INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team HELIX

Informatics and genomics

Grenoble - Rhône-Alpes

THEME BIO

Table of contents

1. Team

Texte a mettre

Head of the team

Alain Viari [Research Director (DR) Inria]

Project assistant

Françoise de Coninck [Secretary (SAR) Inria]

Research scientists (Inria)

Hidde de Jong [Research Director (DR) Inria, HdR]
François Rechenmann [Research Director (DR) Inria, HdR]
Delphine Ropers [Research Associate (CR) Inria]
Marie-France Sagot [Research Director (DR) Inria, HdR]
Eric Tannier [Research Associate (CR) Inria]
Alain Viari [Research Director (DR) Inria]

Research scientists (external)

Sandrine Charles [Associate Professor, University Claude Bernard, HdR]
Vincent Daubin [Research Associate (CR) Cnrs]
Laurent Duret [Research Director (DR) Cnrs, HdR]
Christian Gautier [Professor, University Claude Bernard, HdR]
Johannes Geiselmann [Professor, University Joseph Fourier, HdR]
Philippe Genoud [Associate Professor, University Joseph Fourier]
Manolo Gouy [Research Director (DR) Cnrs, HdR]
Laurent Guéguen [Associate Professor, University Claude Bernard]
Daniel Kahn [DR Inra, HdR]
Jean Lobry [Professor, Universite Claude Bernard, HdR]
Christelle Lopes [ATER, Universite Claude Bernard, started October 2007]
Gabriel Marais [Research Associate (CR) Cnrs]
Dominique Mouchiroud [Professor, University Claude Bernard, HdR]
Sylvain Mousset [Associate Professor, University Claude Bernard]
Michel Page [Associate Professor, University Mendes France]
Guy Perrière [Research Director (DR) Cnrs, HdR]
Franck Picard [Research Associate (CR) Cnrs, started October 2007]
Raquel Tavares [Associate Professor, University Claude Bernard]
Jean Thioulouse [Research Director (DR) Cnrs, HdR]
Danielle Ziébelin [Associate Professor, University Joseph Fourier, HdR]

External members

Eric Coissac [Associate Professor, University Joseph Fourier, HdR]
Corinne Lachaize [Project technical staff, Swiss Institute of Bioinformatics]
Anne Morgat [Project technical staff, Swiss Institute of Bioinformatics]

Visitors

José Luis Aguirre [Professor, Technologico de Monterrey, Mexico, 8 months]
Ana Teresa Freitas [Instituto Superior Técnico, Lisbon, Portugal, 1 week]
Fábio Martinez [Federal University of Mato Grosso do Sul, Brazil, 2 weeks]
Christelle Melo de Lima [University of Lyon II, 9 months]
Alair Pereira do Lago [University of São Paulo, Brazil, 3 months]
Pierre Peterlongo [INRIA Symbiose, various short visits]
Paulo Gustavo Soares da Fonseca [Federal University of Pernambuco, Recife, Brazil, 2 years]

Technical staff

Bruno Besson [CDD European project Cobios]
Edouard Blondeau [CDD Aventis-Pasteur since August 2007]

Ludovic Cottret [CDD ANR]
Estelle Dumas [CDD Graduate Engineer INRIA]
Jean-Francois Gout [CDD ANR]
Sophie Huet [CDD Fondation Rhône-Alpes Futur]
Nathalie Lépy [CDD Fondation Rhône-Alpes Futur]
Vincent Lombard [CDD ANR]
Adrien Maudet [Associate Engineer]
Emmanuel Prestat [CDD ANR]

Post-doctoral fellows

Valentina Baldazzi [scholarship INRIA, European project EC-MOAN, since May 2007]
Cinzia Pizzi [scholarship INRIA, until December 1, 2007]
Adrien Richard [scholarship INRIA, with POPART and HELIX, until October 2007, now research associate at CNRS]
Patricia Thébault [scholarship INRIA, until September 2007, now Associate Professor at University of Bordeaux]
Augusto F. Vellozo [scholarship University Lyon Claude Bernard, started July 2007]

PHD Students

Sophie Abby [scholarship Ministère de la Recherche, supervisors: Vincent Daubin and Manolo Gouy]
Vicente Acuña [scholarship Conicyt (Chile) and INRIA, supervisors: Marie-France Sagot and Christian Gautier]
Anne-Muriel Arigon [scholarship Ministère de la Recherche, supervisors: Manolo Gouy and Guy Perrière]
Elise Billoir [scholarship Ministère de la Recherche, supervisor: Sandrine Charles]
Bastien Boussau [scholarship BDI, CNRS, supervisor: Manolo Gouy]
Yves-Pol Deniélou [scholarship ENS, supervisors: Alain Viari and Marie-France Sagot]
Marc Deloger [scholarship BDI, CNRS, supervisors: Cristina Vieira (LBBE) and Marie-France Sagot]
Marília Dias Vieira Braga [scholarship AlBan, supervisors: Marie-France Sagot and Eric Tannier]
Samuel Druhle [scholarship ENS Cachan, supervisors: Giancarlo Ferrari-Trecate (University of Pavia) and Hidde de Jong]
Claire Guillet [scholarship ENS Lyon, supervisor: Laurent Duret]
Janice Kielbassa [scholarship Ministère de la Recherche European Mobility, supervisor: Sandrine Charles, started September 2007]
Claire Lemaitre [scholarship Ministère de la Recherche, supervisors: Marie-France Sagot and Christian Gautier]
Yann Letrillard [scholarship Institut National du Cancer, supervisor: Guy Perrière]
Nuno Mendes [scholarship Portuguese Ministry of Research (FCT), supervisors: Ana Teresa Freitas (IST, Lisbon, Portugal) and Marie-France Sagot]
Pedro Monteiro [scholarship Portuguese Ministry of Research (FCT), Supervisors: Ana Teresa Freitas (IST Lisbon, Portugal), Hidde de Jong and Radu Mateescu]
Anamaria Necsulea [scholarship Ministère de la Recherche, supervisor: Jean Lobry]
Alexandra Popa [scholarship Ministère de la Recherche, supervisors: Christian Gautier and Dominique Mouchiroud, started September 2007]
Clément Rezvoy [scholarship Ministère de la Recherche, supervisors: Frédéric Vivien, LIP-ENS Lyon and Daniel Kahn]
Anne-Sophie Sertier [scholarship Ministère de la Recherche, supervisors: Daniel Kahn and Vincent Daubin]
Patrícia Simões [scholarship Portuguese Ministry of Research (FCT), supervisors: Sylvain Charlat (LBBE) and Marie-France Sagot, started September 2007]

Former PHD Students

Vincent Lacroix [scholarship BDI, CNRS, supervisor: Marie-France Sagot, PhD defended in October 2007, currently postdoc at Universitat Pompeu Fabra, Barcelona, Spain]
Leonor Palmeira [scholarship Ministère de la Recherche, supervisors: Laurent Guéguen and Jean Lobry, PhD defended July 2007, currently postdoc at ENS Lyon]

Master Students

Juliet Ansel [Master M2 Pro, University Rouen and University Claude Bernard, supervisors: Marie-France Sagot with Ludovic Cottret]
Priscilla Champelovier [Master M2 Research, University Claude Bernard, supervisors: Gabriel Marais]
Yves Clément [Master M2 Research, University Claude Bernard, supervisor: Vincent Daubin]
Julien Jourde [Master Bioinformatique, University Rouen, supervisor: Eric Coissac]
Emna Marrakchi [Master M2 Research, University Claude Bernard, supervisor: Marie-France Sagot]
Alexandre Moret [Master M2 Research, University Joseph Fourier, supervisor: François Rechenmann]
Julien Soubrier [Master M2 Research, University Claude Bernard, supervisors: Manolo Gouy]

2. Overall Objectives

2.1. Overall Objectives

More than four hundred genomes have already been fully sequenced, among which around forty of eukaryotes including man and mouse. Obtaining the genomic sequences is, however, just a first step towards trying to understand how life develops and is sustained. After the sequencing, it is necessary to interpret the information contained in the genomes. One must identify the genes, that is, the regions coding for proteins, and then understand the function of these proteins and the network of interactions that control the expression of the genes according to the needs of an organism. Beyond that, it is important to understand how all the different structures sustaining life are established and maintained in the course of evolution. This evolutionary perspective cannot be ignored, as it allows us to compare and decipher the function of genes, the modification of metabolic pathways, the preservation and variation of signalling systems. In order to study life, it is essential not to limit oneself to genomic data. Other types of data that have become available recently are of equal importance and the information extracted from them must be compared and confronted with the results obtained from the analysis of genomic sequences. Examples of such data are the experimental data obtained by means of DNA microarrays, 2D gels, and mass spectrometry, as well as data on regulatory interactions extracted from the scientific literature.

Computational Biology (or Bioinformatics) is now recognised to play a key role in the process of turning experimental information into new biological knowledge. The HELIX group conducts research in this field with a rather broad spectrum of activities. The group develops new algorithms and applies them to bioinformatics objects, such as DNA and protein sequences, but also phylogenetic trees, as well as graphs which formalise gene interaction networks or metabolic pathways. From the biological point of view, the emphasis is put on comparative genomics and evolutionary biology.

One of the founding principles of the overall approach of the HELIX group is that every object of interest has to be explicitly represented and described, together with its relations to other objects. The group is thus performing an important activity in knowledge representation. A second founding principle is that the mathematical basis of our approaches should be clearly stated. An important part of the activity of HELIX therefore concentrates on the (re)formulation of biological questions into mathematical forms suitable for computer analysis. The fundamental problem is therefore how to design a model that should be simple enough to be practically useful but not so simple as to miss the subtleties of biological questions. The solution to this problem goes far beyond a simple remote collaboration between computer scientists and biologists and requires a real “symbiosis” between the two cultures.

The activities of HELIX are organised in two main research areas (Comparative and Functional genomics), each of them being divided into sub-topics.

1. Comparative genomics;
    1. Computational analysis of the evolution of species and gene families;
    2. Modelling and analysis of the spatial organisation and dynamics of genomes;

3. Motif search and inference.

2. Functional genomics
    1. Computational proteomics and transcriptomics;
    2. Modelling of metabolism: molecular components, regulation, and pathways;
    3. Modelling and simulation of genetic regulatory networks;
    4. Inter- and intra-chromosomal regulatory networks.

The methodological aspects of the above research areas concern mainly knowledge representation, algorithmics, dynamic systems and statistics.

The HELIX project has the particularity that it bridges two geographical locations and two different bioinformatic cultures. While one group is located in Grenoble and has its origin in computer science, the two other groups reside in Lyon and have their roots in biology and biometry for one of them, and computer science and mathematics for the other. However, a long tradition of collaboration between the three groups confers coherence to the HELIX project, with respect both to computational methods and biological topics. Knowledge representation is certainly the best example of the methodological unity existing between the groups, while comparative genomics is at the heart of their biological concerns. Most of the research areas mentioned above involve HELIX members in both Grenoble and Lyon. In addition, members of other groups in the "Laboratoire de Biométrie et Biologie Évolutive" in Lyon, the associated group Swiss-Prot from the Swiss Institute of Bioinformatics in Geneva and the associated group from the Department of Computer Science of the University of São Paulo, Brazil, contribute to the research activities of HELIX, through co-supervision of PhDs and other forms of collaboration.

Participation in the development of two platforms plays an essential part in the integration of the various biological topics and methods developed in the HELIX project:

- GENOSTAR is a bioinformatics platform for exploratory genomics which integrates methods and tools for modelling genomic data and knowledge developed both within and outside the project (Section 5.11).
- PRABI is a Web server resource (http://www.prabi.fr) providing software which may be downloaded or used through facilities available on the Web. The HELIX group is one of the major participants in the development and maintenance of this platform, which is recognized at the national level as one of the RIO and Genopole platforms. The facilities offered by the PRABI cover such areas as genomics, structural biology, proteomics, health, and ecology. The director of the PRABI (C. Gautier) is a member of HELIX. The PRABI is currently composed of two servers in Lyon (La Doua and Gerland) and one server hosted at INRIA Rhône-Alpes in Grenoble. The latter one was launched in april 2007 and is devoted to proteomic resources (http://www.grenoble.prabi.fr/prabig).

3. Scientific Foundations

3.1. Comparative genomics

Keywords: *Evolution*, *combinatorics*, *data analysis*, *genome dynamics*, *genome organisation*, *inference*, *knowledge bases*, *motifs*, *permutations*, *phylogenetic reconstruction*, *probabilistic modelling*, *search*, *text algorithms*, *tree algorithms*.

Participants: Sophie Abby, Vicente Acuña, Bastien Boussau, Yves Clément, Eric Coissac, Yves-Pol Deniélou, Vincent Daubin, Marc Deloger, Marília Dias Vieira Braga, Laurent Duret, Christian Gautier, Philippe Genoud, Jean-Francois Gout, Manolo Gouy, Laurent Guéguen, Claire Guillet, Daniel Kahn, Claire Lemaitre, Jean Lobry, Gabriel Marais, Dominique Mouchiroud, Sylvain Mousset, Anamaria Necsulea, Leonor Palmeira, Guy Perrière, Alexandra Popa, François Rechenmann, Marie-France Sagot, Anne-Sophie Sertier, Patrícia Simões, Paulo Gustavo Soares da Fonseca, Eric Tannier, Raquel Tavares, Augusto F. Vellozo, Alain Viari, Danielle Ziébelin.

Comparative genomics may be seen as the analysis and comparison of genomes from different species in order to identify important genomic features (genes, promoter and other regulatory sequences, regions homogeneous for some characteristics such as composition etc.), study and understand the main evolutionary forces acting on such genomes, and analyse the general structure of the genomic landscape, how the different features relate to each other and may interact in some life processes.

Computationally speaking, comparative genomics requires expertise with knowledge representation, probabilistic modelling techniques, general data analysis and text algorithmic methods, phylogenetic reconstructions, and combinatorics. All such expertises are present in HELIX as reflected in past and current publications.

3.1.1. Computational analysis of the evolution of species and gene families

Evolution is the main characteristic of living systems. It creates biological diversity that results from the succession of two independent processes: one introducing mutations that allow the genetic information transmitted to a descendant to vary slightly in relation to the genetic information present in the parent organism, and another fixing the mutation, where the frequency of occurrence of a tiny fraction of the errors increases in the population until the errors become the norm.

The analysis of the origin and frequency of mutations, as well as the constraints on their fixation, in particular the effect of natural selection, underlies an important part of the field of molecular computational biology. It therefore appears in almost all research areas developed in the HELIX project.

The comparison of proteic or nucleic sequences allows the *a priori* reconstruction of the whole of the Tree of Life. However, the mathematical complexity of the processes involved requires methods for approximate estimation. Moreover, sequences are not the only source of information available for reconstructing phylogenetic trees. The order of the genes along a genome is undergoing progressive change and the comparison of the permutations observed offers another way of estimating evolutionary distances. The methodological problems encountered are mainly related to the estimation of such distances in terms of the number of elementary (and biologically meaningful) operations enabling one permutation to succeed another. Sophisticated algorithms are required to deal with the problem. Once phylogenetic trees have been constructed, other problems arise that concern their manipulation and interpretation. Currently, more than 6000 families of genes (having more than 4 specimens) are known, and hence can be represented by more than 6000 different trees (HELIX also developed specialized databases to hold this kind of information). The management, comparison and update of these trees represents a challenging computational and mathematical problem.

Another challenge regarding phylogeny concerns the study of co-evolution. Co-evolution refers to the mutual evolutionary influence between two (or more) species. Each party in a co-evolutionary relationship exerts selective pressures on the other, thereby affecting each other's evolution. HELIX is more particularly interested in the co-evolution of a host species and its parasites. One organism of choice is the bacterium *Wolbachia* which infects arthropods, including a high proportion of insects. It is one of the world's most common parasitic microbes and is potentially the most common reproductive parasite in the biosphere. More than 20% of insect species carry this bacterium The question of how tight is the association between these "influential passengers" as they are called and their hosts has received so far only limited attention, although it is crucial to the assessment of the evolutionary consequences such passengers may have, and to an understanding of their long term evolutionary trajectories. More specifically, the questions for which one would like to find an answer are, how often are symbionts horizontally transferred among branches of the host phylogenetic tree, how long do infections persist following the invasion of a new lineage finally, what processes underlie this dynamic gain/loss equilibrium? These questions, crucial to determine the evolutionary trajectories and impact of the infections, remain unanswered. Molecular data is most often used to reconstruct and confront phylogenies for both the symbiont and host cytoplasmic lineages. Existing co-phylogenetic techniques/methods are based on a phylogenetic approach and while the controversy on which specific method, if any, is "the best" remains very much alive, recent results show that all may lead to inaccurate (wrong) results, thereby potentially seriously compromising their interpretation with a view to understanding the evolutionary dynamic of parasites in

communities. This work involves a collaboration with Sylvain Charlat from the LBBE "Génétique et Évolution des interactions Hôtes-Parasites (GEIHP)" team, as well as other experimentalists from the GEIHP.

3.1.2. Modelling and analysis of the spatial organisation and dynamics of genomes

Genomic sequences are characterized by strong biological and statistical heterogeneities in their composition and organisation. In fact, neighbouring genes along a genome often share multiple properties, whose nature is structural (size and number of introns), statistical (base and codon frequencies), and linked to evolutionary processes (substitution rates). In certain cases, such neighbouring structures have been interpreted in terms of biological processes. For instance, in bacteria the spatial organisation of genomes results in part from the mechanism of replication. Other local structures, however, still resist the discovery of a mechanism that could explain their generation and maintenance. The most characteristic example in vertebrates concerns isochores usually defined as regions that are homogeneous in terms of their G+C composition. The identification of isochores is essential for the annotation of sequences as it correlates with various other genomic features (base frequency, gene structure, nature of transposable elements). The analysis of the spatial structure of a genome requires the elaboration of correlation methods (non-parametric correlation determination along a neighbour graph and Markov processes) and of partitioning (or segmentation) techniques.

In the course of evolution, the spatial organisation of a genome undergoes several changes that are the result of biological processes also not yet fully understood, but which generate various types of modifications. Among these changes are permutations between closely located genes, inversion of whole segments, duplication, and other long-range displacements. It is therefore important to be able to define a permutation distance that is biologically meaningful in order to derive true evolutionary scenarios between species or to compare the rates of rearrangements observed in different genomic regions. The HELIX project has been particularly interested in elaborating an operational definition for the notion of synteny in bacteria and in eukaryotes (two completely different notions for the two kingdoms). The elaboration of these definitions, together with their precise mathematical characterisations require expertise both in biology and in computer science.

