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 **1**
2. Overall Objectives **1**
2.1. Overall Objectives 1
3. Scientific Foundations **2**
3.1. Multi-scale Models of Data and Behavior 2
3.2. Algorithms for Sequence Analysis 3
3.3. Data-mining and Classification 3
4. Application Domains **4**
4.1. Comparative Genomics of Yeasts 4
4.2. Construction of Biological Networks 5
4.3. Modeling Biological Systems 5
5. Software **7**
5.1. Magus: Collaborative Genome Annotation 7
5.2. Génolevures On Line: Comparative Genomics of Yeasts 7
5.3. ProViz: Visualization of Protein Interaction Networks 7
5.4. BioRica: Multi-scale Stochastic Modeling 8
6. New Results **8**
6.1. Clustering and fusions 8
6.2. Genome rearrangements 8
6.3. Modeling through comparative genomics 8
6.4. Set automata 9
6.5. Genome annotation methods 9
7. Other Grants and Activities **9**
7.1. International Activities 9
7.1.1. HUPO Proteomics Standards Initiative 9
7.1.2. Génolevures Consortium 9
7.2. European Activities 10
7.2.1. Yeast Systems Biology Network (FP6) 10
7.2.2. ProteomeBinders (FP6) 10
7.2.3. IntAct 11
7.3. National Activities 11
7.3.1. ACI IMPBIO Génolevures En Ligne 11
7.4. Regional Actions 11
7.4.1. Aquitaine Region “Génotypage et génomique comparée” 11
7.4.2. Aquitaine Region “Pôle Recherche en Informatique” 11
8. Dissemination **11**
8.1. Program committees 11
8.2. Seminars and keynotes 11
8.3. Thesis committees 12
8.4. Reviewing 12
8.5. Recruiting committees 12
8.6. Teaching 12
9. Bibliography **12**

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 the team

David James Sherman [Associate Professor (MCF) ENSEIRB, HdR]

Assistante de projet

Brigitte Cournou [Secretary (SAR) Inria]

Team Member from CNRS

Pascal Durrens [Research scientist (CR1) CNRS]
Macha Nikolski [Research scientist (CR2) 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]

Former Ph.D. Students

Roland Barriot [PhD in 2005, currently Postdoc with Prof. Yves Moreau at K.U. Leuven, Belgium]

Associated Researchers

Sandrine Palcy [PostDoc ProteomeBinders, ENSEIRB]
Grégoire Sutre [Research scientist (CR2) CNRS]
Serge Dulucq [Professor Univ. Bordeaux 1]
Isabelle Dutour [Associate Professor (MCF) Univ. Bordeaux 1]
Antoine De Daruvar [Professor Univ. Bordeaux 2]

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.

3. Scientific Foundations

3.1. Multi-scale Models of Data and Behavior

The core motivation for our research program lies in three fundamental and complex questions in the life sciences for which the computer sciences play an essential role: How do genomes evolve? How do gene products cooperate to realize cellular functions? How, on different scales, are organized biomolecules, genomes, cells, multicellular organisms, and populations? By providing reusable methods for getting to the heart of complex data sets, we bring to these questions both practical means, and formal approaches, for constructing multi-scale models of biological systems and for understanding the corresponding complex phenomena.

The techniques that we apply are multi-scale in three different ways. First, the computer representations we construct of living systems are **structurally multi-scale**: a gene has an exon-intron architecture imposed on its primary mRNA transcript that leads to alternate mature transcripts, each of which has a trinucleotide codon structure that determines a peptide (protein) sequence itself composed of functional domains. Similar proteins can be classified into functional classes, and cooperating proteins (which are necessarily dissimilar since they assure different functions) can be organized into networks that describe their interactions. Components of these networks represent core cellular functions that are present or absent in a given species; an organism is thus composed of a set of intersecting biochemical functions that are determined by the genes in its genome and their regulatory relations. The *data (or knowledge) representations* of these systems must take into account the multiple levels of these often complementary scales. These representations are often identified using *data-mining* and *classification* of large complex datasets. *Formal models* of these systems must be hierarchically composed on multiple scales in order that human experts may design, understand, and validate them.

