INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team MAGNOME

Models and Algorithms for the Genome

Futurs

THEME BIO

Table of contents

1. Team

MAGNOME is a joint project of INRIA Futurs and the CNRS, through the Laboratoire Bordelais de Recherche en Informatique (LaBRI, UMR 5800) joint research unit of the CNRS, University Bordeaux 1, ENSEIRB, and University Bordeaux 2.

Head of project-team
David James Sherman [Associate Professor (MCF) ENSEIRB seconded to INRIA, HdR]

Administrative assistant
Marie Sanchez [Secretary (SAR) Inria]

Research scientist
Pascal Durrens [Research scientist (CR1) CNRS, HdR]
Macha Nikolski [Research scientist (CR1) CNRS]
Tiphaine Martin [Research engineer (IR) CNRS]

Ph.D. Students
Emmanuelle Beyne [Ph.D. student Univ. Bordeaux 1]
Florian Iragne [Ph.D. student Univ. Bordeaux 1]
Hayssam Soueidan [Ph.D. student Univ. Bordeaux 1]
Géraldine Jean [Ph.D. student Univ. Bordeaux 1]
Nicolás Loira [Ph.D. student Univ. Bordeaux 1]

Post-doctoral fellows
Adrien Goëffon
Julie Bourbeillon
Sandrine Palcy [PostDoc ProteomeBinders, ENSEIRB]

Technical staff
Cyril Cayla [ANR Contract Univ. Bordeaux 1]
Simon Frey

Student interns
Meryem Mekouar [Univ. Paris 6]

External collaborators
Grégoire Sutre [Research scientist (CR1) CNRS]
Serge Dulucq [Professor Univ. Bordeaux 1, HdR]
Isabelle Dutour [Associate Professor (MCF) Univ. Bordeaux 1]
Antoine De Daruvar [Professor Univ. Bordeaux 2, HdR]

Visiting faculty
Toby Gibson [EMBL Heidelberg]

2. Overall Objectives

2.1. Introduction

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

A large-scale deployment of the MAGUS web-based genome annotation system was used by a network of 20 experts to completely analyze and annotate four complete genomes. We identified 17 500 new genes, completely classified the total set of 48 000 genes in our data warehouse using our consensus clustering methods, and developed a new model of the genome architecture of the ancestral genome.

Using a new approach based on optimization by local search and metaheuristics, we have devised a novel algorithm for computing median genomes and genome rearrangement trees. 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. The algorithm was successfully applied to five complete genomes using markers identified by *in silico* chromosomal painting.

Using improvements to the BioRica platform for hybrid modeling, we developed a new hybrid model for replicative senescence. The BioRica simulation package outperformed Mathematica xCellerator by two orders of magnitude due to the use of a native code compiler, allowing us to exhaustively explore the model by running large simulation batches, even for parameter values that were difficult to handle in the xCellerator package.

3. Scientific Foundations

3.1. Introduction

The development of high-throughput techniques for genomics and post-genomics has considerably changed the way that many biologists do their research. Knowledge of complete genomes and, more recently, metabolic, regulatory, and interaction networks has made it possible to consider a living cell not as a loose collection of individual components but as a *system*. These *global approaches* in biology contribute to deeper understanding of living systems, but produce an accompanying volume of information that only informatic methods can master. Global answers to biological questions are more and more dependent of pluridisciplinary approaches that link biology and bioinformatics. The ultimate goals of computational biology are to extract knowledge from large scale data sets; to build complete representations of cells, organisms, and populations; and to predict computationally complex systems from bodies of less complex data [50]. The inference of the behavior of a living organism at a systems level, based on the knowledge of other living organisms, will be very valuable in medicine and biotechnology. Indeed, a large number of living organisms are out of reach for thorough experimental investigation, either for technical or financial reasons. As the acquisition of genomic sequences is becoming easier and more cost effective, computational biology must fill the gap between the genome and the understanding of a living organism as a system.

Addressing the challenges of systems-level understanding of living organisms requires a three-fold view [36]. The first step is the identification of components constituting the system, starting from the genome. The second step is understanding the function of each component, which in case of biological systems requires the understanding of genome evolution and how these components arose. The third step is the unraveling of the way that these components cooperate, thus realising complex functions at a cellular level. The latter requires both the understanding of the "wiring diagram" between components, as well as the dynamics of the system. This vision in turn presents numerous technical challenges for the information sciences: algorithmic techniques for finding patterns in data, knowledge representation and data integration on a semantic level, algorithmic predictive methods for building hypotheses that can be tested in the laboratory, and formal tools for modeling and simulating complex system behavior.

MAGNOME is an interdisciplinary project that addresses these challenges through a systems approach that draws its strength from close collaborations between computer scientists and biologists. Historically, the members of the MAGNOME team come from varied backgrounds: formal methods and analysis of complex industrial systems, efficient implementation of logic and rewriting systems, and molecular genetics. This historical basis is reflected in the scientific foundations of the MAGNOME team, which are a unique combination of three mutually-reinforcing scientific domains.

- In **comparative genomics** we identify and analyse differences between genomes, in order to understand their past history and current function, and the processes that shape them.
- Our focus in **data-mining and data integration** is both on efficient algorithms for identifying pertinent groupings in complex data sets, and multi-scale representations of those data that admit complex queries and reasoning.
- Our long-standing work in **formal methods** applied to complex systems combines efficient representations of state spaces with model-checking to analyze the realm of system behaviors.

While each of these domains can be studied independantly, we have found that their combination provides a robust approach where each subject is reinforced by the context provided by the other two.

