INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

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 Scientist

David James Sherman [Team leader; INRIA Senior Research Scientist (DR), HdR]
Pascal Durrens [CNRS, Research scientist (CR), HdR]
Macha Nikolski [CNRS, Research scientist (CR), HdR]

Faculty Member

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

External Collaborator

Grégoire Sutre [CNRS, Research scientist (CR)]
Razanne Issa [Syrian exchange teacher at U. Bordeaux]

Technical Staff

Tiphaine Martin [CNRS, Research engineer]
Alice Garcia [Contract engineer for BioRica ADT]

PhD Student

Rodrigo Assar-Cuevas [CORDI-S INRIA, since Oct. 2008]
Natalia Golenetskaya [CORDI-S INRIA, since Oct. 2009]
Nicolás Loira [CONICYT Chile, since Mar. 2007]
Anasua Sarkar [EMMA co-reg. Jadavpur University, since Oct. 2009]
Hayssam Soueidan [MENSR University Bordeaux, since Mar. 2006]

Post-Doctoral Fellow

Adrien Goëffon [INRIA, until August 2009]
Julie Bourbeillon [ATER University Bordeaux, until August 2009]
Géraldine Jean [ATER University Bordeaux]

Visiting Scientist

Nikolai Vyahhi [University of St. Petersburg]

Administrative Assistant

Marie Sanchez [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.

2.2. Highlights of the year

In collaboration with the Génolevures Consortium and Washington University at St. Louis, MAGNOME completed a large-scale study of five complete yeast genomes from the implicated in various biotechnological applications. These species have been baptised "protoploid" because they are the best contemporary genomes representing the ancestral chomosome number for this phylogenetic branch. Nearly all of MAGNOME's core methodologies were brought on line for this study: genome annotation and analysis [18], [15], median and ancestral genome reconstruction [14], [23], data integration and web deployment [17].

In collaboration with the Institute of Wine and Vine Science, MAGNOME improved understanding of the relation between genome variation and efficiency of cell factory microorganisms used in wine making.

Macha Nikolski (CR1 CNRS) of MAGNOME defended her HDR [11].

3. Scientific Foundations

3.1. Scientific Foundations

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 multi-scale modeling and large-scale comparative genomics: data models for knowledge representation, stochastic hierarchical models for behavior of complex systems, algorithms for genome analysis, and data mining and classification. Our research program builds on our experience in comparative genomics, data-mining and classification, and formal methods for multi-scale stochastic modeling of complex systems.

The first overall goal for MAGNOME is to develop **methods for understanding the structure and history of eukaryote genomes**, in order to identify their differences and the link between these differences and the dynamic behavior of these organisms. 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 novel techniques to address fundamental questions of mechanisms of gene dynamics, and the ways that genes and their products are organized at different scales. These results are then combined into integrated models through the organization of these objects into networks and pathways that can be used to predict the dynamic behavior of cells. Through combinatorial optimization we can construct plausible hypotheses about the structure of ancestral genome architectures, which may provide deep insight both into the past histories of particular genomes and the general mechanisms of their formation.

The methods designed by MAGNOME for comparative genome annotation, structured genome comparison, and construction of integrated models are applied on a large scale to yeasts from the hemiascomycete class [50], [51], [52], [31], [36], which provide a unique tool for studying eukaryotic genome evolution over a broad range of distances. 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, depolymerise tannin extracts, and produce hormones and vaccines in industrial quantities through heterologous gene expression. Several yeast species are pathogenic for humans. The hemiascomycetous yeasts represent a homogeneous phylogenetic group of eukaryotes with a relatively large physiological and ecological diversity.

The second overall goal for 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, 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**.

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*[1] 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.

Hierarchical modeling that integrates both genome-scale models of metabolism and fine-grained models of particular processes of interest in a given application is recognized as a major challenge in systems biology both by the European Union (see "Systems biology: a grand challenge for Europe," ESF Grand Challenges, Sept. 2007). Furthermore the NSF in the United States recognized since 2004 that multi-scale modeling that integrates all scales from molecular through population levels, is the way for modeling to impact the understanding of biological processes (see, for example NSF 04-607).

The MAGNOME BioRica system is a high-level modeling framework integrating discrete and continuous multi-scale dynamics within the same semantics domain, while offering a easy to use and computationally efficient numerical simulator. It is based on a generic approach that captures a range of discrete and continuous formalisms and admits a precise operational semantics [56]. On the practical level, 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.

MAGNOME's short- and mid-term objectives can be described as follows:

3.1.1. Comparative genome annotation

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.

Phylogenetic protein families establish relations of conservation and lineage-specific gain and loss that permit the detailed study of adaptation and functional specialization. Algorithmic techniques must be developed to improve precision across million-year phylogenetic ranges. Two challenges must be addressed in the classification methods: better definition of inclusion relations, and incorporation of gene fusion and fission events, which induce reticulate relations between family classifications.

Rather than compare flat sets of genes grouped into functional classes, *structured comparison* explores the topological structure of the graph of relations between genes. Biomolecular networks are one way to perform such comparisons. Using graph theoretic techniques we can assess the relative conservation of networks from one species to another, with the aim of identifying functional differences between the species.

The computational and storage needs of large-scale global comparison of genomes require a dedicated integrated platform for knowledge representation, high-performance computing and software development. A complete analysis chain for new genomes must start from a genome sequence and produce a preliminary annotation, including prediction of genes, putative assignment to protein families, and application of coherency

[1] R. Alur *et al.* Generating embedded software from hierarchical hybrid models. In *Proceedings of LCTES*, pp 171–82, 2003.

rules. These tools must take into account specificities of fungal genomes such as clade-specific gene architectures, lineage-specific protein families and pathways, and known phylogenetic relationships.

We validate these methodological advances through application to sets of species of biotechnological interest, in collaboration with our biological partners. MAGNOME manages a key a comprehensive comparison of eighteen yeast genomes, annotated by the Génolevures consortium [52]. This annotation effort by 40 scientists in France and Belgium has resulted in a complete catalogue of protein-coding genes and other genetic elements, and work by the MAGNOME team has classified these elements into phylogenetic, structural, and functional categories. These analyses must be extended to systematically cover the range of relations defined above, and will constitute a fundamental resource for the development of dynamic models.