3.1.3. Motif search and inference

The term motif is quite general, referring to locally-conserved structures in biological entities. The latter may correspond to biological sequences and 3D structures, or to abstract representations of biological processes, such as evolutionary trees or graphs, and biochemical or genetic networks. When referring to sequences, the term motif must be understood in a broad sense, which covers binding sites in both nucleic and amino acid sequences, but also genes, CpG islands, transposable elements, retrotransposons, etc.

The occurrence of motifs in a sequence provides an indication of the function of the corresponding biological entity. Identifying motifs, whether using a model established from previously-obtained examples of a conserved structure or proceeding *ab initio*, represents therefore an important area of research in computational biology. Motif identification consists of two main parts: 1. feature identification, which aims at finding and precisely mapping the main features of a genome: protein or RNA-coding genes, DNA or RNA sequence or structure signals, satellites (tandem repeats) or transposable elements (dispersed repeats with a specific structure), regulatory regions, etc; 2. relational identification, the goal of which consists in finding relations existing among the features individually characterized in the first step. Such relations are diverse in nature. They may, for instance, concern the participation of various features in a cellular process, or their physical interaction.

Search and inference problems, whether they concern features or relations, are in fact the extremes of a continuum of problems that range from seeking for something well-known to trying to identify unknown objects. The main difficulty lies in the fact that features and the relations holding between them should in general be inferred together. However, the information that must be manipulated in this case (cooperative signals, operons, regulons, reaction pathways or molecular assemblies) is more complex than the initial genome data and thus requires a higher degree of abstraction, and more sophisticated algorithms or statistical approaches. Various search and inference methods have already been developed by HELIX. These include methods for DNA and protein sequence motifs inference, gene finding, satellites and repeats identification and RNA common substructure inference. More recent work concerns the definition of motifs in graphs

representing, for instance, metabolic pathways. In the last year, work has started also and on mixing information from various often quite heterogeneous sources to infer motifs on finding RNA motifs .

The first include for now sequence information with information on gene expression coming from microarray experiments and information provided by the signals evolution imprints into genomes. The final objective is to be able to automatically infer whole cellular modules, that is in fact small or, in the longer term, larger-scale biological networks. This new topic provides a strong link between comparative and functional genomics, molecular biology seen at the linear level of a genome and networks.

The recent discoveries of microRNAs (miRNA), short interference RNAs (siRNA) and other small non-coding RNAs (ncRNA) have revealed the important regulatory role they have in the control of gene expression. Recent results report that the predicted miRNA targets are highly biased towards transcription factors and other regulatory genes. This ability is highly connected to the secondary structure of the molecule, which in turn depends on the presence of specific short conserved regions or motifs. Very recently, the prediction of functional RNA regions involved algorithms that scanned whole genomes for short conserved segments in order to unravel the putative regions that are able to bind to specific targets. These results show that the development of efficient methods to infer functional RNA motifs is crucial to understand gene regulatory networks in depth.

3.2. Functional genomics

Keywords: *Networks*, *combinatorics*, *data analysis*, *dynamical systems*, *evolution*, *functional annotation*, *graph algorithms*, *inference*, *knowledge bases*, *motifs*, *probabilistic modelling*, *search*.

Participants: Vicente Acuña, Valentina Baldazzi, Bruno Besson, Eric Coissac, Ludovic Cottret, Marc Deloger, Hidde de Jong, Samuel Druhle, Estelle Dumas, Laurent Duret, Christian Gautier, Philippe Genoud, Johannes Geiselmann, Manolo Gouy, Laurent Guéguen, Sophie Huet, Daniel Kahn, Corinne Lachaize, Vincent Lacroix, Claire Lemaitre, Pedro Monteiro, Anne Morgat, Dominique Mouchiroud, Michel Page, Guy Perrière, Franck Picard, Emmanuel Prestat, François Rechenmann, Adrien Richard, Delphine Ropers, Marie-France Sagot, Paulo Gustavo Soares da Fonseca, Eric Tannier, Raquel Tavares, Patricia Thébault, Jean Thioulouse, Alain Viari.

Functional genomics refers to arriving at an understanding of the different features of a genome such as genes, non-coding RNAs etc. This requires in general understanding how such features are related to one another, that is understanding the network of relations holding among the different elements of the genomic landscape, and between genomes and their cellular and extra-cellular environment.

Computationally speaking, funtional genomics requires therefore expertise in particular with graph theory and algorithmics (with tree algorithmics as a special case), but also with dynamic systems and, as for comparative genomics, with general data analysis methods (of proteomic, transcriptomic and other "omic" data), knowledge representation, and combinatorics (concerning random graph models more specially). Again, these are expertises well covered within HELIX. Functional genomics requires further good visualisation tools for which HELIX built solid collaborations with outside experts.

3.2.1. Computational proteomics and transcriptomics

By analogy with the term genomics, referring to the systematic study of genes, proteomics is concerned with the systematic study of proteins. More particularly, proteomics aims at identifying the set of proteins expressed in a cell at a given time under given conditions, the so-called proteome. Recent progress in mass spectrometry (MS) has resulted in efficient techniques for the large-scale analysis of proteomes. In particular, the MS/MS technique allows for the determination of complete or partial sequences of proteins from their fragmentation patterns. State-of-the-art mass spectrometers produce large volumes of data the interpretation of which can no longer be carried out manually. In fact, there is a growing need for computer tools allowing for a fully automated protein identification from raw MS/MS data. This has motivated a collaboration between HELIX and the "Laboratoire de Chimie des Proteines" (LCP) at the CEA in Grenoble. The aim of the collaboration is to develop computer tools for the analysis of data produced by the MS/MS approach. In particular, efficient algorithms have been designed for generating partial sequence (Peptide Sequence Tags, PST) MS/MS spectra,

for scanning protein databases in search of sequences matching these PSTs, and for mapping the PSTs on the complete translated genome sequence of an organism. These algorithms have been implemented in a high-throughput software pipeline installed at the LCP in order to provide support to the Genopole proteomic platform.

The dynamic link between genome, proteome and cellular phenotype is formed by the subset of genes transcribed in a given organism, the so-called transcriptome. The regulation of gene expression is the key process for adaptation to changes in environmental conditions, and thus for survival. Transcriptomics describes this process at the scale of an entire genome. There are two main strategies for transcriptome analysis: i) direct sampling (and quantification) of sequences from source RNA populations or cDNA libraries (the most common techniques of this type are ESTs and SAGE) and ii) hybridization analysis with comprehensive non-redundant collections of DNA sequences immobilised on a solid support (the methods most often used in this case are DNA macroarrays, microarrays, and chips). Members of the HELIX project have worked with SAGE, EST and DNA microarray data in particular, to analyse the transcription pattern of transposable elements, improve the inference of sequence motifs and work towards an automatic inference method of small genetic networks, and provide initial links between genetic information and metabolism (and therefore between genotype and phenotype where by genotype one undertands the specific genetic makeup – the specific genome – of an individual, and by phenotype either an individual's total physical appearance and constitution or a specific manifestation of a trait, such as size, eye color, or behaviour that varies between individuals).

With the recruitment in 2007 of Franck Picard as a CNRS Chargé de Recherches, HELIX has started working also with CGH (Comparative Genomic Hybridisation) arrays. CGH is a molecular-cytogenetic method for the analysis of copy number changes (gains / losses) in the DNA content of cells. The resolution of CGHs has been recently greatly improved using microarray technology. Nowadays (2006), it may reach down to segments of 5-10 kilobases. The purpose of array-based Comparative Genomic Hybridisation (array CGH) is to detect and map chromosomal aberrations, on a genomic scale, in a single experiment. CGH arrays have been used in particular for identifying recurrent chromosomal aberrations that occur in some types of cancer tumors. HELIX' work on CGH arrays will also be conducted in collaboration, on one side with the "Santé et Statistique" team of the LBBE, and on the other side, with the members of Franck Picard's PhD InaPG-INRA laboratory, in particular Stéphane Robin and Jean-Jacques Daudin.

3.2.2. Modelling and analysis of metabolism: molecular components, regulation, and pathways

Beyond genomic, proteomic and transcriptomic data, a large amount of information is now available on the molecular basis of cellular processes. Such data are quite heterogeneous, including among other things the organisation of a genome into operons and their regulation, and the chemical transformations occurring in the cell (together with their metabolites). The challenge of biology today is to relate and integrate the various types of data so as to answer questions involving the different levels of structural, functional, and spatial organisation of a cell. The data gathered over the past few decades are usually dispersed in the literature and are therefore difficult to exploit for answering precise questions. A major contribution of bioinformatics is therefore the development of databases and knowledge bases allowing biologists to represent, store, and access data. The integration of the information in the different bases requires explicit, formal models of the molecular components of the cell and their organisation. HELIX is involved in the development of such models and their implementation in object-oriented or relational systems. The contribution of HELIX to this field is twofold: on one hand some HELIX members are interested in the development of knowledge representation systems, whereas other members are interested in putting these systems to work on biological data. In this context, HELIX collaborates tightly with the SwissProt group at SIB in order to set up a database of metabolic pathways (UniPathway).

Another aspect of the activity of HELIX in this field concerns the design of algorithms to reconstruct and analyse metabolic pathways. By contrast to homology-based approaches, we try to tackle the problem of reconstruction in an *ab-initio* fashion. Given a set of biochemical reactions together with their substrates and products, the reactions are considered as transfers of atoms between the chemical compounds. The basic idea is to look for sequences of reactions transferring a maximal (or preset) number of atoms between a given source compound and the sink compound.

In the same vein, several related problems (for instance, comparing biochemical networks to genomic organisation) have been put in the form of a graph-theoretical problem (such as finding common connected components in multigraphs) in order to provide a uniform formalisation. This activity in graph theory applied to biological problems is now conducted in a collaboration between Grenoble and Lyon, in particular through the question of searching and inferring modules in metabolic networks by defining "connected subgraph motifs". Beyond practical applications, this raises interesting and difficult questions in combinatorics and statistics. The combinatoric aspects are adressed in collaboration with the University of São Paulo, Brazil and the statistical aspects are studied in collaboration with Sophie Schbath (INRA, Jouy-en-Josas) and Stéphane Robin (InaPG, Paris).

A simple graph model may be enough to conceive and to apply methods such as the search or inference of motifs but meets its limit as soon as one wishes to push further the analysis of the results obtained. A natural extension consists in representing a metabolic network with an hypergraph instead, which allows to capture in a more realistic way the links between the different metabolites, and therefore to detect finer structural properties. Furthermore, performing structural analyses using such representation enables an interesting parallel with other methods for analysing metabolic networks that are based on a decomposition of the stoichiometric matrix (constraint-based model). A stoichiometric matrix indicates the proportion of each metabolite that participates in a reaction as input or output. HELIX has started working with this hypergraph representation, and with the question of enumerating elementary modes and minimal reaction cuts in a network. An elementary mode may be seen as a set of reactions that, when used together, perform a given task while a minimal reaction cut set is a set of reactions one needs to inhibit to prevent a given task, also called *target reaction*, from being performed. This work is done in collaboration with Alberto Marchetti-Spaccamela from the University of Rome, Italy, and Leen Stougie from the Eindhoven University of Technology and the CWI at Amsterdam, Netherlands.

3.2.3. Modelling and simulation of genetic regulatory networks

All the aforementioned research topics concern, basically, a "static" description of cellular processes. Except for evolution (but on a very different time-scale), time is not explicitely taken into account. To achieve a better understanding of the functioning of an organism, the networks of interactions involved in gene regulation, metabolism, signal transduction, and other cellular and intercellular processes need to be represented and analyzed from a dynamical perspective.

Genetic regulatory networks control the spatiotemporal expression of genes in an organism, and thus underlie complex processes like cell differentiation and development. They consist of genes, proteins, small molecules, and their mutual interactions. From the experimental point of view, the study of genetic regulatory networks has taken a qualitative leap through the use of modern genomic techniques that allow simultaneous measurement of the expression of all genes of an organism, such as the above-mentioned transcriptomics techniques. However, in addition to these experimental tools, mathematical methods supported by computer tools are indispensable for the analysis of genetic regulatory networks. As most networks of interest involve many genes connected through interlocking positive and negative feedback loops, it is difficult to gain an intuitive understanding of their dynamics. Modelling and simulation tools allow the behaviour of large and complex systems to be predicted in a systematic way.

A variety of methods for the modelling and simulation of genetic regulatory networks have been proposed, such as approaches based on differential equations and stochastic master equations. These models provide detailed descriptions of genetic regulatory networks, down to the molecular level. In addition, they can be used to make precise, numerical predictions of the behaviour of regulatory systems. Many excellent examples of the application of these methods to prokaryote and eukaryote networks can be found in the literature. In many situations of biological interest, however, the application of the above models is seriously hampered. In the first place, the biochemical reaction mechanisms underlying regulatory interactions are usually not or incompletely known. In the second place, quantitative information on kinetic parameters and molecular concentrations is only seldom available, even in the case of well-studied model systems.

The aim of the research being carried out in HELIX is to develop methods for the modelling and simulation of genetic regulatory networks that are capable of dealing with the current lack of detailed, quantitative data. In particular, a method for the qualitative simulation of genetic regulatory networks has been developed and implemented in the computer tool GENETIC NETWORK ANALYZER (GNA). The method and the tool have been applied to the analysis of prokaryote regulatory networks in collaboration with experimental biologists at the Université Joseph Fourier (Grenoble) while several other groups have used GNA for similar purposes. Recently, the scope of the research has been enlarged to the reduction, validation and identification of models of genetic regulatory networks.

3.2.4. Inter- and intra-chromosomal regulatory networks

For many years, work in the area of gene regulation concentrated on finding the sequences upstream of genes that could correspond to promoter or other regulatory sequences, that is, to sites where protein, RNA or protein/RNA complexes would bind and thereby initiate, stop, up or down-regulate the level of expression of a given gene. Even the cooperative aspects of such binding was for long fully or partly ignored in the algorithms developed to identify the sites from the genomic sequence alone, or, in rare cases, from the genomic sequence and its inferred structure. Such analyses were essentially based on the simplistic assumption that the sequence of regulatory sites should be more conserved on average than the remaining non-coding sequence.

The advent of high-throughput microarray technologies has added one important level of information to this image, as it enabled to measure the co-expression of sets of genes in a given tissue in given conditions. These studies suggest a correlated action of different elements of the gene regulation machinery and their potential interaction. More recently, the importance and extent of regulation at the epigenetic level started to be fully realised. This refers to heritable changes in gene regulation that occur without a change in the DNA sequence and are therefore not encoded at the genomic level. Epigenetic regulation was nevertheless recognised first at the scale of DNA molecules. This concerns in particular the chromatin structure and the possible chemical modification of some DNA bases. Chromatin is a complex made of DNA and of a special type of protein, called histones, around which DNA molecules densely wound to form a packed structure.

While these aspects of epigenetic regulation remain still largely unexplored by computational biologists, another level of complexity has in the last two decades emerged in the study of gene regulation. It is indeed now increasingly realised that, besides chromatin structure and DNA modification, the spatial arrangement of the chromosomes of a eukaryotic genome inside a cell is related to gene regulation, and possibly also to other important life processes. During interphase, chromosomes thus occupy distinct territories with preferred radial locations inside the nucleus, that is with preferential locations relative to the nuclear center. These preferred locations are cell type-specific and are conserved in the same cell type accross different primates. In the last decade, intra-chromosomal interaction or simple spatial proximity between genetic elements situated at often distant positions along the genome have started revealing their importance in gene expression. More recent work suggests, based upon strong evidence, that chromosome territories also intermingle. The pattern of intermingling appears to correlate with translocation frequencies (translocations are exchanges of genetic material between chromosomes) and to be changed when transcription is blocked. HELIX has started addressing this issue in 2007, in close collaboration with an experimental biologist, Ana Pombo, from the MRC at the Imperial College, London.

4. Application Domains

4.1. Panorama

Keywords: *agriculture*, *medicine*.

Various members of the HELIX project, both in Grenoble and Lyon, are engaged in activities that are oriented either towards the use of internally- or externally-developed software for doing bioanalysis, or to the development of systems that allow the integration of a variety of methods inside a single architecture, and the comparison of the results obtained by different approaches for the same problem. These activities sometimes reflect research topics that do not fall within the research areas outlined above, but that involve groups, either within public organisms or private companies, with whom HELIX collaborates. These collaborations often concern applications in medicine or agriculture.

5. Software

5.1. Alfacinha

Keywords: *knowledge representation*.

Participants: laurent Guéguen, Jean Lobry, Leonor Palmeira [Correspondent].

Alfacinha is a package of Python modules for easy building and simulation of sequence evolution with neighbouring-site dependencies. A formal definition of models of sequence evolution is presented which can incorporate neighbouring-site dependencies and propose an efficient algorithm for their simulation. The algorithm was implemented in a flexible way so that many types of neighbouring-site dependencies can be incorporated in order to answer a wide range of questions. As an example, this application can be used to efficiently simulate methylation dependent substitution mechanisms like the well known CpG effect. This work was done in collaboration with Jean Bérard, from the Institut Camille Jordan, UCBL. It is available on Linux, Unix and MacOS X operating systems and is licensed under the GPL license.

5.2. AROM

Keywords: *knowledge representation*.

Participants: Philippe Genoud, Danielle Ziébelin [Correspondent].

AROM (“Associate Relationships and Objets for Modeling”) is both a knowledge representation formalism and a knowledge base management system that implements this formalism. AROM belongs to the family of Object Oriented Knowledge Representation Systems. The originality of AROM is to explicitly represent relationships between instances of classes by a specific modeling entity called Association. An association can link several (i.e more than two) classes; it is defined by the roles these classes play in the associations and by cardinality constraints. As for Classes, Associations may have attributes and can be organized in specialization hierarchies. AROM is implemented in Java. Its fully documented API makes it easy to integrate in a larger system. The explicit description of associations allows to design easy to read knowledge bases and appears to be particularly adapted for representing biological knowledge. AROM is the very substrate of the GENOSTAR/IOGMA platform. For more information, see: http://www-helix.inrialpes.fr/article221.html

5.3. BaobabLuna

Keywords: *reversal distance*.

