INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team HELIX

Informatics and genomics

Rhône-Alpes

THEME 3A

Table of contents

1. Team **1**
2. 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 5
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 7
3.6. Modelling and simulation of genetic regulatory networks 8
3.7. Bioanalysis and cross-sectional activities 9
4. Software **9**
4.1. Box 9
4.2. FactorTree 9
4.3. FamFetch 10
4.4. GenoExpertBacteria (GEB) 10
4.5. GenoStar 10
4.6. GeM 10
4.7. Genetic Network Analyzer (GNA) 11
4.8. Herbs 11
4.9. Hogenom and Hovergen 11
4.10. Hoppsigen 11
4.11. Identitag 11
4.12. ISee 12
4.13. Mentalign 12
4.14. MicrOBI 12
4.15. Migal 12
4.16. Oriloc 12
4.17. PEPLINE 13
4.18. PhyloJava 13
4.19. Satellites 13
4.20. seqinR 13
4.21. Smile and Riso 13
4.22. Utopia 13
4.23. Other software developed in HELIX 14
5. New Results **14**
5.1. Computational analysis of the evolution of species and gene families 14
5.2. Modelling and analysis of the spatial organisation and dynamics of genomes 15
5.3. Motif search and inference 16
5.4. Computational proteomics and transcriptomics 17
5.5. Modelling and analysis of metabolism: molecular components, regulation, and pathways 17
5.6. Modelling and simulation of genetic regulatory networks 18
5.7. Bioanalysis and cross-sectional activities 18
6. Contracts and Grants with Industry **19**
6.1. Aventis-Pasteur 19
6.2. Genostar 19
6.3. Sanofi 19
7. Other Grants and Activities **19**

7.1. National projects 19
7.2. Projects funded by international organisms or including international teams 21
8. Dissemination **22**
8.1. Talks 22
8.2. Organisation of conferences, workshops and meetings 25
8.3. Editorial and reviewing activities 26
8.4. Administrative activities 27
8.5. Teaching activities 27
9. Bibliography **29**

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 François Rechenmann, 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 created in September 2004 and 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.

Head of project

François Rechenmann [Research director, INRIA]

Administrative assistant

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

Permanent researchers and professors

Vincent Daubin [Research associate, CNRS, since October 2004]
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 associate, INRIA]
Jean Lobry [Associate professor, Université Claude Bernard]
Gabriel Marais [Research associate, CNRS, since October 2004]
Dominique Mouchiroud [Professor, Université Claude Bernard]
Sylvain Mousset [Associate professor, Université Claude Bernard, since October 2004]
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, since October 2004]
Alain Viari [Research director, INRIA]
Danielle Ziébelin [Associate professor, Université Joseph Fourier]

Permanent technical staff

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

Software and system engineers

Pierre-Emmanuel Ciron [Project technical staff, INRIA, until September 2004]
Gilles Faucherand [Project technical staff, INRIA]
Agnès Iltis [Project technical staff, INRIA]
Timothée Silvestre [Project technical staff, CNRS, until July 2004]
Erik Wessel [Junior technical staff, INRIA]

External members

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

Visitors

Cristina Gomes Fernandes [Associate professor, University of São Paulo, Brazil, 6 weeks]
Katia Guimarães [Associate professor, Federal University of Pernanbuco, Brazil, 3 weeks]

Post-doctoral fellows

Abdel Aouacheria [Association pour la Recherche sur le Cancer]
Frédéric Boyer [INRIA, Aventis Pasteur, since October 2004]
Simon Penel [European contract]
Delphine Ropers [INRIA]
Eric Tannier [INRIA, until September 2004]

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]
Frédéric Boyer [scholarship Ministère de la Recherche and ATER Université Joseph Fourier, supervisors: Laurent Trilling, Alain Viari. Post-doctorate since October 2004]
Alexandra Calteau [scholarship Ministère de la Recherche, supervisors: Guy Perrière, Manolo Gouy]
Stéphane Descorps-Declère [CIFRE convention, GENOME express, supervisors: Alain Viari, Pierre Netter, Université Paris 6]
Samuel Drulhe [scholarship ENS Cachan, supervisors: Hidde de Jong, François Rechenmann]
Jean-François Dufayard [scholarship Ministère de la Recherche, supervisors: Manolo Gouy, François Rechenmann. ATER Université Claude Bernard since October 2004]
Adel Khelifi [scholarship Ministère de la Recherche, supervisor: Dominique Mouchiroud]
Vincent Lacroix [scholarship BDI, CNRS, supervisor: Marie-France Sagot]
Christelle Melo de Lima [scholarship Ministère de la Recherche, supervisors: Didier Piau, François Rechenmann]
Julien Meunier [scholarship ENS Lyon, supervisor: Laurent Duret]
Vincent Navratil [scholarship INRA, supervisor: Christian Gautier, Denis Milan, INRA]
Leonor Palmeira [scholarship Ministère de la Recherche, supervisors: Laurent Guéguen, Jean Lobry]
Mi Phi Long Nguyen [assistant professor at HoChiMinh, supervisor: Christian Gautier]
Marie Semon [scholarship Ministère de la Recherche, supervisor: Laurent Duret]

2. Overall Objectives

The information necessary to the development and the maintenance of a living organism is mainly contained in its genome, materialised within each cell by one or more macromolecules of DNA. This molecule is an oriented linear chain of four different types of nucleic acids symbolised by the letters, A, C, G, and T. The information content of a genome can thus, as a first approximation, be represented as a text on a four-letter alphabet.

More than a hundred and sixty genomes have already been fully sequenced, among which around twenty of eukaryotes including man and mouse. The length of this "text" varies from a few million letters to some three billion for *Homo sapiens*. 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 structure, the modification of metabolic pathways, the preservation and variation of signalling systems, *etc.* (Figure 1).

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.

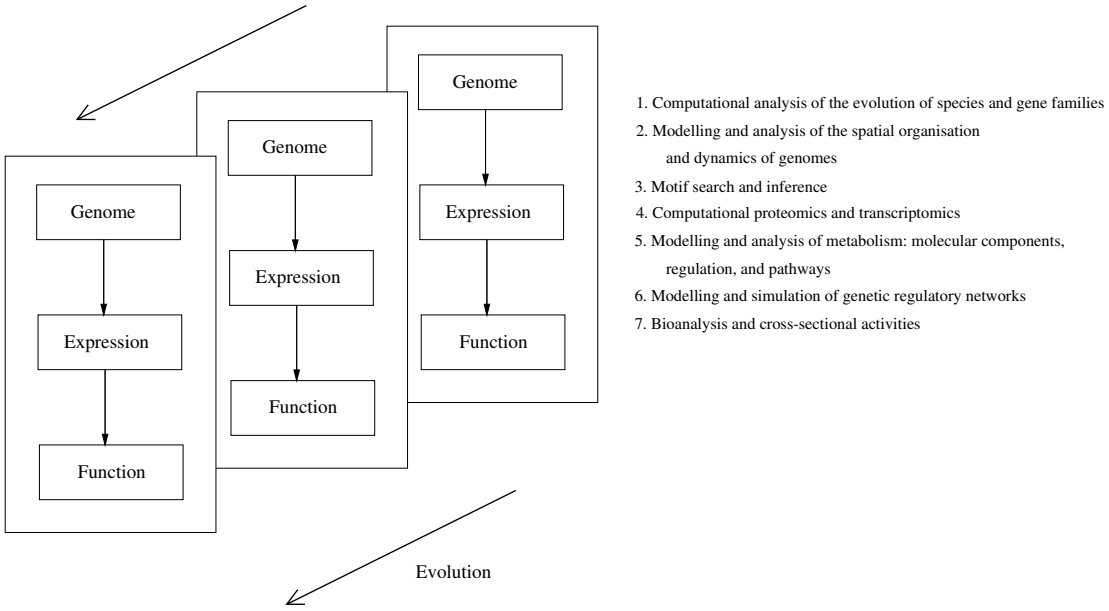

Figure 1. Biology from a double genomic and evolutionary perspective. The research areas of HELIX concern the development of methods and computer tools to study different aspects of this double perpective.

