INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team HELIX

Informatics and genomics

Rhône-Alpes

THEME BIO

Table of contents

1. Team **1**
2. Overall Objectives **2**
2.1. Overall Objectives 2
3. Scientific Foundations **4**
3.1. Computational analysis of the evolution of species and gene families 4
3.2. Modelling and analysis of the spatial organisation and dynamics of genomes 4
3.3. Motif search and inference 5
3.4. Computational proteomics and transcriptomics 6
3.5. Modelling and analysis of metabolism: molecular components, regulation, and pathways 6
3.6. Modelling and simulation of genetic regulatory networks 7
3.7. Bioanalysis and cross-sectional activities 8
4. Software **8**
4.1. AROM 8
4.2. Box 8
4.3. C3P 8
4.4. FactorTree 9
4.5. FamFetch 9
4.6. GenoExpertBacteria (GEB) 9
4.7. GenoStar 9
4.8. GeM 9
4.9. Genetic Network Analyzer (GNA) 10
4.10. Herbs 10
4.11. Hogenom and Hovergen 10
4.12. Hoppsigen 10
4.13. Identitag 10
4.14. ISee 11
4.15. LalnView 11
4.16. Mentalign 11
4.17. MicrOBI 11
4.18. Migal 12
4.19. Oriloc 12
4.20. PEPLINE 12
4.21. PhyloJava 12
4.22. Satellites 12
4.23. seqinR 12
4.24. Smile and Riso 13
4.25. Utopia 13
4.26. Other software developed in HELIX 13
5. New Results **13**
5.1. Computational analysis of the evolution of species and gene families 13
5.2. Modelling and analysis of the spatial organisation and dynamics of genomes 14
5.3. Motif search and inference 15
5.4. Computational proteomics and transcriptomics 15
5.5. Modelling and analysis of metabolism: molecular components, regulation, and pathways 15
5.6. Modelling and simulation of genetic regulatory networks 16
5.7. Bioanalysis and cross-sectional activities 17
6. Contracts and Grants with Industry **17**

6.1. Aventis-Pasteur 17
6.2. Genostar 17
6.3. Sanofi 18
7. Other Grants and Activities 18
7.1. National projects 18
7.2. Projects funded by international organisms or including international teams 19
8. Dissemination 20
8.1. Talks 20
8.2. Organisation of conferences, workshops and meetings 22
8.3. Editorial and reviewing activities 23
8.4. Administrative activities 23
8.5. Teaching activities 24
9. Bibliography 26

1. Team

The HELIX project is located in Montbonnot (Grenoble) and on the Campus of La Doua in Villeurbanne (Lyon). The members of the group in Grenoble, headed by Alain Viari, work in the Rhône-Alpes research unit of INRIA. The members in Lyon are part of two groups within the "Laboratoire de biométrie et de biologie évolutive" (CNRS/Université Claude Bernard de Lyon, UMR 5558), directed by Christian Gautier: the group "Bioinformatics and Evolutionary Genomics" headed by Manolo Gouy, and the group BAOBAB headed by Marie-France Sagot. The SwissProt group, headed by Amos Bairoch within the SIB (Swiss Institute of Bioinformatics) in Geneva, is associated with the HELIX project. This association is named Sibelius. Since January 2005, Helix is also associated with the Departamento de Ciência da Computação within Instituto de Matemática e Estatística, Universidade de São Paulo (USP), Brazil. The association is coordinated, on the Brazilian side, by Prof. Yoshiko Wakabayashi and is named ArcoIris.

Head of project

Alain Viari [Research director, INRIA]

Administrative assistant

Françoise de Coninck [Senior secretary, part-time in the project]

Permanent researchers and professors

Sandrine Charles [Associate Professor, Université Claude Bernard]
Vincent Daubin [Research associate, CNRS]
Laurent Duret [Research director, CNRS]
Christian Gautier [Professor, Université Claude Bernard]
Philippe Genoud [Associate professor, Université Joseph Fourier]
Manolo Gouy [Research director, CNRS]
Laurent Guéguen [Associate professor, Université Claude Bernard]
Hidde de Jong [Research director, INRIA]
Daniel Kahn [Research Director, INRA (detached from INRA to INRIA)]
Jean Lobry [Associate professor, Université Claude Bernard]
Gabriel Marais [Research associate, CNRS]
Dominique Mouchiroud [Professor, Université Claude Bernard]
Sylvain Mousset [Associate professor, Université Claude Bernard]
Michel Page [Associate professor, Université Pierre Mendès-France]
Guy Perrière [Research director, CNRS]
François Rechenmann [Research director, INRIA]
Marie-France Sagot [Research director, INRIA]
Eric Tannier [Research associate, INRIA]
Raquel Tavares [Associate Professor, Université Claude Bernard, since September 2005]
Jean Thioulouse [Research director, CNRS]
Alain Viari [Research director, INRIA]
Danielle Ziébelin [Associate professor, Université Joseph Fourier]

Permanent technical staff

Simon Penel [Technical staff, CNRS]
Stéphane Delmotte [Technical staff, CNRS]
Bruno Spataro [Technical staff, CNRS]

Software and system engineers

Annick Chamontin [CNAM, INRIA, since February 2005]
Ludovic Cottret [Project technical staff, Université Claude Bernard]
Sylvain Duhoo [CNAM, INRIA, since September 2005]
Agnès Iltis [Project technical staff, INRIA, until September 2005]
Règis Monte [Project technical staff, INRIA, until October 2005]

Emmanuel Prestat [Project technical staff, PRABI, Université Claude Bernard]

External members

Eric Coissac [Associate professor, Université Paris 6, in delegation at the INRIA until September 2005]
Corinne Lachaize [Project technical staff, SIB]
Anne Morgat [Project technical staff, Fondation Rhône-Alpes Futur, SIB]

Post-doctoral fellows

Abdel Aouacheria [Association pour la Recherche sur le Cancer, until December 2005]
Frédéric Boyer [INRIA, Aventis Pasteur]
Christelle Melo de Lima [ATER UCBL, from december 2005]
Delphine Ropers [INRIA, CEE]
Patricia Thebault [INRIA, ARC]

PhD students

Anne-Muriel Arigon [scholarship Ministère de la Recherche, supervisors: Manolo Gouy and Guy Perrière]
Grégory Batt [scholarship ENS Lyon, supervisors: Hidde de Jong, François Rechenmann, until September 2005]
Elise Billoir [scholarship Ministère de la Recherche, supervisor : Sandrine Charles, since september 2005]
Bastien Boussau [BDI scholarship, supervisor : Manolo Gouy, since september 2005]
Alexandra Calteau [scholarship Ministère de la Recherche, supervisors: Guy Perrière, Manolo Gouy]
Marilia Dias Vieira Braga [AlBan scholarship, supervisors: Marie-France Sagot and Eric Tannier, since September 2005]
Stéphane Descorps-Declère [CIFRE convention, GENOME express, supervisors: Alain Viari, Pierre Netter, Université Paris 6, until September 2005]
Samuel Drulhe [scholarship ENS Cachan, supervisors: Hidde de Jong, François Rechenmann]
Paulo Gustavo da Fonseca [supervisors : Katia Guimarães (Univ. Fédérale de Pernambuco, Brésil) and Marie-France Sagot, since september 2005]
Adel Khelifi [scholarship Ministère de la Recherche, supervisor: Dominique Mouchiroud, until June 2005]
Vincent Lacroix [scholarship BDI, CNRS, supervisor: Marie-France Sagot]
Claire Lemaitre [scholarship Ministère de la Recherche, supervisors: Marie-France Sagot and Christian Gautier, since september 2005]
Christelle Melo de Lima [scholarship Ministère de la Recherche, supervisors: Didier Piau, François Rechenmann, until November 2005]
Julien Meunier [scholarship ENS Lyon, supervisor: Laurent Duret, until June 2005]
Vincent Navratil [scholarship INRA, supervisor: Christian Gautier, Denis Milan, INRA, until December 2005]
Anamaria Necsulea [scholarship Ministère de la Recherche, supervisor: Jean Lobry, since september 2005]
Mi Phi Long Nguyen [assistant professor at HoChiMinh, supervisor: Christian Gautier]
Leonor Palmeira [scholarship Ministère de la Recherche, supervisors: Laurent Guéguen, Jean Lobry]
Marie Semon [scholarship Ministère de la Recherche, supervisor: Laurent Duret, until September 2005]

2. Overall Objectives

2.1. Overall Objectives

More than three 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 recognized to play a key role in the process of turning these 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 these to bioinformatics objects, such as DNA and protein sequences, but also phylogenic trees, as well as graphs which formalize 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 but requires a real "symbiosis" between the two cultures.

Seven main research areas organize the activities of the project:

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;
4. Computational proteomics and transcriptomics;
5. Modelling of metabolism: molecular components, regulation, and pathways;
6. Modelling and simulation of genetic regulatory networks;
7. Bioanalysis and cross-sectional activities.

