IN PARTNERSHIP WITH:
CNRS

Université de Bordeaux

Activity Report 2014

Project-Team MAGNOME

Models and Algorithms for the Genome

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

RESEARCH CENTER
Bordeaux - Sud-Ouest

THEME
Computational Biology

Table of contents

Project-Team MAGNOME

Keywords: Computational Biology, Genomics, Knowledge Engineering, Modeling, High Performance Computing

Creation of the Project-Team: 2009 July 01, end of the Project-Team: 2014 December 31.

1. Members

Research Scientists
David Sherman [Team leader, Inria, Senior Researcher, HdR]
Pascal Durrens [CNRS, Researcher, HdR]

Engineers
Xavier Calcas [CNRS, until Dec 2014]
Florian Lajus [Inria, until May 2014]

PhD Students
Razanne Issa [Exchange Fellowship Syria]
Anna Zhukova [Inria, until Nov 2014]

Post-Doctoral Fellow
Witold Dyrka [Inria, granted by ANR Mykimum project]

Administrative Assistants
Anne-Laure Gautier [Inria]
Flavie Tregan [Inria, from May 2014 to Sep 2014]

Others
Philippe Chaumeil [INRA, from Sep 2014]
Alain Franc [INRA, from Sep 2014]
Jean-Marc Frigerio [INRA, from Sep 2014]

2. Overall Objectives

2.1. Overall Objectives

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

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

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

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

3. Research Program

3.1. Overview

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

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

- **Genome annotation** for complete genomes, performing *syntactic* analyses to identify genes, and *semantic* analyses to map biological meaning to groups of genes [22], [5], [9], [10], [32], [33].
- **Integration of heterogenous data**, to build complete knowledge bases for storing and mining information from various sources, and for unambiguously exchanging this information between knowledge bases [1], [3], [25], [27], [21].
- **Ancestor reconstruction** using optimization techniques, to provide plausible scenarios of the history of genome evolution [10], [7], [28], [34].
- **Classification and logical inference**, to reliably identify similarities between groups of genetic elements, and infer rules through deduction and induction [8], [6], [9].
- **Hierarchical and comparative modeling**, to build mathematical models of the behavior of complex biological systems, in particular through combination, reutilization, and specialization of existing continuous and discrete models [24], [20][12].

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

3.2. Comparative genomics

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

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

The methods designed by MAGNOME for comparative genome annotation, structured genome comparison, and construction of integrated models are applied on a large scale to: eukaryotes from the hemiascomycete class of yeasts [32], [33], [5], [9], [2], [10] and to lactic bacteria used in winemaking [31], [26]

3.3. Comparative modeling

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

Comparative modeling is also a matter of reconciling experimental data with models [4] [20] and inferring new models through a combination of comparative genomics and successive refinement [29], [30].

4. Application Domains

4.1. Function and history of genomes

Yeasts provide an ideal subject matter for the study of eukaryotic microorganisms. From an experimental standpoint, the yeast *Saccharomyces cerevisiae* is a model organism amenable to laboratory use and very widely exploited, resulting in an astonishing array of experimental results. From a genomic standpoint, yeasts from the hemiascomycete class provide a unique tool for studying eukaryotic genome evolution on a large scale. With their relatively small and compact genomes, yeasts offer a unique opportunity to explore eukaryotic genome evolution by comparative analysis of several species. MAGNOME applies its methods for comparative genomics and knowledge engineering to the yeasts through the ten-year old Génolevures program (GDR 2354 CNRS), devoted to large-scale comparisons of yeast genomes with the aim of addressing basic questions of molecular evolution.

We developed the software tools used by the CNRS's http://www.genolevures.org/ web site. For example, MAGNOME's Magus system for simultaneous genome annotation combines semi-supervised classification and rule-based inference in a collaborative web-based system that explicitly uses comparative genomics to simultaneously analyse groups of related genomes.

4.2. Alternative fuels and bioconversion

Oleaginous yeasts are capable of synthesizing lipids from different substrates other than glucose, and current research is attempting to understand this conversions with the goal of optimizing their throughput, production and quality. From a genomic standpoint the objective is to characterize genes involved in the biosynthesis of precursor molecules which will be transformed into fuels, which are thus not derived from petroleum. MAGNOME's focus is in acquiring genome sequences, predicting genes using models learned from genome comparison and sequencing of cDNA transcripts, and comparative annotation. Our overall goal is to define dynamic models that can be used to predict the behavior of modified strains and thus drive selection and genetic engineering.

4.3. Winemaking and improved strain selection

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

Comparative genomics is a powerful tool for strain selection even when genetic engineering must be excluded. Large-scale comparison of the genomes of experimentally characterized strains can be used to identify quantitative trait loci, which can be used as markers in selective breeding strategies. Identifying individual SNPs and predicting their effect can lead to better understanding of the function of genes implicated in improved strain performance, particularly when those genes are naturally mutated or are the result of the transfer of genetic material from other strains. And understanding the combined effect of groups of genes or alleles can lead to insight in the phenomenon of heterosis.