3.1.2. Genome dynamics and evolutionary mechanisms

We develop algorithms for detecting historical relations between genomes and exploring the concrete events and general mechanisms of molecular evolution, in particular mechanisms of rearrangement and duplication that reshape genomes.

Genome rearrangements on two scales contribute to this systematic comparison. Using a complete analysis of gene fusion events across the yeasts and fungi, we identify small-scale events that lead to the birth of new genes and the acquisition of new or improved functions [37]. On a larger scale, rearrangements of large segments are investigated through a combination of conserved segment identification (*in silico* chromosomal painting using chromosomal homology established using conserved protein families) and combinatorial techniques we have developed for median genome and rearrangement scenario computation[14], [23], [40].

The expected results are a comprehensive view of yeast genome organization and evolution, described at multiple scales.

3.1.3. Hierarchical modeling

We develop practical and semantically rigorous formalisms for constructing hybrid hierarchical models of dynamic, stochastic biological processes, with a particular focus of *model reuse*, and build software tools for simulation and analysis of these models. BioRica [56] is a formalism for hybrid hierarchical modeling developed by MAGNOME and instantiated in a software platform.

Formal analysis of biological models is usually faced with two major challenges: on one hand these models exhibit complex behaviors since they may contain both hybrid and stochastic modeling features, which leads to theoretical limitations (undecidability in general). On another hand, precise models tend to be very large, with thousands of discrete or continuous variables, and moreover with multiple time-scales. This leads in practice to the well-known combinatorial explosion problem. We improve the state-of-the-art by adapting strategies that have led to significative successes in modeling human-engineered systems, in particular extending the reach of abstraction-based formal analysis techniques to these models[22]. Both trace-based abstraction and qualitative abstraction of hybrid stochastic systems will be developed.

Validation of this approach is based on applications in dynamic modeling of fermenting and oleaginous yeasts. We will develop a modeling methodology that will first, advance the state of the art of modular modeling in systems biology, and second, enable mixing phenomena described with different precision within the same framework of stochastic hybrid hierarchical models.

4. Application Domains

4.1. Comparative Genomics of Yeasts

The best way to understand the **structure** and the **evolutionary history** of a genome is to compare it with others. At the level of single genes this is a standard and indeed essential procedure: one compares a gene sequence with others in data banks to identify sequence similarities that suggest homology relations. For most gene sequences these relations are the only clues about gene function that are available. The procedure is

essential because the difference between the number of genes identified by *in silico* sequence analysis and the number that are experimentally characterized is several orders of magnitude. At the level of whole genomes, large-scale comparison is still in its infancy but has provided a number of remarkable results that have led to better understanding, on a more global level, of the mechanisms of evolution and of adaptation.

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. The most well known yeast in the Hemiascomycete class is *S. cerevisiae*, widely used as a model organism for molecular genetics and cell biology studies, and as a cell factory. As the most thoroughly-annotated genome of the small eukaryotes, it is a common reference for the annotation of other species. 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 [31], [33], [45], [44], [35], [47], [36].

The *Génolevures* program is devoted to large-scale comparisons of yeast genomes from various branches of the Hemiascomycete class, with the aim of addressing basic questions of molecular evolution such as the degrees of gene conservation, the identification of species-specific, clade-specific or class-specific genes, the distribution of genes among functional families, the rate of sequence and map divergences and mechanisms of chromosome shuffling.

The differences between genomes can be addressed at two levels: at a molecular level, considering how these differences arise and are maintained; and at a functional level, considering the influence of these molecular differences on cell behavior and more generally on the adaptation of a species to its ecological niche.

4.2. Construction of Biological Networks

Comparative genomics provides the means to identify the set of protein-coding genes that comprise the components of a cell, and thus the set of individual functions that can be assured, but a more comprehensive view of cell function must aim to understand the ways that those components work together. In order to predict how genomic differences influence function differences, it is necessary to develop representations of the ways that proteins cooperate.

One such representation are networks of *protein-protein interactions*. Protein-protein interactions are at the heart of many important biological processes, including signal transduction, metabolic pathways, and immune response. Understanding these interactions is a valuable way to elucidate cellular function, as interactions are the primitive elements of cell behavior. One of the principal goals of proteomics is to completely describe the network of interactions that underly cell physiology.

As networks of interaction data become larger and more complex, it becomes more and more important to develop data mining and statistical analysis techniques. Advanced visualization tools are necessary to aid the researcher in the interpretation of these relevant subsets. As databases grow, the risk of false positives or other erroneous results also grows, and it is necessary to develop statistical and graph-theoretic methods for excluding outliers. Most importantly, it is necessary to build *consensus networks*, that integrate multiple sources of evidence. Experimental techniques for detecting protein-protein interactions are largely complementary, and it is reasonable to have more confidence in an interaction that is observed using a variety of techniques than one that is only observed using one technique.

The ProViz software tool [42] addresses the need for efficient visualization tools, and provides a platform for developing interactive analyses. But the key challenge for comparative analysis of interaction networks is the reliable extrapolation of predicted networks in the absence of experimental data.

A complementary challenge to the network prediction is the extraction of useful summaries from interaction data. Existing databases of protein-protein interactions mix different types too freely, and build graph representations that are not entirely sensible, as well as being highly-connected and thus difficult to interpret. We have developed a technique called *policy-directed graph extraction* that provides a framework for selecting observations and for building appropriate graph representations. A concrete example of graph extraction is *subtractive pathway modeling*, which uses correlated gene loss to identify loss of biochemical pathways.

4.3. Modeling Biological Systems

Realistic, precise simulation of cell behavior requires detailed, precise models and fine-grain interpretation. At the same time, it is necessary that this simulation be computationally tractable. Furthermore, the models must be comprehensible to the biologist, and claims about properties of the model must be expressed at an appropriate level of abstraction. Reaching an effective compromise between these conflicting goals requires that these systems be **hierarchically composed**, that the overall semantics provide means for combining components expressed in **different quantitative or discrete formalisms**, and that the simulation admit **stochastic behavior** and evaluation at **multiple time scales**.

In general, numerical modeling of biological systems follows the process shown below.