Participants: Marília Braga [Correspondent], Marie-France Sagot, Eric Tannier.

BAOBABLUNA is a software for the manipulation of signed permutations in a genomic context. Several routines are implemented, such as the computation of the reversal distance of a permutation, and the equivalence classes of the solution space of sorting by reversals. For more information, see: http://www.geocities.com/mdvbraga/baobabLuna.html

5.4. C3P

Keywords: *graph merging*, *multigraph common connected component*.

Participants: Anne Morgat, Alain Viari [Correspondent].

The C3P package implements a generic approach to merge the information from two or more graphs representing biological data, such as genomes, metabolic pathways or protein-protein interactions, in order to infer functional coupling between them (*e.g.* to find all adjacent genes on a chromosome that encode for enzymes catalysing connected biochemical reactions). The method relies on the computation the Common Connected Components of a multigraph summarising the biological data considered. It was developed with Frédéric Boyer, currently at the CEA Grenoble. The code (in C) is distributed under GPL license. For more information, see: http://www.inrialpes.fr/helix/people/viari/cccpart

5.5. DNA array analysis tool

Keywords: *DNA array analysis*.

Participant: Guy Perrière [correspondent].

In collaboration with the groups of Michel Bihl (University of Basel) and Desmond Higgins (University of Dublin), we have developed a new resampling strategy for the statistical analysis of DNA array data sets. This strategy can be applied with any supervised clustering method used for the analysis of gene expression data. The corresponding software is available as R scripts at http://pulmogene.unibas.ch/articles/optimization.

5.6. Ed'Nimbus

Keywords: *filter for sequence alignment and repeat identification*.

Participant: Marie-France Sagot [Correspondent].

ED'NIMBUS is an algorithm that filters DNA sequences previous to a multiple sequence alignment or repeats detection program. ED'NIMBUS was developed by Pierre Peterlongo during his PhD and is maintained by him at the University of Marne-la-Vallée (http://igm.univ-mlv.fr/~peterlon/officiel/ednimbus/index.php) in collaboration with Nadia Pisanti from the University of Rome, Italy and Alair Pereira do Lago from the University of São Paulo, Brazil.

5.7. FamFetch

Keywords: *database*, *phylogenetic trees*, *tree pattern search*.

Participants: Laurent Duret, Manolo Gouy, Simon Penel, Guy Perrière [Correspondent].

FAMFETCH is a set of tools to search for tree patterns in databases of phylogenetic trees. FAMFETCH is available for download at (http://pbil.univ-lyon1.fr/software/famfetch.html). It was developed with Jean-François Dufayard who did his PhD in HELIX. He is now IR at the LIRMM, Montpellier.

5.8. GeM

Keywords: *comparative genomics*, *database*, *vertebrates*.

Participants: Christian Gautier [Correspondent], Bruno Spataro.

GEM is a project that associates laboratories from the INRIA (HELIX), the CNRS, the University Claude Bernard (LBBE), the INRA and the INSERM to develop and maintain a database for comparative analysis of complete vertebrate genomes. An UML model has been implemented using both PostGres and ACNUC. An interface with R is also provided that allows users to perform complex queries and statistical analyses, and to obtain graphic representations directly from an internet connection. For more information see : http://pbil.univ-lyon1.fr/gem/gem_home.php). Processing the data in the database involves massive computation that is done using the IN2P3 facilities of the CNRS (http://institut.in2p3.fr/). Participated also in the development of this software Gisèle Bronner and Vincent Navratil, who did their PhD in HELIX. Gisèle is now Associate Professor at the University of Clermont-Ferrand and Vincent Navratil is Postdoc at the INRA.

5.9. Genepi (Genepi)

Keywords: *annotation system*, *blackboard architecture*.

Participants: François Rechenmann, Alain Viari, Danielle Ziébelin.

GENEPI is a blackboard framework for developing automatic annotation systems. The system is not bound to any specific annotation strategy. Instead, the user will specify a blackboard structure in a configuration file and the system will instantiate and run this particular annotation strategy. Although the system is robust enough to be used on real-size applications, it is of primary use to bioinformatics researchers who want to experiment with blackboard architectures. The system was developed with Stéphane Descorps-Declere, currently Postdoc at the University of Orsay. For more information, see: http://www.inrialpes.fr/helix/people/viari/genepi

5.10. Genetic Network Analyzer (GNA)

Participants: Bruno Besson, Estelle Dumas, Hidde de Jong [Correspondent], Pedro Monteiro, Michel Page, Delphine Ropers.

GENETIC NETWORK ANALYZER (GNA) is the implementation of a method for the qualitative modelling and simulation of genetic regulatory networks developed in the HELIX project. The input of GNA consists of a model of the regulatory network in the form of a system of piecewise-linear differential equations, supplemented by inequality constraints on the parameters and initial conditions. From this information, GNA generates a state transition graph summarising the qualitative dynamics of the system. GNA is currently distributed by the company Genostar, but remains freely available for academic research purposes. For more information, see http://www-helix.inrialpes.fr/gna.

5.11. GenoStar

Keywords: *bioinformatics environment*.

Participants: Anne Morgat, François Rechenmann [Correspondent], Alain Viari [Correspondent], Danielle Ziébelin.

GENOSTAR is an integrated bioinformatics environment, which was developed by a consortium of four members: INRIA, Institut Pasteur, Hybrigenics and GENOME express. GENOSTAR is made up of several application modules which share data and knowledge management facilities. All data manipulated by the application modules, and all results thus produced, are explicitely represented in an entity-relationship model: AROM. Within a module, the methods are organised into strategies, the execution of which requires complex analysis tasks. The GENOSTAR platform has now been transferred to the Genostar company. Its four modules (GenoAnnot, GenoLink, GenoBool and GenoExpertBacteria) have been integrated in the Iogma bioinformatics environment (see 7.1Genostar), which is based on the same framework. For more information, see (http://www-helix.inrialpes.fr/article121.html)

5.12. Herbs

Keywords: *annotation support*.

Participants: Corinne Lachaize, Anne Morgat, Alain Viari [Correspondent].

HERBS (HAMAP EXPERT RULE BASED SYSTEM) provides computer support for the reannotation of complete bacterial proteomes. It is being developed in collaboration with the Swiss Institute of Bioinformatics (Geneva) in the framework of the HAMAP project. HERBS is able to check the consistency of the annotation of proteins involved in metabolic pathways at the organism level. HERBS consists of an inference engine, based on the system Jess (Java Expert System Shell), and a knowledge base containing the facts and rules of interest. The use of HERBS is facilitated by a graphical user interface. For more information, see: http://www-helix.inrialpes.fr/article542.html.

5.13. Hogenom and Hovergen

Keywords: *databases*, *genomes*.

Participants: Laurent Duret, Manolo Gouy, Simon Penel, Guy Perrière [Correspondent], Dominique Mouchiroud.

HOGENOM is a database of homologous genes in fully-sequenced genomes, structured under the ACNUC sequence database management system. It allows the selection of sets of homologous genes among general or vertebrate species, and to visualise multiple alignments and phylogenetic trees. Thus HOGENOM is particularly useful for comparative sequence analysis, phylogeny and molecular evolution studies. More generally, HOGENOM gives an overall view of what is known about a specific gene family. HOVERGEN is a similar database exclusively dedicated to homologous vertebrate genes. For more information see : (http://pbil.univ-lyon1.fr/databases/hogenom.html)

5.14. Hoppsigen

Keywords: *database*, *pseudogenes*.

Participant: Dominique Mouchiroud [Correspondent].

HOPPSIGEN is a nucleic database of homologous processed pseudogenes. For more information, see http://pbil.univ-lyon1.fr/databases/hoppsigen.html.

5.15. HoSeqI

Keywords: *gene family database*, *sequence identification*.

Participants: Manolo Gouy, Guy Perrière [correspondent].

HOSEQ1 (Automated homologous sequence identification in gene family databases) is a web service available at http://pbil.univ-lyon1.fr/software/HoSeqI The user can position a protein or a DNA sequence relatively to a database of families of homologous sequences and identify the family to which the sequence belongs, as well as its position within the multiple alignment and the evolutionary tree of this family. HOSEQ1 was developed by Anne-Muriel Arigon during her PhD at the University Claude Bernard. She is now ATER at the University of Lyon 2.

5.16. Identitag

Keywords: *SAGE*, *database*.

Participants: Laurent Duret, Dominique Mouchiroud.

IDENTITAG is a relational database for SAGE tag identification and interspecies comparison of SAGE libraries. IDENTITAG has been developed in collaboration with C. Keime, F. Damiola, and O. Gandrillon from the CGMC Lab of the Université Claude Bernard. For more information, see http://pbil.univ-lyon1.fr/software/identitag/.

5.17. ISee

Participants: Annick Chamontin, Philippe Genoud [Correspondent], François Rechenmann, Danielle Ziébelin.

The aim of ISEE (IN SILICO BIOLOGY E-LEARNING ENVIRONNEMENT) is to explain the principles of the main bioinformatics algorithms through interactive graphical user interfaces and to illustrate the application of the algorithms to real genomic data. Written in Java, ISEE defines a generic framework for combining algorithms with courses. More precisely, the environment implements the metaphor of a lab notebook: the left pages present and explain the experiments to be carried out by the student, whereas the right pages display the progress of these experiments, *i.e.* the execution of the associated algorithms. In its present state, the environment offers different algorithmic modules structured into three main chapters: sequence comparison, statistical analysis of DNA sequences for the identification of coding regions, and basic pattern-matching algorithms including the use of regular expressions. These and other algorithms have been integrated in two original practical courses. The first one is an introduction to the statistical analysis of genetic sequences and leads the student to the identification of the origin of replication within bacterial genomes. The second one shows the student how to identify coding regions in bacterial genomes and to characterize their products. The latter course is developed in collaboration with the CCSTI ("Centre de Culture Scientifique Technique et Industrielle") in Grenoble, which uses ISEE for its "École de l'ADN". For more information, see: http://www-helix.inrialpes.fr/article124.html.

5.18. LalnView

Keywords: *local alignment*, *visualizer*.

Participants: Laurent Duret [Correspondent], Jean-Francois Gout.

LALNVIEW is a graphical program for visualising local alignments between two sequences (protein or nucleic acids). Blocks of similarity between the two sequences are colored according to the degree of identity between segments.

The program is also able to display sequence features (active site, domain, motif, propeptide, exon, intron, promoter, etc.) along with the alignment. This allows one to make the link between sequence similarity and known functions. For more information, see : http://pbil.univ-lyon1.fr/software/lalnview.html.

5.19. MareyMap

Keywords: *recombination rate estimator*.

Participants: Delphine Charif, Laurent Guéguen [Correspondent], Gabriel Marais.

MAREYMAP is an meiotic recombination rate estimation program. It is based on R and features a graphical interface in tcl/tk.

5.20. Migal

Keywords: *RNA*, *tree comparison*.

Participant: Marie-France Sagot [Correspondent].

MIGAL is an algorithm that compares two RNA structures. MIGAL was developed by Julien Allali during his PhD and is maintained by him at the University of Marne-la-Vallee (http://www-igm.univ-mlv.fr/~allali/logiciels/index.en.php).

5.21. Motus

Keywords: *reaction motif search and inference*.

Participants: Ludovic Cottret, Vincent Lacroix, Marie-France Sagot [Correspondent].

MOTUS is an algorithm for searching and inferring coloured motifs in metabolic networks. Coloured motifs are multisets of colours (in this case EC numbers). Occurrences are connected sets of nodes whose colours match the motif. The algorithm includes measures for motif over-representation, as well as clustering and drawing options for analysing the results. The algorithm was developed by Vincent Lacroix, and the web interface by Ludovic Cottret and Odile Rogier from the PRABI. The algorithm for drawing motifs and occurrences was done by Fabien Jourdan (INRA Toulouse).

5.22. Njplot

Keywords: *phylogenetic tree drawing*.

Participant: Manolo Gouy [correspondent].

This program for drawing phylogenetic trees has been updated by allowing unresolved trees to be processed, and by adding a PDF graphical output. Available at: http://pbil.univ-lyon1.fr/software/njplot.html.

5.23. OBIWarehouse

Keywords: *database*.

Participants: Eric Coissac [Correspondent], Anne Morgat, Alain Viari.

OBIWAREHOUSE (formerly MicrOBI) is a relational database devoted to microorganisms, integrating and synchronizing heterogeneous data from various public sources: genome data (EBI genome files), proteome data (Swiss-Prot and HAMAP), metabolic data (Enzyme and KEGG) and functional classification (GeneOntology). It has been implemented using PosgreSQL and ZOPE. It acts as a data source for UNIPATHWAY (Section 5.37), but can also be used as a stand-alone database. Since 2006, it is part of a new opensource project called "OBISchema". For more information see: http://www.grenoble.prabi.fr/obiwarehouse

5.24. Oriloc

Keywords: *replication origin and terminus*.

Participant: Jean Lobry [Correspondent].

ORILOC is a program to predict the putative origin and terminus of replication in prokaryotic genomes. The program works with unannotated sequences and therefore uses GLIMMER2 outputs to discriminate between codon positions. For more information see : http://pbil.univ-lyon1.fr/software/oriloc.html

5.25. PepLine

Keywords: *proteomic data analysis*.

Participant: Alain Viari [Correspondent].

PEPLINE is a software pipeline supporting the high-throughput analysis of proteomic data, in particular the identification of proteins from MS/MS spectra. At present, PEPLINE consists of two components: TAGGOR and PMMATCH. TAGGOR generates so-called PSTs (Peptide Sequence Tags) from MS/MS data, while PMMATCH maps the PSTs to sequences in protein databanks, or to the complete translated genome of an organism, thus helping to locate the gene coding for the protein. PEPLINE was developed in close collaboration with the Laboratoire d'Etude de la Dynamique des Protéomes (EDyP) (formerly LCP) headed by J. Garin at the CEA of Grenoble. For more information, see : http://www.grenoble.prabi.fr/protehome/software/pepline.

5.26. PhyloJava

Keywords: *phylogenetic reconstruction*.

Participants: Laurent Duret, Manolo Gouy [Correspondent], Simon Penel.

PHYLOJAVA is a server for phylogenetic reconstruction that is able to distribute a computation on a grid. For more information, see : http://pbil.univ-lyon1.fr/software/phylojava/phylojava.html. PHYLOJAVA was developed also with Timothée Sylvestre.

5.27. ProDom

Keywords: *database*, *protein domain families*.

Participant: Daniel Kahn [correspondent].

PRODOM (2005; *Nucleic Acids Res.* 33(Database issue):212-215) is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TREMBL sequence databases. The analysis of evolutionary scenarios of protein domain families from PRODOM showed that only a small minority of domain families is truly ancestral. Far from being static, the protein domain repertoire undergoes a continuous innovation process. Therefore the tremendous diversity of modular proteins results from both the combinatorial assortment of protein domains and an ongoing process of protein domain innovation. The PRODOM database is available at: http://prodom.prabi.fr/prodom/current/html/home.php.

5.28. PSbR

Keywords: *perfect sorting by reversals*.

Participants: Marie-France Sagot, Eric Tannier [correspondent].

Implemented by Yoan Diekman, based on an algorithm published by Marie-France Sagot and Eric Tannier. It is used to test the compatibility of the parsimony hypothesis for evolution and the preservation of common clusters of genes (http://biomserv.univ-lyon1.fr/~tannier/PSbR/).

5.29. Remote Acnuc Access

Keywords: *access to molecular databases*.

Participant: Manolo Gouy [correspondent].

A network protocol has been developed to allow remote access to biomolecular databases of the PRABI through the internet (http://pbil.univ-lyon1.fr/databases/acnuc/remote_acnuc.html). This protocol has been used by a client retrieval program, query_win (http://pbil.univ-lyon1.fr/software/query_win.html), that was previously able to query local databases only. Three APIs have been developed for the C and Python languages and for the R environment of statistical computing. The last one is at the basis of the database access to the seqinR package.

5.30. Repseek

Keywords: *DNA sequences*, *approximate repeat detection*.

Participants: Eric Coissac [Correspondent], Alain Viari.

REPSEEK is a program for finding approximate repeats in large DNA sequences. While there are several efficient methods for detecting strict (or almost strict) repeats, REPSEEK has been designed to efficiently detect approximate repeats in DNA sequences allowing for deletion and substitution scores. REPSEEK also uses a statistical framework to ascertain the significance of the repeats. REPSEEK is based on a new, space efficient, implementation of the Karp-Miller-Rosenberg algorithm. It has been developed, through a collaboration, by Guillaume Achaz at the Université Paris VI. For more information see: http://wwwabi.snv.jussieu.fr/~public/RepSeek/

5.31. RFDD

Keywords: *creation and update database*, *transcriptomic technique*.

Participant: Guy Perrière [correspondent].

In collaboration with Helene Simonnet (Centre de Genetique Moleculaire et Cellulaire, UMR CNRS 5534, Lyon), we have developed a web service for bioinformatic analysis of RFDD (Restriction Fragment Differential Display) results. RFDD is a transcriptomic technique derived from AFLP (Analysis of Fragment Length Polymorphism). This service uses a relational database of restriction fragments obtained by *in silico* digestion of all human and rat transcribed sequences present in the RefSeq database. We have developed software for the creation and update of this database and for providing web access to it (http://pbil.univ-lyon1.fr/software/RFDD/). A first biological result has been derived from this tool by analysis of transcriptome data from rat under hypoxic conditions.

5.32. Sarment

Keywords: *sequence partitioning*.

Participants: Bastien Boussau, Laurent Guéguen [correspondent], Manolo Gouy.

SARMENT is a source package of object-oriented Python modules for sequence segmentation via HMM analysis and Maximal Predictive Partitioning. Available at: http://pbil.univ-lyon1.fr/software/sarment/.

5.33. SeaView

Keywords: *editor of multiple sequence alignments*.

Participant: Manolo Gouy [correspondent].

This program for editing multiple sequence alignments has been updated again to by allow its interface with any external multiple sequence alignment program. Seaview is diffused by the PRABI (http://pbil.univ-lyon1.fr/software/seaview.html) and as a Debian package.

Available at: http://pbil.univ-lyon1.fr/software/seaview.html.

5.34. SeqinR

Keywords: *analysis and management of biological (DNA and protein) sequences*, *exploration*, *visualization*.