4.4. Knowledge bases for molecular tools

Affinity binders are molecular tools for recognizing protein targets, that play a fundamental in proteomics and clinical diagnostics. Large catalogs of binders from competing technologies (antibodies, DNA/RNA aptamers, artificial scaffolds, etc.) and Europe has set itself the ambitious goal of establishing a comprehensive, characterized and standardized collection of specific binders directed against all individual human proteins, including variant forms and modifications. Despite the central importance of binders, they presently cover only a very small fraction of the proteome, and even though there are many antibodies against some targets (for example, $>$ 900 antibodies against p53), there are none against the vast majority of proteins. Moreover, widely accepted standards for binder characterization are virtually nonexistent. Alongside the technical challenges in producing a comprehensive binder resource are significant logistical challenges, related to the variety of producers and the lack of reliable quality control mechanisms. As part of the ProteomeBinders and Affinomics projects, MAGNOME works to develop knowledge engineering techniques for storing, exploring, and exchanging experimental data used in affinity binder characterization.

5. New Software and Platforms

5.1. Magus: Genome exploration and analysis

Participants: David James Sherman [correspondant], Pascal Durrens, Florian Lajus, Xavier Calcas.

The MAGUS genome annotation system integrates genome sequences and sequences features, *in silico* analyses, and views of external data resources into a familiar user interface requiring only a Web navigator. MAGUS implements annotation workflows and enforces curation standards to guarantee consistency and integrity. As a novel feature the system provides a workflow for simultaneous annotation of related genomes through the use of protein families identified by *in silico* analyses; this results in an n-fold increase in curation speed, compared to curation of individual genes. This allows us to maintain standards of high-quality manual annotation while efficiently using the time of volunteer curators. MAGUS can be used on small installations with a web server and a relational database on a single machine, or scaled out in clusters or elastic clouds using Apache Cassandra for NoSQL data storage and Apache Hadoop for Map-Reduce (figure 1). For more information see the MAGUS Gforge web site. [1] MAGUS 2.0 was developed in an Inria Technology Development Action (ADT) and is distributed with an open-source license.

5.2. Pantograph: Inference of metabolic networks

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

[1] http://magus.gforge.inria.fr

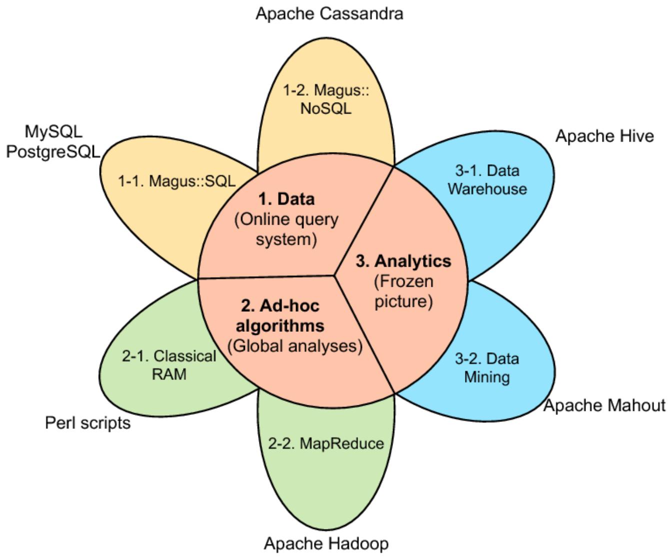


Figure 1. General architecture of the Tsvetok system implemented in MAGUS, showing the role of the NoSQL (Apache Cassandra) and Map-Reduce (Apache Hadoop) paradigms

Pantograph is a software tool developed by Nicolás Loira for his thesis, that infers whole-genome metabolic models for eukaryote cell factories from reference models and genome comparison. A novel feature of Pantograph is that it uses expert knowledge implicitly encoded in the scaffold's gene associations, and explicitly transfers this knowledge to the new model. Pantograph is available under an open-source license. For more information see the Pantograph Gforge web site. [2].

5.3. Mimoza: Generalizing and Visualizing Metabolic Models

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

Mimoza uses metabolic model generalization and cartographic paradigms to allow human experts to explore a metabolic model in a hierarchical manner. The software creates an zoomable representation of a model submitted by the user in SBML [3] format. The most general view represents the compartments of the model; the next view shows the visualization of generalized versions of reactions and metabolites in each compartment (see section 6.3); and the most detailed view visualizes the initial model with the generalization-based layout (where similar metabolites and reactions are placed next to each other). The zoomable representation is implemented using the Leaflet [4] JavaScript library for mobile-friendly interactive maps. Users can click on reactions and compounds to see the information about their annotations. The resulting map can be explored on-line, or downloaded in a COMBINE archive. The software and examples are available at http://mimoza.bordeaux.inria.fr.