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

The HELIX project has the particularity that it bridges two geographical locations and two different bioinformatics 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 de biologie évolutive” in Lyon and of the associated group from the Swiss Institute of Bioinformatics in Geneva contribute to the research activities, through co-supervision of PhD theses 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 4.7).
- PRABI is a Web service resource 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 is a member of HELIX.

3. Scientific Foundations

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

Participants: Anne-Muriel Arigon, Bastien Boussau, Alexandra Calteau, Laurent Duret, Christian Gautier, Manolo Gouy [Correspondent], Laurent Guéguen, Jean Lobry, Julien Meunier, Dominique Mouchiroud, Annamaria Necsulea, Leonor Palmeira, Guy Perrière, Mi Phi Long, Marie-France Sagot, Marie Semon, Eric Tannier.

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 of fixing the mutation, where the frequency of occurrence of a tiny fraction of the errors increases in the population until these 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 within the HELIX project.

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 becomes a computational and mathematical problem that requires the expertise of almost all of the components of the HELIX project.

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

Participants: Eric Coissac, Marilia Dias Vieira Braga, Laurent Duret, Christian Gautier [Correspondent], Laurent Guéguen, Adel Khelifi, Claire Lemaitre, Jean Lobry, Julien Meunier, Anne Morgat, Dominique Mouchiroud, Guy Perrière, Marie-France Sagot [Correspondent], Marie Semon, Bruno Spataro, Eric Tannier, Raquel Tavares, Alain Viari.

Genomic sequences are characterized by strong biological and statistical heterogeneities in their composition and organization. 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 interpretable 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 that is, 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 characterizations require expertise both in biology and in computer science.

3.3. Motif search and inference

Participants: Christian Gautier, Laurent Gueguen, Paulo Gustavo da Fonseca, Vincent Lacroix, Christelle Melo de Lima, Leonor Palmeira, Guy Perrière, Marie-France Sagot [Correspondent], Alain Viari.

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 (see Sections 3.5 and 3.6 for biochemical and 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*, therefore represents an important area of research in computational biology. Motif identification consists of two main parts: *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.* *Relational identification*, the goal of which consists in finding relations existing between 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.

3.4. Computational proteomics and transcriptomics

Participants: Laurent Duret, Christian Gautier, Paulo Gustavo da Fonseca, Dominique Mouchiroud, Guy Perrière [Correspondent for transcriptomics], Simon Penel, Emmanuel Prestat, Raquel Tavares, Jean Thioulouse, Alain Viari [Correspondent for proteomics].

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 fully automated protein identification from raw MS/MS data. This has motivated a collaboration between HELIX and the "Laboratoire de Chimie des Protéines" (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 the genome, the proteome and the 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 on 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 and DNA microarray data. *Serial Analysis of Gene expression (SAGE)* is a method of large-scale gene expression analysis that has the potential to generate the full list of mRNAs present within a cell population at a given time as well as to estimate their frequency. An essential step in the analysis of a SAGE library is the unambiguous assignment of each 14 bp tag to the transcript from which it was derived. Concerning *DNA microarray* data, members of the HELIX project, in collaboration with the Conway Institute in Dublin, have worked on finding attractive and computationally-efficient methods for the discrimination and classification of microarray gene expression profiles, as well as on the integrated analysis of multiple gene expression data sets derived from the same biological material.

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

Participants: Frédéric Boyer, Ludovic Cottret, Vincent Lacroix, Anne Morgat, Marie-France Sagot [Correspondent], Patricia Thebault, Alain Viari [Correspondent], Erik Wessel.

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, the chemical transformations occuring 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 (such as AROM), 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 setup a database of metabolic pathways (UniPathway).

Another aspect of the activity of HELIX in this field concerns the design of algorithms to reconstruct metabolic pathways. By contrast to homology-based approaches, we try to tackle this problem 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. This problem can be formally stated as finding a composition of partial injections which maximizes the image size, a problem of demonstrated high complexity. In the same vein, several related problems (such as comparing biochemical networks to genomic organization) 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 formalization. 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 "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 Sao Paulo, Brazil and the statistical aspects are studied in collaboration with Sophie Schbath (INRA, Jouy-en-Josas) and Stéphane Robin (InaPG, Paris).

3.6. Modelling and simulation of genetic regulatory networks

Participants: Grégory Batt, Samuel Drulhe, Hidde de Jong [Correspondent], Michel Page, Delphine Ropers.

All the aforementionned research topics concern, in some way, "static" data (i.e. the description of the cellular actors, together with their interactions). Except, for evolution (but on a very different time-scale) time is not taken explicitely 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 handled within 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) (Section 4.9). 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) and the École Normale Supérieure (Paris). Recently, the scope of the reseach has been enlarged to the validation and identification of genetic regulatory networks.

3.7. Bioanalysis and cross-sectional activities

Participants: Abdel Aouacheria, Frédéric Boyer, Eric Coissac, Stéphane Descorps-Declère, Christian Gautier [Correspondent], Dominique Mouchiroud, Anne Morgat, Vincent Navratil, Simon Penel, François Rechenmann, Alain Viari [Correspondent].

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 enterprises, with whom HELIX collaborates. These collaborations often concern applications in medicine or agriculture.

4. Software

4.1. AROM

Participants: Philippe Genoud, Danielle Ziebelin [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

4.2. Box

Participants: Anne Morgat, Alain Viari [Correspondent].

The primary objective of BOX, acronym for *Bio Oriel XML-schema*, is to provide an open core of well-defined UML and XML specifications for the dissemination of genomic data. The first release of this core library deals with metabolic data and genome annotation data. It is composed of model specifications, XML-schema implementations, and associated documentation (BOXml). The second release incorporates a Java toolkit (BOXtk) for format transformations (XSLT) and data handling. The final release adds a third component (BOXweb) allowing access to BOXtk through web services (SOAP). BOX was developed with Antoine Brun. For more information, see : http://www-helix.inrialpes.fr/article397.html.

4.3. C3P

Participants: Frédéric Boyer, 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 summarizing the biological data considered. For more information, see : http://www.inrialpes.fr/helix/people/viari/cccpart

4.4. FactorTree

Participant: Marie-France Sagot [Correspondent].

FACTORTREE is an algorithm that builds an index for a text called a k-depth factor tree. This is a tree of all the factors of length at most k of a text. The k-depth factor tree allows to save space and is appropriate when the tree is then used for inferring motifs whose length is no greater than k. The economy in space varies depending on the type of text considered. For k between 10 and 20, the economy ranges from 10-20% for biological sequences to more than 40-50% for texts in a formal language or some texts in natural language. FACTORTREE was developed by Julien Allali during his PhD. The code for FACTORTREE (in C++) is freely-available to academics and non-profit organisations upon request to Julien Allali (allali@univ-mlv.fr) or Marie-France Sagot (Marie-France.Sagot@inria.fr).

4.5. FamFetch

Participants: Jean-François Dufayard, 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)

4.6. GenoExpertBacteria (GEB)

Participants: Frédéric Boyer, Anne Morgat [Correspondent], Alain Viari, Erik Wessel.

GENOEXPERTBACTERIA is an environment for the analysis of genomic and metabolic data in bacteria. It integrates a knowledge base and a graphical user interface facilitating the exploration and analysis of the available data. GEB has now been integrated (under the name "PathwayExplorer") into the Iogma bioinformatics environment developed and distributed by Genostar company (see 6.2). For more information, see : http://www-helix.inrialpes.fr/article141.html

4.7. GenoStar

Participants: Pierre-Emmanuel Ciron, Gilles Faucherand, Agnès Iltis, Anne Morgat, François Rechenmann [Correspondent], Alain Viari [Correspondent].

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 organized into strategies, the execution of which requires complex analysis tasks.

The GENOSTAR platform has now been transferred to the Genostar company. Its three modules (GenoAnnot, GenoLink and GenoBool) have been integrated in the Iogma bioinformatics environment (see 6.2), which is based on the same framework. For more information, see (http://www-helix.inrialpes.fr/article121.html)

4.8. GeM

Participants: Gisèle Bronner [now at Clermont-Ferrand University], Christian Gautier [Correspondent], Vincent Navratil, 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/).

4.9. Genetic Network Analyzer (GNA)

Participants: Grégory Batt, Samuel Drulhe, Hidde de Jong [Correspondent], 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. In addition to HELIX, various other groups are using GNA in their modeling projects. At the time of writing, about 100 copies of the current version 5.5 of the program have been distributed. GNA has now been integrated into the Iogma bioinformatics environment developed and distributed by Genostar company (see 6.2). For more information, see : http://www-helix.inrialpes.fr/article122.html.

4.10. Herbs

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

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.

4.11. Hogenom and Hovergen

Participants: Jean-François Dufayard, 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)

4.12. Hoppsigen

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.

4.13. Identitag

Participants: Laurent Duret, Céline Keime [CGMC, LBBE], 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/.