Second, genomes are **historically multi-scale**. The history of genomes is a complex set of events at multiple evolutionary scales, for which only overlapping and diffuse traces remain in contemporary genomes. These events may segmental-scale or chromosome-scale sequence duplications, small-scale gene fusion or tandem duplications, gene-level deletions, family-level expansions or contractions, and organism-level adoption of new mechanisms or adaptation to specific environments. Investigation of these events cannot be reduced to the sole study of phylogeny. *Combinatorial representations* of combinations of major and minor evolutionary events must take into account these different, interacting, scales.

Third, **multi-scale simulation** is necessary for *in silico* experimentation. Realistic, precise simulation of cell behavior requires detailed formal models and fine-grain interpretation. At the same time, it is necessary that this simulation be computationally tractable. Reaching an effective compromise between these conflicting goals requires techniques at **multiple time scales**, where computational effort can be applied where the behavior of the model is most dynamic (or chaotic) and the needs of precision are greatest. Multi-scale models of cell behavior must take into account these different levels of precision, and provide means for drilling down the scale of simulation in some parts of the model while accelerating or abstracting others.

Instead of simply simulating the behavior of a system, it is possible to study the formal properties of possible executions of the system. Using **formal methods** for models of biological system faces a number of challenges. Among them are the hierarchical structure of these models, their intrinsic stochasticity and the co-existence of multiple time-scales within the same model . Such systems are often computationally intractable for formal verification.

In particular our team considers properties that ought to be satisfied by the model as logical expressions in stochastic temporal logic. We also study other symbolic techniques such as *abstraction* to approximate a model with measurable loss of precision.

Another major obstacle for using formal methods in the context of biological 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 behaviour*, since we can't reasonably consider a fixed bound

on the number of created entities. Set automata provide a formalism able to describe infinite set computations. Our team studies decidable subclasses of set automata for which automatic verification can be expressed in temporal logic.

3.2. Algorithms for Sequence Analysis

Sequence analysis using probabalistic models, notably hidden markov models, are used for syntactic analysis of macromolecular sequences, predicting 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 team develops novel techniques for exploring evolutionary scenarios using signed permutation representations of contemporary genomes, which make it possible to parsimoniously order and date the major rearrangement events that shaped individual lineages. We also develop incremental graph algorithms for the extraction of intelligible subgraphs from protein-protein interaction networks [7].

The first step in decoding a genome sequence is identifying the loci of protein-coding genes. This is a form of syntactic analysis that applies rules derived from models of how the cell's transcriptional machinery recognizes and interprets the DNA sequence. The result of this analysis is a *gene architecture*, that described the structure of the gene in terms of exon, intron, and signal sequences. Unlike higher eukaryotes, yeasts have relatively few intron-bearing genes and those that do appear to have relatively few introns, based on *in silico* analyses [22] and experimental characterization of introns in *S. cerevisiae*. Our own work suggests that detection of yeast gene architectures requires specific rules. We have developed techniques based, on the one hand, on combinatorial search for conserved sequence motifs, and on the other, on supervised classification of candidate gene architectures.

Candidate genes architectures produced by these two intrinsic techniques are then competitively compared with extrinsic data such as sequence alignments and GeneMark predictions of coding potential, using a classification procedure based on decision trees. Ranked candidates can be used automatically or presented to curators for manual annotation.

At a higher level, it is possible in strictly delimited cases to uniquely identify syntenic regions between yeast genomes. Such is the case for *S. cerevisiae* and *C. glabrata*, for example. When these regions of chromosomal homology cover a significant proportion of the individual genomes, it becomes possible to study the set of rearrangement events that occurred during the descent of each species from their common ancestor, under an assumption of parsimony. We have developed a combinatorial method for calculating rearrangement distances using a simple set of operations inspired by [28], [37], but including biologically-inspired constraints such as centromere position and the number of chromosomes. 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. 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 Classification

While the life sciences provide the underlying motivation for our work, as researchers in the computer sciences we address these questions through the development of algorithms and methods in computer science and computational biology.