5.4. Génolevures On Line: Comparative Genomics of Yeasts

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

[2] http://pathtastic.gforge.inria.fr
[3] http://sbml.org
[4] http://leafletjs.com

The Génolevures online database provides archival data for exploring the annotated genome sequences of more than 20 genomes, determined and manually annotated by the Génolevures Consortium to facilitate comparative genomic studies of hemiascomycetous yeasts. Data are presented with a focus on relations between genes and genomes: conservation of genes and gene families, speciation, chromosomal reorganization and synteny. Génolevures online uses our open-source MAGUS system for genome navigation, with project-specific extensions developed by MAGNOME. For more information see the Génolevures web site. [5]

6. New Results

6.1. Highlights of the Year

In collaboration with colleagues from the Institut du Vigne et du Vin (ISVV), Bordeaux and the Universidade Nova de Lisboa, Lisbon we used a population genomics approach to investigate the global phylogeography and domestication fingerprints of winemaking yeasts, using a collection of isolates obtained from fermented beverages and from natural environments on five continents. These results appeared in *Nature Communications* [11].

6.2. A Gondwanan imprint of S. uvarum diversity

Domestication of livestock and crops has been amply demonstrated through historical and archeological records, but domestication of microorganisms is much more difficult to establish. In a large-scale study [11] of the wine and cider yeast *Saccharomyces uvarum* conducted with colleagues from the Institut du Vigne et du Vin, Bordeaux and the Universidade Nova de Lisboa, Lisbon, we found the first indications of its domestication in the transition from its habitat in *Nothofagus* (southern beech) trees on the Gondwana mega-continent, to its present-day diversity in the Holarctic. The global phylogeography of these microorganisms was investigated through genome sequencing and comparison of 54 strains isolated on five continents, resulting in the identification of 10^5 high-quality SNPs and a remarkable pattern of introgressions ([11] figure 3 http://dx.doi.org/10.1038/ncomms5044).

The 54 genomes in this study were isolated, selected, and sequenced, and both assembled and aligned against reference genomes. Phylogenies were based on concatenated SNP alignment of selected chromosomes. The structure of the population was investigated using model-based Bayesian clustering.

In addition to the biological result, this study illustrates the ubiquity of an experimental approach based on large-scale sequencing of highly related genomes, in order to isolate tiny differences linked to a trait of interest. This is in contrast to the strategy that was current eight years ago, based on sequencing of a modest number of genomes spanning a much greater evolutionary range.

6.3. Improving inference of metabolic models

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

The Pantograph approach uses reference model annotated by *gene associations*, and voting between complementary predictions of homology between reference genes and target genes, to decide whether a reaction that is present in the scaffold ought be be present in the target. A gene association implicitly represents expert knowledge about the role of genes in a compact way. If the gene association can be rewritten into a possibly satisfiable formula, then the corresponding reaction is instantiated in the target model.

Historically, gene associations have been used intuitively by experts during the model design and curation process, and are often inconsistent. We have formalized the construction of gene associations based on the semantics of different interpretations, showing how different boolean formulas should be constructed when the application is *i*) metabolic model inference, *ii*) flux-balance analysis, *iii*) hierarchical modeling, or *iv*) dynamic simulation (Razanne Issa, MS in prep.).

[5] http://www.genolevures.org/

Second, we have refined our strategy for inferring metabolic models using abductive logic. We have shown that given a set of genes as observations in the target organism, and rules for rewriting gene associations while respecting integrity constraints for the model, then the reactions in the target model can be abduced as hypotheses that “explain” the presence of a maximial number of genes in the target genome. The advantage of this approach is that it can invent, through specialization, reactions that are not present *per se* in the reference model. Two classes of reactions can be invented: substrate-specific reactions inferred from expansion in gene families, and transport reactions needed to maintain model integrity for constitutive compartments.

6.4. Knowledge-based generalization of metabolic models

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

Large metabolic networks are hard to understand and curate, because the large number of detailed reactions, which are needed for accurate modeling and simulation, obscure the high-level structure of the reaction network. We defined knowledge-based methods that factor similar reactions into “generic” reactions in order to visualize a whole pathway or compartment, while maintaining the underlying model so that the user can later “drill down” to the specific reactions if need be[15], [16] An implementation of this method is available as a Python library (see paragraph 5.3).

Figures 2 and 3 illustrate model generation for *Yarrowia lypolitica* fatty acid oxidation in the peroxisome. Molecular species are represented using SBGN notation: as circular nodes, and the reactions as square ones, connected by edges to their reactants and products. Ubiquitous species are of smaller size and colored gray. Non-ubiquitous species are divided into six equivalence classes, and coloured accordingly. The size of the model does not allow for readability of the species labels, thus we do not show them (figure 2).

The specific model is appropriate for simulation, because it contains all of the precise reactions. The generalized model is suited for a human, because it reveals the main properties of the model and masks distracting details. For example, the generalized model highlights the fact that there is a particularity concerning *C24:0-CoA (stearoyl-CoA)* (yellow): there exists a "shortcut" reaction (orange), producing it directly from another *fatty acyl-CoA* (yellow), avoiding the usual four-reaction beta-oxidation chain, used for other *fatty acyls-CoA*. This shortcut is not obvious in the specific model, because it is hidden among a plethora of similar-looking reactions.

We formally defined the generalization method in [15] and showed how to calculate it using a good approximation to an NP-complete set cover problem. The method was further validated in a collection of 1283 inferred models and revealed, on the one hand, a number of probable errors in the inferred models, and on the other hand, that there exist different families of generalization with a plausible link to different adaptive responses.