Participants: Delphine Charif, Jean Lobry [correspondent], Anamaria Necsulea, Leonor Palmeira.

This program has been updated and some non-parametric statistics for the analysis of dinucleotide over- and under-representation in sequences have been implemented and are now available since version 1.0-5. Available at: http://cran.univ-lyon1.fr/src/contrib/Descriptions/seqinr.html.

5.35. Smile and Riso

Keywords: *inference*, *motifs*, *promoters*, *regulatory sequences*, *word statistics*.

Participant: Marie-France Sagot [Correspondent].

SMILE (http://www-igm.univ-mlv.fr/~marsan/smile_english.html) and RISO (http://algos.inesc-id.pt/~asmc/software/riso.html) are motif inference algorithms that take as input a set of DNA (RNA) or protein sequences. SMILE was developed by Laurent Marsan, now at the University of Versailles. The code (in C) can be freely obtained by academics and non-profit research organisations by simply sending a mail to marsan@univ-mlv.fr or to Marie-France.Sagot@inria.fr. The core of SMILE has been improved and extended into a new algorithm, RISO, by Alexandra Carvalho from the Instituto Superior Tecnico (IST) of Lisbon, Portugal, in a collaboration with researchers from the IST.

5.36. SymBioCyc

Keywords: *database*, *endosymbiont*, *metabolism*.

Participants: Ludovic Cottret, Marie-France Sagot [Correspondent].

SYMBIOCYC is a database of metabolic data dedicated to endosymbiotic organisms (that is, to organisms that live within the body or cells of another organism). SYMBIOCYC follows the same representation model as BioCyc (http://www.biocyc.org/) but the data it contains has been expertised both automatically and manually by Ludovic Cottret. Available at: http://pbil.univ-lyon1.fr/software/symbiocyc.

5.37. UniPathway

Keywords: *database*, *metabolism*.

Participants: Eric Coissac, Anne Morgat [Correspondent], Alain Viari.

UNIPATHWAY is a database of manually curated metabolic pathways developped in collaboration with the Swiss-Prot group at the Swiss Institute of Bioinformatics (http://www.expasy.ch/people/swissprot.html). Its primary goal is to describe the metabolic data of UniProtKB/Swiss-Prot Database entries, but the UniPathway data model can also be used in a more general context. The UNIPATHWAY DB is hosted by the PRABIG Server at INRIA-Rhône-Alpes (Section 2.1). Available at: http://pbil.univ-lyon1.fr/software/symbiocyc.

5.38. WellReader

Participants: Bruno Besson, Johannes Geiselmann, Hidde de Jong [Correspondent], Delphine Ropers.

WELLREADER is a program for the analysis of gene expression data obtained by means of fluorescent and luminescent reporter genes. WellReader reads data files in an XML format or in a format produced by microplate readers, and allows the user to detect outliers, perform background corrections and spline fits, and compute promoter activities and protein concentrations. WellReader has been written in Matlab and will be made available to the academic research community early 2008.

5.39. Other software developed in HELIX

Participants: Manolo Gouy [Correspondent], Alain Viari [Correspondent].

HELIX has contributed to the development of software by other members of the PRABI (Section 2.1). This is in particular the case for:

- ROSO (INSA, N. Raymond), which supports the efficient design of eukaryotic DNA chips;
- RTKDB (CGMC, université Claude Bernard, J. Grassot), which is a database dedicated to the tyrosine kinase receptors. RTKDB uses the FAMFETCH environment (Section 5.7);
- BIBI (LBBE, J.-P. Flandrois), which is a powerful tool for identifying pathogenic bacteria from genomic sequences.

Several other programs have resulted from the activities of HELIX members, but are no longer being actively developed. This concerns the following programs (with the contact person between brackets): ACNUC (Manolo Gouy), ALICE (Marie-France Sagot), COMBI (Marie-France Sagot), COSAMP (Marie-France Sagot), DOMAINPROTEIX (Alain Viari), DRUID (Marie-France Sagot), EMKOV (Alain Viari), FACTORTREE (Marie-France Sagot), GEM (Bruno Spataro), JADIS (Dominique Mouchiroud), MTDP (Alain Viari), SATELLITES (Marie-France Sagot), and SEAVIEW (Manolo Gouy), UTOPIA (Marie-France Sagot).

6. New Results

6.1. Comparative genomics

6.1.1. Computational analysis of the evolution of species, genomes and gene families

Understanding genome evolution requires to focus on individual gene histories. However each gene may have evolved differently through time leading to individual phylogenies that are all distinct. The gene trees obtained may also, and in general will be different from the species tree. The resulting incongruence among gene trees, and between them and the evolutionary history of the species can be turned to good account in order to identify the evolutionary events that led to the discrepancies observed. These could be gene duplication, recombination, horizontal gene transfer etc. By statistically comparing gene trees and so-called phylum reference trees (a phylum is a taxonomic unit in the rank below kingdom and above class) obtained by concatenating genes, an approach by was proposed that allows to analyse numerous gene families (extracted from HOGENOM database). This study is part of the PhD of Sophie Abby. She is particularly interested in one type of evolutionary event that may scramble the phylogenetic evidence, namely horizontal gene transfer (HGT). Her aim is to highlight the role of vertical inheritance relatively to horizontal transfer. HGT was the main topic of a survey paper she co-wrote this year [3]. Despite the noise, some characters are often considered as more reliable for inferring a phylogeny. This is the case of the patterns of insertions and deletions in protein families. How much more reliable these characters are, remains, however, an open issue that has been explored by Yves Clément during his master together with Bastien Boussau as part of his PhD.

Horizontal gene transfer (HGT) is recognised as a major force for bacterial genome evolution. Yet, numerous questions remain about the transferred genes, their function, quantity and frequency. The extent to which genetic transformation by exogenous DNA has occurred over evolutionary time was initially addressed by an *in silico* approach using the complete genome sequence of the *Ralstonia solanacearum* GMI1000 strain [20]. Methods based on phylogenetic reconstruction of prokaryote homologous genes families detected 151 genes (13.3%) of foreign origin in the *R. solanacearum* genome and tentatively identified their bacterial origin. These putative transfers were analysed in comparison to experimental transformation tests involving 18 different genomic DNA positions in the genome as sites for fully or partially homologous recombination. Significant transformation frequency differences were observed among the positions tested. The genomic positions containing the putative exogenous DNA were not systematically transformed at the highest frequencies. These results support the notion that the bacterial cell is equipped with active mechanisms to modulate acquisition of new DNA at different genomic locations. HGTs were also studied using a partitioning algorithm http://pbil.univ-lyon1.fr/software/sarment/ while positionning in the bacterial phylogeny the Aquificales (presently considered to represent the most deeply branching order within the domain of the bacteria).

Besides duplication, recombination and HGT, the evolutionary process may also create new gene families by assembling new genes from copies of pieces of various older genes, rapidly building new functions from a novel collection of already reliable parts. This process was termed “modular evolution” and many proteins are thus organised as a succession of modules called *domains*. Anne-Sophie Sertier has began a PhD on the study of such protein modular evolution. The approach developed uses a bayesian network in order to map the evolution of protein families and protein domains along the tree of life, and to link the evolution of new protein families with both domain innovation and domain shuffling. HELIX is also involved in the development of InterPro (http://www.ebi.ac.uk/interpro), an integrated resource for protein families, domains and functional sites, which integrates different protein signature databases, including ProDom, the database of protein domains established and maintained by Daniel Kahn. Several recent additions to InterPro that are detailed in [36].

The analysis of evolutionary rates is a well-known approach to characterising the effect of natural selection at the molecular level. Sequences contributing to a species adaptation are expected to evolve faster than nonfunctional sequences because favourable mutations have a higher fixation probability than neutral ones. Such an accelerated rate of evolution might be due to factors other than natural selection, in particular GC-biased gene conversion (a form of recombination where the sequence of one gene “converts” the other, that is,

is nonreciprocally transferred to replace the other). This is true of neutral sequences, but also of constrained sequences, which can be illustrated using the mouse Fxy gene. Several criteria can discriminate between natural selection and biased gene conversion models. These criteria suggest that the human regions recently reported as being under accelerated evolution are most likely the result of biased gene conversion. These regions, far from contributing to human adaptation, may therefore represent the Achilles' heel of our genome [22].

When evolutionary processes (mutation and selection pressures) are symmetrical with respect to the two DNA strands, the equimolarities [A]=[T] and [G]=[C] are expected on each strand, at equilibrium. These parity rules are observed on long DNA sequences (for example on whole chromosomes), but local deviations (*i.e* base composition asymmetry) are frequently encountered, in the three kingdoms of life. Among the probable causes of base composition asymmetry are essential cellular mechanisms: DNA replication and transcription. In most bacterial and archaeal genomes protein-coding genes are preferentially situated on the leading strand for replication, in order to avoid head-on collisions between the replication and transcription machineries. This means that nucleotide asymmetry is caused by two superposing mechanisms, because the leading strand for replication is also the coding strand for transcription. As part of her PhD, Anamaria Necsulea developed a computational method for estimating the relative contributions of the two mechanisms (replication and transcription) to the evolution of base composition asymmetry [38], [63].

The non homogeneous evolutionary model of nucleotide sequences and its implementation in the NHPHYML software is being used to estimate the G+C contents of key ancestral sequences: those of all bacteria, all archaea, all eukaryotes, as well as of LUCA, the last universal ancestor. Results show independent enrichment in G+C contents of the bacterial and archaeal lineages, and thus suggest adaptations to warmer environments of these two lineages in their early history. Publication of these results is in preparation in collaboration with colleagues from the LIRMM at the University of Montpellier (Samuel Blanquart and Nicolas Lartillot) who developed similar models applied to protein sequences.

The evolution of nucleic sequences is usually modelled by point substitutions and under the hypothesis that sites evolve independently of each other. This hypothesis is mainly kept for mathematical purposes and has no biological foundation, as it is now clear that molecular substitution mechanisms frequently involve adjacent bases. The most typical example is the highly frequent spontaneous chemical transformation of CpG dinucleotides observed on some sequences (the " CpG" notation is used to distinguish a cytosine C followed by guanine G from a cytosine base paired to a guanine). Leonor Palmeira defended her PhD in July on this topic, and a method for sequence evolution under this kind of model is currently submitted. A program of phylogenic reconstruction, based on a maximum-likelihood approach, under a neighbour-dependent evolution modelling is under development in collaboration with Jean Bérard from the Mathematics Department of the Université Claude Bernard.

Most eukaryotic genes are interrupted by non-coding introns that must be accurately removed from pre-mRNAs to produce translatable mRNAs. Splicing is locally guided by short conserved sequences, but genes typically contain many potential splice sites, and the mechanisms specifying the correct ones remain poorly understood. In most organisms, introns cannot be efficiently predicted on the sole basis of sequence motifs. Hence, bioinformatics methods for gene annotation have to rely on other information (search for open reading frames, cross-species conservation of protein-coding sequences). This raises the question of how eukaryotic cells recognize intron signals within pre-mRNA sequences. The length distribution of introns (experimentally validated) in complete eukaryotic genomes was analysed [25]. In all taxa (protists, plants, fungi and animals) we found a deficit in introns whose length is a multiple of three (3n introns). This deficit is specific of 3n introns that do not contain stop codons in frame with the preceding exon. This suggests that there is a selective pressure to avoid 3n-stopless introns (i.e. translatable introns) within eukaryotic genes. Various hypotheses have been put forward to explain this bias. The study has suggested also that, ironically, gene annotation software, that rely on the search of open-reading frames, appear in fact to mimic the real cellular process of control of splicing accuracy.

Ehrlichia ruminantium is the causative agent of heartwater, a major tick-borne disease of livestock in Africa introduced in the Caribbean and threatening to emerge and spread in the American mainland. HELIX'

biological collaborators from the CIRAD in Montpellier were involved in the complete genome sequencing of two isolates of *E. ruminantium* of differing phenotype, namely the isolates Gardel (Erga) from the Guadeloupe Island and Welgevonden (Erwe) originating from South Africa. A comparative analysis of the genomes was performed and revealed the presence of unique and truncated CDS differentiating Erga from Erwe/Erwo [21], [54]. Some of these species-specific CDSs were further considered as targets for differential diagnosis on several natural isolates of *E. ruminantium*. PCR analysis of these target genes generated strain-specific patterns that can therefore be used as signatures for PCR diagnosis and vaccine management tools.

DNA barcoding should provide a fully automated way to rapidly and accurately identify species by using a standardized DNA region as a tag. Based on sequences available in GenBank and sequences specifically produced for a study conducted by the HELIX team at Grenoble, the resolution power of the whole chloroplast trnL (UAA) intron (254-767 bp) and of a shorter fragment of this intron (the P6 loop, 10-143 bp) amplified with highly conserved primers was evaluated. It was thus shown that, despite a relatively low resolution, the whole trnL intron and one of its loop present many advantages as tags: the primers are highly conserved, and the amplification system is very robust [50]. The loop can even be amplified when using highly degraded DNA from processed food or from permafrost samples, and has the potential to be extensively used in food industry, in forensic science, in diet analyses and in ancient DNA studies.

6.1.2. Modelling and analysis of the spatial organisation and dynamics of genomes

6.1.2.1. Spatial organisation

Since its discovery in 1976, the structure into isochores of some vertebrate genomes has provided a notable example of the relation between genetic information and the molecular functioning of a genome. Defined initially as a regionalisation of the G+C content of a genome, isochores have progressively been associated with numerous other biological properties, to the point where no consensual definition exists anymore. In collaboration with Christelle Melo de Lima, a re-examination was made of this structure which associated approaches as "objective" as possible with models for the relation between structure and a number of biological properties. This work has also in part involved a then Master student, Alexandra Popa, who is currently doing a PhD on the influence of the evolution of recombination rates on the structure of a genome. The work relied on the use of Hidden Markov models (HMMs) specially adapted to take into account many biological properties other than G+C and to analyse complex gene structures with bell-shaped length distributions through the introduction of macro-states. This has enabled to reveal an isochore structure in the chimpanzee genome and even in some fishes (*Tetraodon* and *Danio*) where isochores were believed until recently not to exist. The issue of isochores is however not free from controversy as the signal is often weak (such as in the fishes). Despite this, and the fact that their function and mechanism are yet unknown, it seems clear that isochores are important elements of a genome that require understanding. Among others, this may come from being able to identify what originates them.

The signal is weak also as concerns possible functional characteristics of the distribution of repeated elements, particularly of the transposable element type, along a genome. Transposable elements are sequences of DNA that can move around to different positions within the genome of a single cell, a process called transposition. Transposable elements of a same family are approximate repeats of one another. Those that are still active are transcribed and we know that such transcription is regulated but genome-scale studies of their profile of expression has rarely been attempted. Members of HELIX in collaboration the "Génomes et populations" team of Christian Biémont at the LBBE are addressing this question, using the genome of *Drosophila melagonaster* (the common fly) as model. The activity of a transposable element was measured using ESTs (Expressed Sequence Tags). These are relatively short sub-strings of a transcribed protein-coding or non-protein-coding nucleotide sequence originally intended as a way to identify gene transcripts and available in various public databases. Initial results seem to indicate a correlation between the number of transposable elements in a family to which an EST maps (and that one may therefore assume is expressed) and the number of copies in that family. The latter is in contradiction with previously published studies. The results point also to a different profile of expression of transposable elements in the X chromosome. Further investigation into the correlation between transposable elements, and in particular, their expression and sex evolution as well as

genome organisation, general gene expression and other genomic features is under way and a paper is in preparation.

Transposable elements tend to accumulate in heterochromatic (tightly packed DNA) regions where they are progressively degraded by mutational processes. In a collaboration with the “Génomes et populations” team, at the LBBE, a model for the ecolution of copies of LTR retrotransposons LTR in the heterochromatic regions of *Drosophila'* has been done, together with an estimation of the number of transposable elements inserted in the genome. This has permitted to reveal a recent wave of insertions of such elements in the genome of *Drosophila*. The work has been submitted to *Molecular Biology and Evolution*.

Retrotransposons are ubiquitous in the plant genomes and are responsible for their plasticity. Recently, members of HELIX have described a novel family of retrotransposons, named Retand, in the plant *Silene latifolia* which possesses evolutionary young sex chromosomes of the mammalian type (XY). Long terminal repeats (LTRs) of Retand were analised. A majority of X and Y-derived LTRs formed a few separate clades in phylogenetic analysis reflecting their high intrachromosomal similarity. Moreover, the LTRs localised on the Y chromosome were less divergent than the X chromosome-derived or autosomal LTRs. These data can be explained by a homogenisation process, such as gene conversion, working more intensively on the Y chromosome.

The comparison of genetic and physical maps is still the most widely used approach to estimate genome-wide recombination rates. This comparison relies on an a priori modelling of the process of occurrence of cross-overs (exchange of DNA material between the arms of chromosomes from a same pair during meiosis). However current models do not take into account known results concerning both the molecular processes at play and the chromosomes themselves. A more realistic modelling is being developed as part of the PhD of Alexandra Popa and has already allowed to independently verify a well-known correlation between the recombination rate and the length of the chromosomes arms. Besides this, a software called MareyMap (a Marey map is a graph that relates physical distance on a genome with genetic distance) has been developed based on GNU R and Tcl/Tk [45] (see Software).

Finally, in the context of his PhD, co-supervised between Grenoble and Lyon, Yves-Pol Deniélou successfully extended the "common connected components" (CCC) approach initially developed by Frédéric Boyer during his PhD in HELIX to tackle the problem of looking for conserved gene locations across several bacterial species and allowing for permutations in gene order. The goal was to introduce a notion of quorum in the definition of conserved blocks. For instance, when looking for syntenies (blocks of similar genes which relative location is conserved across species) between n ? 3 genomes, the requirement is that some of the conserved genes may be missing in some species (more precisely they should occur in at least q ? n/2 genomes). The resolution of this problem necessitated a complete re-design of the primary algorithm that could not be extended trivially. It is interesting to note that in the case q=n, the new algorithm turned out to be more efficient than the previous one.