1. Starting from experimental data, sort possible molecular processes and retain the most plausible.
2. Build a schema depicting the overall model and refine it until it is composed of elementary steps.
3. Translate these steps into mathematical expressions using the laws of physics and chemistry.
4. Translate these expressions into time-dependent differential equations quantifying the changes in the model.
5. Analyze the differential system to assess the model.
6. Elaborate predictions based on a more detailed study of the differential system.
7. Test some selected predictions *in vitro* or *in vivo*.

This approach has proven substantial properties of various biological processes, as for example in the case of cell cycle [59]. However, it remains tedious and implies a number of limitations that we shortly describe in this section.

Many biochemical processes can be modeled using continuous domains by employing various kinetics based on the mass action law. However quite a number of biological processes involve small scale units and their dynamics can not be approximated using a global approach and needs to be considered unit-wise.

Some of the biological systems are now known to have a switch-like behavior and can only be specified in a continuous realm by using zero-order ultra-sensitive parametric functions converging to a sharply sigmoid function, which artificially complexifies the system.

The lack of formalized translations between each step makes the whole modeling process error-prone, since immersing the high-level comprehensible cartoon into a low-level differential formalism is completely dependent on the knowledge of the modeler and his/her mathematical skills. Maybe even worse, it blurs the explanatory power of the schema.

As an illustration of the last point it is well-known that the same high level process of the lysis/lysogeny decision in lambda bacteriophage infecting an *E. coli* cell can be specified using different low-level formalisms, each producing unique results contradicting the others.

The assessment step of the modeling process is usually conducted by slow and painful *parameter tinkering*, upon which some artificial integrators and rate constants are added to fit the model to the experimental data without any clue as to what meanings these integrators could have biologically speaking.

Two complementary approaches are necessary for model validation. The first is the validation from the computer science point of view, and is mainly based on intrinsic criteria. The second is the external validation, and in our case requires confirmation of model predictions by biological experiments.

In addition to classic measures such as indexes of cluster validity, our use of instrinsic criteria in comparative genomics depends on treatment of the organism as a system. We define coherency rules for predictions that take into account essential genes, requirements for connectivity in biochemical pathways, and, in the case of genome rearrangements, biological rules for genome construction. These rules are defined at appropriate levels in each application.

Experimental validation is made possible by collaboration with partner laboratories in the biological sciences.

5. Software

5.1. Magus: Collaborative Genome Annotation

Participants: David James Sherman [correspondant], Pascal Durrens, 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 (http://magus.gforge.inria.fr) 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 [58], various biomedical ontologies (http://obo.sf.net), and a web interface implementing a representational state transfer (REST) architecture [39].

See also the web page http://magus.gforge.inria.fr/.

5.2. Faucils: Analyzing Genome Rearrangement

Participants: Macha Nikolski, Adrien Goëffon, Géraldine Jean, 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.

These tools are developed internally on the INRIA Gforge site and are licensed under CeCILL.

5.3. BioRica: Multi-scale Stochastic Modeling

Participants: David James Sherman, Macha Nikolski [correspondant], Hayssam Soueidan, Nicolás Loira, Grégoire Sutre.

Multi-scale modeling provides one avenue to better integrate continuous and event-based modules into a single scheme. The word *multi-scale* itself can be interpreted both at the level of building the model, and at the level of model simulation. At the modeling level, it involves building *modular* and *hierarchical* models. An attractive feature of such modeling is that it provides a systematic means to balance the need for greater biological detail against the need for simplicity. At the execution level, it implies the co-existence of phenomena operating at different time scales in an integrated fashion. This is a very lively research topic by itself, and has promising applications to biology, such as for example in [46].

We are developing *BioRica*, a high-level modeling framework integrating discrete and continuous multi-scale dynamics within the same semantics field. BioRica has been adopted as an INRIA Technology Development Action (ADT).

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. As for the discrete part of the model, its role is that of a controller.

As a means to counteract the over-genericity of re-usable modular models and their underlying simulation complexity, *BioRica* will provide an abstraction module, whose aim is to preserve only the pertinent information for a given task. The soundness of this approach is ensured by a formal study of the operational semantics of *BioRica* models[22] that adopts the theoretical framework of *abstract interpretation* [34].

The current stage of development extends the AltaRica modeling language to Stochastic AltaRica Dataflow [55] semantics, but also provides parsers for widely used SBML [41] data exchange format. The corresponding simulator is easy to use and computationally efficient.

See also the web page http://www.labri.fr/.

5.4. Génolevures On Line: Comparative Genomics of Yeasts

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

The Génolevures online database (http://cbi.labri.fr/Genolevures/) provides tools and data relative to 9 complete and 10 partial genome sequences determined and manually annotated by the Génolevures Consortium, to facilitate comparative genomic studies of hemiascomycetous yeasts. With their relatively small and compact genomes, yeasts offer a unique opportunity for exploring eukaryotic genome evolution. The new version of the Génolevures database provides truly complete (subtelomere to subtelomere) chromosome sequences, 48 000 protein-coding and tRNA genes, and *in silico* analyses for each gene element. A new feature of the database is a novel collection of conserved **multi-species protein families** and their mapping to metabolic pathways, coupled with an advanced search feature. 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.

The focus of the Génolevures database is to describe the relations between genes and genomes. We curate relations of orthology and paralogy between genes, as individuals or as members of protein families, chromosomal map reorganization and gain and loss of genes and functions. We do not provide detailed annotations of individual genes and proteins of *S. cerevisiae* which are already carefully maintained by the MIPS in the CYGD database (http://mips.gsf.de/projects/fungi) [48] in Europe and by the SGD (http://www.yeastgenome.org/) [32] in North America, as well as in general-purpose databases such as UniProtKB [30] and EMBL [43].

While extensive chromosomal rearrangements combined with segmental and massive duplications make comparisons of yeast genome sequences difficult [54], relations of homology between protein-coding genes can be identified despite their great diversity at the molecular level [36]. 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 [49] applied to a complementary set of homeomorphic [sharing full-length sequence similarity and similar domain architectures, see [60]] and nonhomeomorphic systematic Smith-Waterman [53] and Blast [29] sequence alignments. Similar approaches are developed on a wider scale [60] and are complementary to these yeast-specific families.