3.1.1. A Systems Approach

Biological systems are *complex systems* in the sense that their behavior cannot be completely described by the behavior of their individual components. Interaction between simple components leads to complex system behavior.

MAGNOME uses 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. Rather than study individual components of these genomes or individual biochemical reactions, we build views of these organisms as systems of cooperating and competing biological processes.

Our research program develops novel applications in comparative genomics of eukaryotic microorganisms, predictive construction of biological networks such as protein-interaction networks and biochemical pathways, and practical modeling and simulation of biological systems using the BioRica framework. This activity has produced a wide variety of software tools designed for the biological user, developed in through international collaboration with partner laboratories in France and in Europe.

3.2. Comparative Genomics

The goal of comparative genomics is to understand the structure and function of genomes through the comparison of related species. While this goal is inherently biological, the techniques brought into play are inherently informatic and comprise a domain of scientific study in their own right. The research of the MAGNOME team involves three axes within this domain.

Genome annotation is the process of associating biological knowledge to sequences. This involves identification of the genes through analysis of the sequence, clustering the genes and other elements into phylogenetic and functional groups, and integrating heterogenous data sources into efficient software tools for exploration, analysis, and visualization. References [67] and [68] provide an overview of our work.

Sequence analysis using probabalistic models, notably hidden markov models, are used for syntactic analysis of macromolecular sequences, applying rules derived from models of how the cell's transcriptional machinery recognizes and interprets the DNA sequence to predict whether a given sequence code for protein, is intronic, participates in gene regulation, etc. Our team adapts and develops algorithms for predicting gene architectures based on intrinsic evidence (based only on the sequence) and extrinsic evidence (including outside information such as sequence alignments).

Combinatory analysis, including algorithms for permutations and other word problems, and graph algorithms are widely used for biological data. Our own work involves combinatorial methods for calculating rearrangement distances using operations inspired by [47], [64], [77], but including biologically-inspired constraints such as centromere position and a cost model adapted to our models of yeast genome evolution. Formally, each genome is coded by a signed permutation, where each element denotes a syntenic region conserved across species, and the sign of the element indicates its relative orientation along the sense or the antisense strand. Genome rearrangements are thus represented by reversal and translocation operations on these permutations. Optimization in the space of genome rearrangements is accomplished using local search techniques. A key advantage of our approach is that it gives the means to explore rearrangement scenarios that are suboptimal with regard to the mathematical formulation, but possibly more reasonable with regard to biological constraints.

3.3. Data-mining and Data Integration

Broadly speaking, *data-mining* methods seek to find meaningful patterns in volumes of data, ideally patterns that are both previously unknown and useful for some application. We can contrast this with *data integration*, where the goal is to link related information in a semantically coherent way. Both kinds of methods are developed in the MAGNOME team.

Consensus clustering. *Clustering* is a widely used data-mining technique whose goal is to learn a set of classes or categories for the given data, without an predetermined idea of what those classes will be. Its utility for applications in computational biology stems from widespread use of "guilt by association" reasoning: phenomena that appear under the same conditions in a experiment often take part in a common, unknown mechanisms of biological interest. Many varieties of clustering algorithms for biological data have consequently been developed, and in large numbers (see [27] for review), which leads to an important practical problem: how to decide which algorithms, or which learning parameters, to use for a given application?

We have addressed one part of this problem through the development of techniques for clustering ensembles, where the goal is to combine the strengths of a chosen set of different (presumably complementary) clustering techniques. This can be formulated as a search for a median partition Π that minimizes $S = \sum_{i=1}^{k} d(\Pi^i, \Pi)$, given k partitions $\Pi^1, ..., \Pi^k$ and a distance function d. The first mathematical treatment goes back to Régnier [66], and [29] shows that the general problem is NP-complete. If the partition Π of the dataset D, $|D| = n$ to discover is not necessarily one of the original partitions $\Pi_1, ..., \Pi_k$, then the size of the potential search space corresponds to the *Bell numbers* [30]. Heuristic approaches have been developed for this inherently intractible problem: exact methods using cutting planes [46], co-association methods [44], voting approach [79], information-theoretic approach [58], hypergraph partitioning methods [76] and using mixture models [78].

The solution we have developed [16] is tailored to the specific problem of consensus clustering for protein families, where in our application $n = 50\,000$ but singleton families (containing only one protein) are allowed. The approach uses a compact bipartite graph encoding of the confusion matrices of pairwise comparaisons between two input partitions Π^1 and Π^2, where nodes are clusters in one or the other inputs, and edges indicate that the two clusters have an element in common. Choice of a consensus among the k partitions can be made by choosing within the connected components of the confusion matrix, in such a way as to cover all the initial elements. Such a choice can be formulated as an instance of minimum exact cover (MDC), also NP-complete [45]. Since we allow singleton families we can further relax the problem to minimum *inexact* cover. In [16] we define an efficient heuristic running in low-order polynomial time that uses a Condorcet election procedure to choose an inexact cover that minimizes inter-partition distance while maximizing cluster similarity.

Enrichment analysis. Guilt by association methods are widely used to search for enrichment of a query set by use of a statistical model to identify similar target sets (see [54] for review). These methods often involve a large number of target sets, each of which must be stored and compared to the query, and produce large numbers of redundant results that overwhelm the user with non-pertinent information. This is a classic query optimization problem involving a time-space tradeoff and an early pattern evaluation, since ideally we would like to only generate interesting nonredundant targets on the fly from a less explicit representation.

In [28] we developed a guilt by association method for integrating heterogenous data collections using a uniform set-based representation of relations between data items, and a probabalistic measure of similarity between sets. Adopting the Danchin view that biological entities must be understood in terms of their relationships and not only in terms of their individual properties [38], the *BlastSets* system provides a systematic means of representing and querying data banks through the use of gene "neighborhoods." Unfortunately, like many others, this system suffered precisely from the efficiency and redundancy problems described in the preceding paragraph.