The overall objective of HELIX is to develop methods and computer tools to help biologists represent, access, and analyse the huge amounts of data available today. 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 4.5).
- 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, Alexandra Calteau, Jean-François Dufayard, Laurent Duret, Christian Gautier, Manolo Gouy [Correspondent], Laurent Guéguen, Jean Lobry, Julien Meunier, Dominique Mouchiroud, 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. Two topics stand out in particular:

- *Reconstruction of the Tree of Life.* The evolutionary distance between genomes increases with time since speciation. This makes it possible to estimate the topology of the Tree of Life and the distances along its edges.
- *Genome annotation.* The information carried by the genomes of different organisms is quite diverse in nature and varies with the functions that are associated with it and with the way this information is expressed. The process of error fixation, even the process of error itself, depends on these two aspects and leaves traces in the genome that are comprehensible only in the light of evolution.

In an often arbitrary way, biology divides the study of evolution into the study of *patterns* and the study of *processes*. This means that the reconstruction of the Tree of Life (the pattern) is separated from the evolutionary mechanisms themselves (the processes). We have adopted this approach for convenience, even though reconstructing the Tree implies understanding all the mechanisms underlying evolution.

Genomic or protein sequences have the same format, whatever organism they belong to. Therefore, their comparison allows, *a priori*, to reconstruct the whole of the Tree of Life. However, the mathematical complexity of the processes involved requires methods for approximate estimation. Sequences are not the only source of information available for reconstructing a phylogenetic tree. 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 feasible) operations enabling one permutation to succeed another. Sophisticated algorithms are required to deal with the problem.

Phylogenetic reconstruction based on the sequence of a single gene in different organisms leads to a tree. Currently, 6000 families of genes having more than 4 specimens are known, and hence 6000 different trees can be reconstructed. The management and comparison of these trees is a computational and mathematical problem that requires the expertise of the three groups composing the HELIX project and is a good example of their collaboration.

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

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

The founding groups of the HELIX project participated very early (more than 20 years ago) in the discovery of strong biological and statistical heterogeneities in genomic sequences. The modelling and analysis of the *spatial organisation and dynamics of genomes* has continued to be one of its most active research areas. 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 (illustrations of this are given at http://pbil.univ-lyon1.fr/software/Oriloc/). 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, *etc.*).

In the course of evolution, the spatial organisation of a genome undergoes other 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 (see Section 2.1) 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 (in the case of bacteria, the notion of synteny refers to a group of orthologous genes whose spatial organisation is conserved between two species). This may require a better understanding of the biological processes underlying such long-range changes in genome structure, which in turn may be enabled, or at least facilitated, by analysing what is happening around the regions that are subject to breakage and rearrangement. In order to achieve this, one has to be able to precisely locate such regions, which is already a non-trivial problem. Moreover, one is confronted with the problem of finding information without having a clue about the form the information might assume.

Modelling and analysing genomic maps requires expertise in various areas such as knowledge representation, statistics, and algorithmics. 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.

3.3. Motif search and inference

Participants: Christian Gautier, Laurent Gueguen, 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:

1. *Feature identification*, which aims at finding and precisely mapping the main features of a genome: protein or RNA-coding genes, DNA or RNA sequence or structure signals, satellites (tandem repeats) or transposable elements (dispersed repeats with a specific structure), regulatory regions, *etc*.
2. *Relational identification*, the goal of which consists in finding relations existing 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 or little-known objects. The HELIX project is interested in all problems within this continuum, most of which have been unsatisfyingly addressed up to now. This is essentially due to the fact that features and the relations holding between them should in general be inferred simultaneously. 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 require a higher degree of abstraction, and more sophisticated algorithms or statistical approaches.

Our studies concerning problems within the continuum are driven by evolutionary issues in the sense that weak or strong selective pressures acting upon the various important (*i.e.* functional) features in a genome or in a biological process will leave an imprint that may, in some cases, be identified by comparison inside or between different genomes or organisms. This imprint is often modelled in terms of features, or observed relations between features, that occur with an unexpectedly-high or low frequency, or in a very regular fashion.

Various search and inference methods have already been developed by HELIX. These include methods for DNA and protein sequence motifs inference (SMILE and RISO, Section 4.21), gene finding (UTOPIA, Section 4.22; EMKOV, Section 4.23), satellite identification (SATELLITES, Section 4.19), RNA common substructure inference (MIGAL, Section 4.15), *etc.*

3.4. Computational proteomics and transcriptomics

Participants: Laurent Duret, Guy Perrière [Correspondent], Dominique Mouchiroud, Alain Viari [Correspondent].

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. Two main approaches towards protein identification by means of mass spectrometry can be distinguished.

1. The *MS approach* consists in weighing each of the ions obtained by digesting the protein being studied by an enzyme, usually trypsine. The results can be plotted in a mass spectrum, representing a fingerprint of the protein.

2. The *MS/MS* or *tandem MS approach* carries on the MS approach by fragmenting every trypsin ion in the mass spectrum and analysing the resulting peptide fragments by mass spectrometry. From these measurements, so-called *peptide sequence tags (PSTs)* are generated. PSTs are short peptide sequences flanked by two polypeptides of known, measured mass, which can be used to identify the protein.

State-of-the-art mass spectrometers produce large volumes of data. For example, in some experimental settings, the tandem MS approach can yield up to 1500 peptides per day. It is obvious that with these amounts of data the interpretation of the mass spectra 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" at the CEA in Grenoble. The aim of the collaboration is to develop computer tools for the analysis of data produced by the tandem MS approach. In particular, efficient algorithms have been designed for generating PSTs from tandem 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 called PEPLINE (Section 4.17) and installed at the LCP.

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:

1. *direct sampling* (and quantification) of sequences from source RNA populations or cDNA libraries (the most common techniques of this type are ESTs and SAGE);
2. *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. The improvement of this process of *tag-to-gene mapping* has been addressed in the context of the development of the tool IDENTITAG (Section 4.11). 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, Stéphane Bruley, Vincent Lacroix, Anne Morgat [Correspondent], Marie-France Sagot [Correspondent], Alain Viari [Correspondent], Erik Wessel.

Advanced experimental techniques in genomics have produced huge amounts of data on the molecular basis of cellular processes. Such data are quite heterogeneous, including among other things the genomic sequences of organisms as well as information on the organisation of a genome into operons, the regulation of these operons, the structure and function of the proteins encoded by the genes in an operon, and the temporal evolution of the concentration of a protein in response to an environmental stress. Moreover, in the case of metabolism, additional information concerning chemical transformations (biochemical reactions catalysed by enzymes) is also available and should be taken into account. 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 genomic data gathered over the past few decades are usually dispersed in the literature and are therefore difficult to exploit for answering questions of the kind mentioned above. A major contribution of bioinformatics has been 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. The development of such models has been carried out in the context of two complementary systems. The relational database (MICROBI) stores data from, and maintains the consistency between, several public resources devoted to microorganisms (Section 4.14), while GENOEXPERTBACTERIA (GEB), an object-oriented system connected to MICROBI, allows easy browsing and querying of the raw data contained in the latter (Section 4.4). Both systems allow access to:

1. *Genetic elements* (genes, signals, operons, *etc.*);
2. *Proteins* (post-translational modifications, protein complexes, catalytic activities, *etc.*);
3. *Intermediate metabolism* (reactions, substrates, products, *etc.*).

MICROBI and GEB provide a rich source of information to analyse the chromosomal organisation and metabolism of bacteria. This idea is being explored in several directions. First, the metabolic pathways of newly sequenced organisms can be reconstructed from the enzymatic reactions in the knowledge base. Instead of the classical approach towards pathway reconstruction, which checks whether already-characterized pathways occur in the new organism, an alternative approach, called *ab initio* reconstruction, is followed. This approach consists in finding hypothetical metabolic pathways connecting a set of given compounds, using no other information than a set of reactions (and the chemical compounds they involve).

A second use of the information accessible through MICROBI and GEB consists in systematically exploring the metabolic pathways, in order to find motifs and modules preserved by evolution, which may be an indication that they have a particular biological function. Most of the work in this area consists of explorations of the possible definitions of motifs, as well as studies of the complexity of algorithms for searching known motifs and inferring new motifs. The question of the statistical meaning of the motifs that are started to be found is also being addressed. The particular nature of metabolic networks, which is quite different from other known networks (for instance, protein interaction networks or the WWW network), raises non-trivial and interesting problems in graph theory.