The Génolevures database uses a straightforward object model mapped to a relational database. Flexibility in the design is guaranteed through the use of ontologies and controlled vocabularies: the Sequence Ontology [38] for DNA sequence features and GLO, our own ontology for comparative genomics (D. Sherman, unpublished data). Browsing of genomic maps and sequence features is provided by the Generic Genome Browser [58]. The Blast service is provided by NCBI Blast 2.2.6 [29]. The Génolevures web site uses a REST architecture internally [39] and extensively uses the BioPerl package [57] for manipulation of sequence data.

See also the web page http://cbi.labri.fr/Genolevures/.

6. New Results

6.1. Genome annotation of protoploid Saccharomycetacae

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

6.1.1.

Using our whole genome annotation pipeline (defined by David Sherman and Tiphaine Martin), we have successfully realized a complete annotation and analysis of four new genomes, provided to the Génolevures Consortium by the Centre National de Séquençage - Génoscope (Évry) and by the Washington University Genome Sequencing Center (St. Louis, USA). This result required a year of work by a network of 20 experts from 6 partner labs, using the Magus web-based system for collaborative genome annotation, and hundreds of hours of computation on our dedicated 76-core computing cluster. The analysis of these results, performed by members of the Consortium, include identification of 17 500 novel genes, genome comparative cartography and breakpoint analysis, assessment of protein family-specific phylogenetic trees and fast-evolving genes, and definition of a molecular clock through characterization of families of homologous and orthologous protein-coding genes. This major result was published in *Genome Research*.

6.2. Modeling through comparative genomics

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

Using comparative genomics to inform mathematical models of cell function is a central challenge of the MAGNOME research program. Emmanuelle Beyne developed *in silico* methods for predicting protein complexes, one form of protein-protein interaction that provide the building blocks of cell machinery. These predictions were compared to experimental results from gel electrophoresis. This work was extended in a large-scale experimental study using quantitative proteomics and expression data, during a long-term visit to Prof. Steve Oliver's lab at Cambridge University. Florian Iragne has refined his methods for subtractive modeling of biochemical pathways, using his algorithmic framework for policy-directed graph extraction to identify cases of pathway loss through search for correlated gene losses. Nicolás Loira has used a large dataset of protein families from the Génolevures complete genomes and sub-partitioned it through clustering methods to obtain reliable indications of enzyme conservation in nine species. The resulting determination of enzyme conservation is mapped to biochemical reaction models and used to infer stoichiometric models that are currently being evaluated through comparison with experimental results produced by Prof. Nicaud's group at AgroParisTech.

6.3. Analysis of oenological genomes

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

Two activities contributed to improved understanding of the relation between genome variation and efficiency of cell factory microorganisms used in wine making. The first, led by Pascal Durrens, is analysis and mapping of the genomes variations involved in quantitative traits. In collaboration with the ISVV, we detect and map single nucleotide polymorphism (SNP) associated with fermentation parameters during wine fermentation by oenological yeasts. The results will be exploited both in yeast strain improvement (selection of the relevant gene variants) and in modelisation of the fermenting cell (indication of the key metabolic steps).

The second is led by Elisabeth Bon. Through her association with MAGNOME, the team has acquired a new expertise on prokaryotic models, and notably on the non-pathogenic food production bacterium, Oenococcus oeni. This species 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 is in charge of sequencing effort coordination, explorative and comparative genome data analysis, and comparative 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. 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.

6.4. Algorithms for genome rearrangements

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

We developed an improved algorithm, SyDIG, for identifying synteny in distant genomes. It is designed for widespread cases where existing methods, such as filtered genome alignments (e.g. GRIMM-Synteny), or profile-based iterated search (e.g. i-AdHoRe), do not work. This in turn has led to improvements in our method for identifying super-blocks of syntenic segments[14], improving on and building a bridge between competing methods defined by Sankoff and by Bourque and Pevzner. Super-blocks represent the semantics of the ancestral architecture, and provide a piecewise approximation to this architecture that provides a reasonable upper bound on the sum of rearrangement distances between contemporary genomes and the theoretical median. Super-blocks have been successfully identified for a range of species in the Hemiascomycetous yeasts[18].

Using a new formulation in terms of optimization, we devised a new algorithm, FAUCILS, using techniques from optimization by local search and metaheuristics [40]. The algorithm maintains a population of configurations, modified depending on the set of architectures, and evaluated using the rearrangement distance. The result is a robust approach that converges rapidly, and obtains better results that those reported elsewhere. Compared with competing algorithms currently used, this new algorithm takes only a few minutes, compared to several hours; does so on tens of genomes, compared to a maximum of three; and includes biological constraints such as centromere presence and gene super-block conservation, which competing algorithms do not. A follow-up to FAUCILS uses any colony swarming to identify pairwise rearrangement scenarios[23].

6.5. Gene fusion and fission events

Participants: David James Sherman, Pascal Durrens [correspondant], Macha Nikolski, Razanne Issa.

One consequence of genome remodelling in evolution is that these large-scale events can modify genes on the periphery of the duplicated or displaced segment, either by fusion with other genes, or by fission of a gene into several parts. These events produce radical changes in gene content, compared to the more progressive modifications produced by nucleotide substitution, and induce non-treelike, reticulate relations between genes. We have developed a novel algorithmic method for large-scale detection of gene fusion and fission events in fungal genomes, that explicitly uses relations between groups of paralogous genes in order to compensate for genome redundancy. By tracking the mathematical relations between groups of similar genes, rather than between individual genes, we can paint a global picture of remodelling across many species simultaneously.

Indeed, fusion and fission events are landmarks of random remodelling, independent of mutation rate: they define a metric of "recombination distance." This distance lets us build a genome evolution history of species and may well be a better measure than mutation distance of the process of adaptation.

6.6. Definition of the BioRica platform

Participants: David James Sherman [correspondant], Macha Nikolski, Grégoire Sutre, Alice Garcia.

A major development in 2005-9 was the development of BioRica, an extension of the AltaRica modeling language for complex industrial systems. BioRica is a high-level modeling framework integrating discrete and continuous multi-scale dynamics within the same semantics domain, while offering an easy to use and computationally efficient numerical simulator. It is based on a generic formalism that captures a range of discrete and continuous formalisms and admits a precise operational semantics. BioRica models have a corresponding compositional semantics in terms of an extension of Generalized Markov Decision Processes. This semantics allowed us to prove that BioRica models admit an operational semantics in terms of continuous stochastic processes, and that this operational semantics is correctly simulated by the discrete event stepper used during numerical simulation.