Element neighborhoods are defined with respect to each discrete or continuous property stored in the data bank, and in terms of different qualitative similarity thresholds. Sets in a neighborhood are thus overlapping and their elements can be partially ordered by the inclusion relation $\subseteq$. By representing these posets by *Hasse diagrams* we obtain a compact DAG representation of the neighborhood. Since redundancy between two target sets occurs when they have the same common elements or the same differing elements, we say that a target set T in a neighborhood N is *pertinent* for a query set Q iff $T \cap Q \neq \varnothing$ and $\neg\exists T'$ with the same differences w.r.t. Q; without loss of generality this can be defined using the cardinality of these set differences, and can be implemented by a breadth-first bottom-up traversal of the DAG representation [11].

3.4. Modeling and Formal Methods

Early work of members of the MAGNOME team concerned formal methods for the modeling and analysis of complex industrial systems, including model checking and reliability analysis. These problems are generally characterized by an explosion in the size of the state space, whether that space represents the behavior of a system or a truth table encoding a boolean function. An early focus of our work was consequently the efficient encoding of complex sets in systems with uniform sharing of congruent subsets [69], [81], [61], in large part ordered binary decision diagrams, and on semantics-preserving rewriting transformations of these representations to dynamically improve performance [62].

To model systems with stochastic behavior we have extended the AltaRica modeling language [25] with probabalistic choice to define a language whose execution semantics is provided by constraint automata. Constraints between state variables implicitly define transitions, and by assigning weight and durations to transition triggers, generalize both discrete processes such as Markov chains, and continuous stochastic processes such as Markov continuous processes with memory. When exponential laws are used for transition probabilities, the resulting system is a Markov process; accessibility is thus decidable, and model checking can be performed. When other probability laws are used, the system is general stochastic, and only simulation can be used to study mode behavior.

Simulation of complex systems with both continuous and discrete components is hampered by the mix of formalisms and, specifically, by the absence of a formal semantics for combinations of components. We have defined a formal framework for such combinations [72], whose semantics is provided by hybrid automata [48]. An added benefit is that the defined models are hierarchical: each component describes a specific automaton, and components are combined together by composition functions such as parallel composition, connection between Input/Output variables, and synchronization on events. The low-level explicit formalism of BioRica is built upon General Semi-Markovian Decision Processes, an expressive semi-discrete formal model that has been shown [73] to capture most discrete and continuous models while being able to approximate at any precision arbitrary continuous and hybrid processes [72].

Another major challenge for using formal methods for real-world systems is that they lack the ability to reason about the creation and destruction of entities involved. However, this is an essential part of any biological system as can be exemplified by cell division and death, or protein synthesis and degradation. Formally speaking, such models exhibit *infinite behavior*, since we cannot reasonably consider a fixed bound on the number of created entities. Set automata provide a formalism able to describe infinite set computations. In general such systems are indecidable. We have characterized decidable subclasses possessing maximal expressivity. Automatic verification of the expected behaviour of these models can be expressed in the temporal logic *AllTL* [40]. We have extended AllTL to allow for quantification over entities and comparison with automata variables [63].

A distinguishing feature of our approach is the systematic use of *abstraction*. In the case of *AllTL* we define an automatic sound and complete parametrized abstraction that can reduce the infinite state transition system to a finite one. Model checking of such systems is decidable and properties can be verified using standard automata theoretic techniques. In the case of dynamic hybrid systems in BioRica, we use abstraction to reduce an infinite control with finite data, to a finite control on infinite data. The properties of these systems can be studied using counter automata and counter abstraction [60], [65].

4. Application Domains

4.1. Comparative Genomics of Yeasts

Keywords: *bio-technologies*, *biology*, *health*.

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], [35], [53], [52], [39], [56], [41].

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

Keywords: *biology*, *health*, *metabolic pathways*, *protein interaction 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 (see below) 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

Keywords: *bio-technologies*, *biology*, *health*, *stochastic models*.

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 the elementary 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 [80]. 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

Keywords: *collaborative workflows*, *genome annotation*, *in silico analysis*.

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

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 [75], various biomedical ontologies (http://obo.sf.net), and a web interface implementing a representational state transfer (REST) architecture [43].

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

5.2. BioRica: Multi-scale Stochastic Modeling

Keywords: *formal methods*, *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 [55].

We are developing *BioRica*, 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. 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* provides an automatic 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 that adopts, in particular, the theoretical framework of *abstract interpretation* [37].

The current stage of development extends the AltaRica modeling language to Stochastic AltaRica Dataflow [73] semantics, but also provides parsers for widely used SBML [49] data exchange format. The corresponding simulator is easy to use and computationally efficient.

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

5.3. Génolevures On Line: Comparative Genomics of Yeasts

Keywords: *comparative genomics*, *databases*, *knowledge representation and ontologies*, *web design*.

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

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) [59] in Europe and by the SGD (http://www.yeastgenome.org/) [34] in North America, as well as in general-purpose databases such as UniProtKB [26] and EMBL [51].

While extensive chromosomal rearrangements combined with segmental and massive duplications make comparisons of yeast genome sequences difficult [71], relations of homology between protein-coding genes can be identified despite their great diversity at the molecular level [41]. 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 [16] applied to a complementary set of homeomorphic [sharing full-length sequence similarity and similar domain architectures, see [82]] and nonhomeomorphic systematic Smith-Waterman [70] and Blast [24] sequence alignments. Similar approaches are developed on a wider scale [82] 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 [42] 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 [75]. The Blast service is provided by NCBI Blast 2.2.6 [24]. The Génolevures web site uses a REST architecture internally [43] and extensively uses the BioPerl package [74] for manipulation of sequence data.