6.1.2.2. Dynamics

The rearrangement dynamics of eukaryotic chromosomes still lack good models and good methods that would enable to predict the past of genomes with some reliability. Why a rearrangement occurs and why it can be fixed in a population remain widely open questions. The only existing model states that a rearrangement is the result of a chromosome break that can happen uniformly at random on the genome, and has little impact on an organism’s fitness. This model has been much criticised, but no alternative has been formally described so far. As part of her PhD, Claire Lemaitre is participating in the construction of a new model, by studying the relationship between the positions of the chromosome breakpoints, the replication domains and chromatine structure, as well as various other properties such as C+G content, repeats etc. Some of this work, notably on the correlation between breakpoints and replication domains, is done in close collaboration with the group of Alain Arnéodo at the Joliot-Curie lab at the ENS-Lyon. The above model and analysis uses as one of its initial step a method designed by Claire Lemaitre to identify the location of chromosome breakpoints of mammalian genomes with a high precision. The method proceeds by comparing pairwise the genomes that have been sequenced. It then consists in building syntenic blocks, that is, portions of the genomes that have not been separated by a rearrangement, and in refining the extremities of the blocks by performing sequence

alignements and a partition method. A set of coordinates of breakage locations was thus obtained on the human genome, along with information that is partial for now on the position of the breakpoints in the phylogenetic tree of mammals. The method for breakpoint detection and the new rearrangement model are currently been written into two papers to be submitted towards the end of 2007.

In order to understand the mutations that separate several lineages, and how living species evolved each with their own molecular characteristics, it is useful to know the chromosome configurations of their ancestors. As no DNA molecule is known to be conserved after a few hundred thousand years, it is necessary to predict the ancestral configurations by comparing actual genomes, studying what they have in common and what differenciates them. The HELIX group is involved in the methodological study of ancestral chromosome configurations. Several methods are begin investigated, and some of them have already given some results: for example, as the proportion of G and C bases in a sequence is highly correlated with the size of the chromosome it is in, predicting the ancestral G+C proportion, with the method that Bastien Boussau developed during his PhD, may be used as a hint to predict the ancestral chromosome sizes. This has been used to discriminate among several possible configurations of mammalian ancestral chromosomes discussed by biologists, and to identify the most plausible one.

The methods of reconstruction of evolution scenarios by rearrangements are numerous, but almost all of them end up with only one possible solution, discarding a huge number of equivalent ones. A structure of equivalent solutions was given by Bergeron and colleagues, but with no accompanying algorithm for enumerating all such solutions. As part of her PhD, Marília Braga together with an Italian master student, Celine Scornavacca devised in 2006 an algorithm that computes this structure for small permutations [61]. It has now been applied to some examples of chromosomal evolution, notably the mammalian X and Y chromosomes. Indeed, a very special type of evolution is observed on such chromosomes. The process of divergence of the two is at the origin of sexual differentiation – the female XX and the male XY pairs. While female organisation favours the X chromosome conservation, the male XY pair evolution causes a degeneration of the Y chromosome. Rearrangements probably play an important part in this process, but since the two chromosomes are now very different, it is difficult to precisely predict the events that progressively led them to diverge. Thanks to the method described above, Marília Braga computed the structure of all parsimonious scenarios of reversals that may have transformed the chromosome X into the Y. She was then able to assess a probability of occurrence of some reversals that were predicted in the literature. Moreover, Claire Lemaitre has applied her method described above to precisely compute the positions of the Y chromosome breakpoints, and has thus been able to identify some of the reversals that took place thanks to several repeated features in the sequences. A paper is in preparation on this topic.

As part of her PhD, Marília Braga has also continued to study some theoretical aspects of the sorting by reversals problem. When only genomes without gene duplications are considered, there are efficient algorithms to find an optimal solution to this problem. However, it becomes much more complicated when duplications are allowed. In collaboration with members of the Associated Team with the University of São Paulo, Brazil, and some of their ex-PhD students currently holding a position at the Federal University of Mato Grosso do Sul, a new approach was proposed to the sorting by reversals problem with duplicated genes, called the Repetition-free Longest Common Subsequence (RFLCS). Given two strings, the RFLCS is the problem of finding the longest subsequence that contains no repeated symbols. The RFLCS was proved, by polynomial reduction from a particular version of max 2-SAT, to be NP-hard even when the number of occurrences of a symbol in each genome is at most equal to 2 [58]. Three simple approximation algorithms and an integer linear programming formulation (IP) for the RFLCS were also proposed and implemented by the Brazilians and applied to some real examples.

The study of reversals was also applied to the study of some intracellular bacteria that form a symbiotic system with insects and nematodes. This work is done in close collaboration with Fabrice Vavre from the LBBE team “Génétique et Évolution des Interactions Hôtes-Parasites” and the bacterium chosen *Rickettsia*, an evolutionary close relative of the bacterium *Wolbachia* which is the one that interests F. Vavre’s team but for which too few data exist at current time. Among both the *Rickettsias* and the *Wolbachias*, at least one strain shows characteristics that are very different from the characteristics of others. In particular, its genome

contains more repeated elements and has been more rearranged than the genomes of intracellular bacteria that are in general well conserved. One main final objective of the study is to attempt a link between the structure of *Wolbachia*'s genome and its ecology, in particular its capacity to infest numerous host species.

Rearrangements and accelerated mutation rates are observed in the two strains of *Ehrlichia ruminatium* annotated by the group of Roger Frutos (CIRAD Montpellier) with whom members of HELIX collaborate (see Section 6.1.1). A comparative genomic analysis of these strains as well as additional species of *Ehrlichia* was performed. This has suggested the presence of an active and specific mechanism of genomic plasticity, probably following the exposure to a diverse environment (different hosts), which could explain the limited field-efficiency of vaccines against *E. ruminantium*. Further experiments, including attenuated strains are under way to understand this mechanism.

6.1.3. Motif search and inference

Work on sequence analysis, notably on alignment of and long repeats detection in whole chromosomes or genomes, continues with Alair Pereira do Lago from the University of São Paulo, Brasil who is visiting again the French group, this year for a longer period, from September 20 to December 13, 2007. Since last year, the paper that had been submitted with as further co-authors Pierre Peterlongo (currently postdoc at SYMBIOSE) and a long-term Italian collaborator, Nadia Pisanti, has been accepted and is in press [43]. A second paper is about to be submitted (with also a student of Alair, Gustavo Akio Tominaga Sacomoto, as participant) after a delay due to a desire to further improve the performance of the new algorithm, called TUIUIU. The previous work had already led to an algorithm, Ed'Nimbus publicly available (see Software).

On the topic of RNA motifs detection, a collaboration was set up with the group of Eric Westhof at the IBMC at Strasbourg to analyse a dataset of small RNAs that were sequenced in the lab. The primary goal of this project is to predict the most likely origin of each of the small RNAs while also searching for novel miRNA candidates. This project presented itself as an opportunity to test several criteria for miRNA precursor identification. The first task of the project consisted in the identification of all approximate matches of the small RNAs in the genome of our organism. We used approximate rather than exact searches to account for the possibility of sequencing errors, which were estimated not to exceed two per sequence. Following the identification of all genome hits, each was classified with respect to the statistical significance of the match in its genomic context, the annotation of the site, and a number of other criteria such as folding energy, adequate srtuctural features etc. Considering all the aforementioned criteria, 35 good candidates were identified, that are pending further analysis.

HELIX members are also involved in another RNA-motivated project which was funded by the ANR in 2006. The project includes three original parts: the design of sound combinatorial models (based on graph and tree theory) along with efficient algorithms to handle the structural modelling and functional assignment of RNAs, the constitution of annotated benchmark sets for the evaluation and validation on real biological problematics of such algorithms, and the development of freely-distributed software, as well as appropriate visualisation facilities. This project is conducted in close collaboration with Julien Allali, ex-PhD student of Marie-France Sagot and now Associate Professor at the LABRI, University of Bordeaux. He had introduced a new data structure, called MIGAL for "Multiple Graph Layers", composed of various graphs linked together by relations of abstraction/refinement. The new structure has proved useful for representing information that can be described at different levels of abstraction, each level corresponding to a graph. An algorithm was proposed for comparing two MIGALS. This algorithm was improved and a paper submitted in 2006 has now been accepted [5]. MIGAL is currently been compared to other available software on benchmark datasets that we are also helping to set up together with two further collaborators, Claude Thermes and Yves d'Aubenton from the Centre de Génétique Moléculaire (CGM) at Gif-sur-Yvette.

As part of his PhD, Paulo G. Fonseca has been concerned with the identification of transcription regulation modules, *i.e.* groups of co-regulated genes and their regulators. One important distinction of this work in relation to what is available in the literature is that he proposes to identify modules that are evolutionarily conserved. From the biological point of view, the approach is supported by three main premises: 1. co-regulated genes are bound by common regulatory proteins (transcription factors-TFs) and so they must

present common sequence patterns (motifs) in their regulatory regions, which correspond to the binding sites of those TFs; 2. co-regulated genes respond coordinately to certain environmental or growth conditions, and so they must be co-expressed under those conditions; 3. since transcription modules are supposedly responsible for important biological functions, they are more subject to selective pressure and therefore they must be evolutionary conserved. He thus defined the concept of transcriptional regulation metamodules (TRMMs) as groups of genes sharing regulatory motifs and displaying coherent context-specific expression behaviour consistently across species. From the methodological perspective, we note that the incompleteness and elevated noise levels of currently available data impose severe limitations on the reliability of the conclusions that can be drawn through the analysis of one data type in isolation. Therefore he proposed to analyse heterogeneous experimental data concerning several species simultaneously, with emphasis on genomic sequence and gene expression data. A PhD thesis proposition including a bibliographic review, problem formalisation and expected results has been produced and defended (as is the custom in Brazil) in early 2007 by Paulo G. Fonseca at the University of Recife to which he is attached as a PhD student. The PhD defence itself is expected for early 2008. A stand-alone Java application with graphical user interface is being developed as part of the work.

6.2. Functional genomics

6.2.1. Computational proteomics and transcriptomics

Concerning experimental proteomics and mass spectrometry, the PEPLINE software (Section 5.25) has been released to the mass spectrometry community and a paper describing the approach as well as some applications to the chloroplastic membrane of *A. thaliana* has been submitted. Work in experimental proteomics has been continued in two directions. First, several, scoring functions for the mapping of Peptide Sequence Tags (PST) on chromosome(s) have been proposed and implemented by Jérémie Turbet at the Laboratoire d'Etude de la Dynamique des Protéomes (EDyP, CEA Grenoble). In parallel, we tried to extend the original definition of a PST (one sequence tag and two flanking masses) in order to gain more information from the spectra. The new definition now involves several sequence tags separated by unresolved regions where only the mass is known. These extension are now being evaluated at the EDyP laboratory.

SAGE has been widely used to study the expression of known transcripts, but much less to annotate new transcribed regions. LongSAGE produces tags that are sufficiently long to be reliably mapped to a whole-genome sequence. Members of HELIX have used this property to study the position of human LongSAGE tags obtained from all public libraries. The focus was mainly to tags that do not map to known transcripts. The frequency of tags matching once the genome sequence but not in an annotated exon suggests that the human transcriptome is much more complex than shown by the current human genome annotations, with many new splicing variants and antisense transcripts [27].

Microarrays are the technique of choice to measure gene expression differences between sets of biological samples. Many of these will be due to differences in the activities of transcription factors. In principle, these differences can be detected by associating motifs in promoters with differences in gene expression levels between the groups. In practice, this is hard to do. In collaboration with the group of Des Higgins at the University College of Dublin, members of HELIX have combined correspondence analysis, between-group analysis and co-inertia analysis to determine which motifs, from a database of promoter motifs, are strongly associated with differences in gene expression levels. Given a database of motifs and gene expression levels from a set of arrays, the method produces a ranked list of motifs associated with any specified split in the arrays [26].

In a collaboration with Xavier Nesme (INRA), microarrays were also used to study the genes characteristic of genomic species in the cluster of species *"Agrobacterium"*, more specifically the presence/absence of genes in the sister species in order to understand the roles and history of genes specific to this species.

With the advance of microarray technology, several methods for gene classification and prognosis have already been designed. However, under various denominations, some of these methods have similar approaches. In a study performed by Caroline Truntzer, ex-PhD student co-supervised by Christian Gautier, an evaluation was

made of the influence of gene expression variance structure on the performance of methods that describe the relationship between gene expression levels and a given phenotype through projection of data onto discriminant axes [53]. Between-Group Analysis and Discriminant Analysis (with prior dimension reduction through Partial Least Squares or Principal Components Analysis) were thus compared. A geometric approach showed that these two methods are strongly related, but differ in the way they handle data structure. Three main situations may be identified. When the clusters of points are clearly split, both methods perform equally well. When the clusters superpose, both methods fail to give interesting predictions. In intermediate situations, the configuration of the clusters of points has to be handled by the projection to improve prediction. The use of Discriminant Analysis appear recommendable in this case. An original simulation method was devised to conduct the study. It allows to generate the three main structures by modelling different partitions of the whole variance into within-group and between-group variances. These simulated datasets were used as a complement to some well-known public datasets to investigate the behaviour of the methods in a large diversity of situations. To examine the structure of a dataset before analysis and preselect an *a priori* appropriate method for its analysis, a two-graph preliminary visualisation tool was proposed that plots patients on the Between-Group Analysis discriminant axis (x-axis) and the within-group Principal Components Analysis component on the first as well as the second axis (y-axis).

Molecular data on biodiversity both in health and ecology represent new challenges for data analysis. In particular, the large number of probes present in this case on the DNA chips invalidate all discriminant analyses. HELIX has concentrated its effort on addressing this issue. Two main results that have been obtained. The first, by Jean Thioulouse, is that the association between DNA chips devoted to biodiversity analysis and environment data cannot be made by a classical maximization of correlation (canonical correlation analysis); however the use of co-inertia analysis (CIA) that maximizes covariance has continued to prove its efficiency on several studies of soil microbial biodiversity [6] [29] [30] [40] [41] [44]. Besides this, Jean Thioulouse together with Stéphane Dray from the Department of Ecology of the LBBE have presented an overview of ade4TkGUI, a Tcl/Tk graphical user interface for the most essential methods of ade4 [52]. The latter is a multivariate data analysis package for the R statistical environment. Both packages are available on CRAN. The pros and cons of ade4TkGUI this approach are discussed. The conclusion of the study is that command line interfaces (CLI) and graphical user interfaces (GUI) are complementary. ade4TkGUI can be valuable for biologists and particularly for ecologists who are often occasional users of R. It can spare them having to acquire an in-depth knowledge of R, and it can help first time users in a first approach.

Nuclear receptors (NRs) are transcription factors that are implicated in several biological processes such as embryonic development, homeostasis, and metabolic diseases. To study the role of NRs in development, it is critically important to know when and where individual genes are expressed. Although systematic expression studies using reverse transcriptase PCR and/or DNA microarrays have been performed in classical model systems such as *Drosophila* and mouse, no systematic atlas describing NR involvement during embryonic development on a global scale has been assembled. Adopting a systems biology approach, members of HELIX in collaboration with the “Structure and Evolution of Nuclear Hormone Receptors” team headed by Vincent Laudet at the ENS-Lyon have conducted a systematic analysis of the dynamic spatiotemporal expression of all NR genes as well as their main transcriptional coregulators during zebrafish development (101 genes). This extensive dataset establishes overlapping expression patterns among NRs and coregulators, indicating hierarchical transcriptional networks. This complete developmental profiling provides an unprecedented examination of expression of NRs during embryogenesis, uncovering their potential function during central nervous system and retina formation. Moreover, our study reveals that tissue specificity of hormone action is conferred more by the receptors than by their coregulators. Finally, further evolutionary analyses of this global resource led us to propose that neofunctionalisation of duplicated genes occurs at the levels of both protein sequence and RNA expression patterns. Altogether, this expression database of NRs provides novel routes for leading investigation into the biological function of each individual NR as well as for the study of their combinatorial regulatory circuitry within the superfamily.

6.2.2. Modelling and analysis of metabolism: molecular components, regulation, and pathways

Several developments have been carried out in 2007 concerning the work on motif search in metabolic networks which constituted the main part of Vincent Lacroix' PhD, defended in October of this year. Previous work bore exclusively on motif search. Extensions of this work in 2007 include: 1. the possibility to extract (infer) motifs from a graph (only the size of the motif is given), 2. a score reflecting the statistical significance of a motif, 3. a clustering method to group the occurrences that share the same topology (*i.e.e* that are isomorphic), and 4. a viewer to explore the results of the extraction. All of these new functionalities are included in the software MOTUS (see Software). The development of this software involved also the participation of several additional people: Odile Rogier, Ludovic Cottret and Fabien Jourdan from the UMR 1089 Xénobiotiques INRA-ENVT in Toulouse. MOTUS has been used to investigate the relationship between connectivity in metabolism and proximity of the corresponding genes on the genome. One of the results of this study has been that repeated connected sets of enzymes tend to be more organised in operons than non-repeated sets, thereby suggesting a mechanism for the appearance of repetitions. This work continues on eukaryotic genomes while the algorithm for inferring motifs is being improved by Cinzia Pizzi, post-doc in HELIX until December 2007 in a continuing collaboration with Vincent Lacroix, currently post-doc in the group of Roderic Guigó at the Centre for Genomic Regulation, Barcelona, Spain.

During her post-doc in HELIX, Patricia Thébaut also used MOTUS to study the relationship between coexpression of enzymes, sharing of transcription factors and connectivity in the metabolic network. One of the results has been that connected enzymes tend to be more co-expressed than non-connected ones. A paper in currently in preparation to present these results. Several other tests are being developed to evaluate the role of transcription factors in the link between expression and metabolism.

Metabolic networks can be decomposed into pathways. The notion of pathway is usually unclearly defined. Yet, there exists a formal definition of pathway as an elementary mode (denoted by EM). This is a set of enzymes that operate together at steady state. The computation of the elementary modes of a network has been extensively studied in the past years due to the number of applications related to this notion. Yet, all methods rely on linear programming to solve the problem whereas this problem seems to be combinatorial in nature. The goal we wish to achieve is to find a combinatorial algorithm for the calculation of elementary modes. While working in this direction, we believe that a reformulation of related concepts like minimal cut sets in terms of hypergraph problems would be of great help to improve the algorithms that are used for their calculation. Finally, a major issue in the computation of elementary modes and related concepts is the very large size of the output. Enumerating all EMs might not be of great help, but finding a way of grouping them would be very useful. We believe that using a combinatorial framework should facilitate this task. This is work done by the (ex-)PhDs Vincent Lacroix and Vicente Acuña in collaboration with Alberto Marchetti-Spaccamela (University of Rome) and Leen Stougie (Eindhoven University of Technology). A first paper has been submitted to *BioSystems*.