6.5. Characterization of STAND protein families

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

In collaboration with Sven Saupe and Mathieu Paoletti from IBGC Bordeaux (ANR Mykimun), we worked on characterization of the STAND protein family in the fungal phylum. We established an *in silico* screen based on state-of-the-art bioinformatic tools, which – starting from experimentally studied sequences from *Podospora anserina* – allowed us to determine the first systematic picture of fungal STAND protein repertoire (ms. in preparation). Most notably, we found evidence of extensive modularity of domain associations, and signs of concerted evolution within the recognition domain [13]. Both results support the hypothesis that fungal STAND proteins, originally described in the context of vegetative incompatibility, are involved in a general fungal immune system. In addition, we investigated improved protein domain representations and elaborated a grammatical modelling method [23], which will be used to elucidate mechanisms of formation and operation of the STAND proteins.

NLR domains identified in this work have been incorporated into the upcoming release of Pfam [6].

[6] http://pfam.xfam.org

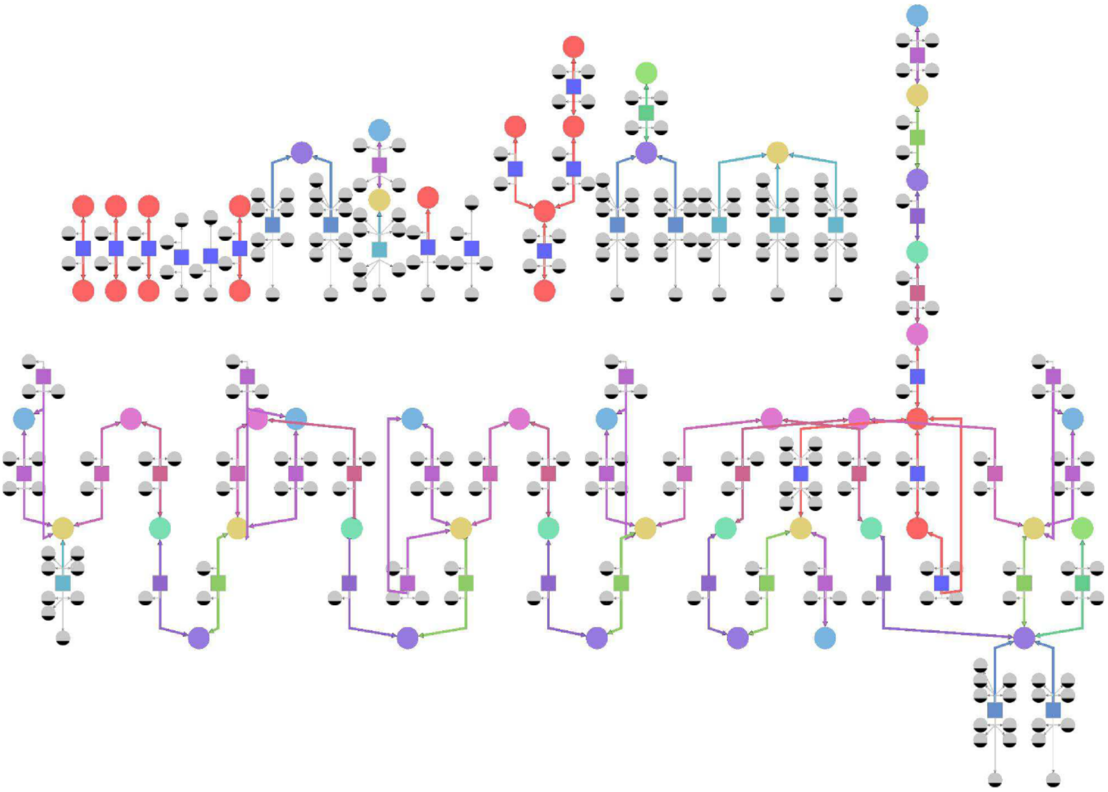


Figure 2. Yarrowia lypolitica fatty acid oxidation model before generalization. Reactions of the specific model are divided into fifteen equivalence classes, represented by different colours. Generally speaking, β-oxidation is a transformation of fatty acyl-CoA (yellow) into dehydroacyl-CoA (violet), then into hydroxyacy fatty acyl-CoA (dark green), 3-ketoacyl-CoA (magenta), and back to fatty acyl-CoA (with a shorter carbon chain); while the specific model describes the same process in more details, specifying those reactions for each of the fatty acyl-CoA species presented in the organisms' cell (e.g. decanoyl-CoA, dodecanoyl-CoA, etc.). This high-level, repetitive structure is obscured by the detail of the individual reactions.