3.6. Modelling and simulation of genetic regulatory networks

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

It is now commonly accepted that most interesting properties of an organism emerge from the interactions between its genes, proteins, metabolites, and other constituents. This implies that, in order to understand 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 elucidated.

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. 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.

In addition to 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 4.7). 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, Stéphane Descorps-Declère, Christian Gautier [Correspondent], Dominique Mouchiroud, Vincent Navratil, François Rechenmann [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. 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). For more information, see http://oriel.inrialpes.fr:8080/box. BOX was developed with Antoine Brun.

4.2. 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 at the University of Marne-la-Vallée. 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).

4.3. FamFetch

Participants: Jean-François Dufayard, Laurent Duret, Manolo Gouy, Simon Penel, Guy Perrière [Correspondent].

FAMFETCH (http://pbil.univ-lyon1.fr/software/famfetch.html) is a set of tools to search for tree patterns in databases of phylogenetic trees.

4.4. 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 (originating from an earlier work on the system Panoramix) and a graphical user interface facilitating the exploration and analysis of the available data. GEB can be run as a stand-alone application or as a GENOSTAR module. In this latter case, GEB can exchange data and results with the other modules of the GENOSTAR environment (Section 4.5). For more information, see http://www-geb.inrialpes.fr/.

4.5. 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 second version (1.2) of GENOSTAR is made up of three application modules: GENOANNOT, GENOLINK, and GENOBOOL which can easily exchange data. GENOANNOT relies on several sequence analysis methods to perform the syntaxic annotation of bacterial genomes. It produces predictions on the position of genes and other pertinent features. By allowing biologists to browse through a network of biological entities and bioinformatics objects, GENOLINK helps them in understanding the function of genes. The links of a network represent different relationships between the entities and the set of their types is easily extendable. GENOBOOL offers several data analysis methods which can be applied to heterogeneous sets of data after they have been adequately coded, for example using boolean coders. GENOBOOL thus allows the user to discover new relationships between properties of biological entities.

However, the attractiveness of GENOSTAR for genomic data analysis goes far beyond the capabilities of these three modules. Its very architecture offers several original features which help the biologist in applying complex and exploratory analysis tasks. For more information, see http://www.genostar.org/english/index.html.

4.6. GeM

Participants: Gisèle Bronner [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 (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/).

4.7. 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. For more information, see http://www-helix.inrialpes.fr/gna.

4.8. Herbs

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

HERBS (HAMAP EXPERT RULE BASED SYSTEM) provides computer support for the reannotation of complete bacterial genomes. 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. This means that it analyses the metabolic pathways and warns the user of 'missing', 'unexpected', 'ambiguous', and 'normal' proteins. 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.

4.9. Hogenom and Hovergen

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

HOGENOM (http://pbil.univ-lyon1.fr/databases/hogenom.html) 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.

4.10. Hoppsigen

Participants: Adel Khelifi, Dominique Mouchiroud [Correspondent].

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

4.11. 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/.

4.12. 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. The current, second version of ISEE has several additional functionalities:

- Specification of the experimental conditions (input data, parameter values), which are saved in the left pages of the lab notebook;
- Management of multiple experiments, notably execution of algorithms under different experimental conditions and comparison of the results.

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".

For more information, see http://www-helix.inrialpes.fr/article124.html.

4.13. 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.

4.14. 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 4.4), but can also be used as a stand-alone database.

4.15. 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).

4.16. Oriloc

Participant: Jean Lobry [Correspondent].

ORILOC (http://pbil.univ-lyon1.fr/software/oriloc.html) 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.

4.17. 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. For more information, see http://www-helix.inrialpes.fr/article228.html. PEPLINE was developed with Estelle Nugues and Erwan Reguer.

4.18. 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.

4.19. 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@inria.fr.

4.20. 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. For more information, see http://pbil.univ-lyon1.fr/software/SeqinR/seqinr_home.php. SEQINR was developed by Delphine Charif during her DESS.

4.21. 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.

4.22. 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@inria.fr.

4.23. Other software developed in HELIX

Participants: François Rechenmann [Correspondent], Manolo Gouy [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 4.3);
- 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 (A. 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

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

Building accurate phylogenetic trees is important for studying the evolution of genomes and gene families (Section 3.1). Obtaining good alignments is a required first step, but severely constrained by the fact that existing alignment and tree-building algorithms are not appropriate for dealing with large families of homologous sequences (that is, consisting of more than a few thousand sequences), because full computation of the alignment and of the tree is not feasible. This presents a problem of practical relevance, since several known families of homologous sequences contain tens of thousand of sequences.

The problem of reconstructing phylogenetic trees of large families of homologous sequences has been addressed in the PhD research of Jean-François Dufayard (defended in December 2004). In particular, he has proposed an incremental algorithm for the multiple alignment and the computation of phylogenies. In this approach, every time when a new sequence is added to a pre-aligned family, the alignment and the tree are modified rather than fully recomputed. The algorithm has been implemented in a user-friendly computer tool called MENTALIGN (Section 4.13), which has been tested on several families of homologous sequences, such as the cytochrome B family. Notably, it allows the alignment of a family with 16,000 members of 1000-character long sequences. A related problem is being studied in the PhD of Anne-Muriel Aragon. The aim of her thesis is to develop a method that identifies the family to which a given sequence belongs, and then aligns it with this family. She will later study the influence of the addition of one or more new sequences to an alignment on the topology of the resulting phylogenetic tree.

Phylogenetic methods have been used by Alexandra Calteau in her PhD on studying *Horizontal Gene Transfers (HGT)* in prokaryotic organisms. The work is motivated by the recent discovery of large transfers occurring in hyperthermophilic bacteria, which are thought to have undergone a large number of HGTs: up to 16% in *Aquifex aeolicus* and 24% in *Thermotoga maritima*. Notably, three putative horizontal transfers between bacterial and archaeal species involving large clusters of genes have been reported recently [27]. For example, an operon of 13 genes, called *mbx*, was probably transferred into the genome of *T. maritima* from a species belonging or close to the *Pyrococcus* genus. All transfers affected operons coding for multisubunits membrane-bound [NiFe]-hydrogenases involved in the energy metabolism of the donor genomes. The existence of such large transfers raises the question which biological mechanisms might have led to their apparition (Section 3.1).

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

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

Isochores (Section 3.1) constitute a feature of the spatial organisation of genomes that has been studied for a long time. Our recent work has led to the important insights that G+C-isochores in mammals are not at evolutionary equilibrium, but tend to disappear [5]. Moreover, we have shown that isochores are likely created by a phenomenon called *biased gene conversion (BGC)*, which tends to increase the G+C content of sequences located in regions of high recombination rate [13]. Another example of a study of the structure and evolution of DNA sequences is the demonstration, by means of the analysis of the distribution of transposable elements in the human genome, that many of the transcription units covering at least 50% of the genome do not correspond to protein-coding genes [18]. In addition, new results on the development of novel functions through insertion events have been presented [6].

The spatial organization and dynamics of genomes have also been studied from a more algorithmic point of view, using probabilistic models, in particular *Hidden Markov Models (HMM)* and *Bayesian analysis*. In the context of her PhD thesis, Christelle Gonindard has built HMM models of whole human chromosomes in order to model isochores classes. The segmentation of chromosomes by isochores classes thus obtained is quite efficient. HMMs and Bayesian analysis will also be used for studying and predicting the influence of neighbour-dependent mutations in all species for which the genome has already been sequenced. This will be the subject of the PhD of Leonor Palmeira. All segmentation algorithms developed are currently been made accessible as a library of Python modules through the PRABI.

Vertebrate genomes are mostly composed of non-coding sequences (95%), many of which are repetitive (60%). There are several classes of repetitive sequences, but for his PhD, Adel Khelifi has been mostly interested in studying *processed pseudogenes*. 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 RNA and translated into proteins makes these genes useful for molecular evolution, in particular for studying substitution patterns. This approach has been applied to the human and mouse genomes, giving rise to a database of annonated pseudogenes, called HOPPSIGEN (Section 4.10), which has been the subject of a recently-accepted paper. This work is being pursued with a study of the impact of recombination on the evolution of processed pseudogenes.

