INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team magnome

Models and Algorithms for the Genome

Bordeaux - Sud-Ouest

Theme : Computational Biology and Bioinformatics

Table of contents

MAGNOME is an INRIA Project-Team joint with University of Bordeaux and CNRS (LaBRI, UMR 5800)

1. Team

Research Scientists

David James Sherman [Team leader; Senior Researcher (DR), HdR]
Pascal Durrens [Junior Researchert (CR), HdR]
Macha Nikolski [Junior Researcher (CR); until 2010-08-31, HdR]

Faculty Member

Elisabeth Bon [University Bordeaux, Associate Professor (MCF)]

Technical Staff

Tiphaine Martin [Research engineer (IR)]
Alice Garcia [Contract engineer for BioRica ADT]
Aurélie Goulielmakis [Contract engineer for ANR DIVOENI]

PhD Students

Rodrigo Assar-Cuevas [CORDI-S INRIA]
Natalia Golenetskaya [CORDI-S INRIA]
Razanne Issa [Exchange Fellowship Syria]
Nicolás Loira [CONICYT Chile]
Anasua Sarkar [EMMA co-reg. Jadavpur University]

Administrative Assistants

Marie Sanchez [until 2010-10-07]
Anne-Laure Gautier [since 2010-10-08]

Other

Anna Zhukova [Masters student intern since 2010-08-10]

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.

2.2. Highlights

As a result of a four-year collaboration under the auspices of the Yeast Systems Biology Network (YSBN), our consortium completed an integrated multilaboratory systems biology study involving ten research centers and 11 experimental techniques. Analysis of this high-quality, systematic dataset revealed insights into protein metabolism, and was published in *Nature Communications*.

In collaboration with Prof Jean-Marc Nicaud's lab at the INRA Grignon, we developed the first functional whole-genome metabolic model for an obligate aerobic, oleaginous yeast, *Yarrowia lipolytica*. This will provide an important mathematical tool in developing high-quality products through bioconversion of industrial wastes.

We successfully completed and released the MIAPAR and PSI-PAR international standards for knowledge representation and data exchange of affinity binder properties, a five-yeasr effort organized as part of the ProteomeBinders and HUPO-PSI consortia. These standards were reported in *Nature Biotechnology* and *Molecular and Cellular Proteomics* to the research community and are being adopted by European bioinformatics data centers and industry.

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 [5], [9], [10], [48], [49].
- **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], [11], [13], [38], [30].
- **Ancestor reconstruction** using optimization techniques, to provide plausible scenarios of the history of genome evolution [10], [7], [40], [56].
- **Classification and logical inference**, to reliably identify similarities between groups of genetic elements, and infer rules through deduction and induction [8], [6], [9].
- **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 [23], [21], [53], [33], [52].

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 a software platform, 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 [48], [49], [5], [9], [2] and to
- prokaryotes from the lactic bacteria used in winemaking [31], [36], [29].

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 certain wealth of models is proposed using specific and *ad hoc* methods. A recurring challenge is that hand-tuned, accurate models tend to be so focused in scope that it is difficult to repurpose them. Instead of modeling individual processes individually *de novo*, we claim that 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.

MAGNOME uses theoretical results from formal methods to define a mathematical framework in which discrete and continuous models can communicate with a clear semantics. We exploit this to develop the BioRica platform [23], [53], a **modeling middleware** in which hierarchical models can be assembled from existing models. Such models are 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 [12], [21] and inferring new models through a combination of comparative genomics and successive refinement [24], [25].

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 [32], [42], [41], [35], [44], [2], [5].

MAGNOME applies its methods for comparative genomics and knowledge engineering to the yeasts through the ten-year old *Génolevures* program, devoted to large-scale comparisons of yeast genomes with the aim of addressing basic questions of molecular evolution. We provide the tools behind the world-renowned genolevures.org web site.

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 partners from the CAER project 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.

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.

The strategy defined with our partners proceeds with the complete sequencing and annotation of selected species. This will entail acquiring genome sequences, predicting gene models, and annotation of the resulting genes. As these yeasts are known to contain multi-intron genes, we will also sequence cDNA transcripts to refine gene models. The Magus system developed by MAGNOME will be used for this annotation and for the constitution of a genome database. The use of Magus will permit us to integrate quantitative transcriptome data, and use it to refine a functional flux-balance analysis model developed for *Yarrowia lipolytica* by the MAGNOME team.

