS" comparative genomics for wine starters

Participants: David James Sherman [correspondant], Elisabeth Bon, Pascal Durrens, Aurélie Goulielmakis, Nicolás Loira, Tiphaine Martin.

This project is a collaboration between the company SARCO, specialized in the selection of industrial yeasts with distinct technological abilities, with the ISVV and MAGNOME. The goal is to use genome analysis to identify molecular markers responsible for different physiological capabilities, as a tool for selecting yeasts and bacteria for wine fermentation through efficient hybridization and selection strategies. This collaboration has obtained the INNOVIN label.

8.2. National Initiatives

8.2.1. ANR MYKIMUN

Participants: Pascal Durrens [correspondant], Witold Dyrka, David James Sherman.

Signal Transduction Associated with Numerous Domains (STAND) proteins play a central role in vegetative incompatibility (VI) in fungi. STAND proteins act as molecular switches, changing from closed inactive conformation to open active conformation upon binding of the proper ligand. Mykimun, coordinated by Mathieu Paoletti of the IBGC (Bordeaux), studies the postulated involvement of STAND proteins in heterospecific non self recognition (innate immune response). MAGNOMEdevelops machine learning techniques for classifying and identifying STAND proteins in fungal genomes, as well as statistical analysis of their genomic neighborhoods.

8.2.2. ANR DIVOENI, 2008-2012

Participants: Elisabeth Bon [correspondant], Aurélie Goulielmakis.

LaBRI, through Elisabeth Bon, is a partner in DIVOENI, a four-year ANR project concerning intraspecies biodiversity of the oenological bacteria *Oenococcus oeni*. Coordinated by Prof. Aline Lonvaud (Univ. Bordeaux Segalen) from the Institute of Vine and Wine Sciences of Bordeaux – Aquitaine, this scientific programme was developed:

1. To evaluate the genetic diversity of a vast collection of strains, to set up phylogenetic groups, then to investigate relationships between the ecological niches (cider, wine, champagne) and the essential phenotypical traits. Hypotheses on the evolution in the species and on the genetic stability of strains will be drawn.
2. To propose methods based on molecular markers to make a better use of the diversity of the species.
3. To measure the impact of the repeated use of selected strains on the diversity in the ecosystem and to draw the conclusions for its preservation.

Elisabeth is in charge of the computational infrastructure dedicated to genomics and post-genomics data storage, handling and analysis. She coordinates collaboration with the CBiB-Centre de Bioinformatique de Bordeaux (Aurélien Barré).

8.3. European Initiatives

8.3.1. FP7 Projects

8.3.1.1. Affinity Proteomics

Participants: David James Sherman [correspondant], Natalia Golenetskaya.

A major objective of the "post-genome" era is to detect, quantify and characterise all relevant human proteins in tissues and fluids in health and disease. This effort requires a comprehensive, characterised and standardised collection of specific ligand binding reagents, including antibodies, the most widely used such reagents, as well as novel protein scaffolds and nucleic acid aptamers. Currently there is no pan-European platform to coordinate systematic development, resource management and quality control for these important reagents.

MAGNOME is an associate partner of the FP7 "Affinity Proteome" project coordinated by Prof. Mike Taussig of the Babraham Institute and Cambridge University. Within the consortium, we participate in defining community for data representation and exchange, and evaluate knowledge engineering tools for affinity proteomics data.

8.3.2. Collaborations with Major European Organizations

Prof. Mike Taussig: Babraham Institute & Cambridge University

Knowledge engineering for Affinity Proteomics

Henning Hermjakob: European Bioinformatics Institute

Standards and databases for molecular interactions

8.4. International Initiatives

8.4.1. Inria Associate Teams

8.4.1.1. CARNAGE

Program: Inria-Russia

Title: CARNAGE: Combinatorics of Assembly and RNA in GEnomes

Inria principal investigator: Mireille Régnier

International Partner (Institution - Laboratory - Researcher):

State Research Institute of Genetics and Selection of Industrial Microorganisms (Russia (Russian Federation)) - Bioinformatics laboratory - Vsevolod Makeev

Duration: 2012–13

See also: http://en.inria.fr/domaines-epi/computational-sciences-for-biology-medicine-and-the-environment

CARNAGE addresses two main issues on genomic sequences, by combinatorial methods.

Fast development of high throughput technologies has generated a new challenge for computational biology. The recently appeared competing technologies each promise dramatic breakthroughs in both biology and medicine. At the same time the main bottlenecks in applications are the computational analysis of experimental data. The sheer amount of this data as well as the throughput of the experimental dataflow represent a serious challenge to hardware and especially software. We aim at bridging some gaps between the new "next generation" sequencing technologies, and the current state of the art in computational techniques for whole genome comparison. Our focus is on combinatorial analysis for NGS data assembly, interspecies chromosomal comparison, and definition of standard pipelines for routine large scale comparison.

This project also addresses combinatorics of RNA and the prediction of RNA structures, with their possible interactions.