4.14. ISee

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

The aim of ISEE (In Silico biology e-learning environment) 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 CCSTI (Centre de Culture Scientifique Technique et Industrielle) in Grenoble, which uses ISEE for its "Ecole de l'ADN". ISee modules can be packaged in applets, examples (sequence alignment with NWS algorithm, DNA-walk, search of coding sequences) are available from web site http://interstices.info/ in a course dedicated to bioinformatics. For more information, see : http://www-helix.inrialpes.fr/article124.html.

4.15. LalnView

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

LALNVIEW is a graphical program for visualizing 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.

4.16. Mentalign

Participants: Jean-François Dufayard, Manolo Gouy [Correspondent], Guy Perrière.

MENTALIGN is an incremental algorithm for performing a multiple alignment and building the phylogenetic tree of members of a same gene family. When a new sequence is added to a pre-aligned family, the alignment and the tree are modified rather than fully recomputed. For more information please contact : mgouy@biomserv.univ-lyon1.fr

4.17. MicrOBI

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

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 and uses trigger mechanisms for automatic updates and data consistency checks.

It acts as a data source for GEB (Section 4.6), but can also be used as a stand-alone database. For more information please contact : Eric.Coissac@inrialpes.fr

4.18. Migal

Participant: Marie-France Sagot [Correspondent].

MIGAL is an algorithm that compares two RNA structures. MIGAL was developed and is maintained by Julien Allali during his PhD at the University of Marne-la-Vallée. The prototypal code for MIGAL (in C++) is freely available to academics and non-profit organisations upon request to Julien Allali (allali@univ-mlv.fr) or Marie-France Sagot (Marie-France.Sagot@inria.fr).

4.19. Oriloc

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 sc Glimmer2 outputs to discriminate between codon positions. For more information see : http://pbil.univ-lyon1.fr/software/oriloc.html

4.20. PEPLINE

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 PEPMAP. TAGGOR generates so-called PSTs (Peptide Sequence Tags) from MS/MS data, while PEPMAP 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 with Estelle Nugues and Erwan Reguer thru a collaboration with the Laboratoire de Chimie des Protéines headed by J. Garin at CEA Grenoble. For more information, see : http://www-helix.inrialpes.fr/article228.html.

4.21. PhyloJava

Participants: Laurent Duret, Manolo Gouy [Correspondent], Timothée Sylvestre.

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.

4.22. Satellites

Participant: Marie-France Sagot [Correspondent].

SATELLITES is an exact algorithm for detecting tandem arrays (that is, series of contiguous repeats) in DNA sequences. A prototypal version for proteins is also available. The repeats are approximate: a maximum number of differences (substitutions, insertions and deletions) is thus allowed. This number is specified by the user. The code (in C) can be freely obtained by academics and non-profit research organisations by sending an email to Marie-France.Sagot@inrialpes.fr.

4.23. seqinR

Participant: Jean Lobry [Correspondent].

SEQINR is a package of functions for the exploratory data analysis and data visualisation of biological sequence (DNA and protein) data. The package also includes utilities for sequence data management under the ACNUC system. Moreover, an integrated environment for sequence multivariate analysis will soon be available as an R package. SEQINR was developed by Delphine Charif during her DESS. For more information, see : http://pbil.univ-lyon1.fr/software/SeqinR/seqinr_home.php.

4.24. Smile and Riso

Participant: Marie-France Sagot [Correspondent].

SMILE is a motif inference algorithm that takes 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. SMILE is currently being improved an extended into a new algorithm, called RISO, by Alexandra Carvalho from the Instituto Superior Técnico (IST) of Lisbon, Portugal, in a collaboration with researchers from the IST.

4.25. Utopia

Participant: Marie-France Sagot [Correspondent].

UTOPIA is a gene inference algorithm using an approach by pure homology. The algorithm performs a doubly-spliced alignment of two genomic sequences using a generic gene model. Frameshifts due to possible sequencing errors are taken into account. The algorithm may infer more than one gene at once. The genes sought must in this case appear in the same order in the two sequences for the algorithm to be able to identify them. UTOPIA was developed by Philippe Blayo during his PhD at the University of Marne-la-Vallée. The current version (in C++) together with scripts for post-processing is freely-available to academics and non-profit research organisations by sending a mail to Marie-France.Sagot@inrialpes.fr.

4.26. 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 4.5);
- 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), GEM (Bruno Spataro), JADIS (Dominique Mouchiroud), MTDP (Alain Viari), and SEAVIEW (Manolo Gouy).

5. New Results

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

Alexandra Calteau defended her Ph-D thesis on Horizontal Gene Transfers in prokaryotic genomes, in November 2005. Her work focused on the mechanism of gene transfers between archae and hyperthermophylic bacteria. From the methodological point of view, she designed new supertree methods to build relevant phylogenies of the bacterial kingdom. She showed, in particular, that hyperthermophilic bacteria have undergone many more transfers than thermophilic ones. This puts some doubts on the hypothesis of an hyperthermophilic character of the last common ancestor of the whole living world. The main part of the transfers imply genes coding for proteins of unknown function. This suggests a potential role of these unknown proteins in the adaptation to high temperatures.

The mechanism and role of horizontal gene transfers have become a major theme in the Lyon part of the project, with the work of Vincent Daubin and Emmanuelle Lerat (recruited in 2004 and 2005). They evidenced, in particular, the role of bacteriophages in gene transfers between bacteria, which suggests that transfers participate in the reproductive success of bacteriophages before the one of bacteria.

Manolo Gouy participated in the phylogenetic analysis of the sequenced genome of *Encephalitoozoon cuniculi*, showing the impact of the relative evolutionary rate of each protein on phylogenetic reconstructions.

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

Participants: Eric Coissac, Marilia Dias Vieira Braga, Laurent Duret, Christian Gautier [Correspondent], Laurent Guéguen, Adel Khelifi, Claire Lemaitre, Jean Lobry, Julien Meunier, Anne Morgat, Dominique Mouchiroud, Guy Perrière, Marie-France Sagot [Correspondent], Marie Semon, Bruno Spataro, Eric Tannier, Raquel Tavares, Alain Viari.

Isochores represent a feature of the spatial organisation of genomes that has been studied for a long time. Christelle Melo de Lima has defended her Ph-D in December 2005, on Hidden Markov Chains to identify isochores along a genome. The model was trained on the human genome and allowed to evidence some isochore organisation on the tetraodon genome, even in the absence of obvious C+G homogeneous regions in this latter case.

In the context of his PhD thesis, defended in September 2005, Julien Meunier has studied the substitution patterns on CpG dinucleotides in primates. He could demonstrate the existence of an homology-dependent mechanism implying hypermethylation of repeated sequences (transposable elements or pseudogenes). This is the first demonstration of the existence of a defense mechanism against the invasion of transposable elements in mammal genomes.

Adel Khelifi has defended his Ph-D thesis in September 2005, about processed pseudo-genes. Processed pseudogenes are generated by the reverse transcription of mRNAs corresponding to functional genes, resulting in genes that are no longer functional. The fact that they are generally no longer transcribed into RNAs and translated into proteins makes these genes useful for studying molecular evolution, in particular for examining substitution patterns. This approach has been applied to the human and mouse genomes, giving rise to a database of annotated pseudogenes, called HOPPSIGEN.

Recombination patterns have been the subject of several studies, related to isochores, and to the structural organisation of genomes. We have recently studied a case of loss of recombination in the scope of the appearance of sexual reproduction in a family of dioecious plants. It appears that suppression of recombination is accompanied by a degeneracy of the genes localised on the Y chromosome, which points out the possibility of a selective advantage for recombination.

Several theoretical results have been obtained on retrieving the evolutionary history from the comparative study of gene order data. A model of conservation and disruption in gene order has been designed and implemented in the framework of a MSc thesis of a student from the University of Marne-la-Vallée. It is used to test the parsimony hypothesis, that is the likelihood of the evolutionary scenario that explains with a minimum number of events the difference between two genomes. It often appears that the most parcimonious scenario is not compatible with the preservation of clusters of genes. A software called CASSIS has been developed by Claire Lemaitre, starting a Ph-D in HELIX, to identify all the regions of a genome that have been broken by large-scale rearrangements.

Marie Semon defended her Ph-D thesis in September 2005, on the analysis of expression data in mammals, compared with gene order data. She showed that gene order is not uniform along genomes, and that there are significant clusters related to the putative function of the genes. However quantifying this effect, she proved that this clustering concerns a very small proportion of the genes on the human genome, and can be explained by neutral processes of evolution.

We also started a collaboration with researchers at the LIRIS laboratory in Lyon, on the subject of genes clusters and their relation to metabolic pathway, or co-expression. This work is an extension of the results of the thesis of Frédéric Boyer in 2004, now a post-doc in the Lyon part of the project.

5.3. Motif search and inference

Participants: Christian Gautier, Laurent Gueguen, Paulo Gustavo da Fonseca, Vincent Lacroix, Christelle Melo de Lima, Leonor Palmeira, Guy Perrière, Marie-France Sagot [Correspondent], Alain Viari.

A large part of the algorithmic studies in the HELIX project concerns the search of regularities at many levels of living systems. They are motifs in DNA sequences, RNA structures or protein structures, as well as motifs in metabolic networks (see also the next section).