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 [47], [46], [45]. 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 O. 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 [31], [36], [29].

MAGNOME works with partners from the Institute for Wine and Vine Sciences in Bordeaux, as well as local industry, to 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 for storing molecular interaction data [39], standards for data exchange between peers [38], [43],[13] and reporting standards [11], [27].

5. Software

5.1. Magus: Collaborative Genome Annotation

Participants: David James Sherman [correspondant], Pascal Durrens, Natalia Golenetskaya, Anna Zhukova, 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 Génolevures annotation workflow 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 [55], various biomedical ontologies (http://obo.sf.net), and a web interface implementing a representational state transfer (REST) architecture [37].

For more information see magus.gforge.inria.fr, the Magus Gforge web site.

5.2. Faucils: Analyzing Genome Rearrangement

Participants: Macha Nikolski, David James Sherman [correspondant], Tiphaine Martin.

The Faucils suite uses evolutionary and combinatory algorithms to facilitate mathematical exploration of eukaryote genome rearrangement. It is composed of a number of cooperating tools: SyDIG, a method for detecting synteny in distantly related genomes; SuperBlocks, a method for computing ancestral superblocks; Faucils, tools for computing median genomes and rearrangement trees using stochastic local search and any colony optimization; and Virage, an tools for interactive visual exploration of divergent rearrangement scenarios.

For more information see faucils.gforge.inria.fr, the Faucils Gforge web site.

5.3. BioRica: Multi-scale Stochastic Modeling

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

BioRica is a high-level modeling framework integrating discrete and continuous multi-scale dynamics within the same semantics field. The co-existence of continuous and discrete dynamics is assured by a pre-computation of the continuous parts of the model. Once computed, these parts of the model act as components that can be queried for the function value, but also modified, therefore accounting for any trajectory modification induced by discrete parts of the model. To achieve this we extensively rely on methods for solving and simulation of continuous systems by numerical algorithms. Discrete nodes in the model act as controllers.

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.

For more information see biorica.gforge.inria.fr, the BioRica Gforge web site. BioRica is an INRIA Technology Development Action (ADT).

5.4. Pathtastic: Inference of whole-genome metabolic models

Participants: David James Sherman, Pascal Durrens, Nicolás Loira [correspondant].

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.

5.5. Génolevures On Line: Comparative Genomics of Yeasts

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

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.

While extensive chromosomal rearrangements combined with segmental and massive duplications make comparisons of yeast genome sequences difficult [51], relations of homology between protein-coding genes can be identified despite their great diversity at the molecular level [5]. Families of homologous proteins provide a powerful tool for appreciating conservation, gain and loss of function within yeast genomes. Génolevures provides a unique collection of paralogous and orthologous protein families, identified using a novel consensus clustering algorithm [8] applied to a complementary set of homeomorphic [sharing full-length sequence similarity and similar domain architectures, see [57]] and nonhomeomorphic systematic Smith-Waterman [50] and Blast [28] sequence alignments. Similar approaches are developed on a wider scale [57] and are complementary to these yeast-specific families.

The Génolevures database uses a bespoke object model mapped to a relational database. Flexibility in the design is guaranteed through the use of ontologies and controlled vocabularie Browsing of genomic maps and sequence features is provided by the Generic Genome Browser [55]. The Blast service is provided by NCBI Blast. The Génolevures web site uses a REST architecture internally [37] and extensively uses the BioPerl package [54] for manipulation of sequence data.

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

6. New Results

6.1. Yeast comparative genomics

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

Using the Magus comparative genome annotation system, we have successfully realized a full annotation and analysis of seven new genomes, provided to the Génolevures Consortium by the CEA - Génoscope (Évry). Two distant genomes from the *Debaryomycetaceae* and *mitosporic Saccharomycetales* clades of the *Saccharomycetales* were annotated using previously published Génolevures genomes[5] as references. A further group of five species, comprised of pathogenic and nonpathogenic species, was analyzed with the goal of identifying virulence determinants. 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. The approximately 40,000 genes from these studies were classified into existing Génolevures families as well as branch-specific families. The results from these two studies will be published in the coming year.