Data mining and rule inference is a general class of data analysis techniques that search for emerging relations or for new rules describing the relations between elements; the latter can be association rules, decision trees, or other logical expressions. The former look for signals or groupings in data and include classification techniques (described below). Our team has notably developed a novel system for large-scale data mining of relations between sets of elements sharing common properties (called "neighborhood relations") that can be used to look for nonrandom coincidences in classifications determined by analysis of experimental data [1]. We also develop graph-based methods for extracting structures summaries from n-dimensional datasets.

Supervised and nonsupervised classification are machine learning techniques that attempt to learn a function that classifies each observation, represented as a vector of attributes, into one of several classes. Supervised classification methods learn the function by analyzing a set of example observations already assigned to classes, called a *training set*. Nonsupervised classification techniques try to separate observations into "natural" classes determined by the structure of relations between observations, typically using some measure of distance. Some methods require that the number of classes be determined in advance. Some methods do not use a set of disjoint classes, but a hierarchical collection of classes related by inclusion. We develop classification methods for reliably identifying relations between genes.

One classification challenge often encountered in our application domains is the sifting of *in silico* predictions. One can formulate this problem as an instance of the NP-complete *consensus clustering* problem [33], [20]. We have designed a novel algorithm that is computationally efficient in practice and produces high quality results, that uses an election method to construct consensus families from competing clustering computations. This method has been used for Génolevures project to compute protein families of Hemiascomycetous yeasts.

Our consensus clustering algorithm is tailored to serve the specific needs of comparative genomics projects. First, it provides a robust means to incorporate results from different and complementary clustering methods, thus avoiding the need for an *a priori* choice that may introduce computational bias in the results. Second, it is suited to large-scale projects due to the practical efficiency. And third, it produces high quality results where families tend to represent groupings by biological function.

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 [21], [24], [30], [29], [19], [31], [25].

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*.

The gap in complexity between limited data and complex behavior of biological systems can be compared to the behavior complexity emerging in dynamical systems studies (e.g. chaotic systems). The latter is a well-known example of complexity emerging from apparently simple systems. As for biological processes, the large number of interacting components produces a particular difficulty of predicting the emerging behavior.

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

1. Starting from experimental data, sort possible molecular processes and retain the most plausible.
2. Build a cartoon 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. 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. In the case of *Y. lipolytica*, for example, Emmanuelle Beyne of our team works with the team of Marc Bonneu at the Bordeaux Functional Genomics Platform to experimentally identify multi-protein complexes predicted by her *in silico* methods.

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.

The MAGUS genome annotation system integrates genome sequences and sequences features, *in silico* analyses, and views of external data resources into a web-based collaborative platform for annotation of eukaryote genomes. 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.

5.2. 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].

The Génolevures online database (http://cbi.labri.fr/Genolevures/) provides tools and data relative to 4 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, 25 000 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 Génolevures database uses a straightforward object model mapped to a relational database. Flexibility in the design is guaranteed through the use of controlled vocabularies: the Sequence Ontology [26] 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 [36]. The Blast service is provided by NCBI Blast 2.2.6. The Génolevures web site uses a REST architecture internally [27] and extensively uses the BioPerl package [35] for manipulation of sequence data.

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

5.3. ProViz: Visualization of Protein Interaction Networks

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

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

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 [32], Osprey [23], 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.

5.4. BioRica: Multi-scale Stochastic Modeling

Keywords: *formal methods*, *stochastic modeling*.

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

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*.

The current stage of development extends the AltaRica modeling language to Stochastic AltaRica Dataflow [34] semantics, but also provides parsers for widely used SBML data exchange format. The corresponding simulator is easy to use and computationally efficient.