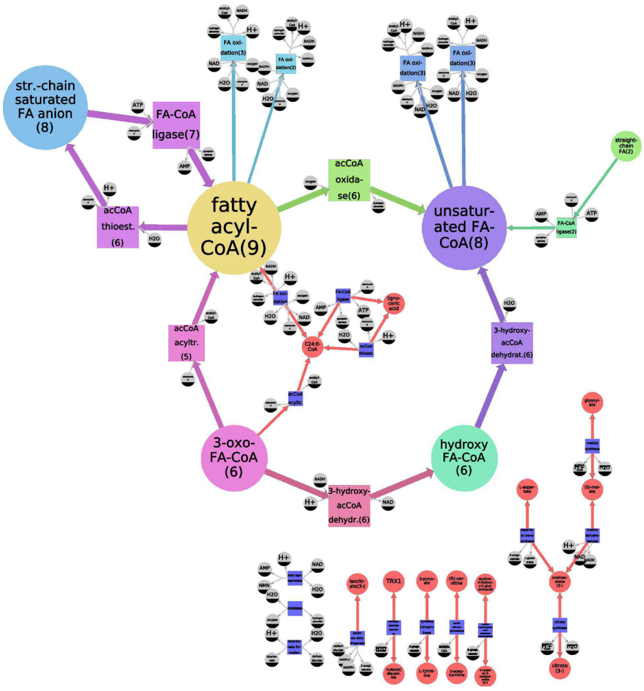


Figure 3. Generalization of the Yarrowia lypolitica fatty acid oxidation model, described as a transformation of fatty acyl-CoA (yellow) into dehydroacyl-CoA (violet), then into hydroxyacy fatty acyl-CoA (dark green), 3-ketoacyl-CoA (magenta), and back to fatty acyl-CoA with a shorter carbon chain. The generalization algorithm identifies equivalent molecular species using an ontology, and groups together reactions that operate on the same abstract species. It finds the greatest generalization the preserves stoichiometry. The generalized model represents quotient species and reactions. For example, the violet dehydroacyl-CoA node is a quotient of hexadec-2-enoyl-CoA, oleoyl-CoA, tetradecenoyl-CoA, trans-dec-2-enoyl-CoA, trans-dodec-2-enoyl-CoA, trans-hexacos-2-enoyl-CoA, trans-octadec-2-enoyl-CoA, and trans-tetradec-2-enoyl-CoA (colored violet in figure 2). In a similar manner, the light-green acyl-CoA oxidase quotient reaction, that converts fatty acyl-CoA (yellow) into dehydroacyl-CoA (violet), generalizes six corresponding light-green reactions of the initial model (figure 2).

To further explor the underlying mechanisms of repeat formation we implemented a stochastic string rewriting system that models the generation process of highly internally conserved repeats. The system is grounded in the biology of the process as it models transformation of repeats through the events of unequal crossing-over and mutation, which are believed to be main mechanisms that produce diversity in repeats. We confirmed that highly variable sites identified on the basis of entropy, are subject to selective pressure towards composition typical for binding sites, which is consistent with the suggested role of recognition epitopes.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

MAGNOME and the company BioLaffort are contracted to develop analyses and tools for rationalizing wine starter strain selection using genomics.

MAGNOME and a consortium of academic (CNRS, INRA, INSA Toulouse) and industrial (Dassault Aviation, Airbus, Turbomeca, SNECMA, Air France, Total) partners coordinated by the French Institute for Petroleum and New Energies are contracted together on a large program of developing and testing alternative fuels for aviation, funded by the Civil Directorate for Aviation. MAGNOME's role is working with biological partners in developing genomic and genetic tools for oleaginous yeasts used in biofuel production.

8. Partnerships and Cooperations

8.1. Regional Initiatives

MAGNOME works with the ISVV and local industry to develop analyses and tools for rationalizing wine starter strain selection using genomics.

8.2. National Initiatives

8.2.1. ANR MYKIMUN.

Signal Transduction Associated with Numerous Domains (STAND) proteins play a central role in vegetative incompatibility (VI) in fungi. STAND proteins act as molecular switches, changing from closed inactive conformation to open active conformation upon binding of the proper ligand. Mykimun, coordinated by Mathieu Paoletti of the IBGC (Bordeaux), studies the postulated involvement of STAND proteins in heterospecific non self recognition (innate immune response).

In MYKIMUN we extend the notion of fungal immune receptors and immune reaction beyond the *P. anserina* NWD gene family. We develop *in silico* machine learning tools to identify new potential PRRs based on the expected characteristics of such genes, in *P. anserina* and beyond in additional sequenced fungal genomes. This should contribute to extend concept of a fungal immune system to the whole fungal branch of the eukaryote phylogenetic tree.

8.3. European Initiatives

8.3.1. FP7 & H2020 Projects

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

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

8.3.2. Collaborations with Major European Organizations

Prof. Mike Taussig: Babraham Institute & Cambridge University

Knowledge engineering for Affinity Proteomics

Henning Hermjakob: European Bioinformatics Institute

Standards and databases for molecular interactions

8.4. International Initiatives

8.4.1. Inria International Partners

8.4.1.1. Informal International Partners

MAGNOME collaborates with Rodrigo Assar Cuevas at the University of Chile, Santiago, Chile and Joaquín Fernandez at the University of Rosario, Rosario, Argentina on hierarchical hybrid modeling using quantized state systems.

MAGNOME collaborates with Nicolás Loira at the University of Chile on methods for inferring genome-scale metabolic models.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Journal

9.1.1.1. Member of the editorial board

Pascal Durrens is :

member of the editorial board of the journal ISRN Computational Biology

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

David Sherman is :

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

9.1.1.2. Reviewer

Pascal Durrens and David Sherman reviewed numerous papers for international journals, including BMC Genomics, Nucleic Acids Research, and Genome Biology and Evolution.