8.4.2. Participation In International Programs

8.4.2.1. Génolevures and Dikaryome Consortia

Participants: David James Sherman [correspondant], Pascal Durrens, Florian Lajus, Tiphaine Martin, Anna Zhukova.

Since 2000 our team is a member of the Génolevures Consortium (GDR CNRS), a large-scale comparative genomics project that aims to address fundamental questions of molecular evolution through the sequencing and the comparison of 14 species of hemiascomycetous yeasts. The Consortium is comprised of 16 partners, in France, Belgium, Spain, the Netherlands (see http://genolevures.org/). Within the Consortium, our team is responsible for bioinformatics, for research in new methods of analysis. Pascal Durrens and Tiphaine Martin of the CNRS are responsible for the development of resources for exploiting comparative genomic data. Pascal Durrens is the editorial manager of the Génolevures on-line resource.

The Dikaryome Consortium is a scientific collaboration between several international partners and the National Center for Sequencing (CEA–Génoscope, Évry) on the sequencing, annotation, and comparative analysis of fungal genomes.

These perennial collaborations continue in two ways. First, a number of new projects are underway, concerning several new genomes currently being sequenced, and new questions about the mechanisms of gene formation. Second, through the development and improvement of the Génolevures On Line database, in whose maintenance our team has a longstanding committment and the improvement of tools like the YAGA software.

8.5. International Research Visitors

8.5.1. Visits of International Scientists

Rodrigo Assar Cuevas was invited for a month in Fall 2012 to work with David James Sherman on Quantized State Systems applied to BioRica hierarchical models.

8.5.2. Visits to International Teams

Anna Zhukova was invited to the Babraham Institute (Babraham, UK) for two week in December, 2012 to work on knowledge engineering for biological networks and visualization.

Pascal Durrens and David James Sherman are invited to the Vavilov Institute for General Genetics in Moscow in December, 2012 to work on regulon identification and analysis in hemisascomycete yeasts.

9. Dissemination

9.1. Scientific Animation

Elisabeth Bon is :

member of the “Comité Technique d’Etablissement” (since 2008 until 2014)

member of the “Comité Hygiène et Sécurité, et Conditions de Travail” (since 2012) at the Univ. of Bordeaux Segalen.

Pascal Durrens is :

leader of the “Comparative Genomics” theme and member of the Scientific Council of the LaBRI UMR 5800/CNRS.

responsible for scientific diffusion for the Génolevures Consortium.

member of the editorial board of the journal ISRN Computational Biology, and was reviewer for the journal BMC Genomics

expert in Genomics for the Fonds de la Recherche Scientifique-FNRS (FRS-FNRS), Belgium

Tiphaine Martin is :

member of the Local Committee and member of Local Committee for Occupational Health and Safety of the Inria Bordeaux Sud-Ouest.

member of the GIS-IBiSA GRISBI-Bioinformatics Grid working group.

member of the Institut de Grilles, and active in the Biology/Health working group.

David Sherman is :

president of the Commission de Jeunes Chercheurs, Inria Bordeaux Sud-Ouest

member of the editorial board of the journal Computational and Mathematical Methods in Medicine

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence : Elisabeth Bon, ICTs-Information & Communication Technologies (basic and advanced sections) and the national "IT and Internet certificate (C2i®, level 1) for the STS- biology variant Licence programs at the Univ. Bordeaux Segalen, and for MISMI Licence program at the Univ. Bordeaux 1. 109h éq. ED.

Licence & Master : Elisabeth Bon, Computer sciences & Bioinformatics-Genomics, Computerised resources, data banks and Methods for the Biology & Healthcare STS (Sciences, Technologies & Sante) bachelor's degrees, research oriented STS licence and master's degrees . 102h éq. ED.

Licence : Elisabeth Bon is responsable for The bachelor's degree "Information Technologies & Internet advanced course", Life Sciences Department, University Bordeaux

Licence : Elisabeth Bon is responsable for The "IT and Internet certificate (C2i®), level 1" at Life Sciences Department, University Bordeaux Segalen

Licence : Elisabeth Bon is responsable for The presidence (2005-2007; since sept. 2011) of the "IT and Internet certificate (C2i, level 1) committee" in charge of the C2i evaluation and certification for students (n=2000), University Bordeaux Segalen

Licence : Laetitia Bourgeade, Informatics for MISMI Licence program, University Bordeaux 1, 43h éq. ED.

Master : Laetitia Bourgeade, Statistics for bioinformaticians, University Bordeaux 1, 16h éq. ED.

Master : Laetitia Bourgeade, Object-oriented Programming, 2ème année Ingénieur, EnseirbMatmeca (Institut Polytechnique de Bordeaux), Bordeaux, 34h éq. ED.