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

5.4. ProViz: Visualization of Protein Interaction Networks

Keywords: *protein-protein interaction networks*, *scientific visualization*.

Participants: David James Sherman [correspondant], Florian Iragne, Simon Frey.

ProViz is a software tool that provides highly interactive visualization of large networks of protein-protein interactions, integrated with the IntAct data model[6]. ProViz is similar in purpose to PIMrider [57], Osprey [33], and other visualization or analysis tools. ProViz improves over existing work by providing a fast, scalable, open tool with extensive plugins, that integrates emerging standards for representing biological knowledge in a biologist-oriented interface.

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

6. New Results

6.1. Algorithms for genome analysis

Keywords: *collaborative workflows*, *genome annotation*, *in silico analysis*.

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

We have developed new data-mining methods for some specific challenges in comparative genomics.

The novel consensus clustering algorithm of Macha Nikolski and David Sherman [16] provides an efficient approximation in low-order polynomial time to the NP-complete Consensus Clustering problem. The algorithm uses a compact coding of the confusion matrix to efficiently identify conflict regions, resolved conflicts using a Condorcet election procedure and relaxation of the problem to computation of a maximal *inexact* set cover, and performs well in applications to protein family definition as described previously. We have refined the algorithm to accurately cluster data over a nonuniform distribution of evolutionary distances. This algorithm was validated on previous and gold-standard datasets, and applied on a large scale to cluster 725×10^6 pairwise gene relations obtained from systematic homeomorphic and nonhomeomorphic Blast and Smith-Waterman alignments. The resulting analysis has formed the basis for a large-scale identification of orthologues and primary homologues, currently being performed within the Génolevures Consortium.

Anastasia Nikolskaya of the Protein Information Resource at Georgetown University (Washington, DC) spent three weeks with us as an invited professor sharing her expertise in protein classification. Together we were able to refine the rules used for consensus computation, and the improved method produces an automatic classification very close to the gold standard provided by manual curation.

The novel algorithm of Pascal Durrens for identifying gene fusion events using a combination of graph-theoretic algorithms and comparison of hidden Markov models, has been extended to take into account congruences in the clustering graph. This leads to better factorization of the graph of fusion/fission event, and a higher-level view of the underlying mechanisms. Through an complete study of 12 fungal genomes, Pascal has identified a variety of competing mechanisms that contribute to such acquisition of new function, and demonstrated that fusion/fission rates define a metric for genome distance.

6.2. Genome rearrangements

Keywords: *algorithmic combinatorics*, *genome architecture*.

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

Macha Nikolski and Géraldine Jean have devised a new combinatorial method for identifying *super-blocks* of syntenic segments, 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. Using *in silico* chromosomal painting in various sets of contemporary genomes, we identified common sets of markers, and used the algorithm to identify super-blocks. Comparisons with previously published data on placental mammals were used to validate the procedure, and new data from five genomes from the *Kluyveromyces* and related clades was analyzed. In the latter, 16 sets of 34-5 super-blocks were found; 29 super-blocks are common to all sets, and 6 can be resolved using a consensus procedure. Analysis of the composition of these super-blocks will prove informative.

6.3. Computation of genome medians

Keywords: *algorithmic combinatorics*, *genome architecture*, *metaheuristic optimization*.

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

Given an encoding of a set of contemporary genomes as signed permutations, the Hannenhalli-Pevzner model defines a rearrangement distance based on the number of inversions necessary to change one permutation into another. A *median genome* for a set of at least three genomes is a permutation that minimizes the sum of rearrangement distances to the contemporary genomes. Current methods for multichromosomal genomes (in particular [32]) use an exact, resource-intensive computation.

Using a new formulation in terms of optimization, Adrien Goëffon with Macha Nikolski and Géraldine Jean devised a new algorithm using techniques from optimization by local search and metaheuristics. 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. The algorithm was successfully applied to five complete genomes using markers identified by *in silico* chromosomal painting

6.4. Modeling through comparative genomics

Keywords: *comparative genomics*, *stochastic modeling*, *systems biology*.

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

Using comparative genomics to inform mathematical models of cell function is a central challenge of the MAGNOME research program.

Nicolás Loira has used a large dataset of protein families from the Génolevures complete genomes and sub-partitionned it through clustering methods to obtain reliable indications of enzyme conservation in nine species. His method computes a consensus from a competition between three complementary clusterings: primary and secondary homologs (collaboration with Fredj Tekaia of the Pasteur Institute, Paris), syntenic homologs (collaboration with Philippe Baret of the UCL, Belgium), and Pfam domain architectures. The resulting determination of enzyme conservation is mapped to biochemical reation models (BIGG, KEGG, BioCyc, YSBN) and used to infer stoichiometric models that are currently being evaluated through simulation.

6.5. Experimental validation of predicted interactions

Keywords: *metabolism*, *systems biology*.

Participants: David James Sherman, Pascal Durrens [correspondant], Macha Nikolski, Emmanuelle Beyne.

Emmanuelle Beyne has extended her previous work on validation of predicted protein-protein interaction networks for *Y. lipolytica*, to a large-scale experimental study using quantitative proteomics and expression data. During a visit to Prof. Steve Oliver's lab at Cambridge University, Emmanuelle performed these experimental analyses based on cultures under different conditions, in order to capture changes in metabolism. These results are being used to validate and refine mathematical models of *Y. lipolytica* metabolism, extrapolated from reference models through comparative genomics. As with Emmanuelle's previous work in predicting and evaluating predictions of protein complex formation, this unique combination of informatic and bench biology techniques will prove highly informative.