Concerning motifs in DNA sequences, Pierre Peterlongo, a Ph-D student co-supervised in HELIX and the University of Marne-la-Vallée, designed an algorithm called NIMBUS for filtering sequences prior to finding repetitions occurring more than twice in a sequence or in more than two sequences. NIMBUS uses gapped seeds that are indexed with a new data structure, called a bi-factor array. Experimental results show that the filter can be very efficient. Said Sadique Adi defended his Ph-D thesis in may 2005, in co-supervision with Carlos Eduardo Fereira, from the University of Sao Paulo, Brazil, about the application of DNA motif search to gene finding.

Concerning RNA secondary structures, Julien Allali, in collaboration with Marie-France Sagot, introduced a new data structure, called MiGaL for "Multiple Graph Layers", that is composed of various graphs linked together by relations of abstraction/refinement. The new structure is useful for representing information that can be described at different levels of abstraction, each level corresponding to a graph. They proposed an algorithm for comparing two MiGaLs. MiGaLs represent a very natural model for comparing RNA secondary structures that may be seen at different levels of detail, going from the sequence of nucleotides, single or paired with another to participate in a helix, to the network of multiple loops that is believed to represent the most conserved part of RNAs having similar function.

Finally, Jeane C. B. de Melo defended her Ph-D thesis in August 2005, in co-supervision with Katia Guimarães, from the Federal University of Pernambuco, in Brazil. Her work is about identifying protein domains using their tertiary structure.

5.4. Computational proteomics and transcriptomics

Participants: Laurent Duret, Christian Gautier, Paulo Gustavo da Fonseca, Dominique Mouchiroud, Guy Perrière [Correspondent for transcriptomics], Simon Penel, Emmanuel Prestat, Raquel Tavares, Jean Thioulouse, Alain Viari [Correspondent for proteomics].

Transcriptomics : MADE4, an R package for multivariate analysis of gene expression data based on the ADE4 package and developed in collaboration with the Conway Institute in Dublin, has been released in september 2005.

Proteomics: This year has seen the completion and installation of the sc PepLine software at the Laboratoire de Chimie des Proteines (LCP) at the CEA of Grenoble and its first application to the genome of *Chlamydomonas reinhardtii* under the supervision of Marianne Tardif. This led us to improve the prediction score of the peptide sequence tag (PST) module (Taggor) as well as the overall speed of the pipeline. In the same vein, we have also started a new project (funded by ACI IMPBIO) to develop a new GENOSTAR module (entitled GenoProteo) dedicated to the manipulation and visualisation of PSTs on complete chromosomes. The aim of the project is to provide the end user with an intuitive and easy to use interface.

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

Participants: Frédéric Boyer, Ludovic Cottret, Vincent Lacroix, Anne Morgat, Marie-France Sagot [Correspondent], Patricia Thebault, Alain Viari [Correspondent], Erik Wessel.

The definition and efficient algorithmic search of motifs and modules in biological networks, and more precisely metabolic networks, has become one of the main research topics of Helix this year, both in Lyon and Grenoble. This is the main topic of the Ph-D thesis of Vincent Lacroix, who has published in collaboration with some members of the University of São Paulo, to which HELIX is associated, a model and a research algorithm for motifs in networks. The originality stands in the fact that in this model, the topology of the motifs is less important than the presence of certain enzymes which catalyse a sub-network of connected reactions. This appeared to be a good starting point to retrieve regular or duplicated motifs in networks. The algorithm, called MOTUS has been implemented and tested to retrieve known metabolic pathways in the KEGG database. An engineer has been recruited to give to this implementation a user-friendly aspect, in order to make it accessible to biologists.

Another important question that has continued to be addressed in 2005 deals with the question of relating metabolic to genomic information (*e.g.*, how genes involved in the catalysis of adjacent biochemical reactions are co-localised on a chromosome). To this purpose, Frédéric Boyer has extended a previous graph-theoretical approach towards finding bacterial syntenies developed in HELIX. In particular, he has elaborated an algorithm that integrates several graphs representing different types of metabolic and genomic information into a single graph that can be analysed for patterns shared by all of the original graphs. The developed approach is fairly general and can be applied to other kinds of biological data that may be represented by graphs (such as protein-protein interactions or regulatory networks).

The collaboration with the Swiss-Prot group headed by Amos Bairoch in Geneva has intensified in 2005. This has led to the development of an ontology dedicated to intermediate metabolism and its implementation within the MICROBI relational database (UniPathway database). The database has been populated with curated data (originating from the Swiss-Prot group) and is now accessible to the Swiss-Prot curators for their daily work. It will be opened to the community in 2006 and will be hosted by the proteomics server, funded by the Réseau National des Génopoles that will be installed in January 2006.

Finally, we would like to stress the fact that the arrival of Daniel Kahn (DR1 INRA) in September 2005 is expected to strongly reinforce our activities in this research area.

5.6. Modelling and simulation of genetic regulatory networks

Participants: Grégory Batt, Samuel Drulhe, Hidde de Jong [Correspondent], Michel Page, Delphine Ropers.

This year much of our efforts have been concentrated on the application of the qualitative simulation tool GENETIC NETWORK ANALYZER (GNA) (section 4.9)) to the modeling of actual genetic regulatory networks. In particular, we study the nutritional stress response in the bacterium *Escherichia coli* in collaboration with experimental biologists in the laboratory of Johannes Geiselmann (Université Joseph Fourier, Grenoble). A model has been developed and published by Delphine Ropers. Within the framework of the European project HYGEIA, we are in the process of extending the model with the help of students from the INSA in Lyon. Moreover, as a visiting scientist in the laboratory of Johannes Geiselmann, Delphine Ropers is actively involved in the experimental validation of the model predictions. Other experiments are being carried out in collaboration with Irina Mihalcescu of the Laboratoire de Spectométrie Physique (Université Joseph Fourier, Grenoble).

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 equilibrium states of the system and analyze the neighbouring states to determine the stability of the equilibrium states. In collaboration with Jean-Luc Gouzé (INRIA Sophia-Antipolis) and Tewfik Sari (Université de Haute-Alsace, Mulhouse) we have worked on the mathematical characterisation of equilibria of piecewise-linear differential equation models, the class of models underlying the qualitative simulation

method. Michel Page and Hidde de Jong have developed a new module of GNA for the search of equilibrium states and the determination of their stability, based on the above-mentioned mathematical study.

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 pursued this approach in collaboration with Radu Mateescu and his colleagues of the VASY project. A refined simulation method has been developed that exploits the concept of discrete abstraction developed in the hybrid systems community. The model-checking approach has been applied to the validation of the *E. coli* nutritional stress response models by means of experimental data available in the literature.

Until now most of our work has focused on the analysis of models obtained through literature study and human expertise. The PhD thesis of Samuel Drulhe, supervised by Hidde de Jong and Giancarlo Ferrari-Trecate (INRIA Rocquencourt and 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 genetic regulatory networks by means of gene expression data, based on existing approaches towards hybrid systems identification.

5.7. Bioanalysis and cross-sectional activities

Participants: Abdel Aouacheria, Frédéric Boyer, Eric Coissac, Stéphane Descorps-Declère, Christian Gautier [Correspondent], Dominique Mouchiroud, Anne Morgat, Vincent Navratil, Simon Penel, François Rechenmann, Alain Viari [Correspondent].

Several activities involving methods and tools developed in the other research areas, often applied to concrete biological problems in collaboration with public organisms or private enterprises, are currently under way.

In collaboration with the CIRAD (Montpellier) and the IGH (Montpellier), the complete genome sequences of two strains of *Ehrlichia ruminantium* have been fully sequenced and annotated, using the GenoStar platform. The closeness of the genomes allowed an in-depth analysis of homologies in coding and non-coding regions, evidencing an active process of genome expansion/contraction targeted at tandem repeats.

Another example of a cross-sectional activity in HELIX is the PhD research of Stéphane Descorps-Declère, who has designed a blackboard architecture to achieve the coordinated application of genome annotation methods. In 2005, the blackboard framework has been instantiated to obtain the prototype of a genome annotation system. This prototype has been tested and its results compared with a validated set of genomic data. This experience confirms the expected positive properties of a blackboard architecture in this context.

6. Contracts and Grants with Industry

6.1. Aventis-Pasteur

Participants: Frédéric Boyer, Anne Morgat, Alain Viari [Correspondent].

In September 2004, HELIX started a two-year contractual relation with the company Aventis-Pasteur located in Lyon. The collaboration concerns the in-depth (re)annotation of pathogenic bacteria of interest to Aventis-Pasteur. In 2005, the (re)annotation of several strains was completed according to the schedule, opening the path to the in-depth analysis of their metabolic capabilities. This work is done by F. Boyer in the context of his post-doc between Lyon and Grenoble.

6.2. Genostar

Participants: Eric Coissac, Anne Morgat, Agnès Iltis, François Rechenmann [Correspondent], Alain Viari.