Master : Laetitia Bourgeade, Methods & Tools for Systems Biology, 2ème année Ingénieur, Enseirb-Matmeca (Institut Polytechnique de Bordeaux), Bordeaux, 22h éq. ED. Tiphaine Martin and Pascal Durrens have :

the supervision of 4 Bioinformatics MSc students from the University of Bordeaux: Master : Development of search tools on Ge´nolevures databases, 6hETC, M1, University Bordeaux 1 and University Bordeaux Segalen, France

Master : Tiphaine Martin, Utilisation of EGEE GRID via virtual organisation GRISBI , 8h, niveau (M2), University Lyon, France

Master : Tiphaine Martin, Utilisation of EGEE GRID via virtual organisation GRISBI, 8h, niveau (M2), INRA Toulouse, France

Master : David Sherman, Web et Interfaces Homme-Machine, 50h, 2ème année Ingénieur, Enseirb-Matmeca (Institut Polytechnique de Bordeaux), Bordeaux

9.2.2. Supervision

PhD in progress : L. Bourgeade, Filtres sur les arborescences modélisant les ARN et plasticité génique, 2011–, E. Bon, P. Ferraro and J. Allali

PhD in progress : N. Golenetskaya, Big Data for comparative genomics, 2009–, D. Sherman

PhD in progress : R. Issa, Analyse symbolique de données génomiques, 2010–, D. Sherman

PhD in progress : A. Zhukova, Knowledge engineering for biological networks, 2011–, D. Sherman

9.2.3. Juries

David Sherman:

was external reviewer and member of the thesis defense jury for Anisah Ghoorah, Nancy.

was a member and president of the jury for the thesis defense of Jean-Paul Soularue, Bordeaux.

10. Bibliography

Major publications by the team in recent years

[1] R. BARRIOT, D. J. SHERMAN, I. DUTOUR. *How to decide which are the most pertinent overly-represented features during gene set enrichment analysis*, in "BMC Bioinformatics", 2007, vol. 8 [*DOI :* 10.1186/1471-2105-8-332], http://hal.inria.fr/inria-00202721/en/.

[2] G. BLANDIN, P. DURRENS, F. TEKAIA, M. AIGLE, M. BOLOTIN-FUKUHARA, E. BON, S. CASAREGOLA, J. DE MONTIGNY, C. GAILLARDIN, A. LÉPINGLE, B. LLORENTE, A. MALPERTUY, C. NEUVÉGLISE, O. OZIER-KALOGEROPOULOS, A. PERRIN, S. POTIER, J.-L. SOUCIET, E. TALLA, C. TOFFANO-NIOCHE, M. WÉSOLOWSKI-LOUVEL, C. MARCK, B. DUJON. *Genomic Exploration of the Hemiascomycetous Yeasts: 4. The genome of Saccharomyces cerevisiae revisited*, in "FEBS Letters", December 2000, vol. 487, n^{o} 1, p. 31-36.

[3] E. BON, A. DELAHERCHE, E. BILHÈRE, A. DE DARUVAR, A. LONVAUD-FUNEL, C. LE MARREC. *Oenococcus oeni genome plasticity is associated with fitness*, in "Applied and Environmental Microbiology", 2009, vol. 75, n^{o} 7, p. 2079-90, http://hal.inria.fr/inria-00392015/en/.

[4] J. BOURBEILLON, S. ORCHARD, I. BENHAR, C. BORREBAECK, A. DE DARUVAR, S. DÜBEL, R. FRANK, F. GIBSON, D. GLORIAM, N. HASLAM, T. HILTKER, I. HUMPHREY-SMITH, M. HUST, D. JUNCKER, M. KOEGL, Z. KONTHUR, B. KORN, S. KROBITSCH, S. MUYLDERMANS, P.-A. NYGREN, S. PALCY, B. POLIC, H. RODRIGUEZ, A. SAWYER, M. SCHLAPSHY, M. SNYDER, O. STOEVESANDT, M. J. TAUSSIG, M. TEMPLIN, M. UHLEN, S. VAN DER MAAREL, C. WINGREN, H. HERMJAKOB, D. J. SHERMAN. *Minimum information about a protein affinity reagent (MIAPAR).*, in "Nature Biotechnology", 07 2010, vol. 28, n^{o} 7, p. 650-3 [*DOI :* 10.1038/NBT0710-650], http://hal.inria.fr/inria-00544750/en.

[5] 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 yeast strains.*, in "Nature Communications", 12 2010, vol. 1, n^{o} 9, 145 [*DOI :* 10.1038/NCOMMS1150], http://hal.inria.fr/inria-00562005/en/.