9.2. Teaching - Supervision - Juries

9.2.1. Supervision

PhD: Anna Zhukova, “Knowledge-based generalization for metabolic models,” 2011-4, Sherman

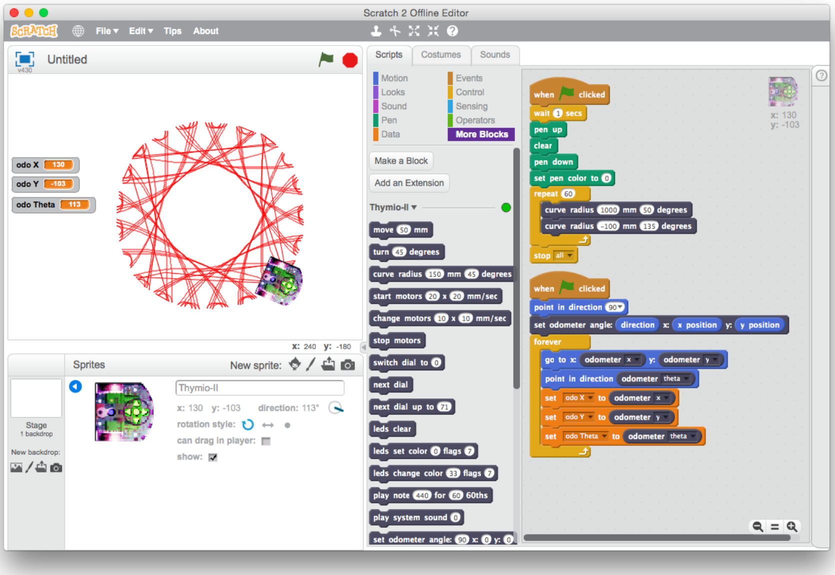
PhD: Razanne Issa, “Analyse symbolique de données génomiques,” 2010–, Sherman

9.2.2. Juries

David Sherman was a member of the juries of:

PhD: Anna Zhukova, “Knowledge-based generalization for metabolic models,” December 18, 2014

HDR: Sofian Maabout, “Contributions à l’optimisation de requêtes multidimensionnelles,” December 12, 2014



```
when [flag] clicked
wait 1 secs
pen up
clear
pen down
set pen color to 0
repeat 60
    curve radius 1000 mm 50 degrees
    curve radius -100 mm 135 degrees
stop all

when [flag] clicked
point in direction 90
set odometer angle: direction x: x position y: y position
forever
    go to x: odometer x y: odometer y
    point in direction odometer theta
    set odo X to odometer x
    set odo Y to odometer y
    set odo Theta to odometer theta
```

Figure 4. Piloting the Thymio-II robot with Scratch 2

9.3. Popularization

David Sherman works with Didier Roy and Pierre-Yves Oudeyer of the Flowers project-team to develop tools and courseware for helping elementary school students explore robotics. David has developed software for communicating between the Scratch 2 visual programming language and the Thymio-II educational robot, and examples for use in the classroom.

10. Bibliography

Major publications by the team in recent years

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

[2] G. Blandin, P. Durrens, F. Tekaia, M. Aigle, M. Bolotin-Fukuhara, E. Bon, S. Casaregola, J. de Montigny, C. Gaillardin, A. Lépingle, B. Llorente, A. Malpertuy, C. Neuvéglise, O. Ozier-Kalogeropoulos, A. Perrin, S. Potier, J.-L. Souciet, E. Talla, C. Toffano-Nioche, M. Wésolowski-Louvel, C. Marck, B. Dujon. *Genomic Exploration of the Hemiascomycetous Yeasts: 4. The genome of Saccharomyces cerevisiae revisited*, in "FEBS Letters", December 2000, vol. 487, n^{o} 1, pp. 31-36

[3] J. Bourbeillon, S. Orchard, I. Benhar, C. Borrebaeck, A. De Daruvar, S. Dübel, R. Frank, F. Gibson, D. Gloriam, N. Haslam, T. Hiltker, I. Humphrey-Smith, M. Hust, D. Juncker,

M. Koegl, Z. Konthur, B. Korn, S. Krobitsch, S. Muyldermans, P.-A. Nygren, S. Palcy, B. Polic, H. Rodriguez, A. Sawyer, M. Schlapshy, M. Snyder, O. Stoevesandt, M. J. Taussig, M. Templin, M. Uhlen, S. Van Der Maarel, C. Wingren, H. Hermjakob, D. J. Sherman. *Minimum information about a protein affinity reagent (MIAPAR)*, in "Nature Biotechnology", 07 2010, vol. 28, n^{o} 7, pp. 650-3 [*DOI :* 10.1038/nbt0710-650]