Genostar S.A. is a private company based in Paris, which will act as the provider of the integrated, and interoperable software environment GenoStar. GenoStar consists of a highly customizable exploratory platform and components for the analysis of genomic and post-genomic data (sequence analysis and whole genome

annotation, proteomics, expression data, *etc.*). The platform and the components have been initially developed from 1999 to 2003 by a public-private consortium (http://www.genostar.org) involving two biotechnology companies (Hybrigenics in Paris, and GENOME express in Grenoble) and two public research organisms (INRIA, and more specifically the HELIX project, and the Institut Pasteur de Paris). The company GenoStar has been launched by Patrice Garnier, who has previously started and run a successful information technology company.

6.3. Sanofi

Participants: Gilles Faucherand, Agnès Iltis, Alain Viari [Correspondent].

In September 2002, HELIX started a three-year contractual relation with the company Sanofi Synthélabo in Toulouse. The collaboration concerns the design of a software environment (MetaProtan) devoted to the analysis of proteomic data.

7. Other Grants and Activities

7.1. National projects

Project name	BacAttract : Analyse théorique et expérimentale d'attracteurs de réseaux de régulation génique : régulation globale de la transcription chez *Escherichia coli* et *Synechocystis* PCC 6803
Coordinators	H. de Jong
HELIX participants	H. de Jong, M. Page, D. Ropers
Type	ACI IMPBio (2003-2006)
Web page	http://impbio.lirmm.fr/PROJETS_ACCEPTES/paper12.html

Project name	Evolrep
Coordinators	J. Pothier
HELIX participants	E. Coissac, A. Morgat
Type	ACI IMPBio (2003-2006)
Web page	http://impbio.lirmm.fr/PROJETS_ACCEPTES/paper72.html

Project name	Flybase
Coordinators	C. Biémont
HELIX participants	P. Genoud, F. Rechenmann, D. Ziébelin
Type	ACI IMPBio (2004-2007)
Web page	http://impbio.lirmm.fr/PROJETS_ACCEPTES_2004/11.html

Project name	GenoProtéome
Coordinators	C. Bruley
HELIX participants	A. Viari
Type	ACI IMPBio (2004-2007)
Web page	http://impbio.lirmm.fr/PROJETS_ACCEPTES_2004/50.html

Project name	Isimod+
Coordinators	Y. Quentin
HELIX participants	D. Ziébelin
Type	ACI IMPBio (2003-2006)
Web page	http://impbio.lirmm.fr/PROJETS_ACCEPTES/paper77.html

Project name	Evolutionary dynamics of global gene regulatory networks in *Escherichia coli*
Coordinators	J. Geiselmann

HELIX participants	H. de Jong, M. Page, D. Ropers
Type	inter-EPST Microbiologie (2004-2006)
Web page	

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	

Project name	IBN (Integrated Biological Networks)
Coordinators	M.F. Sagot
HELIX participants	all HELIX members
Type	ARC INRIA (2005)
Web page	http://www.inrialpes.fr/helix/people/sagot/team/projects/arc2005_2006/arc2005_2006.html

Project name	Inférence de régularités dans les systèmes biologiques
Coordinators	M.F. Sagot
HELIX participants	BAOBAB team
Type	ANR blanc (2005)
Web page	

Project name	Génomique Microévolutive
Coordinators	L. Duret
HELIX participants	Jean Lobry, Gabriel Marais, Sylvain Mousset, Vincent Daubin, Eric Tannier
Type	ANR jeune Chercheur (2005)
Web page	

7.2. Projects funded by international organisms or including international teams

Project name	HYGEIA : Hybrid Systems for Biochemical Network Modeling and Analysis
Coordinators	J. Lygeros, G. Ferrari-Trecate
HELIX participants	S. Druhle, H. de Jong, M. Page, D. Ropers
Type	European Commission, FP6 Priority, NEST-4995
Web page	http://www.hygeiaweb.gr/home.html

Project name	TEMBLOR/Integr8
Coordinators	G. Cameron
HELIX participants	L. Duret, S. Penel, G. Perrière
Type	European Community Contract No. QLRI-CT-2001-00015 under the specific RTD programme 'Quality of Life and Management of Living Resources'
Web page	http://www.ebi.ac.uk/integr8/

Project name	Sibelius
Coordinators	A. Viari (INRIA), A. Bairoch (SIB)
HELIX participants	E. Coissac, C. Lachaize, A. Morgat (SIB), A. Viari
Type	Equipe associée INRIA
Web page	http://www.inrialpes.fr/helix/SIB/sibelius.html

Project name	Arcoiris

Coordinators	M.F. Sagot (INRIA), Y. Wakabayashi (Univ. Sao Paulo)
HELIX participants	All HELIX members
Type	Equipe associée INRIA
Web page	http://www.inrialpes.fr/helix/people/sagot/team/projects/associated_team_usp_helix/

Project name	Séminaire Algorithmique et Biologie
Coordinators	M.-F. Sagot
HELIX participants	M.-F. Sagot (will include around 70% foreign guest speakers)
Type	ACI IMPBio (2003-2006)
Web page	http://www.inrialpes.fr/helix/people/sagot/AlgoBio/index.html

Project name	Pattern inference in computational molecular biology
Coordinators	C. Iliopoulos, M.-F. Sagot
HELIX participants	M.-F. Sagot
Type	Royal Society, UK (2000-...)
Web page	

Project name	Algorithmics and Combinatorics for Molecular Biology
Coordinators	K. Guimarães, M.-F. Sagot
HELIX participants	M.-F. Sagot, E. Tannier
Type	Capes-Cofecub (2003-2005, renewable for two more years)
Web page	http://www.inrialpes.fr/helix/people/sagot/team/projects/brazil_capes_2004/brazil_capes_2004.html

Project name	π-vert
Coordinators	M.-F. Sagot
HELIX participants	almost all members of HELIX, 8 other French partners and 18 European partners
Type	ACI Nouvelles Interfaces des Mathématiques (end 2004-2007)
Web page	under construction

8. Dissemination

8.1. Talks

Grégory Batt

Title	Event and location	Date
Qualitative analysis and verification of hybrid models of genetic regulatory networks : Nutritional stress response in *Escherichia coli*	Eighth International Workshop on Hybrid Systems : Computation and Control (HSCC'05), Zurich, Switzerland	11/03/05
Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in *Escherichia coli*	Workshop on Dynamical Modeling and Analysis of Biological Regulatory Networks, Marseille	10/05/05
Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in *Escherichia coli*	Thirteenth International Conference on Intelligent Systems for Molecular Biology (ISMB'05), Detroit, Etats-Unis	26/06/05
Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in *Escherichia coli*	Journées Ouvertes Biologie, Informatique et Mathématiques (JOBIM'05), Lyon	08/07/05
Analysis and verification of qualitative models of genetic regulatory networks: A model-checking approach	Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, Royaume-Uni	03/08/05
Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in *Escherichia coli*	Workshop on Computation of Biochemical Pathways and Genetic Networks, Heidelberg, Allemagne	13/09/05

Frédéric Boyer

Title	Event and location	Date
Reconstruction de voies métaboliques	Physique du vivant 2005 : Séminaire Dautreppe de la Société Française de Physique, à l'interface Physique Biologie Grenoble	09/30/2005

Alexandra Calteau

Title	Event and location	Date
Extensive phylogenetic study reveal that hyperthermophilic bacteria have been subject to massive horizontal gene transfer from archaea	JOBIM (Journées Ouvertes pour la Biologie, l'Informatique et les Mathématiques, French conference of bioinformatics), Lyon	juillet 2005.
Extensive phylogenetic study reveal that hyperthermophilic bacteria have been subject to massive horizontal gene transfer.	ESEB (European Society of Molecular Biology), Krakow, Pologne	août 2005

Laurent Duret

Title	Event and location	Date
Relationships between genome organization and gene expression in mammals.	"Benelux bioinformatics Conference", Ghent, Belgique	14-15 avril 2005.
Relationships between genome organization and gene expression in mammals.	"Biological Networks: Interaction with Genome and Developmental Evolution", Bertinoro, Italie	21-27 mai 2005.
Relationships between genome organization and gene expression in mammals: selective constraints or neutral evolution?	24th Summer Symposium in "Molecular Biology: comparative and functional genomics", Pennsylvania State University, USA	20-23 juillet 2005.
Patterns of silent substitutions in the human genome.	10th Congress of the "European Society for Evolutionary Biology", Cracovie, Pologne	15-20 août 2005.
Patterns of silent substitutions in the human genome	3rd RECOMB "Comparative Genomics Satellite Workshop", Dublin, Irlande	18-20 septembre 2005.
Relationships between genome organization and gene expression in mammals: selective constraints or neutral evolution?	"Lausanne genomics days", Lausanne, Suisse	6-7 octobre 2005.