Various initial theoretical and practical results have been obtained concerning the dynamics of genomes. The work of HELIX has particularly focused on the analysis of large-scale transformations in mammalian genomes: how to determine long conserved regions in genomes, and how to detect breakpoint regions with the objective of constructing and evaluating models of large-scale evolution. In the framework of a MSc thesis of a student from the University of Marne-la-Vallée, two competing models on the occurrence and fixation of rearrangements have been tested. In addition, methods for precisely locating the breakpoints along a genome and exploring local characteristics of the surrounding regions have been developed. The above work has led

to new algorithmic results [37] and has suggested that existing mathematical methods for deriving optimal rearrangement scenarios and/or detecting conserved segments may be misguided, raising interesting questions for further work.

5.3. Motif search and inference

Participants: Stéphane Descorps-Declère, Christian Gautier, Laurent Gueguen, Vincent Lacroix, Christelle Melo de Lima, Leonor Palmeira, Guy Perrière, Marie-France Sagot [Correspondent], Alain Viari.

Inferring sequence or structural motifs from DNA or protein sequences is one of the oldest research areas in computational biology, and the HELIX project has already much contributed to it. Despite its long history, many fundamental problems remain. HELIX has therefore continued its research efforts in this area, concentrating on the questioning of biological and mathematical models used for inferring motifs, and on the improvement of existing algorithms. The work has also been extended to protein structures.

The HELIX project remains one of the rare research groups that are trying to tackle the inference of *structured motifs*, that is, motifs composed of several parts separated by non-random distances. The implementation of the algorithm developed to find these motifs, SMILE, has been widely distributed in an informal way. It is currently undergoing some major performance improvements in the context of the PhD of Alexandra Carvalho, who is co-supervised by Arlindo Oliveira at the Instituto Superior Técnico (IST) of Lisbon, Portugal. Other collaborators include Laurent Marsan from the Université de Versailles and Nadia Pisanti from the University of Pisa. The collaboration has led to two accepted publications this year [29][28] and to a prototype implementation, called RISO (Section 4.21), which is being used for the analysis of yeast data.

Motifs have continued to be studied from a more theoretical point of view as well, in particular following up on previous work in HELIX. The results concern the notion of a basis of motifs [15] as well as variously constrained versions of motifs with gaps [31] and functionally equivalent blocks [30]. We focus here on a special form of motifs that have an indirect usefulness for two other ancient areas of research in computational biology: *gene finding* and *repeat identification*. The motifs are employed for filtering purposes in sequence analysis, in particular in the context of local alignment. The filters have been based on the idea that similar regions must contain some common exact factors. We are currently working on filters for multiple sequence comparison, which generalizes on existing approaches in the literature. This work is being carried out with Pierre Peterlongo, a PhD student co-supervised by Maxime Crochemore, and in collaboration with Costas Iliopoulos and Nadia Pisanti. It has led to a first publication [35]. Gene finding itself has been addressed this year in the context of the PhD (to be defended in early 2005) of a Brazilian student, Said Sadique Adi, in co-supervision with Carlos Eduardo Ferreira from the University of São Paulo, Brazil.

RNA motifs form another type of functional element in genomes. The role of small RNAs in regulation has been recently shown to be far more prominent than initially believed. Yet current mathematical and computer tools remain mostly inadequate to identify, analyse, and compare RNAs. In the context of the PhD thesis of Julien Allali (defended in December 2004), carried out at the Université of Marne-la-Vallée in co-supervision with Maxime Crochemore, an algorithm for inferring all sub-structures common to two RNA structures has been developed. A prototype implementation of the algorithm, MIGAL (Section 4.15), is now publicly available. It uses a biologically more appropriate distance than simple edit distance between trees, and performs the comparison at different, increasingly more constrained levels of granularity. A first paper on this work has been published [24] and was selected for publication in an extended form in *IEEE/ACM Transactions in Computational Biology and Bioinformatics*. Other work in progress, in collaboration with Stéphane Vialette (Université d'Orsay) and Christine Gaspin and Thomas Schiex (INRA, Toulouse), concerns the inference of RNA common sub-structures from RNA sequences, also called *RNA structural motifs*.

Finally, the problem of *structural motifs in proteins* is being addressed in the context of the PhD thesis (to be defended in early 2005) of another Brazilian student, Jeane C.B. de Melo, in co-supervision with Katia Guimarães, from the Federal University of Pernanbuco, Brazil. The objective is to see whether it is possible to arrive at cleaner definitions of protein domains and to develop automatic methods for their determination.

Recently, two algorithms have been implemented that will be be used to compare the results obtained with manually-and automatically-identified domains. This work is carried out in collaboration with Joël Pothier (Université Paris VI) and with Jean-François Gibrat (INRA, Jouy-en-Josas).

5.4. Computational proteomics and transcriptomics

Participants: Laurent Duret, Guy Perrière [Correspondent], Dominique Mouchiroud, Alain Viari [Correspondent].

In order to provide an answer to the problem of unambiguously assigning tags to transcripts in the SAGE approach discussed in Section 3.4, a tool for tag identification, called IDENTITAG (Section 4.11), has been developed. The tool is based on a relational database structure in order to allow rapid and easy storage and updating of data and, most importantly, in order to be able to finetune the parameters of the identification process. IDENTITAG connects an observed tag to a virtual tag and the sequence corresponding to the latter. Databases for different species can be connected according to orthology relationships, thus allowing the comparison of SAGE libraries between species. This opens up the possibility to apply IDENTITAG to problems in comparative transcriptomic analysis, an emerging field in biology [12]. IDENTITAG has been successfully used to identify SAGE tags in two SAGE libraries constructed in order to characterize the molecular basis underlying self-renewal versus differentiation decision-making processes in primary, non-immortalized erythroid avian progenitor cells (T2EC cells).

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

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

In the framework of the PhD thesis of Frédéric Boyer, a new approach towards *ab initio metabolic pathway reconstruction* has been proposed. 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 order to resolve it, an algorithm to construct an automaton accepting exactly all words corresponding to maximal compositions of partial injections from a source to a sink compound has been developed. This algorithm has been successfully applied to the reconstruction of standard metabolic pathways with acceptable running-times.

The second part of the thesis deals with the question of relating metabolic to genomic information (*e.g.*, how genes involved in the catalysis of adjacent biochemical reactions are co-localized on a chromosome). To this purpose, Frédéric Boyer has extended previous graph-theoretical approach towards finding bacterial syntenies developed in HELIX. In particular, he has developed an algorithm that integrates several graphs representing different types of metabolic and genomic information into a single graph that can be analyzed for patterns shared by all of the original graphs. This makes it possible to answer simple but fundamental biological questions like the one mentioned above. 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).

In a PhD thesis started this year, Vincent Lacroix is addressing the problem of searching for and inferring *motifs and modules in metabolic networks* (Section 3.5). Three increasingly more constrained definitions of a motif have been established in an initial step, with flexible notions of similarity introduced at different levels. Two of these definitions have been explored, essentially in the context of a search problem for the time being. A first prototypal algorithm has been implemented and applied to KEGG's metabolic network (which is an assembly of the networks of different organisms) in the context of: (1) the search for a motif known from the literature to occur in at least three different metabolic pathways, and (2) an exhaustive analysis of small motifs

(comprising two or three reactions) for statistical purposes. Work on the computation of statistical properties of motifs is being conducted in a collaboration with Sophie Schbath (INRA, Jouy-en-Josas) and Stéphane Robin (InaPG, Paris). The efficient search, inference, and clustering of motifs in networks represent other difficult problems that are being addressed in a collaboration with the combinatorics group of the University of São Paulo, Brazil.