6.2. Assembly and annotation of Oenococcus strains

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

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. Elisabeth Bon of MAGNOME coordinates collaboration with Patrick Lucas's lab of the ISVV Bordeaux that is currently proceeding with genome sequencing, explorative and comparative genome data analysis, and comparative genomics. Aurélie Goulielmakis of MAGNOME has completed a detailed manual annotation of a new reference strain of *O. oeni* and used transcriptome analysis to identify genes differentially expressed under different culture conditions. 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. Preliminary results reveal that the optimal adaptation to wine of a strain mostly depends on the presence of key adaptative 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.

6.3. Interlaboratory systems biology

Participants: David James Sherman [correspondant], Macha Nikolski.

As a result of a four-year collaboration under the auspices of the Yeast Systems Biology Network (YSBN), our consortium completed an integrated multilaboratory systems biology study involving ten research centers and 11 experimental techniques. Analysis of this high-quality, systematic dataset revealed insights into protein metabolism, and was published in *Nature Communications* [12].

The study proceeded through four steps. First, to avoid the problems that auxotrophic strains (like the standard reference CEN.PK) have in studying cell physiology, the consortium designed two new prototrophic strains (YSBN1 and YSBN2) that carry drug resistance markers inserted into their genomes at a phenotypically neutral site. Then, CEN.PK and YSBN2 were grown under two conditions in a single lab. Next, a combination of –omics platforms in the ten participating laboratories were used to measure transcriptomes, proteomes, endo-metabolomes, and exo-metabolomes, with the same techniques used in different laboratories with the same or different protocols. Finally, integrated analysis was used to systematically compare the results and assess reproducibility. This study demonstrates that integrative analysis of complementary datasetscan reveal molecular explanations for observed phenotypes, which would not be possible from a single –omics data set. As an added benefit, this comparison across labs and across methods provides a very useful reference dataset.

6.4. Functional model for Y. lipolytica

Participants: David James Sherman, Pascal Durrens, Macha Nikolski, Nicolás Loira [correspondant].

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.

Results: We produced an accurate functional model for *Y. lipolytica*, iNL705, that includes 705 genes in six compartments[24], [25]. This model is made available in SBML format. We were further able to assess the role of expansion of paralogous families in robustness against gene knockouts.

6.5. Hierarchical modeling with BioRica

Participants: David James Sherman [correspondant], Pascal Durrens, Alice Garcia, Rodrigo Assar-Cuevas.

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.

The BioRica high-level hierarchical modeling framework expresses each existing model as a BioRica node, which are hierarchically composed to build a BioRica system. 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 [23]. The BioRica model compiler and associated tools are available from biorica.gforge.inria.fr.

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.

These approaches are being applied to consensus models of alcohol fermentation developed by Rodrigo Assar [21].

6.6. New Tsvetok data storage model

Participants: David James Sherman, Pascal Durrens, Natalia Golenetskaya [correspondant], Anna Zhukova.

The hundred- to thousand-fold decrease in sequencing costs we have seen in the past four years presents significant challenges for data management and large-scale data mining. The Tsvetok project in MAGNOME, led by Natalia Golenetskaya, targets improvements in the capacity for handling large volumes of data, to permit more automatic analysis of projects of the "comparative genomics of related species" type, where a set of genomes is sequenced and analyzed as part of the same process. This will in turn permit the routine use of this new genome sequencing strategy by smaller groups who are focused on the biological goal and not interested in the development of new in silico analysis methods. The industrial context provided by the CAER and SAGESSE projects will help us refine and validate our new methods. Tsvetok specifically addresses "scaling out," where resources are added by installing additional computation nodes, rather than by adding more resources to existing hardware. Scaling out adds capacity to the resource, of course, but also adds redundancy and hence robustness. To cite a common example, if we assume an installation of 1000 nodes and a mean time to failure of approximately three years, then statistically we should expect one machine to fail every day. A robust architecture minimizes the effect of such a failure, by enforcing data redundancy between nodes, and by reassigning computations to existing nodes as needed.

Natalia Golenetskaya has designed and implemented a NoSQL prototype through the identification of standard queries, definition of the appropriate query-oriented storage schema, and mapping of structured values to this schema. This prototype is being tested on an Apache Cassandra ring deployed in MAGNOME's dedicated computing cluster.