[6] B. DUJON, D. J. SHERMAN, G. FISCHER, P. DURRENS, S. CASAREGOLA, I. LAFONTAINE, J. DE MONTIGNY, C. MARCK, C. NEUVÉGLISE, E. TALLA, N. GOFFARD, L. FRANGEUL, M. AIGLE, V. ANTHOUARD, A. BABOUR, V. BARBE, S. BARNAY, S. BLANCHIN, J.-M. BECKERICH, E. BEYNE, C. BLEYKASTEN, A. BOISRAMÉ, J. BOYER, L. CATTOLICO, F. CONFANIOLERI, A. DE DARUVAR, L. DESPONS, E. FABRE, C. FAIRHEAD, H. FERRY-DUMAZET, A. GROPPI, F. HANTRAYE, C. HENNEQUIN, N. JAUNIAUX, P. JOYET, R. KACHOURI-LAFOND, A. KERREST, R. KOSZUL, M. LEMAIRE, I. LESUR, L. MA, H. MULLER, J.-M. NICAUD, M. NIKOLSKI, S. OZTAS, O. OZIER-KALOGEROPOULOS, S. PELLENZ, S. POTIER, G.-F. RICHARD, M.-L. STRAUB, A. SULEAU, D. SWENNEN, F. TEKAIA, M. WÉSOLOWSKI-LOUVEL, E. WESTHOF, B. WIRTH, M. ZENIOU-MEYER, I. ZIVANOVIC, M. BOLOTIN-FUKUHARA, A. THIERRY, C. BOUCHIER, B. CAUDRON, C. SCARPELLI, C. GAILLARDIN, J. WEISSENBACH, P. WINCKER, J.-L. SOUCIET. *Genome evolution in yeasts*, in "Nature", 07 2004, vol. 430, n^{o} 6995, p. 35-44 [*DOI :* 10.1038/NATURE02579], http://hal.archives-ouvertes.fr/hal-00104411/en/.

[7] P. DURRENS, M. NIKOLSKI, D. J. SHERMAN. *Fusion and fission of genes define a metric between fungal genomes.*, in "PLoS Computational Biology", 10 2008, vol. 4, e1000200 [*DOI :* 10.1371/JOURNAL.PCBI.1000200], http://hal.inria.fr/inria-00341569/en/.

[8] A. GOËFFON, M. NIKOLSKI, D. J. SHERMAN. *An Efficient Probabilistic Population-Based Descent for the Median Genome Problem*, in "Proceedings of the 10th annual ACM SIGEVO conference on Genetic and evolutionary computation (GECCO 2008)", Atlanta United States, ACM, 2008, p. 315-322, http://hal.archives-ouvertes.fr/hal-00341672/en/.

[9] M. NIKOLSKI, D. J. SHERMAN. *Family relationships: should consensus reign?- consensus clustering for protein families*, in "Bioinformatics", 2007, vol. 23, p. e71–e76 [*DOI :* 10.1093/BIOINFORMATICS/BTL314], http://hal.inria.fr/inria-00202434/en/.

[10] D. J. SHERMAN, T. MARTIN, M. NIKOLSKI, C. CAYLA, J.-L. SOUCIET, P. DURRENS. *Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes.*, in "Nucleic Acids Research (NAR)", 2009, p. D550-D554 [*DOI :* 10.1093/NAR/GKN859], http://hal.inria.fr/inria-00341578/en/.

[11] J.-L. SOUCIET, B. DUJON, C. GAILLARDIN, M. JOHNSTON, P. 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 LOUIS, M. LEMAIRE, J. DE MONTIGNY, C. NEUVÉGLISE, A. THIERRY, I. BLANC-LENFLE, C. BLEYKASTEN, J. DIFFELS, E. FRITSCH, L. FRANGEUL, A. GOËFFON, 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, p. 1696-1709, http://hal.inria.fr/inria-00407511/en/.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[12] N. LOIRA. *Scaffold-based Reconstruction Method for Genome-Scale Metabolic Models*, Université Sciences et Technologies - Bordeaux I, January 2012, http://hal.inria.fr/tel-00678991.

Articles in International Peer-Reviewed Journals

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

[14] F. EL GHARNITI, M. DOLS-LAFARGUE, E. BON, O. CLAISSE, C. MIOT-SERTIER, A. LONVAUD-FUNEL, C. LE MARREC. *IS30 elements are mediators of genetic diversity in Oenococcus oeni.*, in "International Journal of Food Microbiology", August 2012, vol. 158, n^{o} 1, p. 14-22 [*DOI :* 10.1016/J.IJFOODMICRO.2012.06.009], http://hal.inria.fr/hal-00722965.

[15] V. LEH LOUIS, L. DESPONS, A. FRIEDRICH, T. MARTIN, P. DURRENS, S. CASAREGOLA, C. NEUVÉGLISE, C. FAIRHEAD, C. MARCK, J. A. CRUZ, M.-L. STRAUB, V. KUGLER, C. SACERDOT, Z.

UZUNOV, A. THIERRY, S. WEISS, C. BLEYKASTEN, J. DE MONTIGNY, N. JACQUES, P. JUNG, M. LEMAIRE, S. MALLET, G. MOREL, G.-F. RICHARD, A. SARKAR, G. SAVEL, J. SCHACHERER, M.-L. SERET, E. TALLA, G. SAMSON, C. JUBIN, J. POULAIN, B. VACHERIE, V. BARBE, E. PELLETIER, D. J. SHERMAN, E. WESTHOF, J. WEISSENBACH, P. BARET, P. WINCKER, C. GAILLARDIN, B. DUJON, J.-L. SOUCIET. *Pichia sorbitophila, an Interspecies Yeast Hybrid, Reveals Early Steps of Genome Resolution After Polyploidization*, in "Genes|Genomes|Genetics", February 2012, vol. 2, n^{o} 2, p. 299-311, the Consortium National de Recherche en Génomique (CNRG) to Génoscope, CNRS (GDR 2354, Génolevures), The computer framework was supported by the funding of the University of Bordeaux 1, the Aquitaine Région in the program "Génotypage et Génomique Comparée," and the ACI IMPBIO "Génolevures En Ligne." [*DOI :* 10.1534/G3.111.000745], http://hal.inria.fr/hal-00674198.