5.6. Modelling and simulation of genetic regulatory networks

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

This year has seen the completion of the first phase of the development of the method for the qualitative modelling and simulation of genetic regulatory networks (Section 3.6). The mathematical and computational basis of the method has been published [22], while version 5.5 of the accompanying computer tool GENETIC NETWORK ANALYZER (GNA) (Section 4.7), having a strongly-improved user interface, has been released in the summer. At the time of writing, more than 50 groups have asked a copy of the new version of the program and the paper on GNA is among the most downloaded in the journal *Bioinformatics* over the last five years (http://www3.oup.co.uk/jnls/list/cabios/special/15/default.html). The transfer of GNA to an industrial partner is currently in preparation. The above work has been summarized by Hidde de Jong at the occasion of his public defense for obtaining a 'Habilitation à diriger des recherches'

In the wake of the application of the method and the tool to a paradigm example of prokaryote development [20], work has continued on the analysis of systems that are currently less understood by biologists. Delphine Ropers, in the framework of her post-doctorate at INRIA, has developed a model of the nutritional stress response in *Escherichia coli*, in collaboration with experimental biologists in the laboratory of Johannes Geiselmann (Université Joseph Fourier, Grenoble) and a student from the INSA in Lyon [43]. We are currently in the process of experimentally validating the predictions of the model. In collaboration with Eric Coissac and Jean Houmard (École Normale Supérieure, Paris), a student of Uppsala University has developed a model of the stimulation of motility by light in the cyanobacterium *Synechocystis* PCC 6803.

On the methodological side, we have continued to work on the mathematical characterisation of attractors of piecewise-linear differential equation models of genetic regulatory networks, the class of models employed by the qualitative simulation method. This research is carried out in collaboration with Richard Casey and Jean-Luc Gouzé (INRIA Sophia-Antipolis). A paper describing the results has been submitted and is also available as an INRIA technical report [41]. In parallel, Michel Page and Hidde de Jong are developing a new module of GNA for the search of equilibrium points and the determination of their stability, exploiting among other things the results of the above-mentioned mathematical study.

Another extension is the subject of the PhD research of Grégory Batt, who focuses on the validation of qualitative models of genetic regulatory networks by means of gene expression data. In order to deal with the validation problem, an approach based on model-checking techniques has been chosen and is currently pursued in collaboration with Radu Mateescu and his colleagues of the VASY project. Much of the work this year has concerned the development of a refined version of the method for qualitative simulation, better adapted to the available experimental data. The results, inspired by work on the analysis of hybrid systems through discrete abstractions, have been accepted for publication and are available in extended form [40]. Furthermore, an exploratory study of the coupling between GNA and the model checker CADP developed in the VASY project has been presented at the *SPIN* conference [25].

A quite recent methodological extension is the work on the identification of genetic regulatory networks by means of gene expression data, carried out in the framework of the PhD thesis of Samuel Drulhe. In collaboration with Giancarlo Ferrari-Trecate (INRIA Rocquencourt), the use of methods for the identification of piecewise-linear differential equation models is currently being explored.

5.7. Bioanalysis and cross-sectional activities

Participants: Abdel Aouacheria, Stéphane Descorps-Declère, Christian Gautier [Correspondent], Dominique Mouchiroud, Vincent Navratil, François Rechenmann [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 the context of his postdoc, funded by the "Association pour la Recherche contre le Cancer" and supervised by Manolo Gouy, Abdel Aouacheria is working on an *in silico* screening approach for proposing new candidate genes and polymorphisms associated with a carcinogenic tumoral phenotype. In particular, he aims also at elucidating the origin and evolution of the apoptose program in *Metazoa* using a comparative approach of the proteins of the Bcl-2 family and modelling the development of the process of apoptosis under physiological and pathological conditions.

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. Today, a multitude of sequence analysis exists, which vary in the nature of the signals they detect (ribosome binding sites, promoters, operators, coding sequences, *etc.*) as well as in the particular identification procedure they use. The blackboard architecture proposed by Stéphane Descorps-Declère provides a hierarchical approach towards the storage of the results obtained by means of different sequence analysis methods and their synthesis on a higher level of abstraction.

6. Contracts and Grants with Industry

6.1. Aventis-Pasteur

Participants: Frédéric Boyer, 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.

6.2. Genostar

Participants: 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	Validation de modèles de réseaux de régulation génique : Régulation globale de la transcription chez *Escherichia coli* et *Synechocystis* PCC 6803
Coordinators	H. de Jong, J. Geiselmann
HELIX participants	G. Batt, H. de Jong, A. Morgat, M. Page, D. Ropers, M.-F. Sagot
Type	inter-EPST Bioinformatique (2002-2004)

Project name	Comparative genomics
Coordinators	T. Faraud, D. Mouchiroud
HELIX participants	C. Gautier, L. Duret, D. Mouchiroud, L. Guéguen, A. Morgat, F. Rechenmann, V. Navratil, B. Spataro, M.-F. Sagot
Type	inter-EPST Bioinformatique (2002-2004)
Web page	

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	GDyn : Analyse dynamique de réseaux de régulation génique
Coordinators	J.-L. Gouzé, H. de Jong
HELIX participants	H. de Jong, M. Page
Type	ARC INRIA (2002-2004)
Web page	http://www-sop.inria.fr/comore/arcgdyn/arcgdyn-eng.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	

7.2. Projects funded by international organisms or including international teams

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	Algorithms for Modelling, Search and Inference Problems in Molecular Biology
Coordinators	M.-F. Sagot
HELIX participants	almost all members of HELIX and six European partners
Type	inter-EPST Bioinformatique (2002-2004)

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/projects/brazil_capes_2004/brazil_capes_2004.html

Project name	Oriel
Coordinators	European Molecular Biology Organisation (EMBO)
HELIX participants	A. Brun, A. Viari
Type	European Commission as ORIEL, contract no. IST-2001-32688, under Key Action 3 of the IST Programme (Multimedia Content and Tools)
Web page	http://www.oriel.org/

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

Abdel Aouacheria

Title	Event and location	Date
Recherche des origines évolutives et de nouveaux modulateurs du programme d'apoptose : Étude de la famille Bcl-2	Laboratoire de RMN de la Matière Condensée, Université Louis Pasteur, Strasbourg	07/07/04

Grégory Batt

Title	Event and location	Date
Vérification de modèles continus de réseaux de régulation génique à l'aide de techniques de model checking : Application à la réponse aux stress nutritionnels chez *Escherichia coli*	Groupe de travail VICANNE, Evry	13/09/04
Qualitative modeling and simulation of genetic regulatory networks	Sixteenth International Symposium on Mathematical Theory of Networks and Systems (MTNS2004), Leuven, Belgium	07/07/04
Model checking genetic regulatory networks using GNA and CADP	11th International SPIN Workshop on Model Checking of Software (SPIN2004), Barcelona, Spain	02/04/04

Frédéric Boyer

Title	Event and location	Date
Reconstruction *ab initio* de voies métaboliques : formalisation et approches combinatoires	Soutenance de thèse, Université Joseph Fourier, Grenoble	09/07/04
Méthodes *in silico* pour la reconstruction de voies métaboliques	Génoscope, Evry	13/10/04
Reconstruction *ab initio* de voies métaboliques : formalisation et approches combinatoires	LORIA, Nancy	03/12/04

Alexandra Calteau

Title	Event and location	Date
Transferts horizontaux entre bactéries et archées hyperthermophiles : le cas des hydrogénases	Rencontres Alphy, UCBL, Lyon	15-16/01/04
Super-tree Approach for Studying the Phylogeny of Prokaryotes: New Results on Completely Sequenced Genomes	International Conference of Computational Science, Kraków, Poland	08/06/04
Horizontal transfer of two operons coding for hydrogenases between bacteria and archaea	Conference of the Society of Molecular Biology and Evolution, Penn State College, USA	19/06/04

Eric Coissac

Title	Event and location	Date
RepSeek: Recherche de répétitions génériques dans de grandes séquences d'ADN	JOBIM, Montreal, Canada	29/06/04

Jean-François Dufayard

Title	Event and location	Date
Incremental algorithms for alignments and phylogenies of large homologous sequence families: Implementation in Mentalign	JOBIM, Montreal, Canada	29/06/04
Algorithmes incrémentaux pour l'alignement et la phylogénie de grandes familles de séquences homologues	Soutenance de thèse, Université Joseph Fourier, Grenoble	14/12/04