Large-scale data-mining such as that required for comparative genomics is fundamentally *data-parallel*: an initial transformation is applied to every data object of a given type (such as genes or even individual nucleotides), then a statistical machine learning procedure is applied to the transformed data to produce a summary or to learn a classification function. Analyses of this kind are the design goal of the Map-Reduce algorithmic technique [34]. Using Tsvetok as a generator for Apache Hadoop, Natalia is designing Map-Reduce solutions for the principal whole-genome analyses used by MAGNOME for eukaryote and prokaryote comparative genomics.

6.7. Standards for affinity binders

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

We successfully completed and released the MIAPAR and PSI-PAR international standards for knowledge representation and data exchange of affinity binder properties, a five-year effort organized as part of the ProteomeBinders and HUPO-PSI consortia. These standards were reported in *Nature Biotechnology* and *Molecular and Cellular Proteomics* to the research community [11], [13], [27] and are being adopted by European bioinformatics data centers and industry. They are extension to previous work such as [38], [43].

7. 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 and its rich experience in comparative analysis of related genomes, SARCO will acquire competence in biological analysis of genomic sequences. At the same time, MAGNOME 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, Turboméca, 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 works with partners from Grignon and Toulouse on the genomics of highly-performant oleaginous yeasts.

8. Other Grants and Activities

8.1. Regional Initiatives

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

Participants: David Sherman [correspondant], Pascal Durrens.

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 DIVOENI

Participant: Elisabeth Bon [correspondant].

Elisabeth Bon of MAGNOME is a partner in DIVOENI, a four-year ANR project concerning intraspecies biodiversity of *Oenococcus oeni*, a lactic acid bacterium of wine. Coordinated by Prof. Aline Lonvaud of the Université Victor Ségalen Bordeaux 2, the aims of the programme are: 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 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.

8.2.2. INRA-INRIA Oleaginous Yeasts

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

We have been woking with the research teams of Cécile Neuvéglise and Jean-Marc Nicaud at the INRIA Grignon, on analysis and modeling of oleaginous yeasts and their genomes. We have performed genome sequence surveys of several related species and are developing a consensus metabolic model for species in the *Yarrowia* clade. These activities will continue in the context of the CAER (Alternative Fuels for Aeronautics) project funded by the French DGAC.

8.3. European Initiatives

8.3.1. ProteomeBinders (FP6) and Affinomics (FP7)

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.

The ProteomeBinders FP6 Coordination Action (`proteomebinders.org`) coordinated 26 European partners and two in the USA, several of which operate infrastructures or large scale projects in aspects including cDNA collections, protein production, polyclonal and monoclonal antibodies. Together the consortium members provide a critical mass of leading expertise in binder technology, protein expression, binder applications and bioinformatics. Many have tight links to SMEs in binder technology, as founders or advisors. The consortium will organisse the resource by integrating the existing infrastructures, reviewing technologies and high throughput production methods, standardising binder-based tools and applications, assembling the necessary bioinformatics and establishing a database schema to set up a central binders repository. The followup project Affinomics in FP7 puts this infrastructure into production.

Within the consortium, we are responsible for formalizing an ontology of binder properties and a set of requirements for data representation and exchange, and generally involved in the standards-making process.

8.3.2. COPY, Comparative Genomics of Yeasts

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

With Teun Boekhout of the CBS (Utrecht) and Toni Gabaldón of the CRG (Barcelona), we are organizing COPY, an European consortium to build an efficient platform for training Europe-wide best practices for sequencing, annotating, analysing and modelling of biotechnologically and medically interesting microorganisms. The proposed team builds on longstanding European strengths in yeast comparative genomics that has been developed by world class academic and industrial partners. To facilitate inter-disciplinary training and diversify future career options, COPY is organized around a core methodological backbone (sequencing technologies, bioinformatics analyses, translation of research), and two applied pillars that focus on particular groups of clinically and industrially relevant fungi. The COPY consortium builds on existing European relations through the Génolevures and Dikaryome consortia, as well as a regular series of biannual conferences organized through EMBO.

8.4. International Initiatives

8.4.1. HUPO Proteomics Standards Initiative

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

We participate actively in the Proteomics Standards Initiative (PSI) of the Human Proteome Organization (HUPO), and international structure for the development and the advancement of technologies for proteomics. The HUPO PSI develops quality and representation standards for proteomic and interactomic data. The principal standards and PSI-MI, for molecular interactions, and PSI-MS, for mass spectometric data. These standards were presented in the journal *Nature Biotechnology*. Our project ProteomeBinders (see below) is a HUPO PSI working group.