The simulation schema for a given BioRica node is given by a hybrid algorithm that deals with continuous time and allows for discrete events that **roll back** the time according to these discrete interruptions. Time advances optimally either by the maximal step size defined by an adaptive integration algorithm, or by discrete jumps defined by the minimal delay necessary for firing a discrete event.

BioRica is instantiated in a software platform for modeling and simulation, that has recently been adopted by the INRIA through a Technology Development Action (ADT).

6.7. Transient Behavior in Parametrized Dynamic Models

Participants: Macha Nikolski [correspondant], Hayssam Soueidan, Grégoire Sutre.

Dynamic models in System Biology rely on kinetic parameters to represent the range of possible behaviors of when enzymatic information is incomplete. Analysis of these parametrized models aims at identifying either parameter ranges yielding similar qualitative behaviors, or parameter values yielding a given behavior of interest. Qualitative transient behavior can be successfully analyzed by model checking algorithms applied on models admitting a computable path semantics. However, in Systems Biology, state explosion and negative decidability results limit the scope of model checking to a certain subset of models. Moreover, some published and curated Systems Biology models lack explicit semantics, and for these "black box" models, not much can be assumed, except the possibility of generating simulation results. Mining these simulation results to identify parameter regions yielding similar behaviors is hindered by the size of the parameter space to explore, numerical artifacts and the lack of formal definition of what it means for simulation results to be similar.

We introduce *Qualitative Transition Systems* (QTS) and define their probabilistic semantics[22]. A novel abstraction operation is defined in with the goal of building QTSs from simulation results. We show that when constructing a QTS from an ODE, the QTS construction can be made independent of the numerical integration scheme. Trajectory comparison using QTS can be made more resistant to noise by detecting points of interest (extremums and inflection) through the construction of a piecewise linear approximation (PLA). We have validated our approach on a large set of SBML models from the BioModels database, including:

- The cell cycle model of Tyson *et al.* (1991) based on interactions between Cdc2 and cyclin, where we investigate "similar" oscillatory behaviors with different transient behaviors.
- The MAPK cascade model with negative feedback of Kholodenko (2000), in which we can compute the probability of oscillatory behavior in a large parameter subspace.
- The model of crosstalk between an extracellular signal regulated kinase ERK and the Wnt pathway of Kim, Rath *et al.* (2007), successfully detecting the irreversible pathological response in the oncogenic positive feedback loop.

7. Other Grants and Activities

7.1. International Activities

7.1.1. HUPO Proteomics Standards Initiative

Participants: David James Sherman [correspondant], Julie Bourbeillon.

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 reference [5] in the journal *Nature Biotechnology*. Our project ProteomeBinders (see below) has been accepted as a HUPO PSI working group.

7.1.2. 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, and England (see http://cbi.labri.fr/Genolevures/). 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.

7.1.3. Systems Biology Markup Language

Participant: Macha Nikolski [correspondant].

Macha Nikolski has recently been implicated in the standards process for version 3 of the SMBL standard, in particular in defining a rigorous mathematical semantics for timed events and hierarchical compositions.

7.2. European Activities

7.2.1. Yeast Systems Biology Network (FP6)

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

Our team is actively involved in the Yeast Systems Biology Network (YSBN) Coordinated Action, sponsored by the EU sixth framework programme. The allocated budget is 1.3 million Euros. The CA is coordinated by Prof. Jens Nielsen (Technical University of Denmark) and involves 17 European universities and 2 start-up biotech companies: InNetics AB and Fluxome Sciences A/S.

The activities of this CA aim at facilitating and improving research in yeast systems biology. The EU team creates standardised methods for research, reference databases, develops inter-laboratory benchmarking, and organizes a international conference, a number of PhD courses, and workshops.

The project involves most of the best EU academic centres in this field of science: Biozentrum University of Basel, Bogazici University Istanbul, Budapest University of Technology and Economics and Hungarian Academy of Sciences, CNSR/LaBRI University Bordeaux, ETH Zurich, Gothenburg University, Manchester University, Lund University, Max Plank Institute of Molecular Genetics, Medical University Vienna, Stuttgart University, Technical University of Denmark, Technical University Delft, University of Milano Bicocca, Virje University Amsterdam, VTT Technical Research Centre Finland.

7.2.2. ProteomeBinders (FP6)

Participants: David James Sherman [correspondant], Julie Bourbeillon.

The ProteomeBinders Coordination Action, sponsored by the EU sixth framework programme, coordinates the establishment of a European resource infrastructure of binding molecules directed against the entire human proteome. The allocated budget is 1.8 million Euros. The CA is coordinated by Prof. Mike Taussig of the Babraham Institute in the UK.

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 Coordination Action (`proteomebinders.org`) coordinates 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. They 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 CA will organise 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. A proteome binders resource will have huge benefits for basic and applied research, impacting on healthcare, diagnostics, discovery of targets for drug intervention and therapeutics. It will thus be of great advantage to the research and biotechnology communities.

Within ProteomeBinders, our team is responsible for formalizing an ontology of binder properties and a set of requirements for data representation and exchange, and for developing a database schema based on these specifications that could be used to set up a central repository of all known ligand binders against the human proteome. The adoption of the proposed standards by the scientific community will determine the success of this activity.

7.2.3. IntAct

Participants: David James Sherman [correspondant], Julie Bourbeillon.

The IntAct project, led by the European Bioinformatics Institute (EBI) within the framework of the European project TEMBLOR (The European Molecular Biology Linked Original Resources), develops a federated European database of protein-protein interactions and their annotations. IntAct partners develop a normalized representation of annotated protein interaction data and the necessary ontologies, a protocol for data exchange between the nodes of the federated database, and a software infrastructure for the installation of these local nodes. In this infrastructure, a large number of software tools have been realized to aid biological user exploit these data reliably and efficiently. Our own tool Proviz is part of this set of tools. Curator annotation, optimization, and quality control tools have also been developed [6]. We also submit experimental data to the repository.