[16] N. LOIRA, T. DULERMO, J.-M. NICAUD, D. J. SHERMAN. *A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica.*, in "BMC Systems Biology", May 2012, vol. 6, n^{o} 1, 35 [*DOI :* 10.1186/1752-0509-6-35], http://hal.inria.fr/hal-00724286.

Invited Conferences

[17] T. MARTIN. *Retour d'expérience : assemblage*, in "Ecole NGS 2012 : De la production à l'analyse des données " Next Generation Sequencing "", Lyon, France, January 2012, http://hal.inria.fr/hal-00734501.

International Conferences with Proceedings

[18] R. ASSAR, M. A. MONTECINO, D. J. SHERMAN. *Stochastic Modeling of Complex Systems and Systems Biology: From Stochastic Transition Systems to Hybrid Systems*, in "XII Latin American Congress of Probability and Mathematical Statistics", Viña del Mar, Chile, March 2012, http://hal.inria.fr/hal-00686072.

[19] 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), Lecture Notes in Computer Science, Springer, 2012, vol. 6479, p. 68–83 [*DOI :* 10.1007/978-3-642-28067-2_6], http://hal.inria.fr/inria-00541215.

[20] L. BOURGEADE, T. MARTIN, E. BON. *PSEUDOE: A computational method to detect Psi-genes and explore PSEUDome dynamics in wine bacteria from the Oenococcus genus.*, in "JOBIM2012- 13ème Journées Ouvertes en Biologie, Informatique et Mathématiques", Rennes, France, F. COST, D. TAGU (editors), SFBI, Inria, July 2012, p. 435-436, http://hal.inria.fr/hal-00722968.

[21] A. GOULIELMAKIS, J. BRIDIER, A. BARRÉ, O. CLAISSE, DAVID JAMES. SHERMAN, P. DURRENS, A. LONVAUD-FUNEL, E. BON. *How does Oenococcus oeni adapt to its environment? A pangenomic oligonucleotide microarray for analysis O. oeni gene expression under wine shock.*, in "OENO2011- 9th International Symposium of Oenology", Bordeaux, France, P. DARRIET, L. GENY, P. LUCAS, A. LONVAUD-FUNEL, G. DE REVEL, P. TEISSEDRE (editors), Dunod, Paris, April 2012, p. 358-363, http://hal.inria.fr/hal-00646867.

[22] T. MARTIN, P. DURRENS. *Feedback in the management of genomic data in the GDR Génolevures*, in "ICY2012 13th International Congress on Yeasts", Madison, United States, August 2012, http://hal.inria.fr/hal-00734502.

National Conferences with Proceeding

[23] L. BOURGEADE, T. MARTIN, A. GOULIELMAKIS, A. LONVAUD-FUNEL, P. LUCAS, E. BON. *Cracking the Pseudome Code: Inside the "silent" Psi-genes language to reconstruct Oenococcus oeni evolutionary*

adaptation to wine, in "18th CBL-Club des Bactéries Lactiques Meeting", Clermont-Ferrand, France, UR CALITYSS - VETAGROSUP (editor), May 2012, page 82, http://hal.inria.fr/hal-00722971.

[24] T. MARTIN, P. DURRENS. *Génolevures : Politique d'annotation automatique des génomes de levures (YAGA)*, in "10e rencontre Levures Modèles & Outils", Toulouse, France, Jean-Luc PARROU, April 2012, http://hal.inria.fr/hal-00734500.

[25] T. MARTIN. *Retour d'expérience dans la gestion des données génomiques au sein du GDR Génolevures.*, in "10e rencontre Levures Modèles & Outils", Toulouse, France, Jean-Luc PARROU, April 2012, http://hal.inria.fr/hal-00734499.

References in notes

[26] R. ASSAR, A. GARCIA, D. J. SHERMAN. *Modeling Stochastic Switched Systems with BioRica*, in "Journées Ouvertes en Biologie, Informatique et Mathématiques JOBIM 2011", Paris, France, Institut Pasteur, July 2011, p. 297–304, http://hal.inria.fr/inria-00617419/en.

[27] R. ASSAR, F. VARGAS, D. J. SHERMAN. *Reconciling competing models: a case study of wine fermentation kinetics*, in "Algebraic and Numeric Biology 2010", Austria Hagenberg, K. HORIMOTO, M. NAKATSUI, N. POPOV (editors), Research Institute for Symbolic Computation, Johannes Kepler University of Linz, 08 2010, p. 68–83, http://hal.inria.fr/inria-00541215/en.