The tools that are available to draw, and to manipulate the drawings of metabolism are usually restricted to metabolic pathways. This limitation becomes problematic when studying processes that span several pathways. In collaboration with Fabien Jourdan (INRA Toulouse), Romain Bourqui and David Auber (LABRI, University of Bordeaux), Vincent Lacroix and Ludovic Cottret participated in the development of a method which enables to draw the entire metabolic network while also taking into account its structuration into pathways [12].

Another way of investigating the metabolic network of an organism consists in determining which metabolites it is able to produce and those it needs to find in its environment. This is a particularly crucial issue in the case of symbiosis, in particular endosymbiosis. Endosymbionts are organisms that live within the tissues of a host, either in the intracellular space or extracellularly. An example of the first is the case of aphids that share a mutually beneficial relationship with the bacterium *Buchnera aphidicola*. The aphids in symbiosis with *Buchnera* rely on the bacteria to synthetise some essential amino acids that it cannot obtain on its own (they are absent in the phloem sap that the aphids eat). On its turn, *Buchnera* lacks, for instance, the genes to make the proteins that are needed for membrane construction, and thus depends on the aphid for shelter and protection against harsh environments and predators. One classical way to answer to the first question (what are the metabolites that an endosymbiont is capable of producing?) is to determine, for each such

metabolite, whether the organism is capable of synthetising the enzymes that catalyse the reactions involved in the reference pathways that have the metabolite as one of their end product. However, the reference is in general to one specific organism, namely *E. coli*. The pathway used for the synthesis of the metabolite may be different in another organism. Furthermore, alternative paths for obtaining the metabolite may exist that cross over different such reference pathways. A so-called "network expansion method" was developed by Handorf *et al.* (*J Mol Evol.*, 2005) that tried to address the problem of which metabolites in general an organism is capable of producing by identifying the subnetwork that may be reconstructed from the reference one starting from the metabolites the organism can obtain from its environment (input metabolites) and using all the reactions known (from genomic analysis) to be available in the organism. In the case of endocytobiotic bacteria, that is of bacteria permanently living in the cells of another organism, this method however can not in general be applied because it is very difficult to determine the set of nutrients the endocytobiont may obtain from its environment. This comes from the fact that endocytobionts can not be cultured outside the host and there is therefore no possibility to control the nutritional medium. This leads then to the second question: which metabolites does the endosymbiont require from its environment (in this case, the host) to produce one or a set of metabolites (called target(s)) that the host in turn may need in order to survive but is unable anymore to produce on its own? The input metabolites are called in this case "precursors". Identifying all such precursors, in fact minimal sets thereof, remains, as far as we know, a largely open question. Ludovic Cottret and Vicente Acuña have started addressing this problem with a first publication accepted [62]. The work will continue in collaboration also with Fábio Martinez and his student, Paulo Vieira Milreu from the Federal University of Campo Grande do Sul, Brazil, in the context of a STIC-AmSud project accepted in 2007 and involving teams from Brazil and Chile.

Anne Morgat, from the Swiss-Prot group at the Swiss Institute for Bioinformatics, has continued her work on the Unipathway project in the framework of the BioSapiens NOE and the UniProt grants. The project aims at providing a standardized representation of metabolic data in the UniProtKB/Swiss-Prot database. These metabolic data are explicitely represented and stored into a relational database (UniPathwayDB). They are hierarchically decomposed into super-pathways, pathways, linear sub-pathways and reactions (steps). The development of UniPathwayDB (using postgreSQL) was performed through a collaboration with Eric Coissac at the Université Joseph Fourier. The database is populated with manually expertised metabolic data (from the Swiss-Prot group) and public data (UniProtKB/Swiss-Prot, complete proteomes (UniProtKB/Swiss-Prot and UniProtKB/TrEMBL), GenomeReview complete genomes, Enzyme). By the end of year 2007, more than 404 pathways were manually curated, representing about 600 distinct biochemical reactions. This covers more than 30 000 Swiss-Prot entries (about 70% of the total number of entries related to metabolism). The database was made available in 2007 through a web site hosted at the INRIA Rhône-Alpes (http://www.grenoble.prabi.fr/obiwarehouse/unipathway). The server, as well as one full time engineer (Sophie Huet) who was hired in october 2006, have been provided by the PRABI (Génopole Rhône-Alpes) to this purpose.

6.2.3. Modelling and simulation of genetic regulatory networks

The group of Hidde de Jong has continued its efforts on the development and application of the qualitative simulation tool GENETIC NETWORK ANALYZER(GNA) (see Software). The mathematical foundations of GNA, based on the PhD thesis of Grégory Batt, have been published in *Automatica* this year [9], while an overview of the tool has appeared in a special issue on systems biology of *Technique et Science Informatique* [8].

In collaboration with experimental biologists in the laboratory of Johannes Geiselmann (Université Joseph Fourier, Grenoble, on leave in HELIX), we have continued to study the nutritional stress responses of the bacterium *Escherichia coli* [48], [47]. Delphine Ropers has further extended her original model with global transcriptional regulators of the bacterium [49] and, together with Valentina Baldazzi, she has worked on a method for the systematic reduction of nonlinear kinetic models to the piecewise-linear models used in GNA. In order to assess the effect of the approximations involved, Monte-Carlo simulation studies involving the efficient, unbiased sampling of the parameter space have been carried out. Two publications on this work are in preparation. The projects EC-MOAN (funded in the framework of the FP6 NEST programme of the European Commission), and MetaGenoReg (funded by the ANR in the framework of the BioSys programme),

will allow us to maintain and extend these modeling activities. In particular, we have started to develop models of bacterial regulatory networks that integrate gene regulation and metabolism, together with Daniel Kahn and members of the COMORE project-team at INRIA Sophia-Antipolis.

The *E. coli* stress response model has given rise to predictions that cannot be tested by currently available experimental data. This has motivated an experimental programme carried out in the laboratory of Johannes Geiselmann, using fluorescent and luminescent gene reporter systems to obtain precise measurements with a high sampling density. Several members of HELIX have contributed to the design of the experiments and the development of appropriate data analysis procedures, work that has been submitted for publication. In parallel, Bruno Besson has continued the implementation of the procedures for the analysis of reporter gene data in WELLREADER (see Software) and improved the graphical user interface of this program. The systematic comparison of the experimental results and the model predictions is currently under way. Other experiments are being carried out in collaboration with Irina Mihalcescu of the Laboratoire de Spectométrie Physique (Université Joseph Fourier, Grenoble).

In addition to HELIX, various other groups are using GNA in their modeling projects. In a number of cases, we have been actively involved in the formulation of the biological problem and the actual application of the tool, for instance in collaborations with Ana Teresa Freitas in Lisbon, Ga?´lle Lelandais in Paris, and Irene Cantone in Naples. The current version 6.0 of GNA has been deposited at the APP and is distributed by the company Genostar. It has also been integrated in the Iogma platform for exploratory genomics developed by the company Genostar (see Software). The European project Cobios (FP6 NEST) has provided additional support to achieve this integration. Bruno Besson has developed an SBML model exchange functionality, while members of HELIX have also contributed to the development by Genostar of a network construction module, a module that will ultimately be coupled to GNA.

As the size and complexity of the genetic regulatory networks under study increase, it becomes more difficult to use GNA. For large and complex models, the state transition graph generated by the program, summarising the qualitative dynamics of the system, may consist of thousands of states and is therefore difficult to analyse by visual inspection alone. In order to cope with this problem, we have followed two approaches .

First of all, instead of generating the entire state transition graph, it is often sufficient to compute the steady states of the system and to analyze the neighbouring states in order to determine the stability of the steady states. Based on the mathematical characterisation of equilibria of piecewise-linear differential equation models and their stability, developed previously, Michel Page and Hidde de Jong have developed an attractor search module for GNA. This module transforms the search of steady states into a SAT problem and exploits existing, efficient SAT solvers to find all steady states of networks of more than thousand genes [57].

A second solution for the upscaling problems consists in the use of model-checking techniques for the automated verification of properties of state transition graphs. In the framework of his PhD thesis, Grégory Batt has initiated this approach in collaboration with Radu Mateescu and his colleagues of the VASY project. This work is now being carried on in several directions. Estelle Dumas has developed a web service for connecting GNA to the model checker CADP, and she has started to implement the operators of a branching-time temporal logic well-adapted for network analysis applications. The theory underlying this temporal logic has been developed by Pedro Monteiro in the framework of his PhD thesis, and he has also started to formulate a high-level specification languages for helping the user to express biological properties the model has to satisfy. Adrien Richard, in collaboration with Gregor Goessler (POP-ART), has proposed a method to complete models with missing information, exploiting the connection between GNA and model-checking tools.

The above-mentioned work has focused on the analysis of models obtained through literature study and human expertise. The PhD of Samuel Drulhe, supervised by Hidde de Jong and Giancarlo Ferrari-Trecate (University of Pavía) within the framework of the European project HYGEIA, takes a different direction. It concerns the development of methods for the identification of piecewise-linear differential equation models of genetic regulatory networks from gene expression data, adapting existing methods for the identification of hybrid systems. A paper describing the threshold reconstruction step has been accepted for a special issue on systems biology of *IEEE Transactions on Automatic Control/IEEE Circuits and Systems I* [18], while a second paper

describing the entire identification chain is in preparation. Shortly, the application of the method to gene reporter data on the *E. coli* nutritional stress response will be undertaken.

7. Contracts and Grants with Industry

7.1. Genostar

Participant: François Rechenmann.

Genostar, an INRIA start-up created in 2004, is a company developing software and solutions for the management and analysis of genomic and post-genomic data. The software has been developed, from 1999 to 2004, by the Genostar consortium (INRIA, Institut Pasteur, and the two biotech companies Genome Express and Hybrigenics) and by the HELIX research team. It includes several modules originally developed by Helix (Section 5.11) as well as GNA, developped by the group of H. de Jong, and the OBIWAREHOUSE database, developed by E. Coissac and A. Morgat. F. Rechenmann is scientific consultant of the company and A. Viari is member of the scientific advisory board.

7.2. Sanofi Pasteur

Participants: Edouard Blondeau, Alain Viari.

In 2007, HELIX started a one-year contractual relation with the Sanofi Pasteur (the vaccine division of the Sanofi Aventis group) located near Lyon. This collaboration is a follow-up of our previous contract on the (re)annotation and comparative analysis of pathogenic bacteria of interest to Sanofi Pasteur. It extends the analysis to expression data provided by Sanofi Pasteur and aims at understanding some global regulation processes involved in pathogenicity.

8. Other Grants and Activities

8.1. National projects

Project name	Caractérisation et modélisation de la “fonction symbiotique” de *Buchnera aphidicola* chez le puceron du pois *Acyrthosiphon pisum*
Coordinator	H. Charles (INSA-INRA Lyon)
HELIX participants	L. Cottret, V. Lacroix, M.-F. Sagot
Type	Projet AgroBi INRA (2006-2008)
Web page	Not yet available

Project name	DUPLIGEN: Conséquences structurales et fonctionnelle des duplications globales de génomes: étude chez le modèle *Paramecium tetraurelia*
Coordinator	J. Cohen (CGM, Gif)
HELIX participants	L. Duret, V. Daubin
Type	ANR Programme blanc (BLAN) NT05-2_41522 (2005-2007)
Web page	Not available for now

Project name	Genomicro
Coordinator	L. Duret
HELIX participants	L. Duret, V. Daubin, G. Marais, S. Mousset, E. Tannier, J. Lobry, V. Lombard
Type	ANR Jeunes chercheurs (2006-2008)
Web page	http://www.agence-nationale-recherche.fr/documents/aap/2005/finances/financeJCBIOLOGIE2005.pdf

Project name	Genomique comparative des recepteurs nucleaires d'hormones
Coordinators	V. Laudet
HELIX participants	G. Perrière
Type	INRA/MRT pour le reseau de recherche et d'innovation technologiques "Genomique des Animaux d'Elevage"
Web page	http://www.ens-lyon.fr/LBMC/laudet/nurebase/nurebase.html

Project name	VICANNE: Modélisation dynamique et simulation des systèmes biologiques
Coordinators	J.-P. Mazat, V. Norris, A. Siegel
HELIX participants	H. de Jong and other HELIX members
Type	ACI IMPBio (2004-2007)
Web page	http://vicanne.inrialpes.fr/

Project name	MetaGenoReg
Coordinator	D. Kahn
HELIX participants	J. Geiselmann, H. de Jong, D. Kahn, D. Ropers
Type	ANR BIOSYS (2006-2009)
Web page	Not available for now

Project name	ECONOMIC
Coordinator	L. Ranjard
HELIX participants	J. Thioulouse
Type	ANR RMQS (2006-2009)
Web page	not available for now

Project name	REGLIS
Coordinator	M.-F. Sagot
HELIX participants	V. Acuña, M. D. V. Braga, L. Canet, L. Cottret, M. Deloger, H. de Jong, C. Gautier, L. Guéguen, V. Lacroix, C. Lemaitre, L. Palmeira, E. Prestat, M.-F. Sagot, P. G. S. Fonseca, E. Tannier, P. Thébault, A. Viari
Type	ANR Blanc (2006-2008)
Web page	http://biomserv.univ-lyon1.fr/baobab/rubrique.php3?id_rubrique=72

Project name	Pathos
Coordinator	S. Nazaret
HELIX participants	J. Thioulouse
Type	ANR RMQS (2007-20010)
Web page	not available for now

8.2. European projects

Project name	EMBRACE. A European Model for Bioinformatics Research and Community Education
Coordinator	G. Cameron
HELIX participants	D. Kahn, A. Laugraud
Type	FP6 Network of excellence LHSG-CT-2004-512092 (2005-2010)
Web page	http://www.embracegrid.info/

Project name	HYGEIA: Hybrid systems for biochemical network modeling and analysis
Coordinators	J. Lygeros, G. Ferrari-Trecate
HELIX participants	B. Besson, S. Druhle, H. de Jong, M. Page, D. Ropers
Type	European Commission, FP6 NEST-4995 (2004-2007)
Web page	http://www.hygeiaweb.gr/home.html

Project name	ChromoNet
Coordinator	M.-F. Sagot
HELIX participants	V. Acuña, L. Cottret, M. Deloger, C. Gautier, V. Lacroix, C. Lemaitre, E. Prestat, M.-F. Sagot, P. G. S. Fonseca, E. Tannier, P. Thébault, A. Viari
European Partner	Ana Pombo, MRC Imperial Collge, London, UK
Type	ARC Inria (2007-2008)
Web page	http://biomserv.univ-lyon1.fr/baobab/rubrique.php3?id_rubrique=82

Project name	ChromoNet
Coordinator	H. Charles (France)
HELIX participants	V. Acuña, C. Gautier, D. Kahn, C. Pizzi, M.-F. Sagot
European Partner	Angela Douglas, York University, UK
Type	ANR BioSys (2007-2010)
Web page	Not yet available

Project name	EC-MOAN: Scalable modeling and analysis techniques to study emergent cell behavior: Understanding the *E. coli* stress response
Coordinators	J. van der Pol
HELIX participants	E. Dumas, J. Geiselmann, H. de Jong, D. Kahn, P. Monteiro, D. Ropers
Type	European Commission, FP6 NEST (2006-2009)
Web page	Not available for now

Project name	COBIOS: Engineering and Control of Biological Systems: A New Way to Tackle Complex Diseases and Biotechnological Innovation
Coordinators	D. di Bernardo
HELIX participants	B. Besson, H. de Jong, M. Page, F. Rechenmann, D. Ropers
Type	European Commission, FP6 NEST (2006-2009)
Web page	http://www.cobios.net

8.3. International projects

Project name	An integrated experimental-computational approach to modeling cellular networks and its application to analyzing the LDB-based transcription complex
Coordinator	R. Sharan (Israel) and M.-F. Sagot (France)
HELIX participants	Various members of HELIX
Type	French-Israel Project (2007-2008)
Web page	Not yet available

Project name	ArcoIris
Coordinator	M.-F. Sagot and Y. Wakabayashi
HELIX participants	V. Acuña, M. D. V. Braga, L. Canet, L. Cottret, M. Deloger, C. Gautier, L. Guéguen, V. Lacroix, C. Lemaitre, L. Palmeira, E. Prestat, M.-F. Sagot, P. G. S. Fonseca, E. Tannier, P. Thébault, A. Viari
Type	Associated Team INRIA-USP (2005-2007)
Web page	http://biomserv.univ-lyon1.fr/~sagot/team/projects/associated_team_usp_helix/associated_team_usp_helix.html

Project name	BemTeVi
Coordinators	C. E. Ferreira and M.-F. Sagot
HELIX participants	V. Acuña, M. D. V. Braga, L. Canet, L. Cottret, M. Deloger, C. Gautier, L. Guéguen, V. Lacroix, C. Lemaitre, L. Palmeira, E. Prestat, M.-F. Sagot, P. G. S. Fonseca, E. Tannier, P. Thébault, A. Viari
Type	Project FAPESP, Brazil (2005-2007)
Web page	http://biomserv.univ-lyon1.fr/~sagot/team/projects/associated_team_usp_helix/associated_team_usp_helix.html

Project name	π-vert
Coordinator	M.-F. Sagot
HELIX participants	V. Acuña, M. D. V. Braga, L. Canet, L. Cottret, M. Deloger, C. Gautier, L. Guéguen, V. Lacroix, C. Lemaitre, L. Palmeira, E. Prestat, M.-F. Sagot, P. G. S. Fonseca, E. Tannier, P. Thébault, A. Viari
Type	ACI Nouvelles Interfaces de Mathematiques (2005-2007)
Web page	http://biomserv.univ-lyon1.fr/baobab/rubrique.php3?id_rubrique=62

Project name	Patagonia
Coordinator	M.-F. Sagot
HELIX participants	V. Acuña, L. Cottret, M. Deloger, C. Gautier, V. Lacroix, C. Lemaitre, E. Prestat, M.-F. Sagot, P. G. S. Fonseca, E. Tannier, P. Thébault, A. Viari
Other Partners	Brazil and Chile
Type	Stic AmSud (2007-2009)
Web page	http://biomserv.univ-lyon1.fr/~sagot/team/projects/amsud/amsud.html