6.6. Genome annotation

Keywords: *algorithmic combinatorics*, *genome architecture*, *metaheuristic optimization*.

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

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 54-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 will be published in the beginning of next year.

6.7. New hybrid model for cell senescence

Keywords: *formal methods*, *parameter estimation*, *systems biology*.

Participants: Hayssam Soueidan, Macha Nikolski [correspondant], David Sherman.

Hayssam Soueidan, in collaboration with Marija Cvijovic of MPI Berlin, used her work on models for yeast senescence as a test case for comparing the xCellerator (Mathematica) and BioRica (Bordeaux) simulation tools.

Most insightful xCellerator models are comprised of a continuous part defined by an Ordinary Differential Equation and a discrete part defined by triggers. Formally, these are "hybrid" models mixing continuous and discrete behaviors. BioRica is explicitly designed for hierarchical modeling and simulation of hybrid models.

We evaluated hybrid modeling in the two systems by translating a yeast replicative senescence model defined in xCellerator into a BioRica model. While producing precisely similar results, the BioRica simulation package outperformed xCellerator by two orders of magnitude due to the use of a native code compiler. This gain thus allowed us to exhaustively explore the initial model by running large automatic simulation batches, even for parameter values that were problematic in the xCellerator package.

However, lack of a fully-functional SBML importer for BioRica was revealed as a real hindrance, and such importing was identified as a key element for tool compatibility. Priority developement of such an importer will permit the existing components of the BioRica toolbox to be connected to existing frameworks and combine the ease of use of other products with the simulation performace of BioRica.

6.8. System identification and parameter estimation methods adapted to properties of biological data

Keywords: *formal methods*, *parameter estimation*, *systems biology*.

Participants: Hayssam Soueidan, Macha Nikolski [correspondant], David Sherman.

Many complex biological process, such as the cell division cycle, involve replicative behaviors where a process can evolve and create another process. Since the initial values of the latter depend on the process state of the former, simulation of such hierarchical systems requires parameter computation and estimation at simulation time.

To this end, we exploited the object-oriented nature of BioRica models by using parallel composition and node instantiation to describe dynamical hierarchical systems. This approach clearly minimizes the modelisation overhead needed to transform a unitary process into a hierarchical system. For example, only three additional lines were needed to transform a single-cell model of replicative senescence into a model for yeast cell populations. This enriched model allowed for the prediction of previously uncomputable behaviors. Simulation data produced from this model are currently being compared to experimental data obtained by Thomas Nystrom's group (Göteborg).

7. Other Grants and Activities

7.1. International Activities

7.1.1. HUPO Proteomics Standards Initiative

Participants: David James Sherman [correspondant], Sandrine Palcy, 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, Cyril Cayla.

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.

7.2. European Activities

7.2.1. Yeast Systems Biology Network (FP6)

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

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], Sandrine Palcy, 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[17] 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. ACI IMPBIO Génolevures En Ligne

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

Génolevures On Line is a public database and a collection of tools for comparative genome analysis, made available to the international community by means of a web site maintained at Bordeaux. In the context of the ACI IMPBIO national program, we develop new resources for Génolevures On Line and maintain existing services, in order to best exploit existing data and efficiently support our common scientific projects for multi-criteria genome comparison. This ACI IMPBIO project involves Jean-Luc Souciet (Strasbourg), Bernard Dujon (Institut Pasteur), Claude Gaillardin (INA-PG), Jean Weissenbach (Génoscope Évry), and the MAGNOME team.

7.3.2. ANR GENARISE

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

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.

7.4. Regional Actions

7.4.1. Aquitaine Region "Génotypage et génomique comparée"

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

In collaboration with the team of Prof. Benoît Arveiler (*Pathologie moléculaire et thérapie génique*) at Universite Victor Ségalen Bordeaux 2 and the Platform for Functional Genomics, we develop new algorithmic techniques and scientific resources for genome sequence analysis. This action provides hardware support for our data warehouse and high-performance computing cluster.

7.4.2. Aquitaine Region "Pôle Recherche en Informatique"

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

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.

7.4.3. 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).

8.2. Memberships and Responsabilities

David Sherman is head of Bioinformatics for the Génolevures Consortium.

Tiphaine Martin and David Sherman are members of the *Institut de Grilles*

David Sherman is member of the steering committee for the Center for Bioinformatics at Bordeaux (CBiB)

8.3. Recruiting committees

Macha Nikolski is external member of the *Commission de spécialistes section 27* of the University Évry Val d'Essonne.

David Sherman is external member of the *Commission de spécialistes section 27* of the University Bordeaux 3.

Pascal Durrens is external member of the *Commission de spécialistes section 65* of the University Victor Ségalen Bordeaux 2.

8.4. Visiting Faculty

Four week stay in Bordeaux of Toby Gibson of the EMBL Heidelberg as invited professor of the University Bordeaux 1.

Three week stay in Bordeaux of Anastasia Nikolskaya of Georgetown University (Washington, DC) as invited professor of the INRIA.

8.5. Participation in colloquia, seminars, invitations

European Conference on Computational Biology (ECCB): Eilat, Israel, January 2007. David Sherman, Géraldine Jean.
David Sherman presented "Family relationships: should consensus reign?"

Marija Cvijovic of Edda Klipp's group at the Max Planck Institute, Berlin, was invited for a short stay in Bordeaux to work with the team on large-scale simulation of a model of cell aging.