[4] A. B. Canelas, N. Harrison, A. Fazio, J. Zhang, J.-P. Pitkänen, J. Van Den Brink, B. M. Bakker, L. Bogner, J. Bouwman, J. I. Castrillo, A. Cankorur, P. Chumnanpuen, P. Daran-Lapujade, D. Dikicioglu, K. Van Eunen, J. C. Ewald, J. J. Heijnen, B. Kirdar, I. Mattila, F. I. C. Mensonides, A. Niebel, M. Penttilä, J. T. Pronk, M. Reuss, L. Salusjärvi, U. Sauer, D. J. Sherman, M. Siemann-Herzberg, H. Westerhoff, J. De Winde, D. Petranovic, S. G. Oliver, C. T. Workman, N. Zamboni, J. Nielsen. *Integrated multilaboratory systems biology reveals differences in protein metabolism between two reference yeast strains*, in "Nature Communications", 12 2010, vol. 1, n^{o} 9, Article number : 145 [*DOI :* 10.1038/ncomms1150], http://hal.inria.fr/inria-00562005/en/

[5] B. Dujon, D. J. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuvéglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J.-M. Beckerich, E. Beyne, C. Bleykasten, A. Boisramé, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri-Lafond, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J.-M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G.-F. Richard, M.-L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wésolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, J.-L. Souciet. *Genome evolution in yeasts*, in "Nature", 07 2004, vol. 430, n^{o} 6995, pp. 35-44 [*DOI :* 10.1038/nature02579], http://hal.archives-ouvertes.fr/hal-00104411/en/

[6] P. Durrens, M. Nikolski, D. J. Sherman. *Fusion and fission of genes define a metric between fungal genomes*, in "PLoS Computational Biology", 10 2008, vol. 4 [*DOI :* 10.1371/journal.pcbi.1000200], http://hal.inria.fr/inria-00341569/en/

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

[8] M. Nikolski, D. J. Sherman. *Family relationships: should consensus reign?- consensus clustering for protein families*, in "Bioinformatics", 2007, vol. 23 [*DOI :* 10.1093/bioinformatics/btl314], http://hal.inria.fr/inria-00202434/en/

[9] D. J. Sherman, T. Martin, M. Nikolski, C. Cayla, J.-L. Souciet, P. Durrens. *Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes*, in "Nucleic Acids Research (NAR)", 2009, pp. D550-4 [*DOI :* 10.1093/nar/gkn859], http://hal.inria.fr/inria-00341578/en/

[10] J.-L. Souciet, B. Dujon, C. Gaillardin, M. Johnston, P. Baret, P. Cliften, D. J. Sherman, J. Weissenbach, E. Westhof, P. Wincker, C. Jubin, J. Poulain, V. Barbe, B. Ségurens, F. Artiguenave, V. Anthouard, B. Vacherie, M.-E. Val, R. S. Fulton, P. Minx, R. Wilson, P. Durrens, G. Jean, C. Marck, T. Martin, M. Nikolski, T. Rolland, M.-L. Seret, S. Casaregola, L. Despons, C. Fairhead, G. Fischer, I. Lafontaine, V. Leh Louis, M. Lemaire, J. De Mon-

TIGNY, C. NEUVÉGLISE, A. THIERRY, I. BLANC-LENFLE, C. BLEYKASTEN, J. DIFFELS, E. FRITSCH, L. FRANGEUL, A. GOËFFON, N. JAUNIAUX, R. KACHOURI-LAFOND, C. PAYEN, S. POTIER, L. PRIBYLOVA, C. OZANNE, G.-F. RICHARD, C. SACERDOT, M.-L. STRAUB, E. TALLA. *Comparative genomics of protoploid Saccharomycetaceae*, in "Genome Research", 2009, vol. 19, pp. 1696-1709, http://hal.inria.fr/inria-00407511/en/

Publications of the year

Articles in International Peer-Reviewed Journals

[11] P. ALMEIDA, C. GONÇALVES, S. TEIXEIRA, D. LIBKIND, M. BONTRAGER, I. MASNEUF-POMARÈDE, W. ALBERTIN, P. DURRENS, D. J. SHERMAN, P. MARULLO, C. TODD HITTINGER, P. GONÇALVES, J. P. SAMPAIO. *A Gondwanan imprint on global diversity and domestication of wine and cider yeast Saccharomyces uvarum*, in "Nature Communications", 2014, vol. 5, Article number: 4044 [*DOI :* 10.1038/NCOMMS5044], https://hal.inria.fr/hal-01002466

[12] R. ASSAR, M. A. MONTECINO, A. MAASS, D. J. SHERMAN. *Modeling acclimatization by hybrid systems: Condition changes alter biological system behavior modelsf*, in "BioSystems", June 2014, vol. 121, pp. 43-53 [*DOI :* 10.1016/J.BIOSYSTEMS.2014.05.007], https://hal.inria.fr/hal-01002987

[13] W. DYRKA, M. LAMACCHIA, P. DURRENS, B. KOBE, A. DASKALOV, M. PAOLETTI, D. J. SHERMAN, S. J. SAUPE. *Diversity and variability of NOD-like receptors in fungi*, in "Genome Biology and Evolution", November 2014, forthcoming [*DOI :* 10.1093/GBE/EVU251], https://hal.inria.fr/hal-01083450