[28] A. ATHANE, E. BILHÈRE, E. BON, P. LUCAS, G. MOREL, A. LONVAUD-FUNEL, C. LE HÉNAFF-LE MARREC. *Characterization of an acquired-dps-containing gene island in the lactic acid bacterium Oenococcus oeni*, in "Journal of Applied Microbiology", 2008, Received 22 October 2007, revised 8 April 2008 & Accepted 8 May 2008 (In press), http://hal.inria.fr/inria-00340058/en/.

[29] R. BARRIOT, J. POIX, A. GROPPI, A. BARRE, N. GOFFARD, D. J. SHERMAN, I. DUTOUR, A. DE DARUVAR. *New strategy for the representation and the integration of biomolecular knowledge at a cellular scale*, in "Nucleic Acids Research (NAR)", 2004, vol. 32, p. 3581-9 [*DOI :* 10.1093/NAR/GKH681], http://hal.inria.fr/inria-00202722/en/.

[30] G. BLANDIN, P. DURRENS, F. TEKAIA, M. AIGLE, M. BOLOTIN-FUKUHARA, E. BON, S. CASAREGOLA, J. DE MONTIGNY, C. GAILLARDIN, A. LÉPINGLE, B. LLORENTE, A. MALPERTUY, C. NEUVÉGLISE, O. OZIER-KALOGEROPOULOS, A. PERRIN, S. POTIER, J.-L. SOUCIET, E. TALLA, C. TOFFANO-NIOCHE, M. WÉSOLOWSKI-LOUVEL, C. MARCK, B. DUJON. *Genomic Exploration of the Hemiascomycetous Yeasts: 4. The genome of Saccharomyces cerevisiae revisited*, in "FEBS Letters", December 2000, vol. 487, n^{o} 1, p. 31-36.

[31] E. BON, S. CASAREGOLA, G. BLANDIN, B. LLORENTE, C. NEUVÉGLISE, M. MUNSTERKOTTER, U. GULDENER, H.-W. MEWES, J. VAN HELDEN, B. DUJON, C. GAILLARDIN. *Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns.*, in "Nucleic Acids Research", February 2003, vol. 31, n^{o} 4, p. 1121-35, http://hal.inria.fr/inria-00339881/en/.

[32] E. BON, C. GRANVALET, F. REMIZE, D. DIMOVA, P. LUCAS, D. JACOB, A. GROPPI, S. PENAUD, C. AULARD, A. DE DARUVAR, A. LONVAUD-FUNEL, J. GUZZO. *Insights into genome plasticity of the wine-making bacterium Oenococcus oeni strain ATCC BAA-1163 by decryption of its whole genome.*, in "9th Symposium on Lactic Acid Bacteria", Egmond aan Zee Netherlands, 2008, http://hal.inria.fr/inria-00340073/en/.

[33] P. CLIFTEN, P. SUDARSANAM, A. DESIKAN, L. FULTON, R. S. FULTON, J. MAJORS, R. WATERSTON, B. A. COHEN, M. JOHNSTON. *Finding functional features in Saccharomyces genomes by phylogenetic footprinting*, in "Science", 2003, vol. 301, p. 71–76.

[34] M. CVIJOVIC, H. SOUEIDAN, D. J. SHERMAN, E. KLIPP, M. NIKOLSKI. *Exploratory Simulation of Cell Ageing Using Hierarchical Models*, in "19th International Conference on Genome Informatics Genome Informatics", Gold Coast, Queensland Australia, J. ARTHUR, S.-K. NG (editors), Genome Informatics, Imperial College Press, London, 2008, vol. 21, p. 114–125, EU FP6 Yeast Systems Biology Network LSHG-CT-2005-018942, EU Marie Curie Early Stage Training (EST) Network “Systems Biology”, ANR-05-BLAN-0331-03 (GENARISE), http://hal.inria.fr/inria-00350616.

[35] F. S. DIETRICH, ET AL.. *The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome*, in "Science", 2004, vol. 304, p. 304-7.

[36] D. DIMOVA, E. BON, P. LUCAS, R. BEUGNOT, M. DE LEEUW, A. LONVAUD-FUNEL. *The whole genome of Oenococcus strain IOEB 8413*, in "9th Symposium on Lactic Acid Bacteria", Egmond aan Zee Netherlands, 2008, http://hal.inria.fr/inria-00340086/en/.

[37] A. ENACHE-ANGOULVANT, J. GUITARD, F. GRENOUILLET, T. MARTIN, P. DURRENS, C. FAIRHEAD, C. HENNEQUIN. *Rapid Discrimination between Candida glabrata, Candida nivariensis, and Candida bracarensis by Use of a Singleplex PCR*, in "Journal of Clinical Microbiology", September 2011, vol. 49, n^{o} 9, p. 3375-3379 [*DOI :* 10.1128/JCM.00688-11], http://hal.inria.fr/inria-00625115/en.

[38] P. FICKERS, A. MARTY, J.-M. NICAUD. *The lipases from Yarrowia lipolytica: genetics, production, regulation, biochemical characterization and biotechnological applications*, in "Biotechnol Adv.", Nov-Dec 2011, vol. 29, n^{o} 6, p. 632–44.