8.4.2. Génolevures Consortium

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

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, Germany, and England (see http://genolevures.org/). Within the Consortium our team is responsible for bioinformatics, both for the development of resources for exploiting comparative genomic data and for research in new methods of analysis.

In 2004 this collaboration with the 60+ biologists of the Consortium realized the complete genomic annotation and global analysis of four eukaryotic genomes sequenced for us by the National Center for Sequencing (Génoscope, Évry). This annotation consisted in: the *ab initio* identification of candidate genes and gene models though analysis of genomic DNA, the determination of genes coding for proteins and pseudo-genes, the association of information about the supposed function of the protein and its relations phylogenetics. For this global analysis in particular we developed a novel method for constructing multi-species protein families and detailled analyses of the gain and loss of genes and functions throughout evolution.

This perennial collaboration continues 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.

9. Dissemination

9.1. Animation of the scientific community

David Sherman is member of the editorial board of the journal *Computational and Mathematical Methods in Medicine*, and reviewer for several in the bioinformatics field.

David Sherman was external reviewer and member of the thesis defense jury for Pierre Blavy, Rennes, 2010-03-12.

Pascal Durrens is responsible for scientific diffusion, and David Sherman is head of Bioinformatics, for the Génolevures Consortium.

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

Tiphaine Martin is member of the Local Committee of the INRIA Bordeaux Sud-Ouest.

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

Tiphaine Martin and David Sherman are members of the *Institut de Grilles*, and Tiphaine is active in the Biology/Health working group.

9.2. Teaching

Elisabeth Bon is on the faculty of the Université Victor Ségalen Bordeaux 2 and teaches courses in bioinformatics and cellular biology.

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. CASARÉGOLA, 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] 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.

[4] 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/.

[5] 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/.

[6] 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/.

[7] 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/.

[8] 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/.

[9] 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/.

[10] J.-L. SOUCIET, B. DUJON, C. GAILLARDIN, M. JOHNSTON, P. V. BARET, P. CLIFTEN, D. J. SHERMAN, J. WEISSENBACH, E. WESTHOF, P. WINCKER, C. JUBIN, J. POULAIN, V. BARBE, B. SÉGURENS, F. ARTIGUENAVE, V. ANTHOUARD, B. VACHERIE, M.-E. VAL, R. S. FULTON, P. MINX, R. WILSON, P. DURRENS, G. JEAN, C. MARCK, T. MARTIN, M. NIKOLSKI, T. ROLLAND, M.-L. SERET, S. CASAREGOLA, L. DESPONS, C. FAIRHEAD, G. FISCHER, I. LAFONTAINE, V. LEH, M. LEMAIRE, J. DE MONTIGNY, C. NEUVEGLISE, A. THIERRY, I. BLANC-LENFLE, C. BLEYKASTEN, J. DIFFELS, E. FRITSCH, L. FRANGEUL, A. 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

Articles in International Peer-Reviewed Journal

[11] 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.

[12] 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/.

[13] D. E. GLORIAM, S. ORCHARD, D. BERTINETTI, E. BJÖRLING, E. BONGCAM-RUDLOFF, C. A. K. 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. HILTKE, 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.

[14] C. HÖDAR, R. ASSAR, M. COLOMBRES, A. ARAVENA, L. PAVEZ, M. GONZÁLEZ, S. MARTÍNEZ, N. C. INESTROSA, A. MAASS. *Genome-wide identification of new Wnt/beta-catenin target genes in the human genome using CART method.*, in "BMC Genomics", 2010, vol. 11, 348 [*DOI :* 10.1186/1471-2164-11-348], http://hal.inria.fr/inria-00547316/en.

[15] U. MAULIK, A. SARKAR. *Evolutionary Rough Parallel Multi-Objective Optimization Algorithm*, in "Fundamenta Informaticae", 01 2010, vol. 99, p. 13-27, http://hal.inria.fr/inria-00563544/en/.

Invited Conferences

[16] P. DURRENS. *The Génolevures database*, in "10th anniversary of Génolevures", Paris France, P. B. DUJON (editor), Académie des Sciences de l'Institut de France, 11 2010, http://hal.inria.fr/inria-00539198/en.