7.3. National Activities

7.3.1. ANR GENARISE

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

GENARISE is a four-year ANR project that explores the question of how genes arise and die. Coordinated by Prof. Bernard Dujon of the Pasteur Institute, this pluridisciplinary project uses an original combination of complementary experimental and informatic techniques to answer specific questions about the mechanisms of genome dynamics. The MAGNOME team contributes much of the informatics expertise in this project and is in particular plays a role as a resource for *in silico* techniques.

7.3.2. 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.

7.3.3. 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.

7.4. Regional Activities

7.4.1. Aquitaine Region "Services robustes pour les réseaux dynamiques (SR2D)"

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

In the wider context of the regional project supporting a research pole in informatics, we work with other experts in data-mining and visualization on the application of these techniques to genomic data. In particular we have develop novel methods for constructing summaries of large data sets, that are coupled with graph visualization techniques in the Tulip platform.

7.4.2. Aquitaine Region "Identification de nouveaux QTL chez la levure pour la sélection de levains œnologiques"

Participant: Pascal Durrens [correspondant].

This project is a collaboration between the company SARCO, specialized in the selection of industrial yeasts with distinct technological abilities, the FCBA technology institute, and the CNRS. The goal is to use genome analysis to identify chromosomal regions (QTLs) responsible for different physiological capabilities, as a tool for selecting yeasts for wine fermentation through efficient crossing strategies. Pascal Durrens is leading the bioinformatic analysis of the genomic and experimental data.

8. Dissemination

8.1. Reviewing

David Sherman was reviewer for the journal *Bioinformatics* (Oxford University Press).

David Sherman was reviewer for the journal *BMC Bioinformatics* (BioMed Central).

David Sherman was reviewer for the journal *Nucleic Acids Research* (Oxford University Press).

David Sherman was a reviewer for the national program GIS IBiSA.

Pascal Durrens was reviewer for the journal *PLoS Computational Biology* (Public Library of Science)

Pascal Durrens was part of the thesis jury for Nicolas Jauniaux at the University of Strasbourg.

Macha Nikolski was a member of the program committee for WABI'09.

8.2. Memberships and Responsabilities

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

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.

David Sherman is member of the Comité Consultatif Régional de Recherche et de Développement Technologique (CCRRDT) de la Région Aquitaine : Commission 3 "Sciences biologiques, médicales et de la santé" (*suppléant* of Claude Kirchner)

David Sherman is member of the Scientific Council of the LaBRI UMR 5800/CNRS

8.3. Recruiting committees

Macha Nikolski was member of the CR recruiting committee of the INRIA Saclay.

Pascal Durrens was member of the selection committee for the University of Strasbourg.

Elisabeth Bon was member of the selection committee for the University of Strasbourg.

Tiphaine Martin was member of the IR selection committee for the INRA Toulouse.

8.4. Visitors

Nikolai Vyahhi of St. Petersburg University, Russia, was invited for three months as a visiting researcher.

8.5. Participation in colloquia, seminars, invitations

David Sherman
23/01/2009 Paris Génolevures
03/02/2009–04/02/2009 Grignon Collaboration INRA
26/02/2009–27/02/2009 Paris
22/03/2009–24/03/2009 Alpbach Austria ProteomeBinders
09/04/2009–22/04/2009 Chicago Invitation U. Chicago
23/04/2009–24/04/2009 Paris Génolevures
26/05/2009 Paris Génolevures
27/05/2009–30/05/2009 Strasbourg Collaboration ULP
21/06/2009–23/06/2009 Paris INRIA
14/09/2009–15/09/2009 Paris INRIA
18/09/2009 Paris GENARISE
08/10/2009 Paris INRIA Evaluation Seminar 06/11/2009 Paris Génolevures
27/11/2009 Paris DIKARYOME
17/12/2009-18/12/2009 Paris INRIA
Pascal Durrens
22/02/2009–27/02/2009 Marseille Ecole d'hiver au CIRM (Modelisation mathématique du cancer)

05/03/2009–06/03/2009 Paris Génolevures
23/03/2009–24/03/2009 Paris Lancement du projet Nakaseomycetes au Génoscope
11/05/2009–11/05/2009 Strasbourg Comité de sélection MdC
14/05/2009–15/05/2009 Paris Génolevures
02/06/2009–03/06/2009 Strasbourg Comité de sélection MdC
25/06/2009–26/06/2009 Paris Génolevures
08/10/2009–09/10/2009 Paris Génolevures
17/10/2009–22/10/2009 Sant Feliu Conference "Comparative Genomics of Eukaryotic Microorganisms"

Tiphaine Martin

23/01/2009 Paris Génolevures
24/02/2009 Strasbourg Génolevures
06/03/2009 Paris Génolevures
10/03/2009–11/03/2009 Lyon Grisbi (grille bioinformatique)
27/03/2009 Paris Thèse Célia Payen
06/04/2009–10/04/2009 Nancy École Grid5000
23/04/2009–24/04/2009 Montpellier Jury IR INRA
04/05/2009–06/05/2009 Roscoff Grisbi (grill bioinformatique)
15/05/2009 Paris Génolevures
26/05/2009–27/05/2009 Lyon Kick-off Grisbi
28/05/2009-30/05/2009 Strasbourg “2nd Greman/French/European Meeting on Yeast and Filamentous Fungi”
03/06/2009–05/06/2009 Cambridge Réunion Grisbi-EMI
08/06/2009–12/06/2009 Nantes JOBIM
26/06/2009 Paris Génolevures
29/06/2009–01/07/2009 Toulouse Jury concours IE INRA
18/09/2009 Paris GENARISE
07/10/2009 Paris Génolevures (site public), groupe restreint
08/10/2009 Paris INRIA Evaluation Seminar
17/10/2009–22/10/2009 San Feliu/Spain; EMBO "Comparative Genomics of Eukaryotic Microorganisms"
06/11/2009 Paris Génolevures
04/12/2009 Paris Génolevures
07/12/2009 Paris Evolution et Biodiversité Bactérienne : impact du séquençage de nouvelle génération, conference
Invitation round table “Promouvoir les filières scientifiques auprès des jeunes filles,” Infosup Dordogne carrière, Perigueux
Conference organization for Biograle 2009, Rennes 24/11/2009–25/11/2009