ProteomeBinders annual meeting: Alpbach, Austria, March 2007. David Sherman.

Yeast Systems Biology Network semi-annual meeting: Gosau, Austria, March 2007. David Sherman.

INRIA Life Sciences Computing Day: Paris, France, April 2007. David Sherman

FungEffector Conference: Bordeaux, France, May 2007. Pascal Durrens, David Sherman.
Pascal Durrens was invited to present "La comparaison de génomes de levures illustre les mécanismes d'évolution et d'adaptation."
David Sherman was invited to present "Bases de données et fouille de données chez les levures hémiascomycètes."

Organization of the EMBRACE European course "Exploring Modular Protein Architecture," taught by Toby Gibson and colleagues.

Conference INRIA: Lyon, France, June 2007. Nicolás Loira.

JOBIM national conference Satellite day: Marseille, France, July 2007. Géraldine Jean. She presented "Méthode *i*n silico pour la reconstruction d'une architecture ancestrale de génome."

Meeting of experts on protein families: Bordeaux, France, June 2007. David Sherman, Anastasia Nikolskaya (prof invité Georgetown University), Fredj Tekaia (Pasteur Institute), Philippe Baret (UCL Belgium).

XXIIIrd International Conference on Yeast Genetics and Molecular Biology: Melbourne, Australia, July 2007. Tiphaine Martin, Pascal Durrens.
Pascal Durrens presented "Gene Fusion and Fission Events in Fungal Genomes."

Otto Warburg International Summer School on Computational Systems Biology: Berlin, Germany, August 2007. Hayssam Soueidan.

Hayssam Soueidan was invited for a week to work in the laboratory of Edda Klipp at the Max Planck Institute, Berlin. August 2007.

Foundations of Systems Biology in Engineering: Stuttgart, Germany, September 2007. Hayssam Soueidan.

Yeast Systems Biology Network semi-annual meeting: Milano, Italy, September 2007. Macha Nikolski, Nicolás Loira.

ESF-EMBO Colloquium on Comparative Genomics of Microorganisms: Sant Feliu de Guixols, Spain, October 2007: David Sherman, Pascal Durrens, Géraldine Jean, Nicolás Loira, Emmanuelle Beyne.
David Sherman chaired the session on databases and bioinformatics tools.
Géraldine Jean presented "Reconstruction and visualization of genome rearrangements within the Kluyveromyces."

Troisième journées **Alta Rica**: Bordeaux, France. Hayssam Soueidan, Macha Nikolski.
Hayssam Soueidan presented "Extensions Alta Rica pour la modélisation en biologie intégrative."

Emmanuelle Beyne was invited for 6 weeks to work in the laboratory of Prof. Steve Oliver, Cambridge University (UK). December 2007.

Inaugural meeting of the **Institut de Grilles**: Paris, France, December 2007. Tiphaine Martin.

Génolevures and ANR GENARISE monthly meetings: Paris, France, February-December 2007. David Sherman, Pascal Durrens, Tiphaine Martin, Cyril Cayla, Emmanuelle Beyne, Florian Iragne, Géraldine Jean, Nicolás Loira, Adrien Goëffon.

8.6. Teaching

David Sherman is on the faculty of the École Nationale Supérieure d'Informatique, Électronique et Radiocommunication de Bordeaux (ENSEIRB) and teaches in the first, second, and third years. In 2004-2006 he was seconded to the CNRS, and is currently seconded to INRIA. This year he taught the Data-mining course in the Master 2 in Informatics at the University Bordeaux 1, and supervised two students independent research projects. He also supervised the summer internship of Meryem Mekouar.

Pascal Durrens teaches the Bioinformatics class in the Master of Bioinformatics program, co-listed with the University Bordeaux 1 (Sciences and Technologies) and the University Victor Ségalen Bordeaux 2 (Medical School).

Macha Nikolski supervised with David Sherman a project in the second year of the ENSEIRB engineering school.

All of the doctoral students in MAGNOME have teaching duties as teaching assistants, in the Universities Bordeaux 1 and Victor Ségalen Bordeaux 2.

9. Bibliography

Major publications by the team in recent years

[1] R. BARRIOT, J. POIX, A. GROPPI, A. BARRÉ, N. GOFFARD, D. SHERMAN, I. DUTOUR, A. D. DARUVAR. *New strategy for the representation and the integration of biomolecular knowledge at a cellular scale*, in "Nucleic Acids Res.", vol. 32, 2004, p. 3581–3589.

[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. SHERMAN, G. FISCHER, P. DURRENS, S. CASARÉGOLA, 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. BOIRAMÉ, J. BOYER, L. CATTOLICO, F. CONFANIOLERI, A. D. DARUVAR, L. DESPONS, E. FABRE, C. FAIRHEAD, H. FERRY-DUMAZET, A. GROPPI, F. HANTRAYE, C. HENNEQUIN, N. JAUNIAUX, P. JOYET, R. KACHOURI, 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. SWENNENE, 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, 2004, p. 35–44.

[4] G. FISCHER, C. NEUVÉGLISE, P. DURRENS, C. GAILLARDIN, B. DUJON. *Evolution of gene order in the genomes of two related yeast species*, in "Genome Res.", vol. 11, 2001, p. 2009–2019.

[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. 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. SHERMAN, R. APWEILER. *IntAct: an open source molecular interaction database*, in "Nucleic Acids Res.", vol. 32, Jan. 2004, p. D452-5.

[7] F. IRAGNE, M. NIKOLSKI, B. MATHIEU, D. AUBER, D. SHERMAN. *ProViz: protein interaction visualization and exploration*, in "Bioinformatics", Advance Access Publication 3 September 2004, vol. 21, n^{o} 2, 2005, p. 272-4.