Hidde de Jong

Title	Event and location	Date
Modélisation et simulation de réseaux de régulation génique	Séminaire de Bioinformatique de la Génopole Montpellier Languedoc-Roussillon, Montpellier	28/02/05
Analysis and verification of qualitative models of genetic regulatory networks: A model-checking approach	19th International Workshop on Qualitative Reasoning, QR'03, Graz, Austria	18/05/05
Qualitative modeling and simulation of genetic regulatory networks	ESF Workshop on Transcription Networks, Madrid, Spain	28/05/05
Modélisation et simulation de réseaux de régulation génique	Journée ACI IMPBio, Lyon	05/07/05
Qualitative modeling and simulation of genetic regulatory networks	Taiwanese Bioinformatics conference BIT2005, Tainan, Taiwan	09/09/05
Modélisation mathématique et analyse expérimentale de réseaux de régulation bactériens	Séminaire Daniel Dautreppe 2005 : Physique du Vivant, Grenoble (with Johannes Geiselmann)	30/09/2005
Modélisation et simulation de réseaux de régulation génique : La réponse au stress nutritionnel chez *E. coli*	4ème Ecole Thématique CNRS/INRA de Biologie Végétale : Biologie intégrative, Batz sur Mer (with Delphine Ropers)	06/10/05
Qualitative modeling and simulation of genetic regulatory networks	2005 DIA Statistical Methodology in Pharmaceutical R&D, Nice	28/10/05

Vincent Lacroix

Title	Event and location	Date
Motifs dans les réseaux métaboliques	INRA - Ferme du Moulon	22 mars 2005
Network biology	Cours Informatique en Biologie de l'Institut Pasteur, Paris	23 mars 2005

Jean Lobry

Title	Event and location	Date
Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria	XXXVIII Conference of Polish Society for Biochemistry, Wroclaw, Poland	18-22 September 2002
Life history traits and genome structure : aerobiosis and G+C content in bacteria	International Conference on Computational Science (ICCS 2004), Krakow, Poland	7-9th June, 2004
seqinR: a contributed package to the R project for statistical computing devoted to biological sequences retieval and analysis	Structural approaches to sequence evolution: Molecules, networks, populations. Dreden, Germany	5-10th july 2004
Unusual geometry in some genomes of the Trypanosomatidae familly	Workshop on Geometry of Genome: visualization of structures hidden in genomic sequences. University of Leicester (Centre for Mathematical Modelling) together with IHES - Institut des Hautes Études Scientifiques	22-24 september 2005
Codon usage and UV exposure in bacteria	Workshop on Physics in Biology: Genes, Genomes and Populations Evolution. Wroclaw University, Poland	29 September - 2 October 2005
Codon usage and optimal growth temperature in Prokaryotes	6th Anton Dohrn workshop: evolutionary genomics, Ischia, Naples, Italy	31 october - 2 november 2005
Revisiting the directional mutation pressure theory: a particular genomic structure in Leishmania major	6th Anton Dohrn workshop: evolutionary genomics, Ischia, Naples, Italy	31 october - 2 november 2005

Leonor Palmeira

Title	Event and location	Date
Robustesse des méthodes de reconstruction phylogénétiques face à un écart à l?hypothèse d?indépendance entre site	ALPHY. Montpellier	18 janvier 2005

Delphine Ropers

Title	Event and location	Date
Analyse de la réponse au manque de carbone chez Escherichia coli : utilisation des gènes rapporteurs *gfp* et *lux*	JOBIM satellite meeting, Lyon	09/07/05
Piecewise-linear models of genetic regulatory networks: Application to the analysis of the carbon starvation response in *E. coli*	European Conference on Mathematical and Theoretical Biology ECMTB 2005, Dresden, Germany	21/07/05
Genetic Network Analyzer: A tool for the qualitative modeling and simulatin of genetic regulatory networks	Taiwanese Bioinformatics conference BIT2005, Tainan, Taiwan	08/09/05
Modélisation et simulation de réseaux de régulation génique : La réponse au stress nutritionnel chez *E. coli*	4ème Ecole Thématique CNRS/INRA de Biologie Végétale : Biologie intégrative, Batz sur Mer (with Hidde de Jong)	06/10/05

Eric Tannier

Title	Event and location	Date
Conservation and Rearrangements in Genomes	Université de Sao Paulo	6 mai 2005
Remaniements et conservation dans les génomes	Laboratoire LIRMM, Montpellier	14 mars 2005
Comparaison des génomes et tri des permutations	Laboratoire Leibniz, Grenoble	3 février 2005
Premiers coups d'oeil à la bioinformatique	Séminaire sport-études de l'ENS, Le Pleynet	15 janvier 2005

8.2. Organisation of conferences, workshops and meetings

Laurent Duret

Type	Location	Date
Conférence IPG (Integrative Post Genomics)	Lyon	1 et 2 décembre 2005

Hidde de Jong

Type	Location	Date

JOBIM satellite meeting on Dynamical Modelling of Biological Regulatory Networks (with C. Chaouiya)	Lyon	09/07/05

Marie-France Sagot

Type	Location	Date
Séminaire Algorithmique et Biologie : Série 17 Génomique Évolutive	Banyuls sur Mer	2-28 october 2005
Biological Networks : Interaction with Genome and Developmental Evolution	Bertinoro, Italie	
CompBioNets 2005	Lyon, France	5-7 december 2005

Guy Perrière

Type	Location	Date
JOBIM (Journées Ouvertes : Biologie Informatique et Mathématiques)	Technopole Lyon-Gerland	6-8 juillet 2005

8.3. Editorial and reviewing activities

Manolo Gouy

Type	Journal or conference
Editorial Board	*Molecular Biology and Evolution*, OUP

Hidde de Jong

Type	Journal or conference
Editorial Board	ACM/IEEE Transactions on Computational Biology and Bioinformatics
Program Committee	CompBioNets, QR, GENSIPS, ISMB, JOBIM, IPG
Scientific Committee	Working group VICANNE (“Modélisation dynamique et simulation des systèmes biologiques”)

François Rechenmann

Type	Journal or conference
Editorial Board	*Bioinformatics*, OUP

Marie-France Sagot

Type	Journal or conference
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 Bioinformatics*, BioMed Central
Editorial Board	*BMC Algorithms for Molecular Biology*, BioMed Central
Editorial Board	*Computational Biology and Chemistry*, Elsevier
Program Committee	RECOMB satellite conference on Comparative Genomics, ECCB-JBI, ISMB, SPIRE, WABI, PSW, CompBioNets (co-chair)

8.4. Administrative activities

Laurent Duret is member of the “Conseil de l’UFR de biologie de l’Université Claude Bernard Lyon (UCBL)”, of the “Commission de spécialistes biologie de l’ENS Lyon” and of the scientific committee of the ACI IMPBio.

Christian Gautier is director of the LBBE (UCBL, UMR 5558), deputy director of the UFR of Biology of the UCBL, chair of the section 29 of the CNRS and responsible for the National Network of the Bioinformatics Platforms.

Manolo Gouy is is a member of the Conseil National des Universités, section 67, of the scientific committee of the “Institut Français de la Biodiversité”, of the promotion committee for CNRS Research Engineers in biology and of the selection committee for the ATIP CNRS “Biodiversité”, the CNRS IE478 and the CNRS IR476 (chair).

Hidde de Jong is a member of the recruiting committee for research associates at INRIA Rocquencourt. In addition, he participates in the International Relations working group of the Conseil d’Orientation Scientifique et Technologique (COST) of INRIA. He is a member of the board of the Société Française de Bioinformatique (SFBI).

Marie-France Sagot is a member of the course “Informatique en Biologie” of the Institut Pasteur in Paris and of the course on Computational Biology of the University of Chile in Santiago, Chile. She is a deputy member of the Evaluation Committee of INRIA. She has participated in the recruiting committee for research associate positions in the track “Modélisation du Vivant” at INRIA. She has participated in the reviewing process of projects or candidates for a research position for the FCI (Canada), the EPSRC (UK), the Technion (Haifa, Israel) and the Netherlands Genomics Initiative (NGI).

François Rechenmann is President of the recruiting committee for CR2 positions at the INRIA Rhone-Alpes in 2005. He is also a member of the scientific committee of the Interstices website (http://interstices.info/). Interstices offers pedagogical presentations of research themes and activities in the computer science domain.

Alain Viari is a member of the "Commission de spécialistes" section 65 at Université de Paris 6. He is president of the "Comité des Emplois Scientifiques" at the INRIA Rhône-Alpes and member of the COST-AE (INRIA).

8.5. Teaching activities

Seven members of the HELIX project, four in Lyon and three in Grenoble, are professors or assistant professors at, respectively, the Université Claude Bernard in Lyon (UCBL) and the Université Joseph Fourier and the Université Pierre Mendès-France in Grenoble. They therefore have a full teaching load of at least 192 hours.

Over the years various members of the project have developed courses in biometry, bioinformatics and evolutionary biology at all levels at universities as well as at the “École Normale Supérieure” (ENS) of Lyon and the “Institut National de Sciences Appliquées” (INSA) in Lyon. One strong motivation of these teaching activities 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 also contributes bioinformatics courses at the level of a “Magistère” at the ENS.

As part of the LMD system that is currently being set up at all universities in France, members of the project have created a complete interdisciplinary module in Lyon offering training in biology, mathematics and computer science. The module is called “Approches Mathématique et Informatique du Vivant” (AMIV), and leads to Master degrees in the scientific and medical fields.