[39] R. FIELDING, R. TAYLOR. *Principled design of the modern Web architecture*, in "ACM Trans. Internet Technol.", 2002, vol. 2, p. 115–150.

[40] A. GARCIA, D. J. SHERMAN. *Mixed-formalism hierarchical modeling and simulation with BioRica*, in "11th International Conference on Systems Biology (ICSB 2010)", United Kingdom Edimbourg, 10 2010, P02.446, Poster, http://hal.inria.fr/inria-00529669/en.

[41] D. GLORIAM, S. ORCHARD, D. BERTINETTI, E. BJÖRLING, E. BONGCAM-RUDLOFF, C. BORREBAECK, J. BOURBEILLON, A. R. M. BRADBURY, A. DE DARUVAR, S. DÜBEL, R. FRANK, T. J. GIBSON, L. GOLD, N. HASLAM, F. W. HERBERG, T. HILTKER, J. D. HOHEISEL, S. KERRIEN, M. KOEGL, Z. KONTHUR, B. KORN, U. LANDEGREN, L. MONTECCHI-PALAZZI, S. PALCY, H. RODRIGUEZ, S. SCHWEINSBERG, V. SIEVERT, O. STOEVESANDT, M. J. TAUSSIG, M. UEFFING, M. UHLÉN, S. VAN DER MAAREL, C. WINGREN, P. WOOLLARD, D. J. SHERMAN, H. HERMJAKOB. *A community standard format for the representation of protein affinity reagents.*, in "Mol Cell Proteomics", 01 2010, vol. 9, n^{o} 1, p. 1-10 [*DOI :* 10.1074/MCP.M900185-MCP200], http://hal.inria.fr/inria-00544751/en.

[42] N. GOLENETSKAYA, D. J. SHERMAN. *Assessing ”last mile” tools for affinity binder databases*, in "5th ESF Workshop on Affinity Proteomics: Ligand Binders against the Human Proteome", Alpbach, Austria, March 2011, http://hal.inria.fr/hal-00653518/en.

[43] A. GOULIELMAKIS, J. BRIDIER, A. BARRÉ, O. CLAISSE, D. J. SHERMAN, P. DURRENS, A. LONVAUD-FUNEL, E. BON. *How does Oenococcus oeni adapt to its environment? A pangenomic oligonucleotide*

microarray for analysis O. oeni gene expression under wine shock., in "OENO2011- 9th International Symposium of Oenology", Bordeaux, France, Dunod, Paris, January 2011, http://hal.inria.fr/hal-00646867/en.

[44] H. HERMJAKOB, L. MONTECCHI-PALAZZI, G. BADER, J. WOJCIK, L. SALWINSKI, A. CEOL, S. MOORE, S. ORCHARD, U. SARKANS, C. VON MERING, B. ROECHERT, S. POUX, E. JUNG, H. MERSCH, P. KERSEY, M. LAPPE, Y. LI, R. ZENG, D. RANA, M. NIKOLSKI, H. HUSI, C. BRUN, K. SHANKER, S. GRANT, C. SANDER, P. BORK, W. ZHU, A. PANDEY, A. BRAZMA, B. JACQ, M. VIDAL, D. J. SHERMAN, P. LEGRAIN, G. CESARENI, I. XENARIOS, D. EISENBERG, B. STEIPE, C. HOGUE, R. APWEILER. *The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data*, in "Nat. Biotechnol.", Feb. 2004, vol. 22, n^{o} 2, p. 177-83.

[45] H. HERMJAKOB, L. MONTECCHI-PALAZZI, C. LEWINGTON, S. MUDALI, S. KERRIEN, S. ORCHARD, M. VINGRON, B. ROECHERT, P. ROEPSTORFF, A. VALENCIA, H. MARGALIT, J. ARMSTRONG, A. BAIROCH, G. CESARENI, D. J. SHERMAN, R. APWEILER. *IntAct: an open source molecular interaction database*, in "Nucleic Acids Res.", Jan. 2004, vol. 32, p. D452-5.

[46] G. JEAN, D. J. SHERMAN, M. NIKOLSKI. *Mining the semantics of genome super-blocks to infer ancestral architectures*, in "Journal of Computational Biology", 2009, http://hal.inria.fr/inria-00414692/en/.

[47] M. KELLIS, B. BIRREN, E. LANDER. *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae*, in "Nature", 2004, vol. 428, p. 617-24.

[48] M. KELLIS, N. PATTERSON, M. ENDRIZZI, B. BIRREN, E. LANDER. *Sequencing and comparison of yeast species to identify genes and regulatory elements*, in "Nature", 2003, vol. 423, p. 241–254.