9. Dissemination

9.1. Talks

Sophie Abby

Title	Event and location	Date
Thousands Of Gene Trees To Reconstruct The Tree Of Life	Rencontres Alphy 2007, Montpellier, France	Jan. 2007
Phylogenomics of alphaproteobacteria : Hundreds of gene trees to reconstruct the history of alphaproteobacteria	160th meeting of the Society for General Microbiology, Manchester, England	Jan. 2007

Valentina Baldazzi

Title	Event and location	Date
Qualitative simulation of nutritional stress responses in *E. coli*	Poster MPG-CNRS Workshop on Systems Biology, Berlin (Germany)	Sep. 2007

Estelle Dumas

Title	Event and location	Date
Vérification formelle des propriétés de réseaux de régulation génique : interopérabilité GNA/CADP	Journée du développement logiciel, INRIA, Grenoble	Jun. 2007

Laurent Duret

Title	Event and location	Date
Analyse comparative des génomes de primates: mais où est donc passée la sélection naturelle ?	ACI-IMPBio GDR Bim, Paris,	Oct. 2007
Since when does Xist exist ?	Réunion Alphy, Montpellier	Feb. 2007
Séquençage et analyse du génome de *Paramecium tetraurelia*: 2 x 2 x 2 = 2.	ENS (Lyon)	Feb. 2007
Recombination hotspots: the Achille's heel of our genome.	International SMBE meeting", Halifax (Canada)	June 2007
The impact of whole genome duplications: insights from *Paramecium tetraurelia*	Conférence Jacques Monod "Population and Evolutionary Genomics", Roscoff	May 2007
The impact of whole genome duplications: insights from *Paramecium tetraurelia*	Gregor-Mendel-Network 1st Conference on Bioinformatics, Brno, Czek Republic	Jun. 2007
The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein- Coding Gene	11th Congress of the European Society for Evolutionary Biology, Uppsala, Sweden	Aug. 2007

Samuel Drulhe

Title	Event and location	Date
Identification de modèles affines par morceaux de réseaux de régulation génique à partir des données expérimentales	Journée satellite JOBIM 2007, Marseille	Jul. 2007
Identification procedure for piecewise-affine models of genetic regulatory networks	Poster Eight International Conference on Systems Biology (ICSB'07), Long Beach, CA (USA)	Oct. 2007
Identification procedure for piecewise-affine models of genetic regulatory networks	Poster Workshop Toward Systems Biology, Grenoble	Oct. 2007

Jean-François Gout

Title	Event and location	Date
Translational control of splicing in eukaryotes: Lessons from the *Paramecium genome*	Conférence Jacques Monod "Population and Evolutionary Genomics", Roscoff	May 2007
Translational control of splicing in eukaryotes: Lessons from the *Paramecium genome*	IPG, Lyon	Nov. 2007

Manolo Gouy

Title	Event and location	Date
The ACNUC Sequence Retrieval System	European Bioinformatics Institute, Hinxton, UK	Mar. 2007
Apports de la bioinformatique et de la génomique comparative à l'étude de l'évolution des mitochondries chez les protistes	JOBIM 2007	Jul. 2007
Apports de la bioinformatique et de la génomique comparative à l'étude de l'évolution des mitochondries chez les protistes	Université Bordeaux 2	Oct. 2007

Hidde de Jong

Title	Event and location	Date
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Workshop Mathematical Modeling and Analysis of Biological Networks, Lorentz Center, Leiden, the Netherlands	Feb. 2007
Réseaux de régulation génique	Journée PRABI, Lyon	Feb. 2007
Piecewise-linear differential equations and application to the regulatory network controlling the stress response in *E. coli*	Spring School on Dynamical Modelling of Biological Regulatory Networks, Les Houches (with D. Ropers)	Apr. 2007
Qualitative simulation of the carbon starvation response in *Escherichia coli*	School on Biological Complexity and Modeling, Evry	May 2007
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Colloque IXXI de l'Institut Rhonalpin de Systèmes Complexes, Lyon	Jun. 2007
Qualitative analysis and verification of piecewise affine models of genetic regulatory networks	HYGEIA summer school on Hybrid Systems Biology, Siena, Italy	Jul. 2007
Qualitative simulation of bacterial stress responses	Workshop Toward Systems Biology, Grenoble	Oct. 2007
Qualitative simulation of bacterial stress responses	Séminaire de Biostatistique, Centre Hospitalier Lyon Sud, Lyon	Oct. 2007
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Special Semester on Quantitative Biology, Workshop on Systems Biology, Linz, Austria	Nov. 2007
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Workshop Control Theory for Systems Biology, Groningen, the Netherlands	Nov. 2007
Modeling of bacterial regulatory networks	Colloque Modélisation du Cluster Recherche Environnement, Grenoble	Nov. 2007

Vincent Lacroix

Title	Event and location	Date
Motif search in metabolic networks	Centre for Genomic Regulation, Barcelona	Jan. 2007
Recherche de motifs dans les réseaux métaboliques	Université Bordeaux I	Feb. 2007

Delphine Ropers

Title	Event and location	Date
Piecewise-linear differential equations and application to the regulatory network controlling the stress response in *E. coli*	Spring School on Dynamical Modelling of Biological Regulatory Networks, Les Houches (with D. Ropers)	Apr. 2007
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Technical University UNESC-ID, Lisbonne (Portugal)	May 2007
Development and experimental validation of piecewise affine models of carbon starvation response in *Escherichia coli*	HYGEIA summer school on Hybrid Systems Biology, Siena, Italy	Jul. 2007
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Seminar Orsay	Sep. 2007
Modélisation de réseaux de régulation bactériens	Séminaire IXXI "Vers une science des systèmes complexes", Grenoble	Nov. 2007

Eric Tannier

Title	Event and location	Date
Ancestral Genome Reconstructions : the two parsimonies	Conférence Canadam, Banff, Canada	May 2007
Reconstructions de chromosomes ancestraux : les deux parcimonies	Laboratoire Bibop-Casys, Grenoble	November 2007
Reconstitution de scénarios évolutifs et contraintes de parcimonie	Laboratoire G-Scop, Grenoble	November 2007

9.2. Editorial and reviewing activities

Laurent Duret

Type	Journal or conference
Member Steering Committee	French national conference on Bioinformatics, Jobim
Editorial Board	*Systematic biology*
Co-organisation	Conference Integrative Post-Genomics, Lyon
Co-organisation	Alphy Meeting, Montpellier

Manolo Gouy

Type	Location
Editorial Board	*Molecular Biology and Evolution*

Daniel Kahn

Type	Location
Editorial Board	*Biology Direct*
Faculty Member	Faculty of 1000

Hidde de Jong

Type	Journal or conference
Editorial Board	ACM/IEEE Transactions on Computational Biology and Bioinformatics, Journal of Mathematical Biology
Program Committee	QR 07, IPG 07, RIAMS 07, AIME 07, CMSB 08
Scientific Committee	Working group VICANNE (Modélisation dynamique et simulation des systèmes biologiques)
Coordinator (with S. Robin)	Working group on Transcriptome, protéome, modélisation, inférence et analyse des réseaux biologiques of GDR CNRS 3003 Bioinformatique moléculaire
Co-organiser C. Chaouiya	Modélisation dynamique et simulation des réseaux biologiques : Où en est-on ? Journée satellite JOBIM 2007, Marseille
Co-organiser with B. Fernandez and D. Thieffry	Spring School on Dynamical Modelling of Biological Regulatory Networks, Les Houches
Co-organiser with O. Maler	Workshop Towards Systems Biology, Grenoble

Delphine Ropers

Type	Journal or conference
Member Program Committee	JOBIM 07
Member Organization Committee	SeMoVi (Séminaire de Modélisation du Vivant)

Marie-France Sagot

Type	Journal or conference
Member Steering Committee	European Conference on Computational Biology (ECCB)
Editorial Board	*Journal of Discrete Algorithms*, Elsevier
Editorial Board	*Research in Microbiology*, Elsevier
Editorial Board	*Lecture Notes in BioInformatics*, Springer Verlag
Editorial Board	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*, IEEE and ACM Press
Editorial Board	*BMC Algorithms for Molecular Biology*, BioMed Central
Editorial Board	*BMC Bioinformatics*, BioMed Central
Member Program Committee	RECOMB, ECCB (steering), APBC, BSB (co-chair), IEEE-BIBM, SPIRE, PSW, WABI, RECOMB Satellite Conf. on Comparative Genomics, APBC

Eric Tannier

Type	Location
Co-organiser	Groupe de Travail en Génomique Comparative, Marseille

Jean Thioulouse

Type	Journal or conference
Editorial Board	*Ecology*, Ecological Society of America
Editorial Board	*Journal of Tropical Ecology*, Cambridge Journals Online
Editorial Board	*Journal of Classification*, Springer Verlag

9.3. Administrative activities

L. Duret is member of the scientific committee of the ANR “Biologie des systemes” and of the scientific committee of the “programme fédérateur INRA de biologie intégrative animale, végétale et microbienne (agroBI)”

C. Gautier is deputy director of the IFR of Biology of the UCBL, chair of the section 29 of the CoNRS, and director of the PRABI.

M. Gouy is member of the "Comité National des Universités", section 67 (Ecology & Evolution), of the selection committee of the CNRS ATIP Biodiversity, of the review panel of the German DFG priority program "Deep metazoan phylogeny" and of the Scientific Advisory Board of the Swiss Institute of Bioinformatics.

Hidde de Jong is a member of the International Relations working group of the Conseil d'Orientation Scientifique et Technologique (COST) of INRIA as well as Europe and International Relations correspondent of INRIA Rhône-Alpes. He has participated in the reviewing process of projects for U.S.-Israel Binational Science Foundation, ANR, LIGUE Nationale contre le Cancer, and Taylor & Francis. In addition, he has been on the PhD thesis committee of Céline Kuttler (rapporteur, University of Lille), Sophie Lebre (University of Evry-Val d'Essonne), Chafika Chettaoui (rapporteur, University of Evry-Val d'Essonne), and Mark Musters (Eindhoven Technical University, the Netherlands), and on the HdR committee of Claudine Chaouiya (University of Grenoble).

D. Kahn is member of the scientific committee of the French National Sequencing Centre (Génoscope, Evry) and of the Pasteur Institute Genopole (Paris) as well as a member of the advisory board of the Jena Centre for Bioinformatics.

D. Mouchiroud is director of the LBBE (UCBL, UMR 5558).

G. Perrière is President of the "Société Française de Bioinformatique" (http://www.sfbi.fr).

François Rechenmann is a member of the editorial committee of the Interstices website (http://interstices.info). Interstices offers pedagogic presentations of research themes and activities in the computer science domain.

Marie-France Sagot is a member of the section 44 of the CoNRS, and of the scientific committee of the course "Informatique en Biologie" of the Institut Pasteur in Paris. She was until September 2007 Director of the PhD Program on Computational Biology, Instituto Gulbenkian de Ciencia, Lisbon, Portugal. She participated in the reviewing process of candidates for a research position for the Universities of Princeton (USA), City New York (USA), McGill (Canada), Ben-Gourion (Israel), Tel Aviv (Israel), Uppsala (Sweden), and of projects for the "Fund for Scientific Research" (FWO) (Flanders, Belgium), the Horizon program (Holland), the BBSRC (UK), the Welcome Trust (UK).

Alain Viari is a member of the "Commission de spécialistes" section 65 at Université de Paris 6 and of the scientific advisory board of the MIA (Mathematics and Applied Mathematics) at the INRA. He is co-responsible of the Bioinformatics program of the "Haut Conseil pour la coopération scientifique et technologique entre la France et Israel". Since February 2007, he is the scientific delegate of the INRIA Rhône-Alpes Research Center.

9.4. Teaching

Ten members of the HELIX project, seven in Lyon and three in Grenoble, are professors or assistant professors at, respectively, the University Claude Bernard in Lyon and the Universities Joseph Fourier and Pierre Mendès-France in Grenoble. They therefore have a full teaching service (at least 192 hours).

Various members of the project have developed over the years courses in biometry, bioinformatics and evolutionary biology at all levels of the University as well as at the "École Normale Supérieure" (ENS) of Lyon and the INSA ("Institut National de Sciences Appliquées"). One strong motivation is the need to provide training to biologists having a good background in mathematics and computer science. The group has thus participated in the creation (in 2000) at the INSA of a new module at the Department of Biochemistry called "Bioinformatics and Modelling". This module is open for students entering the third year of the INSA, and covers 1700 hours of courses over 5 semesters. The project contributes also bioinformatic courses at the level of a "Magistère" at the ENS.

As part of the LMD system that was set up at all Universities in France in 2005, members of the project have created a complete interdisciplinary module of the LMD offering training in biology, mathematics and computer science. The module is called “Approches Mathématique et Informatique du Vivant” (AMIV http://miv.univ-lyon1.fr/fr/). It leads to Master’s diplomas in the scientific and medical fields.

A second important educational activity of the project concerns not disconnecting biology from the teaching of mathematics to biologists. To this purpose, various members of the project work in the context of an INCA (‘Initiative Campus Action’) project together with other Universities in the Rhône-Alpes region to maintain a web site (http://nte-serveur.univ-lyon1.fr/nte/mathsv/) dedicated to the teaching of mathematics to biologists using the latest technologies. The main originality of the site rests upon the complementary balance maintained between the methodological and the biological courses. The first covers biostatistics, biomathematics and bioinformatics while the second concern general and population genetics, and molecular evolution.

Finally, members of the project have participated in, or sometimes organised numerous courses or teaching modules including at the international level, such as, for instance, the creation and support of a Master’s course in Ho-Chi-Minh, Vietnam, and the creation and direction of a PhD Program in Computational Biology in Lisbon, Portugal (http://bc.igc.gulbenkian.pt/pdbc/).

Besides the full time professors in HELIX, the following non professor members have contributed the following courses during the year.

Laurent Duret

Subject	Year	Location	Hours
Bioinformatique	3 to 5	INSA Lyon	16
Bioinformatique	3 to 5	ENS Lyon	7
Bioinformatique	3 to 5	UCBL	14

Hidde de Jong

Subject	Year	Location	Hours
Modelling and simulation of genetic regulatory networks	4	INSA, Lyon	14
Modelling and simulation of genetic regulatory networks	5	Instituto Gulbenkian de Cienca, Lisbon, Portugal	7
Modelling and simulation of genetic regulatory networks	5	University Denis Diderot (Paris 7)	3

Manolo Gouy

Subject	Year	Location	Hours
Molecular phylogeny	3	ENS Lyon	6
Comparative genomics	4	ENS Lyon	2
Molecular phylogeny	5	UCBL	10
Molecular phylogeny	4	INSA, Lyon	6
Molecular phylogeny	-	INRA	2

Daniel Kahn

Subject	Year	Location	Hours
Bioinformatique	3 to 5	INSA Lyon	4
Bioinformatique	3 to 5	INA Paris	2

Pedro Monteiro

Subject	Year	Location	Hours
Modelling and simulation of genetic regulatory networks	5	Instituto Gulbenkian de Cienca, Lisbon, Portugal	7

Guy Perrière

Subject	Year	Location	Hours
Molecular phylogeny	5	Univ. Rouen	11
Databases and alignments	3-5	UCBL	8
Molecular phylogeny	5	UCBL	2
Bacterial genomes plasticity	3-5	UCBL	8
Horizontal gene transfers	5	INSA, Lyon	8

Delphine Ropers

Subject	Year	Location	Hours
Modelling and simulation of genetic regulatory networks	5	Instituto Gulbenkian de Cienca, Lisbon, Portugal	7
Modelling and simulation of genetic regulatory networks	4	UJF, Grenoble	15

Marie-France Sagot

Subject	Year	Location	Hours
Algorithmics for biology	4	INSA Lyon	4

Eric Tannier

Subject	Year	Location	Hours
Algorithms for biology	4	INSA, Lyon	16

10. Bibliography

Year Publications

Books and Monographs

[1] J.-G. DUMAS, J.-L. ROCH, E. TANNIER, S. VARETTE. *Théorie des codes: compression, cryptage, correction*, Dunod, 2007.

[2] M.-F. SAGOT, E. TANNIER. *Sorting permutations by reversals*, in press, Encyclopedia of Algorithms, 2007.

Articles in refereed journals and book chapters

[3] S. ABBY, V. DAUBIN. *Comparative genomics and the evolution of prokaryotes.*, in "Trends in Microbiology", vol. 15, n^{o} 3, 2007, p. 135-141.

[4] G. ACHAZ, F. BOYER, E. P. C. ROCHA, A. VIARI, E. COISSAC. *Repseek, a tool to retrieve approximate repeats from large DNA sequences.*, in "Bioinformatics", vol. 23, n^{o} 1, 2007, p. 119–121.

[5] J. ALLALI, M.-F. SAGOT. *A multiple layer model to compare RNA secondary structures*, in "Software Practice and Experience", in press, 2007.

[6] Z. ANDRIANJAKA, R. BALLY, M. LEPAGE, J. THIOULOUSE, G. COMTE, R. DUPONNOIS. *Biological control of Striga hermonthica by Cubitermes termite mound powder amendment in sorghum culture*, in "Applied Soil Ecology", vol. 37, 2007, p. 175-183.

[7] A. AOUACHERIA, V. NAVRATIL, R. LOPEZ-PEREZ, N. C. GUTIERREZ, A. CHURKIN, D. BARASH, D. MOUCHIROUD, C. GAUTIER. *In silico whole-genome screening for cancer-related single-nucleotide polymorphisms located in human mRNA untranslated regions*, in "BMC Genomics", vol. 8, n^{o} 2, 2007, p. 1-13.

[8] G. BATT, R. CASEY, H. DE JONG, J. GEISELMANN, J.-L. GOUZÉ, M. PAGE, D. ROPERS, T. SARI, D. SCHNEIDER. *Analyse qualitative de la dynamique de réseaux de régulation génique par des modèles linéaires par morceaux*, in "Technique et Science Informatiques", vol. 26, n^{o} 1-2, 2007, p. 11-45.

[9] G. BATT, H. DE JONG, M. PAGE, J. GEISELMANN. *Symbolic reachability analysis of genetic regulatory networks using discrete abstractions*, in "Automatica", In press, 2007.