Laurent Duret

Title	Event and location	Date
How much of the human genome is transcribed?	University of Arizona, Tucson, USA	28/01/04
Evolution of isochores in mammalian genomes	Evolutionary Biology Center, Uppsala, Sweden	16/03/04
When is a new gene a gene? The human ORFan ERVWE1-env is a bona fide gene that derives from a virus	Department of Biochemistry & Molecular Biophysics, University of Arizona, Tucson, USA	21/01/04
Usage des codons synonymes chez les eucaryotes multicellulaires: mais où est donc passée la sélection naturelle ?	Laboratoire Arago, Banyuls-sur-Mer	17/06/04
Impact de la recombinaison sur l'évolution des génomes de vertébrés	Séminaires d'Evolution de Montpellier, Montpellier	08/06/04
The evolution of isochores in vertebrates	Conf. "Structural approaches to sequence evolution: Molecules, networks, populations", Dresden, Germany	05-09/07/04
Recombination drives the evolution of isochores in mammals	Evolutionary Genomics Conference, Tucson, USA	15-17/01/04

Hidde de Jong

Title	Event and location	Date
Qualitative modeling and simulation of genetic regulatory networks	CIMPA-UNESCO Summer School on Mathematical and Computational Methods in Biology, Valdivia, Chili	13-16/01/04
Modélisation et simulation qualitative de réseaux de régulation génique	Institut de Mathématiques de Bourgogne, Dijon	29/01/04
Qualitative modeling and simulation of genetic regulatory networks	Centrum voor Wiskunde en Informatica (CWI), Amsterdam, the Netherlands	06/04/04
Modeling and simulation of genetic regulatory networks	EUNITE/JCB Spring School on BioData Mining, Jena, Germany	13-14/05/04
Modélisation et simulation qualitative de réseaux de régulation génique	Laboratoire TIMC, Institut d'Informatique et de Mathématiques Appliquées (IMAG), Grenoble	26/05/04
Qualitative modeling and simulation of genetic regulatory networks	Second Bertinoro Computational Biology Meeting, Bertirono, Italy	17/06/04
Modeling and simulation of genetic regulatory networks	Coupled Map Lattices, Institut H. Poincaré, Paris	24/06/04
Qualitative modeling and simulation of genetic regulatory networks	Department of Physiology, McGill University, Montreal, Canada	29/06/04
Qualitative modeling and simulation of genetic regulatory networks	Symposium on Integrative Bioinformatics: Aspects of the Virtual Cell, Dagstuhl, Germany	07/07/04
Qualitative modeling and simulation of genetic regulatory networks	Sixth World Congress of the Bernoulli Society, Barcelona, Spain	27/07/04
Modeling and simulation of genetic regulatory networks	Marie Curie Training Course on Multiple Aspects of DNA and RNA: From Biophysics to Bioinformatics, Les Houches	25-26/08/04
Modeling and simulation of genetic regulatory networks	European School on Nanosciences & Nanotechnologies, Grenoble	27/08/04
Qualitative modeling and simulation of genetic regulatory networks	Environmental and Therapeutical Applications of Nanobiotechnology and Genomics, Sophia-Antipolis	09/09/04
Qualitative modeling and simulation of genetic regulatory networks	CIGENE symposium in honour of René Thomas, Averøy, Norway	16/09/04
Modélisation et simulation qualitative de réseaux de régulation génique : des gènes rapporteurs aux solutions de Filippov	Soutenance d'HDR, Université Joseph Fourier, Grenoble	30/09/04
Qualitative modeling and simulation of genetic regulatory networks	ICSB Workshop on Dynamical Modeling of Genetic and Metabolic Networks, Heidelberg, Germany (with D. Ropers)	09/10/04

Modeling and simulation of genetic regulatory networks	Workshop genomics programme INRA, Paris (with D. Ropers)	02/12/04

Vincent Lacroix

Title	Event and location	Date
Motifs in metabolic networks	Seminar of the Combinatorics Group of the Instituto de Matemática e Estatística, University of São Paulo, Brazil	10/12/04

Jean Lobry

Title	Event and location	Date
SeqinR: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis	Max Planck Institut für physik komplexer system, Dresden, Germany	05-10/07/04
Life history traits and genome structure: aerobiosis and G+C content in bacteria	International Conference on Computational Science, Kraków, Poland	06-09/06/04

Christelle de Melo de Lima

Title	Event and location	Date
Modélisation de la distribution de longueur de exons par une somme de lois géométriques pour l'utilisation de chaînes de Markov cachées	Séminaire Modélisation mathématique de la biologie, INSA et Laboratoire Maply de Lyon	04/04
Modélisation markovienne de la structure des gènes : exons de lois non géométriques et influence du taux de G+C	Séminaire du LaPCS, Université Claude Bernard	06/06/04
Modélisation Markovienne de la structure des gènes. Analyse de la structure en isochore	Séminaires Mathématiques pour le génome, Laboratoire de statistique et génome Évry	08/06/04

Vincent Navratil

Title	Event and location	Date
Modélisation des connaissances en cartographie comparée des génomes de mammifères : Application à l'étude de la variabilité génétique	Séminaire des Thésards du département Génétique Animale INRA Tours	05/04

Guy Perrière

Title	Event and location	Date
Transferts horizontaux entre bactéries et archées hyperthermophiles : le cas des hydrogénases	11èmes Rencontres Alphy, UCBL, Lyon	15-16/01/04
Horizontal transfer of two operons coding for hydrogenases between bacteria and archaea	5èmes Journées Ouvertes : Biologie Informatique et Mathématiques, JOBIM 2004, Montreal Univ., Canada	28-30/06/04
Intégration de données hétérogènes dans la banque de données de récepteurs nucléaires NuReBase	Forum BioTech, Lyon	17/06/04
Application of multivariate statistics to the analysis of genomic data	Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, USA	25/08/04

François Rechenmann

Title	Event and location	Date
Introduction to the session "High throughput biological data production, management and analysis"	ESONN'04 : School on nanosciences - nanotechnologies, Grenoble	24/08/04
ISee - A la découverte des génomes et de la bioinformatique	"Le goût des sciences ?" Séminaire sur les pratiques pédagogiques, CCSTI-Grenoble	15/11/04

Delphine Ropers

Title	Event and location	Date
Qualitative simulation of the nutritional stress response in *Escherichia coli*	Summer School on Systems Biology, McGill University, Montreal, Canada	26/05/04
Qualitative simulation of the nutritional stress response in *Escherichia coli*	JOBIM satellite meeting on Dynamical Modelling of Biological Regulatory Networks, Montreal, Canada	01/07/04

Qualitative modeling and simulation of genetic regulatory networks	ICSB Workshop on Dynamical Modeling of Genetic and Metabolic Networks, Heidelberg, Germany (with H. de Jong)	09/10/04
Qualitative simulation of the nutritional stress response in *Escherichia coli*	Séminaire Algorithmique et Biologie, Lyon	22/11/04
Modeling and simulation of genetic regulatory networks	Workshop genomics programme INRA, Paris (with H. de Jong)	02/12/04

Marie-France Sagot

Title	Event and location	Date
Sorting by reversals in subquadratic time	Meeting “Phylogenetic combinatorics and applications”, Uppsala University, Sweden	06/07/04
On bases of motifs	Journées Montoises, Liège, Belgium	11/09/04
Some questions around genome rearrangements	Conference ALGO 2004, Bergen, Norway,	16/09/04
Motifs in metabolic networks	Seminar of the Vlaams interuniversitair Instituut voor Biotechnologie, Gent, Belgium	27/09/04
Some questions around genome rearrangements	2nd RECOMB Comparative Genomics Satellite Workshop, University of Bologna Residential Center, Bertinoro, Italy	17/10/04
Some questions around genome rearrangements	Seminar of the Combinatorics Group of the Instituto de Matemática e Estatística, University of São Paulo, Brazil	08/12/04

Marie Semon

Title	Event and location	Date
No evidence of selection for the clustering of co-expressed genes in vertebrate genomes	Conference of the Society of Molecular Biology and Evolution, Penn State College, USA	19/06/04
No evidence of selection for the clustering of co-expressed genes in vertebrate genomes	IPG, Lyon	14/10/04

Eric Tannier

Title	Event and location	Date
The combinatorics of genome rearrangements	8th Aussois Workshop on Combinatorial Optimization, France	5-9/01/04
Une méthode rapide pour le tri par inversions	Institut Gaspard Monge, Université de Marne-la-Vallée	27/01/04
Les réarrangements génomiques en optimisation combinatoire	Laboratoire LaPCS, Université Claude Bernard	18/02/04
Une méthode rapide pour le tri par inversions	Laboratoire LIAFA, Université de Paris 7	27/03/04
Fast genome rearrangement by reversals	1st French Taiwanese Conference in Information Technologies, École Polytechnique, Palaiseau, France	14-16/04/04
Sorting by Reversals in Subquadratic Time	Combinatorial Pattern Matching 2004, Istanbul, Turkey	05/07/04