[8] M. NIKOLSKI, D. SHERMAN. *Family relationships: should consensus reign?- consensus clustering for protein families*, in "Bioinformatics", vol. 23, 2007, p. e71–e76.

[9] S. ORCHARD, H. HERMJAKOB, R. JULIAN, K. RUNTE, D. SHERMAN, J. WOJCIK, W. ZHU, R. APWEILER. *Common interchange standards for proteomics data: Public availability of tools and schema*, in "Proteomics", vol. 4, 2004, p. 490-1.

[10] D. SHERMAN, P. DURRENS, F. IRAGNE, E. BEYNE, M. NIKOLSKI, J.-L. SOUCIET. *Génolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts*, in "Nucleic Acids Res.", vol. 34, 2006, p. D432–435.

Year Publications

Articles in refereed journals and book chapters

[11] R. BARRIOT, D. SHERMAN, I. DUTOUR. *How to decide which are the most pertinent overly-represented features during gene set enrichment analysis*, in "BMC Bioinformatics", vol. 8, n^{o} 332, 2007.

[12] F. IRAGNE, M. NIKOLSKI, D. SHERMAN. *Extrapolation of metabolic pathways as an aid to modelling completely sequenced non-Saccharomyces yeasts*, in "FEMS Yeast Res.", Epub ahead of print, 2007.

[13] G. JEAN, M. NIKOLSKI. *Genome rearrangements: a correct algorithm for optimal capping*, in "Information Processing Letters", vol. 104, n^{o} 1, 2007, p. 14–20.

[14] 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 Saccharomyces cerevisiae strains*, in "FEMS Yeast Res.", vol. 7, n^{o} 6, 2007, p. 941–52.

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

[16] M. NIKOLSKI, D. SHERMAN. *Family relationships: should consensus reign?- consensus clustering for protein families*, in "Bioinformatics", vol. 23, 2007, p. e71–e76.

[17] 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. 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° 1, 2007, p. 13–17.

Publications in Conferences and Workshops

[18] G. JEAN. *Reconstruction and visualization of genome rearrangements within the Kluyveromyces*, in "ESF-EMBO Symposium on Comparative Genomics of Eukaryotic Microorganisms, Sant Feliu de Guixols, Spain", E. J. LOUIS, T. BOEKHOUT (editors), October 2007.

[19] H. SOUEIDAN, M. NIKOLSKI, D. SHERMAN. *BioRica: A multi model description and simulation system*, in "Proceedings of Foundations of Systems Biology in Engineering (FOSBE), Stuttgart, Germany", F. ALLGÖWER, M. REUSS (editors), ISBN 978-3-8167-7436-5, September 2007.

Miscellaneous

[20] P. DURRENS. *La comparaison de génomes de levures illustre les mécanismes d'évolution et d'adaptation*, Invited talk at the FungEffector conference, Bordeaux, France, May 2007.

[21] P. DURRENS, M. NIKOLSKI, D. SHERMAN. *Gene Fusion and Fission Events in Fungal Genomes*, Presented at the XXIIIrd International Conference on Yeast Genetics and Molecular Biology, Melbourne, Australia, July 2007.

[22] G. JEAN. *Méthode in silico pour la reconstruction d'une architecture ancestrale de génome*, Talk at the JOBIM Satellite meeting, July 2007.

[23] D. SHERMAN. *Bases de données et fouille de données chez les levures hémiascomycètes*, Invited talk at the FungEffector conference, Bordeaux, France, May 2007.

References in notes

[24] 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.

[25] A. ARNOLD, G. POINT, A. GRIFFAULT, A. RAUZY. *The AltaRica formalism for describing concurrent systems*, in "Fundam. Inf.", vol. 40, n° 2-3, 1999, p. 109–124.

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

[27] P. BALDI, S. BRUNAK. *Bioinformatics: The Machine Learning Approach*, Adaptive Computation and Machine Learning Series, MIT Press, Cambridge, Massachusetts, 1998.

[28] R. BARRIOT, J. POIX, A. GROPPI, A. BARRÉ, N. GOFFARD, D. SHERMAN, I. DUTOUR, A. D. DARUVAR. *New strategy for the representation and the integration of biomolecular knowledge at a cellular scale*, in "Nucleic Acids Res.", vol. 32, 2004, p. 3581–3589.

[29] J.-P. BARTHÉLEMY, B. LECLERC. *The median procedure for partitions*, in "DIMACS Series in Discrete Mathematics and Theoretical Computer Science", 1995.

[30] E. BELL. *Exponential numbers*, in "Amer. Math. Monthly", vol. 41, 1934, p. 411–419.

[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] G. BOURQUE, P. PEVZNER. *Genome-scale evolution: reconstructing gene orders in ancestral species*, in "Genome Res.", vol. 12, 2002, p. 9748-9753.

[33] B. BREITKREUTZ, C. STARK, M. TYERS. *Osprey: a network visualization system*, in "Genome Biology", vol. 4, n^{o} 3, 2003, R22.

[34] 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.

[35] 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.

[36] F. S. COLLINS, E. D. GREEN, A. E. GUTTMACHER, M. S. GUYER. *A vision for the future of genomics research*, in "Nature", vol. 422, April 2003, p. 835–847.

[37] 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.

[38] A. DANCHIN. *La Barque de Delphes*, Éditions Odile Jacob, 1998.

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

[40] D. DISTEFANO. *On model checking the dynamics of object based software*, Ph. D. Thesis, University of Twente, 2003.

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

[42] 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.