[10] S. BERTRAND, B. THISSE, R. TAVARES, L. SACHS, A. CHAUMOT, P.-L. BARDET, H. ESCRIVA, M. DUFFRAISSE, O. MARCHAND, R. SAFI, C. THISSE, V. LAUDET. *Unexpected Novel Relational Links Uncovered by Extensive Developmental Profiling of Nuclear Receptor Expression*, in "PloS Genetics", vol. 3, n^{o} 11, 2007, p. e188-e188.

[11] E. BILLOIR, A. PÉRY, S. CHARLES. *Integrating the lethal and sublethal effects of toxic compounds into the population dynamics of Daphnia magna : A combinaison of the DEBtox and matrix population models*, in "Ecological Modelling", vol. 203, 2007, p. 207-214.

[12] R. BOURQUI, L. COTTRET, V. LACROIX, D. AUBER, P. MARY, M.-F. SAGOT, F. JOURDAN. *Metabolic network visualization eliminating node redundance and preserving metabolic pathways*, in "BMC Systems Biology", vol. 1, n^{o} 29, 2007, p. 1-19.

[13] K. BUCHET-POYAU, J. COURCHET, H. L. HIR, B. SERAPHIN, J. Y. SCOAZEC, L. DURET, C. DOMON-DELL, J. N. FREUND, M. BILLAUD. *Identification and characterization of human Mex-3 proteins, a novel family of evolutionarily conserved RNA-binding proteins differentially localized to processing bodies*, in "Nucleic Acids Research", vol. 35, n^{o} 4, 2007, p. 1289-1300.

[14] E. CAMBOUROPOULOS, M. CROCHEMORE, C. S. ILIOPOULOS, M. MOHAMED, M.-F. SAGOT. *All maximal-pairs in step-leap representation of melodic sequence*, in "Inf. Sci.", vol. 177, n^{o} 9, 2007, p. 1954-1962.

[15] D. CHARIF, J. LOBRY. *SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis*, in "Structural approaches to sequence evolution: Molecules, networks, populations, New-York, USA", U. BASTOLLA, M. PORTO, H. ROMAN, M. VENDRUSCOLO (editors), L1, Springer Verlag, 2007, p. 207-232.

[16] A. CHAUMOT, S. CHARLES. *Pollution, stochasticity and spatial heterogeneity in the dynamics of an age-structured population of brown trout living in a river network.*, in "Population-level Ecotoxicological Risk Assessment: Case Studies", R. AKCAKAYA (editor), vol. in press, in press, 2007.

[17] Y. DIEKMANN, M.-F. SAGOT, E. TANNIER. *Evolution under Reversals: Parsimony and Conservation of Common Intervals*, in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", vol. 4, n^{o} 2, 2007, p. 301-309.

[18] S. DRULHE, G. FERRARI-TRECATE, H. DE JONG. *Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks*, in "IEEE Transactions on Automatic Control/IEEE Transactions on Circuits and Systems I", In press, 2007.

[19] V. DUCROT, A. PÉRY, R. MONS, S. CHARLES, J. GARRIC. *Dynamic engergy budgets as a basis to model population-level effects of zinc-spiked sediments in the gastropod Valvata piscinalis*, in "Environmental Toxicology and Chemistry", vol. 26, n^{o} 8, 2007, p. 1774-1783.

[20] S. FALL, A. MERCIER, F. BERTOLLA, A. CALTEAU, L. GUÉGUEN, G. PERRIÈRE, T. VOGEL, P. SIMONET. *Horizontal Gene Transfer Regulation in Bacteria as a Spandrel of DNA Repair Mechanisms*, in "PLoS One", L1, vol. 2, n^{o} e1055, 2007, p. 1-11.

[21] R. FRUTOS, A. VIARI, N. VACHIERY, F. BOYER, D. MARTINEZ. *Ehrlichia ruminantium: genomic and evolutionary features*, in "Trends Parasitol.", vol. 23, n^{o} 9, 2007, p. 414-419.

[22] N. GALTIER, L. DURET. *Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution*, in "Trends in Genetics", vol. 23, n^{o} 6, 2007, p. 273-277.

[23] N. GALTIER, L. DURET. *Biased gene conversion and its impact on human genome evolution*, in "Handbook of Human Molecular Evolution", D. N. COOPER, H. KEHRER-SAWATZKI (editors), vol. In press, John Wiley and Sons, 2007.

[24] F. GROGNARD, H. DE JONG, J.-L. GOUZÉ. *Piecewise-linear models of genetic regulatory networks: Theory and example*, in "Biology and Control Theory: Current Challenges, Berlin", I. QUEINNEC, S. TARBOURIECH, G. GARCIA, S. NICULESCU (editors), Lecture Notes in Control and Information Science, vol. 357, Springer-Verlag, 2007, p. 137-159.

[25] O. JAILLON, K. BOUHOUCHE, J. GOUT, J. M. AURY, B. NOEL, M. NOWACKI, V. SERRANO, B. M. PORCEL, B. SEGURENS, A. LE MOUEL, G. LEPERE, V. SCHACHTER, M. BETERMIER, J. COHEN, P. WINCKER, L. SPERLING, L. DURET, E. MEYER. *Translational control of intron splicing in eukaryotes*, in "Nature", vol. In Press, 2007.

[26] I. JEFFERY, S. MADDEN, P. MCGETTIGAN, G. PERRIÈRE, A. CULHANE, D. HIGGINS. *Integrating transcription factor binding site information with gene expression datasets*, in "Bioinformatics", vol. 23, n^{o} 3, 2007, p. 298-305.

[27] C. KEIME, M. SÉMON, D. MOUCHIROUD, L. DURET, O. GANDRILLON. *Unexpected observations after mapping LongSAGE tags to the human genome*, in "BMC Bioinformatics", vol. 8, n^{o} 154, 2007, p. 1-11.

[28] E. KEJNOVSKY, R. HOBZA, Z. KUBAT, A. WIDMER, G. MARAIS, B. VYSKOT. *High intrachromosomal similarity of retrotransposon long terminal repeats: Evidence for homogenization by gene conversion on plant sex chromosomes?*, in "Gene", vol. 390, n^{o} 1-2, 2007, p. 92-97.

[29] M. KISA, A. SANON, J. THIOULOUSE, K. ASSIGBETSE, S. SYLLA, L. DIENG, J. BERTHELIN, Y. PRIN, A. GALIANA, M. LEPAGE, R. DUPONNOIS. *Arbuscular mycorrhizal symbiosis can counterbalance the negative influence of the exotic tree species Eucalyptus camaldulensis on the structure and functioning of soil microbial communities in a sahelian soil*, in "FEMS Microbiology Ecology", vol. 62, 2007, p. 32-44.

[30] F. LASNE, J. THIOULOUSE, L. MARTIN, J. DE CEAURRIZ. *Detection of recombinant human erythropoietin in urine for doping analysis: interpretation of isoelectric profiles by discriminant analysis*, in "Electrophoresis", vol. 28, 2007, p. 1875-1881.

[31] C. LEMAITRE, M.-F. SAGOT. *A Small Trip in the Untranquil World of Genomes*, in "Theoretical Computer Science", in press, 2007.

[32] G. MARAIS. *Sex chromosomes and mitochondrial DNA polymorphism in birds: The Hill Robertson effects extend from nucleus to mitochondria*, in "Heredity", vol. 99, n^{o} 4, 2007, p. 357-358.

[33] C. MELODELIMA, C. GAUTIER, D. PIAU. *A Markovian approach for the prediction of mouse isochore*, in "Journal of Mathematical Biology", vol. 55, n^{o} 3, 2007, p. 353-364.

[34] C. MELODELIMA, L. GUÉGUEN, C. GAUTIER, D. PIAU. *A Markovian approach for the analysis of the gene structure*, in "International Journal of Foundations of Computer Science", vol. In press, 2007.

[35] C. MELODELIMA, L. GUÉGUEN, D. PIAU, C. GAUTIER. *Segmentation of the chimpanzee genome using a HMM model*, in "Lecture notes in Bioinformatics", LNCS, vol. 4414, 2007, p. 251-262.

[36] N. J. MULDER, R. APWEILER, T. ATTWOOD, A. BAIROCH, A. BATEMAN, D. BINNS, P. BORK, V. BUILLARD, L. CERUTTI, R. COPLEY, E. COURCELLE, U. DAS, L. DAUGHERTY, M. DIBLEY, R. FINN, W. FLEISCHMANN, J. GOUGH, D. HAFT, N. HULO, S. HUNTER, D. KAHN, A. KANAPIN, A. KEJARIWAL, A. LABARGA, P. S. LANGENDIJK-GENEVAUX, D. LONSDALE, R. LOPEZ, I. LETUNIC, M. MADERA, J. MASLEN, C. MCANULLA, J. MCDOWALL, J. MISTRY, A. MITCHELL, A. N. NIKOLSKAYA, S. ORCHARD, C. ORENGO, R. PETRYSZAK, J. D. SELENGUT, C. J. SIGRIST, P. D. THOMAS, F. VALENTIN, D. WILSON, C. H. WU, C. YEATS. *New developments in the InterPro database*, in "Nucleic Acids Research", vol. 35, n^{o} Database issue, 2007, p. D224-D228.

[37] M. MUSTERS, H. DE JONG, P. VAN DEN BOSCH, N. VAN RIEL. *Qualitative analysis of nonlinear biochemical networks with piecewise-affine functions*, in "Hybrid Systems: Computation and Control (HSCC 2007), Berlin", A. BEMPORAD, A. BICCHI, G. BUTTAZZO (editors), Lecture Notes in Computer Science, vol. 4416, Springer-Verlag, 2007, p. 727-730.

[38] A. NECSULEA, J. LOBRY. *A New Method for Assessing the Effect of Replication on DNA Base Composition Asymmetry*, in "Molecular Biology and Evolution", L1, vol. 24, n^{o} 10, 2007, p. 2169-2179.

[39] P. NORMAND, P. LAPIERRE, L. TISA, J. GOGARTEN, N. ALLOISIO, E. BAGNAROL, C. BASSI, A. BERRY, D. BICKHART, N. CHOISNE, A. COULOUX, B. COURNOYER, S. CRUVEILLER, V. DAUBIN, N. DEMANGE, M. FRANCINO, E. GOLTSMAN, Y. HUANG, O. KOPP, L. LABARRE, A. LAPIDUS, C. LAVIRE, J. MARECHAL, M. MARTINEZ, J. MASTRONUNZIO, B. MULLIN, J. NIEMANN, P. PUJIC, T. RAWNSLEY, Z. ROUY, C. SCHENOWITZ, A. SELLSTEDT, F. TAVARES, J. TOMKINS, D. VALLENET, C. VALVERDE, L. WALL, Y. WANG, C. MEDIGUE, D. BENSON. *Genome characteristics of facultatively symbiotic Frankia sp.strains reflect host range and host plant biogeography*, in "Genome Research", 2007, p. 7-15.

[40] L. OUAHMANE, M. HAFIDI, J. THIOULOUSE, M. DUCOUSSO, M. KISA, Y. PRIN, A. GALIANA, A. BOUMEZZOUGH, R. DUPONNOIS. *Improvement of Cupressus atlantica Gaussen growth by inoculation with native arbuscular mycorrhizal fungi*, in "Journal of Applied Ecology", vol. 103, 2007, p. 683-690.

[41] L. OUAHMANE, J. THIOULOUSE, M. HAFIDI, Y. PRIN, A. GALIANA, C. PLENCHETTE, M. KISA, R. DUPONNOIS. *Soil functional diversity and P solubilization from rock phosphate after inoculation with native or allochtonous arbuscular mycorrhizal fungi*, in "Forest Ecology and Management", vol. 241, 2007, p. 200-208.

[42] P. PETERLONGO, J. ALLALI, M.-F. SAGOT. *The Gapped-Factor Tree*, in "International Journal of Foundations of Computer Science", in press, 2007.

[43] P. PETERLONGO, N. PISANTI, F. BOYER, A. P. DO LAGO, M.-F. SAGOT. *Lossless filter for multiple repetitions*, in "Journal of Discrete Algorithms", in press, 2007.

[44] N. RAMANANKIERANA, M. DUCOUSSO, N. RAKOTOARIMANGA, Y. PRIN, J. THIOULOUSE, E. RANDRIANJOHANY, L. RAMAROSON, M. KISA, A. GALIANA, R. DUPONNOIS. *Arbuscular mycorrhizas and ectomycorrhizas of Uapaca bojeri L. (Euphorbiaceae): sporophore diversity, patterns of root colonization, and effects on seedling growth and soil microbial catabolic diversity*, in "Mycorrhiza", vol. 17, n^{o} 3, 2007, p. 195-208.

[45] C. REZVOY, D. CHARIF, L. GUÉGUEN, G. MARAIS. *MareyMap: an R-based tool with graphical interface for estimating recombination rates*, in "Bioinformatics", vol. 23, n^{o} 16, 2007, p. 2188-2189.

[46] E. M. RODRIGUES, M.-F. SAGOT, Y. WAKABAYASHI. *The Maximum Agreement Forest Problem: approximation algorithms and computational experiments*, in "Theoretical Computer Science", vol. 374, 2007, p. 91-110.

[47] D. ROPERS, H. DE JONG, J. GEISELMANN. *Modélisation de la réponse au stress nutritionnel de la bactérie Escherchia coli*, in "Biofutur", vol. 275, 2007, p. 36-39.

[48] D. ROPERS, H. DE JONG, J. GEISELMANN. *Mathematical modeling of genetic regulatory networks: Stress responses in Escherichia coli*, in "Systems Biology and Synthetic Biology", P. Fu, M. Letterich, S. Panke, New Jersey, 2007.

[49] D. ROPERS, H. DE JONG, J.-L. GOUZÉ, M. PAGE, D. SCHNEIDER, J. GEISELMANN. *Piecewise-linear models of genetic regulatory networks: Analysis of the carbon starvation response in Escherichia coli*, in "Mathematical Modeling of Biological Systems", vol. 1, A. Deutsch and L. Brusch and H. Byrne and G. de Vries and H.-P. Herzel, Boston, 2007, p. 85-98.

[50] P. TABERLET, E. COISSAC, F. POMPANON, L. GIELLY, C. MIQUEL, A. VALENTINI, T. VERMAT, G. CORTHIER, C. BROCHMANN, E. WILLERSLEV. *Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding*, in "Nucleic Acids Res.", vol. 35, n^{o} 3, 2007, e14.

[51] E. TANNIER, A. BERGERON, M.-F. SAGOT. *Advances on sorting by reversals*, in "Discrete Applied Mathematics", vol. 155, n^{o} 6-7, 2007, p. 881-888.

[52] J. THIOULOUSE, S. DRAY. *Interactive Multivariate Data Analysis in R with the ade4 and ade4TkGUI Packages*, in "Journal of Statistical Software", vol. 22, n^{o} 5, 2007, p. 1-14.

[53] C. TRUNTZER, C. MERCIER, J. ESTÈVE, C. GAUTIER, P. ROY. *Importance of data structure in comparing two dimension reduction methods for classification of microarray gene expression data*, in "BMC Bioinformatics", L1, vol. 8, 2007, p. 1-12.

[54] N. VACHIERY, G. MAGANGA, T. LEFRANCOIS, Y. KANDASSAMY, F. STACHURSKI, H. ADAKAL, C. FERRAZ, A. MORGAT, A. BENSAID, E. COISSAC, F. BOYER, J. DEMAILLE, A. VIARI, D. MARTINEZ, R. FRUTOS. *Differential strain-specific diagnosis of the heartwater agent: Ehrlichia ruminantium*, in "Infect Genet Evol.", in press, 2007.

[55] E. VAUTRIN, S. CHARLES, S. GENIEYS, F. VAVRE. *Evolution and invasion dynamics of multiple infections with Wolbachia investigated using matrix based models.*, in "Journal of Theoretical Biology", vol. 245, n^{o} 2, 2007, p. 197-209.

[56] S. VAUTRIN, S. CHARLES, F. VAVRE. *Do vertically-transmitted symbionts co-existing in a single host compete or cooperate?*, in "Journal of Evolutionary Biology", vol. in press, 2007.

[57] H. DE JONG, M. PAGE. *Search for steady states of piecewise-linear differential equation models of genetic regulatory networks*, in "ACM/IEEE Transactions on Computational Biology and Bioinformatics", In press, 2007.

Publications in Conferences and Workshops

[58] S. S. ADI, M. BRAGA, C. FERNANDES, C. FERREIRA, F. MARTINEZ, M.-F. SAGOT, M. A. STEFANES, C. TJANDRAATMADJA, Y. WAKABAYASHI. *Repetition-free LCS with few reversals*, in "IV Latin-American Algorithms, Graphs and Optimization Symposium (LAGOS'07)", in press, Electronic Notes in Discrete Mathematics, 2007.

[59] J. AGUIRRE CERVANTES, F. RECHENMANN, J. DOMINGUEZ SANCHEZ. *A Computer Environment for Learning about Reactive Systems Processes: a Multi-agent Based Simulation Approach*, in "MICAI 2007, Aguascalientes, Mexico November 2007", 2007.

[60] J. AGUIRRE CERVANTES, F. RECHENMANN, J. DOMINGUEZ SANCHEZ. *Learning About Molecular Biology in a Multi-Agent Based Simulation Environment*, in "Learning With Games 2007, Sophia-Antipolis, France September 2007", 2007.

[61] M. BRAGA, M.-F. SAGOT, C. SCORNAVACCA, E. TANNIER. *The solution space of sorting by reversals*, in "International Symposium on Bioinformatics Research and Applications", vol. 4463, Lecture Notes in Bioinformatics, 2007, p. 293-304.

[62] L. COTTRET, V. ACUÑA, H. CHARLES, M.-F. SAGOT. *Recherche de précurseurs dans un réseau métabolique*, in "RIAMS'07", J.-P. COMET, F. QUESSETTE, S. VIAL (editors), 2007.

[63] A. NECSULEA, J. LOBRY. *A New Method for Assessing the Effect of Replication on DNA Base Composition Asymmetry*, in "JOBIM'07", C. BRUN, G. DIDIER (editors), L1, vol. 7, 2007, p. 379-381.

Internal Reports

[64] G. BATT, D. ROPERS, H. DE JONG, M. PAGE, J. GEISELMANN. *Symbolic reachability analysis of genetic regulatory networks using qualitative abstractions*, Technical report, n^{o} RR-6136, INRIA, 2007, http://hal.inria.fr/inria-00133991.