Macha Nikolski

09/04/2009–22/04/2009 Chicago Invitation U. Chicago
09/07/2009–10/07/2009 Grenoble Seminar IBIS
14/09/2009–15/09/2009 Hinxton Collaboration N. Le Novère
08/10/2009 Paris INRIA Evaluation Seminar 06/12/2009–23/12/2009 Moscow Building international relations

Hayssam Soueidan

09/07/2009–10/07/2009 Grenoble Seminar IBIS
30/08/2009–02/09/2009 Bologna Italy Computational Methods in Systems Biology (CMSB’09)
14/09/2009–15/09/2009 Hinxton Collaboration N. Le Novère
20/10/2009–31/10/2009 Moscow Building international relations
06/12/2009–23/12/2009 Moscow Building international relations

Nicolás Loira

15/05/2009 Grignon
7/06/2009–12/06/2009 Nantes JOBIM 2009
27/07/2009–31/07/2009 Grignon YALI genome-scale metabolic model

10/10/2009–11/10/2009 Paris; "Challenges in experimental data integration within genome-scale metabolic models" (Université Pierre et Marie Curie)
17/10/2009–22/10/2009 San Feliu/Spain; EMBO "Comparative Genomics of Eukaryotic Microorganisms"
30/11/2009–1/12/2009 Grignon YALI genome-scale metabolic model
Anasua Sarkar
17/10/2009–22/10/2009 San Feliu/Spain; EMBO "Comparative Genomics of Eukaryotic Microorganisms"

8.6. Teaching

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

All of the doctoral students in MAGNOME have teaching duties as teaching assistants, in the Universities Bordeaux 1 and Victor Ségalen Bordeaux 2, or the ENSEIRB. Post-doc Julie Bourbeillon teachines bioinformatics and statistics at the Université Victor Ségalen Bordeaux 2.

9. 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", vol. 8, 2007, 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", vol. 487, n^{o} 1, December 2000, p. 31-36.

[3] 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", vol. 430, n^{o} 6995, 07 2004, p. 35-44, http://hal.archives-ouvertes.fr/hal-00104411/en/.

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

[5] 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.", vol. 22, n^{o} 2, Feb. 2004, p. 177-83.

[6] 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.", vol. 32, Jan. 2004, p. D452-5.

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

[8] 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, http://hal.inria.fr/inria-00341578/en/.

[9] 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, epub ahead of print, http://hal.inria.fr/inria-00407511/en/US.

[10] M. TAUSSIG, O. STOEVESANDT, C. BORREBAECK, A. BRADBURY, D. CAHILL, C. CAMBILLAU, A. DE DARUVAR, S. DUEBEL, J. EICHLER, R. FRANK, T. GIBSON, D. GLORIAM, L. GOLD, F. HERBERG, H. HERMJAKOB, J. HOHEISEL, T. JOOS, O. KALLIONIEMI, M. KOEGLL, Z. KONTHUR, B. KORN, E. KREMMER, S. KROBITSCH, U. LANDEGREN, S. VAN DER MAAREL, J. MCCAFFERTY, S. MUYLDERMANS, P.-A. NYGREN, S. PALCY, A. PLUECKTHUN, B. POLIC, M. PRZYBYLSKI, P. SAVIRANTA, A. SAWYER, D. J. SHERMAN, A. SKERRA, M. TEMPLIN, M. UEFFING, M. UHLEN. *ProteomeBinders: planning a European resource of affinity reagents for analysis of the human proteome*, in "Nature Methods", vol. 4, n^{o} 1, 2007, p. 13–17.

Year Publications

Doctoral Dissertations and Habilitation Theses

[11] M. NIKOLSKI. *From Genomic to Functional Models*, Université Bordeaux 1 Sciences et Technologies, 10 2009, http://www.labri.fr/perso/macha, Habilitation à diriger des recherches, Ph. D. Thesis.

[12] H. SOUEIDAN. *Discrete event modeling and analysis for Systems Biology models*, Université Bordeaux 1 Sciences et Technologies, 12 2009, http://www.labri.fr/perso/soueidan, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

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

[14] 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/.

[15] C. PAYEN, G. FISCHER, C. MARCK, C. PROUX, D. J. SHERMAN, J.-Y. COPPÉE, M. JOHNSTON, B. DUJON, C. NEUVÉGLISE. *Unusual composition of a yeast chromosome arm is associated with its delayed replication.*, in "Genome Research", 2009, epub ahead of print, http://hal.inria.fr/inria-00407518/en/US.

[16] D. J. SHERMAN. *Minimum information requirements : neither bandits in the Attic nor bats in the belfry*, in "New Biotechnology", vol. 25, n^{o} 4, 2009, p. 173-4, http://hal.inria.fr/inria-00407505/en/.

[17] 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, http://hal.inria.fr/inria-00341578/en/.

[18] 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, epub ahead of print, http://hal.inria.fr/inria-00407511/en/US.

Invited Conferences

[19] R. ASSAR, H. SOUEIDAN, D. J. SHERMAN. *Hierarchical study of Guyton Circulatory Model*, in "Les Journées Ouvertes en Biologie, Informatique et Mathématiques JOBIM 2009, France Nantes", I. R. ERIC RIVALS (editor), Eric Rivals, Irena Rusu, 2009, http://hal.inria.fr/inria-00404135/en/.

[20] P. LUCAS, E. BON, V. RENOUF, M. DOLS-LAFARGUE, C. LE MARREC, A. LONVAUD-FUNEL. *Oenococcus oeni genomic adaptation to the wine environment*, in "SGM-Society for General Microbiology Autumn 2009 Meeting, Royaume-Uni Edinburg", 2009, http://hal.inria.fr/inria-00395530/en/.

International Peer-Reviewed Conference/Proceedings

[21] T. MARTIN, M. NIKOLSKI, D. J. SHERMAN, J.-L. SOUCIET, P. DURRENS. *The Génolevures online database*, in "Second German/ French/European/ Meeting Yeast and Filamentous Fungi, France Strasbourg", 2009, http://hal.inria.fr/inria-00409534/en/.

[22] H. SOUEIDAN, G. SUTRE, M. NIKOLSKI. *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", vol. 5688, Springer Verlag, 2009, p. 313–327, http://hal.archives-ouvertes.fr/hal-00408909/en/.

[23] 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, Canada Montréal", F. ROTHLAUF (editor), ACM, ACM SIGEVO, 2009, http://hal.inria.fr/inria-00407508/en/RU.