[49] S. KERRIEN, S. ORCHARD, L. MONTECCHI-PALAZZI, B. ARANDA, A. QUINN, N. VINOD, G. BADER, I. XENARIOS, J. WOJCIK, D. J. SHERMAN, M. TYERS, J. SALAMA, S. MOORE, A. CEOL, A. CHATR-ARYAMONTRI, M. OESTERHELD, V. STUMPFLEN, L. SALWINSKI, J. NEROTHIN, E. CERAMI, M. CUSICK, M. VIDAL, M. GILSON, J. ARMSTRONG, P. WOOLLARD, C. HOGUE, D. EISENBERG, G. CESARENI, R. APWEILER, H. HERMJAKOB. *Broadening the Horizon - Level 2.5 of the HUPO-PSI Format for Molecular*, in "BMC Biology", 10 2007, vol. 5, 9;5(1):44, http://hal.archives-ouvertes.fr/hal-00306554/en/.

[50] R. KOSZUL, S. CABURET, B. DUJON, G. FISCHER. *Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments*, in "EMBO Journal", 2004, vol. 23, n^{o} 1, p. 234-43.

[51] N. LOIRA, T. DULERMO, M. NIKOLSKI, J.-M. NICAUD, D. J. SHERMAN. *Genome-scale Metabolic Reconstruction of the Eukaryote Cell Factory Yarrowia Lipolytica*, in "11th International Conference on Systems Biology (ICSB 2010)", United Kingdom Edimbourg, 10 2010, P02.602, Poster, http://hal.inria.fr/hal-00652922.

[52] N. LOIRA, D. J. SHERMAN, P. DURRENS. *Reconstruction and Validation of the genome-scale metabolic model of Yarrowia lipolytica iNL705*, in "Journée Ouvertes Biologie Informatique Mathématiques, JOBIM 2010", France Montpellier, 09 2010, http://www.jobim2010.fr/?q=fr/node/55.

[53] P. MARULLO, C. MANSOUR, M. DUFOUR, W. ALBERTIN, D. SICARD, M. BELY, D. DUBOURDIEU. *Genetic improvement of thermo-tolerance in wine Saccharomyces cerevisiae strains by a backcross approach*, in "FEMS Yeast Res", 12 2009, vol. 9, n^{o} 8, p. 1148–60.

[54] P. MARULLO, G. YVERT, M. BELY, I. MASNEUF-POMARÈDE, P. DURRENS, M. AIGLE. *Single QTL mapping and nucleotide-level resolution of a physiologic trait in wine Sacchar omyces cerevisiae strains*, in "FEMS Yeast Res.", 2007, vol. 7, n^{o} 6, p. 941–52.

[55] I. MASNEUF-POMARÈDE, C. LEJEUNE, P. DURRENS, M. LOLLIER, M. AIGLE, D. DUBOURDIEU. *Molecular typing of wine yeast strains Saccharomyces uvarum using microsatellite ma rkers*, in "Syst. Appl. Microbiol.", 2007, vol. 30, n^{o} 1, p. 75–82.

[56] D. J. SHERMAN, P. DURRENS, E. BEYNE, M. NIKOLSKI, J.-L. SOUCIET. *Génolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts.*, in "Nucleic Acids Research (NAR)", 2004, vol. 32, p. D315-8, GDR CNRS 2354 "Génolevures" [*DOI :* 10.1093/NAR/GKH091], http://hal.inria.fr/inria-00407519/en/.

[57] D. J. SHERMAN, P. DURRENS, F. IRAGNE, E. BEYNE, M. NIKOLSKI, J.-L. SOUCIET. *Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts.*, in "Nucleic Acids Res", 01 2006, vol. 34, n^{o} Database issue, p. D432-5 [*DOI :* 10.1093/NAR/GKJ160], http://hal.archives-ouvertes.fr/hal-00118142/en/.

[58] D. J. SHERMAN, N. GOLENETSKAYA. *Databases and Ontologies for Affinity Binders*, 05 2010, Overview of advances in defining ontologies and building knowledge bases for affinity binders, over the four years of the ProteomeBinders project. Presented at the Affinomics/ProteomeBinders workshop at the Møller Center, Churchill College, Cambridge University., http://hal.inria.fr/inria-00563531/en/.

[59] H. SOUEIDAN, M. NIKOLSKI, G. SUTRE. *Qualitative Transition Systems for the Abstraction and Comparison of Transient Behavior in Parametrized Dynamic Models*, in "Computational Methods in Systems Biology (CMSB'09)", Italie Bologna, Springer Verlag, 2009, vol. 5688, p. 313–327, http://hal.archives-ouvertes.fr/hal-00408909/en/.

[60] H. SOUEIDAN, D. J. SHERMAN, M. NIKOLSKI. *BioRica: A multi model description and simulation system*, in "F0SBE", Allemagne, 2007, p. 279-287, http://hal.archives-ouvertes.fr/hal-00306550/en/.

[61] L. D. STEIN. *The Generic Genome Browser: A building block for a model organism system database*, in "Genome Res.", 2002, vol. 12, p. 1599-1610.

[62] N. VYAHHI, A. GOËFFON, D. J. SHERMAN, M. NIKOLSKI. *Swarming Along the Evolutionary Branches Sheds Light on Genome Rearrangement Scenarios*, in "ACM SIGEVO Conference on Genetic and evolutionary computation", F. ROTHLAUF (editor), ACM, 2009, http://hal.inria.fr/inria-00407508/en/.