A second important educational activity of the project concerns the tying together of biology and mathematics teaching to biologists. To this end, 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 concerns general and population genetics, and molecular evolution.

Finally, members of the project have participated in, or sometimes organised, numerous courses or teaching modules at the national and international level (such as, for instance, the creation and support of a Master course in Ho-Chi-Minh, Vietnam). Besides the full-time professors, the following members of HELIX have contributed to courses this year.

Laurent Duret

Subject	Year	Location	Hours
Bioinformatique	3 to 5	INSA Lyon, UCBL	40

Manolo Gouy

Subject	Year	Location	Hours
Molecular phylogeny	3 to 5	UCBL, ENS Lyon, INSA Lyon	33
Molecular phylogeny	-	Atelier INSERM, La Londe les Maures	-
Molecular phylogeny	-	École de biologie moléculaire, IFREMER, Banyuls	-

Hidde de Jong

Subject	Year	Location	Hours
Modelling and simulation of genetic regulatory networks (with D. Ropers)	5	UCBL	8
Modelling and simulation of genetic regulatory networks (with G. Batt)	4	INSA, Lyon	14
Modelling and simulation of genetic regulatory networks	5	ENS, Paris	2
Modelling and simulation of genetic regulatory networks	5	Universitat Pompeu Fabra, Universitat de Barcelona	2

Guy Perrière

Subject	Year	Location	Hours
Bioinformatics	3	ENS Lyon	17
Horizontal gene transfer	4	INSA Lyon	8
Introduction to bioinformatics	5	Faculté de Médecine Laënnec	7
Introduction to bioinformatics	4-5	UCBL	15
Phylogeny	5	Institut Universitaire Européen de la Mer	13

François Rechenmann

Subject	Year	Location	Hours
Knowledge modelling	4	Université Joseph Fourier, Grenoble	14
Bioinformatics	4	Université Joseph Fourier, Grenoble	9
Bioinformatics : modeling and analysis of genomic and post-genomic data	5	ENS Lyon	2.5

Delphine Ropers

Subject	Year	Location	Hours
Modelling and simulation of genetic regulatory networks (with H. de Jong)	5	UCBL	8
Modelling and simulation of genetic regulatory networks	4	UJF, Grenoble	14

Marie-France Sagot

Subject	Year	Location	Hours
Algorithmics for biology	5	Master AMIV, Université Claude Bernard, Lyon	10
Algorithmics for biology	5	BIM, INSA Lyon	20
Algorithmics for biology	5	Université de Marne-la-Vallée	8
Motif inference	5	Université de Marne-la-Vallée	10
Algorithmic complexity and NP-completeness	5	Pasteur Institute, Paris	3
Genome rearrangements	5	Pasteur Institute, Paris	3

Eric Tannier

Subject	Year	Location	Hours
Statistics	2	ESQESE, UCL	30

9. Bibliography

Doctoral dissertations and Habilitation theses

[1] E. COISSAC. *La fluididté des génomes*, HDR, Université Paris VI - Pierre et Marie Curie, 2005.

Articles in refereed journals and book chapters

[2] J. ALLALI, M.-F. SAGOT. *A New Distance for High Level RNA Secondary Structure Comparison.*, in "IEEE/ACM Trans. Comput. Biology Bioinform.", vol. 2, n° 1, 2005, p. 3-14, http://doi.acm.org/10.1145/1057653.

[3] A. AOUACHERIA, V. NAVRATIL, W. WEN, M. JIANG, D. MOUCHIROUD, C. GAUTIER, M. GOUY, M. ZHANG. *In silico whole-genome scanning of cancer-associated nonsynonymous SNPs and molecular characterization of a dynein light chain tumour variant*, in "Oncogene", vol. 24, n° 40, Sep 2005, p. 6133–6142.

[4] H. AURELL, P. FARGE, H. MEUGNIER, M. GOUY, F. FOREY, G. LINA, F. VANDENESCH, J. ETIENNE, S. JARRAUD. *Clinical and environmental isolates of Legionella pneumophila serogroup 1 cannot be distinguished by sequence analysis of two surface protein genes and three housekeeping genes*, in "Appl Environ Microbiol", vol. 71, n° 1, Jan 2005, p. 282–289.

[5] G. BATT, D. ROPERS, H. DE JONG, J. GEISELMANN, R. MATEESCU, M. PAGE, D. SCHNEIDER. *Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in Escherichia coli*, in "Bioinformatics", vol. 21 Suppl 1, Jun 2005, p. i19-i28.

[6] E. BAZIN, L. DURET, S. PENEL, N. GALTIER. *Polymorphix: a sequence polymorphism database*, in "Nucleic Acids Res", vol. 33, n° Database issue, Jan 2005, p. 481–484.

[7] B. BONNAUD, J. BELIAEFF, O. BOUTON, G. ORIOL, L. DURET, F. MALLET. *Natural history of the ERVWE1 endogenous retroviral locus*, in "Retrovirology", vol. 2, Sep 2005, 57.

[8] F. BOYER, A. MORGAT, L. LABARRE, J. POTHIER, A. VIARI. *Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data*, in "Bioinformatics", vol. 21, n° 23, Dec 2005, p. 4209–4215.

[9] A. CALTEAU, M. GOUY, G. PERRIERE. *Horizontal transfer of two operons coding for hydrogenases between bacteria and archaea*, in "J Mol Evol", vol. 60, n° 5, May 2005, p. 557–565.

[10] A. M. CARVALHO, A. T. FREITAS, A. L. OLIVEIRA, M.-F. SAGOT. *An Efficient Algorithm for the Identification of Structured Motifs in DNA Promoter Sequences*, in "to appear in IEEE/ACM Trans. Comput. Biology Bioinform.", 2005.

[11] C. CHAOUIYA, H. DE JONG, D. THIEFFRY. *Introduction to special issue on dynamical modeling of biological regulatory networks*, in "Biosystems", Nov 2005.

[12] D. CHARIF, J. THIOULOUSE, J. LOBRY, G. PERRIERE. *Online synonymous codon usage analyses with the ade4 and seqinR packages*, in "Bioinformatics", vol. 21, n° 4, Feb 2005, p. 545–547.

[13] D. CHARLESWORTH, B. CHARLESWORTH, G. MARAIS. *Steps in the evolution of heteromorphic sex chromosomes*, in "Heredity", vol. 95, n° 2, Aug 2005, p. 118–128.

[14] A. C. CULHANE, J. THIOULOUSE, G. PERRIERE, D. G. HIGGINS. *MADE4: an R package for multivariate analysis of gene expression data*, in "Bioinformatics", Evaluation Studies, vol. 21, n° 11, Jun 2005, p. 2789–2790.

[15] G. DECELIERE, S. CHARLES, C. BIEMONT. *The dynamics of transposable elements in structured populations*, in "Genetics", vol. 169, n° 1, Jan 2005, p. 467–474.

[16] J.-F. DUFAYARD, L. DURET, S. PENEL, M. GOUY, F. RECHENMANN, G. PERRIERE. *Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases*, in "Bioinformatics", Evaluation Studies, vol. 21, n° 11, Jun 2005, p. 2596–2603.

[17] R. DUPONNOIS, A. COLOMBET, V. HIEN, J. THIOULOUSE. *The mycorrhizal fungus Glomus intraradices and rock phosphate amendment influence plant growth and microbial activity in the rhizosphere of Acacia holosericea*, in "Soil Biology and Biochemistry", vol. 37, n° 8, 2005, p. 1460-1468.

[18] L. DURET. *Les trésors cachés du génome humain*, in "Pour la Science", jan 2005, p. 16-21.

[19] A. GINOLHAC, C. JARRIN, P. ROBE, G. PERRIERE, T. M. VOGEL, P. SIMONET, R. NALIN. *Type I polyketide synthases may have evolved through horizontal gene transfer*, in "J Mol Evol", vol. 60, n° 6, Jun 2005, p. 716–725.

[20] L. GUEGUEN. *Sarment: Python modules for HMM analysis and partitioning of sequences*, in "Bioinformatics", vol. 21, n° 16, Aug 2005, p. 3427–3428.

[21] V. HUGUET, M. GOUY, P. NORMAND, J. F. ZIMPFER, M. P. FERNANDEZ. *Molecular phylogeny of Myricaceae: a reexamination of host-symbiont specificity*, in "Mol Phylogenet Evol", vol. 34, n° 3, Mar 2005, p. 557–568.

[22] C. S. ILIOPOULOS, J. MCHUGH, P. PETERLONGO, N. PISANTI, W. RYTTER, M.-F. SAGOT. *A first approach to finding common motifs with gaps*, in "International Journal of Foundations of Computer Science", vol. 16, n° 6, 2005, p. 1145-1154.