National Peer-Reviewed Conference/Proceedings

[24] E. BON, A. DELAHERCHE, E. BILHERE, C. MIOT-SERTIER, P. DURRENS, A. DE DARUVAR, A. LONVAUD-FUNEL, C. LE MARREC. *Oenococcus oeni genome plasticity associated with adaptation to wine, an extreme ecological niche*, in "JOBIM-10èmes Journées Ouvertes en Biologie, Informatique et Mathématiques, France Nantes", E. RIVALS, I. RUSU (editors), 2009, p. 121-122, http://hal.inria.fr/inria-00392020/en/.

[25] E. BON, C. MIOT-SERTIER, M. DOLS-LAFARGUE, G. MOREL, A. LONVAUD-FUNEL, C. LE MARREC. *A 24 kb-genomic island contributing to Oenococcus oeni adaptation to wine can excise from the chromosome*, in "16th CBL-Club des Bactéries Lactiques Meeting, France Toulouse", 2009, http://hal.inria.fr/inria-00392022/en/.

[26] F. EL GARNITI, C. MIOT-SERTIER, M. DOLS-LAFARGUE, E. BON, A. LONVAUD-FUNEL, C. LE MARREC. *IS30 elements as mediators of strain diversity in Oenococcus oeni*, in "16th CBL- Club des Bactéries Lactiques Meeting, France Toulouse", 2009, http://hal.inria.fr/inria-00396032/en/.

[27] T. MARTIN, D. J. SHERMAN, M. NIKOLSKI, J.-L. SOUCIET, P. DURRENS. *Base de données Génolevures : génomique comparative des Hemiascomycetes*, in "Journée Ouvertes Biologie Informatique Mathématiques, JOBIM 2009, France Nantes", E. RIVALS, I. RUSU (editors), 2009, p. 181-182, http://hal.inria.fr/inria-00401915/en/.

Other Publications

[28] H. SOUEIDAN. *Extending Discrete Event Systems for the Hierarhical Specification, Analysis and Simulation of Systems Biology Models*, in "Séminaire équipe INRIA IBIS, France", 2009, http://hal.archives-ouvertes.fr/hal-00407522/en/.

References in notes

[29] 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.", vol. 25, 1997, p. 3389–3402.

[30] A. BAIROCH, R. APWEILER, C. WU, W. BARKER, B. BOECKMANN, E. AL.. *The Universal Protein Resource (UniProt)*, in "Nucleic Acids Res.", vol. 33, 2005, p. D154–D159.

[31] 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", vol. 487, n^{o} 1, December 2000, p. 31-36.

[32] J. CHERRY, C. ADLER, C. BALL, S. CHERVITZ, S. DWIGHT, E. HESTER, Y. JIA, G. JUVIK, T. ROE, M. SCHROEDER, S. WENG, D. BOTSTEIN. *SGD: Saccharomyces Genome Database*, in "Nucleic Acids Res.", vol. 26, 1998, p. 73–79.

[33] 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", vol. 301, 2003, p. 71–76.

[34] P. COUSOT, R. COUSOT. *Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints*, in "Conference Record of the Fourth ACM Symposium on Principles of Programming Languages", January 1977, p. 238–252.

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

[36] B. DUJON, D. J. SHERMAN, E. AL.. *Genome Evolution in Yeasts*, in "Nature", vol. 430, 2004, p. 35–44.

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

[38] K. EILBECK, S. LEWIS, C. MUNGALL, M. YANDELL, L. STEIN, R. DURBIN, M. ASHBURNER. *The Sequence Ontology: a tool for the unification of genome annotations*, in "Genome Biology", vol. 6, 2005, R44.

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

[40] 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/.

[41] M. HUCKA, E. AL.. *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models*, in "Bioinformatics", vol. 19, n^{o} 4, 2003, p. 524-31.

[42] F. IRAGNE, M. NIKOLSKI, B. MATHIEU, D. AUBER, D. J. SHERMAN. *ProViz: protein interaction visualization and exploration*, in "Bioinformatics", vol. 21, 2005, p. 272-274, http://hal.inria.fr/inria-00202436/en/.

[43] C. KANZ, P. ALDEBERT, N. ALTHORPE, W. BAKER, A. BALDWIN, K. BATES, E. AL.. *The EMBL Nucleotide Sequence Database*, in "Nucleic Acids Res.", vol. 33 database issue, 2005, p. D29–D33.

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

[45] 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", vol. 423, 2003, p. 241–254.

[46] E. KOROBKOVA, T. EMONET, J. VILAR, T. SHIMIZU, P. CLUZEL. *From molecular noise to behavioural variability in a single bacterium.*, in "Nature", vol. 428, 2004, p. 574–578.

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

[48] H. MEWES, D. FRISCHMAN, U. GULDENER, G. MANNHAUPT, K. MAYER, M. MOKREJS, B. MORGENSTERN, M. MUNSTERKOTTER, S. RUDD, B. WEIL. *MIPS: a database for genomes and protein sequences*, in "Nucleic Acids Res.", vol. 30, n^{o} 1, January 2002, p. 31–34.

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

[50] 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)", vol. 32, 2004, p. D315-8, http://hal.inria.fr/inria-00407519/en/, GDR CNRS 2354 "Génolevures".

[51] 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", vol. 34, n^{o} Database issue, 01 2006, p. D432-5, http://hal.archives-ouvertes.fr/hal-00118142/en/.

[52] 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)", 11 2008, http://hal.inria.fr/inria-00341578/en/, epub ahead of print.

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

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

[55] H. SOUEIDAN, M. NIKOLSKI, G. SUTRE. *Syntaxe, Sémantique et abstractions de programmes AltaRica Dataflow*, Université de bordeaux 1, 2005, http://www.labri.fr/~soueidan/, Masters thesis.

[56] 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/.

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

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

[59] J. TYSON, K. C. CHEN, L. CALZONE, A. CSIKASZ-NAGY, F. R. CROSS, B. NOVAK. *Integrative Analysis of Cell Cycle Control in Budding Yeast*, in "Mol. Biol. Cell", vol. 15, n^{o} 8, 2004, p. 3841-3862, http://www.molbiolcell.org/cgi/content/abstract/15/8/3841.

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