Alain Viari

Title	Event and location	Date
Répétitions et duplication intrachromosomiques	IRISA - Rennes	30/04/04
Bioinformatique pour la protéomique	Université de Strasbourg	30/06/04
Introduction to bioinformatics	ESONN’04 : School on nanosciences - nanotechnologies, Grenoble	25/08/04
Syntactic and semantic consistency in biological databases	Oriel Workshop, Hixton, UK	14/10/04

8.2. Organisation of conferences, workshops and meetings

Laurent Duret

Type	Location	Date
Organization of the “Atelier de formation INSERM: Génomique comparative des vertébrés : concepts et outils bioinformatiques”	La Londe Les Maures	27-28/05/04 & 21-23/06/04

Hidde de Jong

Type	Location	Date
Working group meeting on Integrated Models of Gene Regulation and Metabolism (with D. Ropers)	Grenoble	06-07/05/04
JOBIM satellite meeting on Dynamical Modelling of Biological Regulatory Networks (with C. Chaouiya, S. Provencher, D. Thieffry)	Montreal, Canada	01/07/04
ICSB Workshop on Dynamical Modeling of Genetic and Metabolic Networks (with C. Chaouiya, D. Ropers, D. Thieffry, and others)	Heidelberg, Germany	09/10/04

François Rechenmann

Type	Location	Date
Table ronde de la Journée Bioinformatique, “Bioinformatique - Quels besoins ? Quelles actions ?”	CEA-Grenoble DRDC	27/01/04
Journée technique du Réseau National Genopole “Quelle informatique et bioinformatique pour la protéomique ?”	Grenoble, INRIA Rhône-Alpes	01/06/04
Session “High throughput biological data production, management and analysis”	ESONN’04: School on nanosciences - nanotechnologies, Grenoble	24/08/04

Delphine Ropers

Type	Location	Date
Working group meeting on Integrated Models of Gene Regulation and Metabolism (with H. de Jong)	Grenoble	06-07/05/04
ICSB Workshop on Dynamical Modeling of Genetic and Metabolic Networks (with C. Chaouiya, H. de Jong, D. Thieffry, and others)	Heidelberg, Germany	09/10/04

Marie-France Sagot

Type	Location	Date
CompBioNets (Algorithms and Computational Methods for Biochemical and Evolutionary Networks) 2004	Recife, Brazil	15-18/12/04
BCB (Second Bertinoro Computational Biology Meeting) 2004	University of Bologna Residential Center, Bertinoro, Italy	12-19/06/04
15th series of the Seminar Algorithmics and Biology: Genomic Rearrangements	CNRS Amphitheater, La Doua, Lyon	02-04/02/04
16th series of the Seminar Algorithmics and Biology: Genomes, Chromosomes, Cells and Developmental Evolution	CNRS Amphitheater, La Doua, Lyon	22-24/11/04

8.3. Editorial and reviewing activities

Laurent Duret

Type	Journal or conference
Steering Committee	French national conference on Bioinformatics, Jobim

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 Algorithms for Molecular Biology*, BioMed Central
Program Committee	RECOMB, ECCB-ISMB (steering), SPIRE, GCB, JOBIM (co-chair), ESA, 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 lecturers (section 27) at the Université de Provence (Marseille) and a member of the recruiting committee for research associates at INRIA Rocquencourt.

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 2004. 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.

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	4	INSA, Lyon	14
Modelling and simulation of genetic regulatory networks	5	ENS, Paris	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	Université Claude Bernard, Grenoble	8

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] F. BOYER. *Reconstruction ab initio de voies métaboliques : formalisation et approches combinatoires*, Ph. D. Thesis, Université Joseph Fourier, Grenoble, 2004.

[2] J.-F. DUFAYARD. *Algorithmes incrémentaux pour l'alignement et la phylogénie de grandes familles de séquences homologues*, Ph. D. Thesis, Université Joseph Fourier, Grenoble, 2004.

[3] H. DE JONG. *Modélisation et simulation qualitative de réseaux de régulation génique*, HDR, Université Joseph Fourier, Grenoble, 2004.

Articles in referred journals and book chapters

[4] A. AOUACHERIA, C. CLUZEL, C. LETHIAS, M. GOUY, R. GARRONE, J. EXPOSITO. *Invertebrate data predict an early emergence of vertebrate fibrillar collagen clades and an anti-incest model*, in "J. Biol. Chem.", vol. 279, 2004, p. 47711-47719.

[5] E. BELLE, L. DURET, N. GALTIER, A. EYRE-WALKER. *The Decline of Isochores in Mammals: An Assessment of the GC Content Variation Along the Mammalian Phylogeny*, in "J. Mol. Evol.", vol. 58, 2004, p. 653-660.

[6] B. BONNAUD, O. BOUTON, G. ORIOL, V. CHEYNET, L. DURET, F. MALLET. *Evidence of Selection on the Domesticated ERVWE1 env Retroviral Element Involved in Placentation*, in "Mol. Biol. Evol.", vol. 21, 2004, p. 1895-1901.

[7] N. BONNEFOY-BERARD, A. AOUACHERIA, C. VERSCHELDE, L. QUEMENEUR, A. MARÇAIS, J. MARVEL. *Control of proliferation by Bcl-2 family members*, in "Biochim. Biophys. Acta", vol. 644, 2004, p. 159-168.

[8] D. CHARIF, J. THIOULOUSE, J. LOBRY, G. PERRIÈRE. *Online Synonymous Codon Usage Analyses with the ADE-4 and seqinR packages*, in "Bioinformatics", vol. 20, 2004, p. 1-3.

[9] D. EVEILLARD, D. ROPERS, H. DE JONG, C. BRANLANT, A. BOCKMAYR. *A multi-scale constraint programming model of alternative splicing regulation*, in "Theoretical Computer Science", vol. 325, n° 1, 2004, p. 3-24.

[10] A. GINOLHAC, C. JARRIN, B. GILLET, P. ROBE, P. PUJIC, K. TUPHILE, H. BERTRAND, T. VOGEL, G. PERRIÈRE, P. SIMONET, R. NALIN. *Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones*, in "Applied Env. Microbiol.", vol. 70, 2004, p. 5522-5527.

[11] N. JACQ, C. BLANCHET, C. COMBET, E. CORNILLOT, L. DURET, K. KURATA, H. NAKAMURA, T. SILVESTRE, V. BRETON. *Grid as a bioinformatic tool*, in "Parallel Computing", vol. 30, 2004, p. 1093-1107.

[12] C. KEIME, F. DAMIOLA, D. MOUCHIROUD, L. DURET, O. GANDRILLON. *Identitag a relational database for SAGE tag identification and interspecies comparison of SAGE libraries*, in "BMC Bioinformatics", vol. 5, 2004, 143.

[13] J. MEUNIER, L. DURET. *Recombination drives the evolution of GC-content in the human genome*, in "Mol. Biol. Evol.", vol. 21, 2004, p. 984-990.

[14] L. MICHON, M. MORLE, L. BOZON, L. DURET, J. ZECH, J. GODET, H. PLAUCHU, P. EDERY. *Physical and transcript map of the autosomal dominant colobomatous microphthalmia locus on chromosome 15q12-q15 and refinement to a 4.4 Mb region*, in "Eur. J. Hum. Genet.", vol. 12, 2004, p. 574-578.

[15] N. PISANTI, M. CROCHEMORE, R. GROSSI, M.-F. SAGOT. *A comparative study of bases for motif inference*, in "String Algorithmics", NATO Book series, KCL Press, 2004.

[16] D. ROPERS, L. AYADI, R. GATTONI, S. JACQUENET, L. DAMIER, M. STOLTZFUS, C. BRANLANT, J. STÉVENIN. *Differential effects of the SR proteins 9G8, SC35, ASF/SF2 and SRp40 on the utilisation of the A1 to A5 splicing sites of HIV-1 RNA*, in "Journal of Biological Chemistry", vol. 29, n° 279, 2004, p. 29963-29973.