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

6. New Results

6.1. Clustering and fusions

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

Macha Nikolski and David Sherman have devised a novel consensus clustering algorithm. While the general problem is NP-complete, we obtain good practical results in low-order polynomial time through use of a heuristic (based on a Condorcet election procedure) and through relaxation to maximal *inexact* set cover. The paper was accepted to ECCB 2006 and will appear in the journal *Bioinformatics*. This algorithm has been successfully applied to the complete proteomes of five hemiascomycete yeast species and complete results including validation by human curators is available on the Génolevures web site at http://cbi.labri.fr/Genolevures/fam/index.html.

Pascal Durrens has developed a new clustering algorithm for identifying gene fusion events, using a combination of graph-theoretic algorithms and comparison of hidden Markov models. Candidate fusion events are classified based on topology and manually curated. This method was applied to the complete genomes of 10 fungal species.

6.2. Genome rearrangements

Macha Nikolski and Géraldine Jean have devised a new combinatorial method for calculating median genomes under a uniform-cost rearrangement model, improving on algorithms defined respectively by Sankoff and by Bourque/Pevzner. This algorithm was applied in a large-scale simulation of segmental genome rearrangement on five genomes from the *Kluyveromyces* and related clades, and the results have proved informative about the architectures of putative ancestral genomes.

6.3. Modeling through comparative genomics

Using comparative genomics to inform mathematical models of cell function is a central challenge of the MAGNOME research program. This year we made two importants steps in that direction.

Emmanuelle Beyne develops *in silico* methods for predicting protein complexes in yeast genomes. Protein complexes are one form of protein-protein interaction that provide the building blocks of cell machinery. Working with the laboratory of Prof. Marc Bonneu in the Bordeaux Functional Genomics Platform, Emmanuelle has refined experimental methods using gel electrophoresis to validate her predictions. This unique combination of informatic and bench biology techniques will prove highly informative for models of yeast metabolism under different growth conditions.

Florian Iragne has refined his methods for subtractive modeling of biochemical pathways, using his algorithmic framwork for policy-directed graph extraction. By combining gene homolog information with reference pathways, Florian can identify cases of pathway loss through search for correlated gene losses. These techniques has been applied systematically to five whole proteomes from the Hemiascomycete yeasts, and are available on the Génolevures public web site.

6.4. Set automata

Set automata provide a formalism able to describe infinite set computations. In general such systems are indecidable. Hayssam Soueidan, Macha Nikolski, and Grégoire Sutre have succeeded in characterizing decidable subclasses possessing maximal expressivity. Automatic verification of the expected behaviour of these models can be expressed in the temporal logic *AllTL*. We have extended AllTL to allow for quantification over entities and comparison with automata variables. We defined an automatic sound and complete parametrized abstraction that can reduce the infinite state transition system to a finite one, that can be verified using standard automata theoretic techniques.

6.5. Genome annotation methods

David Sherman and Tiphaine Martin have defined an improved whole genome annotation system that provides a complete analysis pipeline for automatic pre-annotation of uncharacterized genomes. The core method integrates gene calling, putative association with Génolevures protein families, and automatic classification of predicted mRNA transcripts. These elements form the basis of the Magus genome annotation system, described above, and are currently being used to annotate three new genomes that were sequences for us by the Centre National de Séquençage - Génoscope, Évry.

7. Other Grants and Activities

7.1. International Activities

7.1.1. HUPO Proteomics Standards Initiative

Participant: David James Sherman.

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, 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.

7.2. European Activities

7.2.1. Yeast Systems Biology Network (FP6)

Participants: David James Sherman, Macha Nikolski.

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)

Participant: David James Sherman.

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 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

Participant: David James Sherman.

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, Pascal Durrens, Macha Nikolski.

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.

7.4. Regional Actions

7.4.1. Aquitaine Region "Génotypage et génomique comparée"

Participants: David James Sherman, Pascal Durrens.

7.4.2. Aquitaine Region "Pôle Recherche en Informatique"

Participants: David James Sherman, Pascal Durrens, Macha Nikolski.