[23] P. KERSEY, L. BOWER, L. MORRIS, A. HORNE, R. PETRYSZAK, C. KANZ, A. KANAPIN, U. DAS, K. MICHOUD, I. PHAN, A. GATTIKER, T. KULIKOVA, N. FARUQUE, K. DUGGAN, P. MCLAREN, B. REIMHOLZ, L. DURET, S. PENEL, I. REUTER, R. APWEILER. *Integr8 and Genome Reviews: integrated views of complete genomes and proteomes*, in "Nucleic Acids Res", vol. 33, n° Database issue, Jan 2005, p. 297–302.

[24] A. KHELIFI, L. DURET, D. MOUCHIROUD. *HOPPSIGEN: a database of human and mouse processed pseudogenes*, in "Nucleic Acids Res", vol. 33, n° Database issue, Jan 2005, p. 59–66.

[25] E. LERAT, V. DAUBIN, H. OCHMAN, N. A. MORAN. *Evolutionary origins of genomic repertoires in bacteria*, in "PLoS Biol", vol. 3, n° 5, May 2005, e130.

[26] C. LOPES, A. PÉRY, A. CHAUMOT, S. CHARLES. *Ecotoxicology and Population DynamicsÊ: on the use of DEBtox models in a Leslie modeling approach*, in "Ecological Modelling", vol. 188, n° 1, 2005, p. 30-40.

[27] G. MARAIS, P. NOUVELLET, P. D. KEIGHTLEY, B. CHARLESWORTH. *Intron size and exon evolution in Drosophila*, in "Genetics", vol. 170, n° 1, May 2005, p. 481–485.

[28] J. MEUNIER, A. KHELIFI, V. NAVRATIL, L. DURET. *Homology-dependent methylation in primate repetitive DNA*, in "Proc Natl Acad Sci U S A", vol. 102, n° 15, Apr 2005, p. 5471–5476.

[29] A. MORGAT, E. COISSAC. *La gestion des connaissances biologiques*, in "Biofutur", vol. 251, 2005, p. 20–22.

[30] M. NICOLAS, G. MARAIS, V. HYKELOVA, B. JANOUSEK, V. LAPORTE, B. VYSKOT, D. MOUCHIROUD, I. NEGRUTIU, D. CHARLESWORTH, F. MONEGER. *A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants*, in "PLoS Biol", vol. 3, n° 1, Jan 2005.

[31] H. OCHMAN, V. DAUBIN, E. LERAT. *A bunch of fun-guys: the whole-genome view of yeast evolution*, in "Trends Genet", vol. 21, n° 1, Jan 2005, p. 1–3.

[32] H. OCHMAN, E. LERAT, V. DAUBIN. *Examining bacterial species under the specter of gene transfer and exchange*, in "Proc Natl Acad Sci U S A", vol. 102 Suppl 1, May 2005, p. 6595–6599.

[33] N. PISANTI, M. CROCHEMORE, R. GROSSI, M.-F. SAGOT. *Bases of Motifs for Generating Repeated Patterns with Wild Cards.*, in "IEEE/ACM Trans. Comput. Biology Bioinform.", vol. 2, n° 1, 2005, p. 40-50, http://doi.acm.org/10.1145/1057657.

[34] N. PISANTI, M.-F. SAGOT. *Network expression inference*, J. BERSTEL, D. PERRIN (editors). , chap. 4.7, Cambridge University Press, 2005, p. 227-250.

[35] C. RANQUET, A. TOUSSAINT, H. DE JONG, G. MAENHAUT-MICHEL, J. GEISELMANN. *Control of bacteriophage mu lysogenic repression*, in "J Mol Biol", vol. 353, n° 1, Oct 2005, p. 186–195.

[36] F. RECHENMANN, H. DE JONG. *Le vivant en équations*, in "La recherche", vol. 383, 2005, p. 31–37.

[37] F. RECHENMANN. *L'informatique n'est qu'un outil !*, in "Biofutur", vol. 251, 2005, 1.

[38] F. RECHENMANN. *Rechercher un gène dans une botte de lettres*, in "La recherche", vol. 388, 2005, p. 96–97.

[39] D. ROPERS, H. DE JONG, M. PAGE, D. SCHNEIDER, J. GEISELMANN. *Qualitative simulation of the carbon starvation response in Escherichia coli*, in "Biosystems", Nov 2005.

[40] M. SEMON, D. MOUCHIROUD, L. DURET. *Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance*, in "Hum Mol Genet", vol. 14, nº 3, Feb 2005, p. 421–427.

[41] E. TANNIER, A. BERGERON, M.-F. SAGOT. *Advances on sorting by reversals*, in "Discrete Applied Mathematics", vol. to appear, 2005.

[42] O. ZAGORDI, J. R. LOBRY. *Forcing reversibility in the no-strand-bias substitution model allows for the theoretical and practical identifiability of its 5 parameters from pairwise DNA sequence comparisons*, in "Gene", vol. 347, nº 2, Mar 2005, p. 175–182.

[43] H. DE JONG, J. GEISELMANN. *Genomic Signal Processing and Statistics*, J. CHEN, E. DOUGHERTY, I. SHMULEVICH, Z. WANG (editors). , chap. Modeling and simulation of genetic regulatory networks by ordinary differential equations, Hindawi Publishing Corporation, 2005, p. 201-239.

[44] H. DE JONG, D. ROPERS, C. CHAOUIYA, D. THIEFFRY. *Modélisation, analyse et simulation de réseaux de régulation génique*, in "Biofutur", vol. 252, 2005, p. 36–40.

[45] H. DE JONG, D. THIEFFRY. *Multiple Aspects of DNA and RNA: From Biophysics to Bioinformatics*, D. CHATENAY, S. COCCO, R. MONASSON, D. THIEFFRY, J. DALIBARD (editors). , chap. Modeling, analysis, and simulation of genetic regulatory networks: From differential equations to logical models, Elsevier, 2005, p. 325-351.

Publications in Conferences and Workshops

[46] J. ALLALI, M.-F. SAGOT. *A Multiple Graph Layers Model with Application to RNA Secondary Structures Comparison.*, in "SPIRE", Lecture Notes in Computer Science, vol. 3772, 2005, p. 348-359.

[47] G. BATT, D. ROPERS, H. DE JONG, J. GEISELMANN, R. MATEESCU, M. PAGE, D. SCHNEIDER. *Analysis and Verification of Qualitative Models of Genetic Regulatory Networks: A Model-Checking Approach.*, in "IJCAI", Professional Book Center, 2005, p. 370-375.

[48] G. BATT, D. ROPERS, H. DE JONG, J. GEISELMANN, R. MATEESCU, M. PAGE, D. SCHNEIDER. *Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in Escherichia coli*, in "Proceedings of the Workshop on Computation of Biochemical Pathways and Genetic Networks", 2005.

[49] G. BATT, D. ROPERS, H. DE JONG, J. GEISELMANN, R. MATEESCU, M. PAGE, D. SCHNEIDER. *Validation of qualitative models of genetic regulatory networks by model checking: Analysis of the nutritional stress response in Escherichia coli*, in "Proceedings of the Workshop on Dynamical Modeling and Analysis of Biological Regulatory Networks", 2005.

[50] G. BATT, D. ROPERS, H. DE JONG, J. GEISELMANN, M. PAGE, D. SCHNEIDER. *Qualitative Analysis and Verification of Hybrid Models of Genetic Regulatory Networks: Nutritional Stress Response in.*, in "HSCC", Lecture Notes in Computer Science, vol. 3414, 2005, p. 134-150.

[51] A. M. CARVALHO, A. T. FREITAS, A. L. OLIVEIRA, M.-F. SAGOT. *A highly scalable algorithm for*

the extraction of CIS-regulatory regions., in "Proceedings of 3rd Asia-Pacific Bioinformatics Conference", Imperial College Press, London, 2005, p. 273-282.

[52] R. CASEY, H. DE JONG, J.-L. GOUZÉ. *Stability of equilibria for piecewise-linear models of genetic regulatory networks*, in "to appear in the Proceedings of 44th IEEE Conference on Decision and Control (CDC) and European Control Conference (ECC)", 2005.

[53] V. LACROIX, C. G. FERNANDES, M.-F. SAGOT. *Reaction Motifs in Metabolic Networks.*, in "WABI", Lecture Notes in Computer Science, vol. 3692, 2005, p. 178-191.

[54] P. PETERLONGO, N. PISANTI, F. BOYER, M.-F. SAGOT. *Lossless Filter for Finding Long Multiple Approximate Repetitions Using a New Data Structure, the Bi-factor Array.*, in "SPIRE", Lecture Notes in Computer Science, vol. 3772, 2005, p. 179-190.

[55] M.-F. SAGOT, E. TANNIER. *Perfect Sorting by Reversals.*, in "COCOON", Lecture Notes in Computer Science, vol. 3595, 2005, p. 42-51.

[56] H. DE JONG, R. LIMA. *Modeling the dynamics of genetic regulatory networks: Continuous and discrete approaches*, in "Dynamics of coupled map lattices and of related spatially extended systems", J.-R. CHAZOTTES, B. FERNANDEZ (editors). , Lecture Notes in Physics, n° 671, Springer-Verlag, 2005, p. 307-340.