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

[44] A. FRED, A. JAIN.. *Data clustering using evidence accumulation*, in "In Proc. of the 16th Intl. Conference on Pattern Recognition (ICPR 2002)", 2002, p. 276–280.

[45] M. GAREY, D. JOHNSON. *Computers and Intractability; A Guide to the Theory of NP-Completeness*, W. H. Freeman & Co., 1979.

[46] M. GRÖTSCHEL, Y. WAKABAYASHI. *A cutting plane algorithm for a clustering problem*, in "Mathematical Programming B", vol. 59–96, 1989.

[47] S. HANNENHALLI, P. PEVZNER. *Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals)*, in "Proc. 27th Annual ACM-SIAM Symposium on the Theory of Computing", 1995, p. 178–189.

[48] T. HENZINGER. *The theory of hybrid automata*, in "Proceedings of the 11th Annual IEEE Symposium on Logic in Computer Science, New Jersey", 1996, p. 278–292.

[49] M. HUCKA, ET 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.

[50] M. KANEHISA, P. BORK. *Bioinformatics in the post-sequence era*, in "Nature Gen.", Review, vol. 33, March 2003, p. 305–310.

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

[52] 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.

[53] 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.

[54] P. KHATRI, S. DRAGHICI. *Ontological analysis of gene expression data: current tools, limitations, and open problems*, in "Bioinformatics", vol. 21, n^{o} 18, 2005, p. 3587–3595.

[55] 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.

[56] 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.

[57] P. LEGRAIN, J. WOJCIK, J. GAUTHIER. *Protein-protein interaction maps: a lead towards cellular functions*, in "Trends in Genetics", vol. 17, 2001.

[58] M. MEILA. *Comparing Clusterings by the Variation of Information*, in "Proceeding of COLT'2003", 2003, p. 173–187.

[59] 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.

[60] M. MINSKY. *Computation: Finite and Infinite Machines*, Prentice-Hall, 1967.

[61] M. NIKOLSKAIA, L. NIKOLSKAIA. *Size of OBDD representation of 2-level redundancies functions*, in "Theoretical Computer Science", vol. 255, n^{o} 1-2, 2001, p. 615-625.

[62] M. NIKOLSKAÏA, A. RAUZY, D. J. SHERMAN. *Almana: A BDD Minimization Tool Integrating Heuristic and Rewriting Methods*, in "Formal Methods in Computer-Aided Design, Second International Cnference, FMCAD'98, Palo Alto, California", G. GOPALAKRISHNAN, P. WINDLEY (editors), Springer-Verlag LNCS 1522, November 1998.

[63] M. NIKOLSKI, H. SOUEIDAN, G. SUTRE. *Decidability of Model Checking Set Automata*, Submitted for publication, October 2006.

[64] P. PEVZNER, G. TESLER. *Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes*, in "Genome Research", vol. 13, 2002, p. 37–45.

[65] A. PNUELI, J. XU, L. D. ZUCK. *Liveness with (0, 1, infty)-Counter Abstraction*, in "CAV '02: Proceedings of the 14th International Conference on Computer Aided Verification, London, UK", Springer-Verlag, 2002, p. 107–122.

[66] S. RÉGNIER. *Sur quelques aspects mathématiques des problèmes de classification automatique*, in "ICC Bulletin", vol. 4, 1965, p. 175–191.

[67] D. SHERMAN, P. DURRENS, E. BEYNE, M. NIKOLSKI, J.-L. SOUCIET, GŃOLEVURES CONSORTIUM. *Génolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts*, in "Nucleic Acids Res.", vol. 32, 2004, p. D315–D318.

[68] D. SHERMAN, P. DURRENS, F. IRAGNE, E. BEYNE, M. NIKOLSKI, J.-L. SOUCIET, GŃOLEVURES CONSORTIUM. *Génolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts*, in "Nucleic Acids Res.", vol. 34, 2006, p. D432–435.

[69] D. J. SHERMAN, N. MAGNIER. *Factotum: Automatic and Systematic Sharing Support for Systems Analyzers*, in "Tools and Algorithms for the Construction and Analysis of Systems (TACAS'98), Lisbon, Portugal", B. STEFFEN (editor), Springer-Verlag LNCS 1384, March-April 1998.

[70] T. F. SMITH, M. WATERMAN. *Identification of common molecular subsequences*, in "Journal of Molecular Biology", 1981.

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

[72] H. SOUEIDAN, M. NIKOLSKI. *BioRica: Continuous and discrete modular models*, Submitted for publication, 2006.

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

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

[75] 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.

[76] A. STREHL, J. GHOSH. *Cluster ensembles – a knowledge reuse framework for combining multiple partitions*, in "The Journal of Machine Learning Research archive", vol. 3, 2003, p. 583–617.

[77] G. TESLER. *Efficient Algorithms for multichromosomal genome rearrangements*, in "J. Comp. Sys. Sci.", vol. 65, 2002, p. 587–609.

[78] A. TOPCHY, A. JAIN, W. PUNCH. *A Mixture Model for Clustering Ensembles*, in "Proc. SIAM Conf. on Data Mining", 2004, p. 379-390.

[79] A. TOPCHY, M. LAW, A. JAIN, A. FRED. *Analysis of Consensus Partition in Cluster Ensemble*, in "Proc. IEEE International Conference on Data Mining (ICDM'04)", 2004, p. 225–232.

[80] 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.

[81] P. WILLIAMS, M. NIKOLSKAÏA, A. RAUZY. *Bypassing BDD construction for reliability analysis*, in "Information Processing Letters", vol. 75, n^{o} 1–2, 2000, p. 85–89.

[82] 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.", vol. 32, 2004, p. D315–D318.