[17] D. J. SHERMAN. *Methods for understanding function and history in small eukaryote genomes*, in "Colloquium LIX", Palaiseau France, M. RÉGNIER (editor), École Polytechnique, 11 2010, http://hal.inria.fr/inria-00563529/en/.

[18] D. J. SHERMAN. *Two examples of evolutionary algorithms in reconstructing genome evolution*, in "Workshop on Evolutionary Algorithms - Challenges in Theory and Practice", Bordeaux France, INRIA Bordeaux Sud-Ouest, 03 2010, http://hal.inria.fr/inria-00563519/en/.

[19] D. J. SHERMAN, N. LOIRA, N. GOLENETSKAYA. *High-performance comparative annotation*, in "Bioinformatics after next-generation sequencing", Zvenigorod Russian Federation, V. MAKEEV, G. KUCHEROV (editors), Russian Academy of Sciences, 06 2010, http://hal.inria.fr/inria-00563533/en/.

International Peer-Reviewed Conference/Proceedings

[20] R. ASSAR, H. CHRISTIAN, C. MARCELA. *Genome-wide identification of new Wnt/beta-catenin target genes in the human genome using CART method*, in "Statistical and dynamical models in biology and medicine", Germany Heidelberg, Gmds/IBS Working Groups 'Statistical methods in bioinformatics' and 'Mathematical models in medicine' (Tim Beissbarth, Universität Göttingen; Julien Gagneur, EMBL Heidelberg; Nicole Radde, Universität Stuttgart; Ingo Röder, TU Dresden), 10 2010, http://www.ams.med.uni-goettingen.de/modelling_workshop_2010/WorkshopAbstracts.pdf, http://hal.inria.fr/inria-00541263/en.

[21] 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.

[22] T. MARTIN, P. DURRENS. *Génolevures : Policy for automated annotation of genome sequences*, in "Levures, Modèles et Outils IX", France Strasbourg, 08 2010, http://hal.inria.fr/inria-00524385/en.

Workshops without Proceedings

[23] 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.

[24] 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.

[25] 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.

[26] T. MARTIN, D. J. SHERMAN, P. DURRENS. *Génolevures, knowledge base and annotation of hemiascomycete yeast genomes*, in "Journée Ouvertes Biologie Informatique Mathématiques, JOBIM 2010", France Montpellier, 09 2010, http://hal.inria.fr/inria-00524386/en.

Other Publications

[27] 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/.

References in notes

[28] S. F. ALTSCHUL, T. L. MADDEN, A. A. SCHÄFFER, J. ZHANG, Z. ZHANG, W. MILLER, D. J. LIPMAN. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, in "Nucleic Acids Res.", 1997, vol. 25, p. 3389–3402.

[29] 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/.

[30] 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/.

[31] 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/.

[32] P. CLIFTEN, P. SUDARSANAM, A. DESIKAN, L. FULTON, B. 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.

[33] 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.

[34] J. DEAN, S. GHEMAWAT. *MapReduce: simplified data processing on large clusters*, in "Proceedings of the 6th conference on Symposium on Operating Systems Design and Implementation (OSDI'04)", San Francisco, CA, 2004.

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

[38] 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.

[39] 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.

[40] 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/.

[41] 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.

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

[43] 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/.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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/.

[49] 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/.

[50] T. F. SMITH, M. WATERMAN. *Identification of common molecular subsequences*, in "Journal of Molecular Biology", 1981, vol. 147, p. 195–197.

[51] J.-L. SOUCIET, ET AL.. *FEBS Letters Special Issue: Génolevures*, in "FEBS Letters", December 2000, vol. 487, n^{o} 1.

[52] 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/.

[53] 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/.

[54] J. STAJICH, D. BLOCK, K. BOULEZ, S. BRENNER, S. CHERVITZ, ET AL.. *The BioPerl Toolkit: Perl modules for the life sciences*, in "Genome Res.", 2002, vol. 12, p. 1611-18.

[55] 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.

[56] 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/.

[57] C. WU, A. NIKOLSKAYA, H. HUANG, L. YEH, D. NATALE, C. VINAYAKA, Z. HU, R. MAZUMDER, S. KUMAR, P. KOURTESIS, ET AL.. *PIRSF: family classification system at the Protein Information Resource*, in "Nucleic Acids Res.", 2004, vol. 32, p. D315–D318.