[14] G. KUNZE, C. GAILLARDIN, M. CZERNICKA, P. DURRENS, T. MARTIN, E. BÖER, T. GABALDÓN, J. CRUZ, E. TALLA, C. MARCK, A. GOFFEAU, V. BARBE, P. BARET, K. BARONIAN, S. BEIER, C. BLEYKASTEN, R. BODE, S. CASAREGOLA, L. DESPONS, C. FAIRHEAD, M. GIERSBERG, P. P. GIERSKI, U. HÄHNEL, A. HARTMANN, D. JANKOWSKA, C. JUBIN, P. JUNG, I. LAFONTAINE, V. LEH-LOUIS, M. LEMAIRE, M. MARCET-HOUBEN, M. MASCHER, G. MOREL, G.-F. RICHARD, J. RIECHEN, C. SACERDOT, A. SARKAR, G. SAVEL, J. SCHACHERER, D. SHERMAN, N. STEIN, M.-L. STRAUB, A. THIERRY, A. TRAUTWEIN-SCHULT, B. VACHERIE, E. WESTHOF, S. WORCH, B. DUJON, J.-L. SOUCIET, P. WINCKER, U. SCHOLZ, C. NEUVÉGLISE. *The complete genome of Blastobotrys (Arxula) adeninivorans LS3 - a yeast of biotechnological interest*, in "Biotechnology for Biofuels", 2014, vol. 7, n^{o} 1, 66 p. [*DOI :* 10.1186/1754-6834-7-66], https://hal-pasteur.archives-ouvertes.fr/pasteur-00988609

[15] A. ZHUKOVA, D. J. SHERMAN. *Knowledge-based generalization of metabolic models*, in "Journal of Computational Biology", 2014, vol. 21, n^{o} 7, pp. 534-47 [*DOI :* 10.1089/CMB.2013.0143], https://hal.inria.fr/hal-00925881

[16] A. ZHUKOVA, D. J. SHERMAN. *Knowledge-based generalization of metabolic networks: a practical study*, in "Journal of Bioinformatics and Computational Biology", 2014, vol. 12(2), n^{o} 1441001 [*DOI :* 10.1142/S0219720014410017], https://hal.inria.fr/hal-00906911

[17] A. ZIMMER, C. DURAND, N. LOIRA, P. DURRENS, D. J. SHERMAN, P. MARULLO. *QTL dissection of Lag phase in wine fermentation reveals a new translocation responsible for Saccharomyces cerevisiae adaptation to sulfite*, in "PLoS ONE", 2014, vol. 9, n^{o} 1, e86298 [*DOI :* 10.1371/JOURNAL.PONE.0086298], https://hal.inria.fr/hal-00986680

Scientific Popularization

[18] A. ZHUKOVA, D. J. SHERMAN. *Three-level representation of metabolic networks*, November 2014, Journées EDMI, https://hal.inria.fr/hal-01081711

Other Publications

[19] W. DYRKA, P. DURRENS, M. PAOLETTI, S. J. SAUPE, D. J. SHERMAN. *Deciphering the language of fungal pathogen recognition receptors*, October 2014, https://hal.inria.fr/hal-01083421

References in notes

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

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

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

[23] W. DYRKA, J.-C. NEBEL, M. KOTULSKA. *Probabilistic grammatical model for helix-helix contact site classification*, in "Algorithms for Molecular Biology", 2013, vol. 8, 31 p. [*DOI :* 10.1186/1748-7188-8-31], http://hal.inria.fr/hal-00923291

[24] A. GARCIA, D. J. SHERMAN. *Mixed-formalism hierarchical modeling and simulation with BioRica*, in "11th International Conference on Systems Biology (ICSB 2010)", United Kingdom Edimbourg, 10 2010, Poster

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

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

[27] H. HERMJAKOB, L. MONTECCHI-PALAZZI, G. BADER, J. WOJCIK, L. SALWINSKI, A. CEOL, S. MOORE, S. ORCHARD, U. SARKANS, C. VON MERING, B. ROECHERT, S. POUX, E. JUNG, H. MERSCH, P. KERSEY, M. LAPPE, Y. LI, R. ZENG, D. RANA, M. NIKOLSKI, H. HUSI, C. BRUN, K. SHANKER, S.

GRANT, C. SANDER, P. BORK, W. ZHU, A. PANDEY, A. BRAZMA, B. JACQ, M. VIDAL, D. J. SHERMAN, P. LEGRAIN, G. CESARENI, I. XENARIOS, D. EISENBERG, B. STEIPE, C. HOGUE, R. APWEILER. *The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data*, in "Nat. Biotechnol.", Feb. 2004, vol. 22, n^{o} 2, pp. 177-83

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

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

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

[31] A. ROMANO, H. TRIP, H. CAMPBELL-SILLS, O. BOUCHEZ, D. SHERMAN, J. S. LOLKEMA, P. M. LUCAS. *Genome Sequence of Lactobacillus saerimneri 30a (Formerly Lactobacillus sp. Strain 30a), a Reference Lactic Acid Bacterium Strain Producing Biogenic Amines*, in "Genome Announcements", January 2013, vol. 1, n^{o} 1, e00097-12 [*DOI :* 10.1128/GENOMEA.00097-12], http://hal.inria.fr/hal-00863284

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

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

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