8. Dissemination

8.1. Program committees

Pascal Durrens was co-president on the Program committee for JOBIM (Journées Ouvertes Biologie, Informatique, Mathématiques) 2006.

David Sherman was a member of the Program committee for JOBIM 2006.

8.2. Seminars and keynotes

David Sherman was a keynote speaker at the Yeast Systems Biology Network 1st Workshop in Vienna, November, 2006. His speech was entitled "Mining the Hemiascomycete Yeasts."

David Sherman was invited to Georgetown University in July, 2006 and gave a seminar entitled “Protein Families in Yeasts.”

Macha Nikolski was invited to the Mathematics department of St. Petersburg State University in Russia, where she gave a seminar entitled “Consensus clustering for protein families.”

David Sherman, Macha Nikolski, and Géraldine Jean were each invited to speak in the Belgian Academy of Arts and Sciences, Brussels, in September 2006 for the “Yeast Genome Tenth Anniversary.” DS spoke about knowledge representation in the Génolevures comparative genomics database, MN spoke about consensus clustering, and GJ spoke about computation of median ancestral genomes.

8.3. Thesis committees

David Sherman was an external reporter for the thesis of Matthieu Defrance, defended at the University Lille 1 in December, 2006.

8.4. Reviewing

Macha Nikolski was reviewer for the journal *Techniques et Science Informatiques* (Hermes Lavoisier), special issue “Modélisation et simulation pour la post-génomique.”

David Sherman was reviewer for the journal *Yeast* (Wiley).

8.5. Recruiting committees

Macha Nikolski is external member of the *Commission de spécialistes section 27* of the University Évry Vall 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.

8.6. 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. The previous two years (2004-2006) he was seconded to the CNRS.

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 taught an Algorithmics and Data Structures class at the undergraduate level at the University Bordeaux 1.

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] 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.

[9] 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

[10] J. LASSERRE, E. BEYNE, S. PYNDIAH, D. LAPAILLERIE, S. CLAVEROL, M. BONNEU. *A complexomic study of Escherichia coli using two-dimensional blue native/SDS polyacrylamide gel electrophoresis*, in "Electrophoresis", vol. 27, n^{o} 16, August 2006, p. 3306-21.

[11] 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.", Epub ahead of print, 2006.

[12] M. NIKOLSKI, D. SHERMAN. *Family relationships: should consensus reign?*, in "Bioinformatics", To appear, 2006.

[13] 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.

Publications in Conferences and Workshops

[14] A. BARRÉ, V. JOUFFE, M. NIKOLSKI, A. D. DARUVAR, A. BLANCHARD, P. SIRAND-PUGNET. *Annotation transfer based on orthology relationships: reannotation of mycoplasma genomes from the Pneumoniae group*, 2006.

[15] E. BEYNE. *Identification and Comparison of Yeast Complexomes*, Oral presentation, Yeast Systems Biology 1st Workshop, Vienna, November 2006.

[16] H. SOUEIDAN, M. NIKOLSKI. *BioRica: Continuous and discrete modular models*, 2006.

[17] H. SOUEIDAN, M. NIKOLSKI, G. SUTRE. *Model Checking AllTL Properties for Set Automata*, in "MOVEP 2006, Bordeaux, France", 2006.

[18] H. SOUEIDAN. *Formal Verification of Biological Systems with Highly Dynamic Creation and Destruction*, Oral presentation, Yeast Systems Biology 1st Workshop, Vienna, November 2006.

References in notes

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

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

[21] 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.

[22] E. BON, S. CASARÉGOLA, G. BLANDIN, B. LLORENTE, C. NEUVÉGLISE, M. MUNSTERKOTTER, U. GULDENER, H. W. MEWES, J. V. HELDEN, B. DUJON, C. GAILLARDIN. *Molecular evolution of eukaryotic*

genomes: hemiascomycetous yeast spliceosomal introns, in "Nucleic Acids Res.", vol. 31, n^{o} 4, 2003, p. 1121-35.

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

[24] 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.

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

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

[34] 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/.

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

[36] 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.

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