[17] D. RUAU, J. DUARTE, T. OURJDAL, G. PERRIÈRE, V. LAUDET, M. ROBINSON-RECHAVI. *Update of NuReBase : nuclear hormone receptor functional genomics*, in "Nucleic Acids Res.", vol. 32, 2004, p. 165-167.

[18] M. SEMON, L. DURET. *Evidence that functional transcription units cover at least half of the human genome*, in "Trends Genet.", vol. 20, 2004, p. 229-232.

[19] J. WUYTS, G. PERRIÈRE, Y. VAN DE PEER. *The European ribosomal RNA database*, in "Nucleic Acids Res.", vol. 32, 2004, p. 101-103.

[20] H. DE JONG, J. GEISELMANN, G. BATT, C. HERNANDEZ, M. PAGE. *Qualitative simulation of the initiation of sporulation in B. subtilis*, in "Bulletin of Mathematical Biology", vol. 66, nº 2, 2004, p. 261-300.

[21] H. DE JONG, J. GEISELMANN. *Modélisation et simulation de réseaux de régulation génique par des équations différentielles ordinaires*, in "Informatique pour l'analyse du transcriptome, Paris", J.-F. BOULICAUT, O. GANDRILLON (editors)., Hermès, 2004, p. 143-185.

[22] H. DE JONG, J.-L. GOUZÉ, C. HERNANDEZ, M. PAGE, T. SARI, J. GEISELMANN. *Qualitative simulation of genetic regulatory networks using piecewise-linear models*, in "Bulletin of Mathematical Biology", vol. 66, nº 2, 2004, p. 301-340.

Publications in Conferences and Workshops

[23] G. ACHAZ, F. BOYER, E. ROCHA, A. VIARI, E. COISSAC. *RepSeek: recherche de répétitions génériques dans de grandes séquences d'ADN*, in "Recueil des Actes des 5ème Journées Ouvertes Biologie Informatique Mathématiques, JOBIM 2004, Montréal, Canada", 2004, 78.

[24] J. ALLALI, M.-F. SAGOT. *Novel tree edit operations for RNA secondary structure comparison*, in "Proceedings of the 4th Workshop on Algorithms in BioInformatics (WABI)", Lecture Notes in Computer Science, vol. 3240, Springer-Verlag, 2004, p. 412-425.

[25] G. BATT, D. BERGAMINI, H. DE JONG, H. GAVAREL, R. MATEESCU. *Model checking genetic regulatory networks using GNA and CADP*, in "Eleventh International SPIN Workshop on Model Checking of Software, SPIN 2004, Berlin", S. GRAF, L. MOUNIER (editors)., Lecture Notes in Computer Science, vol. 2989, Springer-Verlag, 2004, p. 158-163.

[26] A. CALTEAU, V. DAUBIN, G. PERRIÈRE. *Super-tree approach for studying the phylogeny of prokaryotes: new results on completely sequenced genomes*, in "International Conference of Computational Science (ISCC)", Lecture Notes in Computer Science, vol. 3039, Springer-Verlag, 2004, p. 700-708.

[27] A. CALTEAU, M. GOUY, G. PERRIÈRE. *Horizontal transfer of two operons coding for hydrogenases between bacteria and archaea*, in "Recueil des Actes des 5eJournées Ouvertes Biologie Informatique Mathématiques, JOBIM 2004", 2004.

[28] A. CARVALHO, A. T. FREITAS, A. L. OLIVEIRA, M.-F. SAGOT. *A parallel algorithm for the extraction of structured motifs*, in "Proceedings of 19th ACM Symposium on Applied Computing (SAC)", 2004, p. 147-153.

[29] A. CARVALHO, A. T. FREITAS, A. L. OLIVEIRA, M.-F. SAGOT. *Efficient Extraction of Structured Motifs Using Box-links*, in "Proceedings of the 11th Conference on String Processing and Information Retrieval (SPIRE)", Lecture Notes in Computer Science, vol. 3246, Springer-Verlag, 2004, p. 267-268.

[30] M. CROCHEMORE, R. GIANCARLO, M.-F. SAGOT. *Longest motifs with a functionally equivalent block*, in "Proceedings of the 11th Conference on String Processing and Information Retrieval", Lecture Notes in Computer Science, vol. 3246, Springer-Verlag, 2004.

[31] M. CROCHEMORE, C. ILIOPOULOS, M. MOHAMED, M.-F. SAGOT. *Longest Repeated Motif with k Don't Cares*, in "Theoretical Informatics: 6th Latin American Symposium (LATIN)", Lecture Notes in Computer

Science, vol. 2976, Springer-Verlag, 2004, p. 271-278.

[32] J.-F. DUFAYARD, G. PERRIÈRE, M. GOUY. *Incremental algorithms for alignments and phylogenies of large homologous sequence families: Implementation in Mentalign*, in "Recueil des Actes des 5eJournées Ouvertes Biologie Informatique Mathématiques, JOBIM 2004, Montréal, Canada", 2004, 62.

[33] D. EVEILLARD, A. LARHLIMI, D. ROPERS, S. BILLAUT, S. PEYREFITTE. *KOALAB: A new method for regulatory motif search. Illustration on alternative splicing regulation in HIV-1*, in "Recueil des Actes des 5eJournées Ouvertes Biologie Informatique Mathématiques, JOBIM 2004, Montréal, Canada", 2004, 81.

[34] C. GAUTIER, V. NAVRATIL, C. M. DE LIMA. *Genomic Information along the Genome: Modeling Statistical analysis and Evolutionary Implication*, in "Proceedings 18th IFIP World Computer Congress", 2004.

[35] C. S. ILIOPOULOS, J. MCHUGH, P. PETERLONGO, N. PISANTI, W. RYTTER, M. F. SAGOT. *A First Approach to Finding Common Motifs with Gaps*, in "Proceedings of the 2004 Prague Stringology Conference (PSC'04)", 2004.

[36] J. LOBRY. *Life history traits and genome structure: aerobiosis and G+C content in bacteria*, in "International Conference of Computational Science (ISCC)", Lecture Notes in Computer Science, vol. 3039, Springer-Verlag, 2004, p. 679-686.

[37] E. TANNIER, M.-F. SAGOT. *Sorting by reversals in subquadratic time*, in "Proceedings of the 15th Annual Symposium on Combinatorial Pattern Matching (CPM)", Lecture Notes in Computer Science, vol. 3109, Springer-Verlag, 2004, p. 1-13.

[38] 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, Berlin", J.-R. CHAZOTTES, B. FERNANDEZ (editors)., Lecture Notes in Physics, Springer-Verlag, 2004.

Internal Reports

[39] J. ALLALI, M.-F. SAGOT. *The at-most k-deep factor tree*, Technical report, nº RR-IGM 2004-03, Institut Gaspard Monge, 2004.

[40] G. BATT, H. DE JONG, J. GEISELMANN, M. PAGE, D. ROPERS, D. SCHNEIDER. *Symbolic reachability analysis of genetic regulatory networks using qualitative abstraction*, Technical report, nº RR-5362, INRIA, 2004, http://www.inria.fr/rrrt/rr-5362.html.

[41] R. CASEY, H. DE JONG, J.-L. GOUZÉ. *Piecewise-linear models of genetic regulatory networks: Equilibria and their stability*, Technical report, nº RR-5353, INRIA, 2004, http://www.inria.fr/rrrt/rr-5353.html.

[42] M. CROCHEMORE, R. GIANCARLO, M.-F. SAGOT. *Longest Motifs with a Functionally Equivalent Central Block*, Technical report, nº RR-IGM 2004-01, Institut Gaspard Monge, 2004.

[43] D. ROPERS, H. DE JONG, M. PAGE, D. SCHNEIDER, J. GEISELMANN. *Qualitative simulation of the nutritional stress response in Escherichia coli*, Technical report, nº RR-5412, INRIA, 2004, http://www.inria.fr/rrrt/rr-5412.html.

[44] H. DE JONG. *Qualitative simulation and related approaches for the analysis of dynamical systems*, Technical report, nº RR-5128, INRIA, 2004, http://www.inria.fr/rrrt/rr-5128.html.