INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team HELIX

Informatics and genomics

Rhône-Alpes

THEME 3A

Activity Report

2003

Table of contents

1. Team 1

2. Overall Objectives 2

3. Scientific Foundations 3

3.1. Evolution of species and of gene families 3

3.2. Modelling and analysis of the spatial organization of genomic information 4

3.3. Motif search and inference 5

3.4. Modelling and analysis of metabolism: molecular components, regulation, and pathways 6

3.5. Modelling and simulation of genetic regulatory networks 7

3.6. Computational proteomics 7

5. Software 8

5.1. Bibi 8

5.2. Box 8

5.3. FactorTree 8

5.4. FamFetch 9

5.5. GenoExpertBacteria (GEB) 9

5.6. GenoStar 9

5.7. Genetic Network Analyzer (GNA) 9

5.8. Herbs 10

5.9. Hogenom and Hovergen 10

5.10. ISee 10

5.11. Migal 10

5.12. Oriloc 10

5.13. PBIL 11

5.14. PepLine 11

5.15. PhyloJava 11

5.16. Phylo_win 11

5.17. Roso 11

5.18. RTKdb 11

5.19. Satellites 11

5.20. Smile 12

5.21. Utopia 12

5.22. Other software developed in HELIX 12

6. New Results 12

6.1. Computational analysis of the evolution of species and gene families 12

6.2. Modelling and analysis of the spatial organization of genomic information 13

6.3. Motif search and inference 14

6.4. Modelling and analysis of metabolism: molecular components, regulation, and pathways 15

6.5. Modelling and simulation of genetic regulatory networks 16

6.6. Computational proteomics 17

7. Contracts and Grants with Industry 17

7.1. Introduction 17

7.2. PepMap 17

7.3. Sanofi 18

8. Other Grants and Activities 18

8.1. National projects 18

8.2. Projects funded by international organisms or including international teams 18

9. Dissemination 19

9.1. Talks 19
9.2. Organization of conferences, workshops and meetings 22
9.3. Editorial and reviewing activities 22
9.4. Administrative activities 23
9.5. Teaching activities 24
10. Bibliography **25**

1. Team

The HELIX project is located in Montbonnot (Grenoble) and on the Campus of La Doua (Villeurbanne, next to 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 the group 'Bioinformatics and Evolutionary Genomics' headed by Manolo Gouy, within the 'Laboratory of Biometry and Biological Evolution' (CNRS/Université Claude Bernard, Lyon, UMR 5558), directed by Christian Gautier. 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

Laurent Duret [Research associate, 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]
Dominique Mouchiroud [Professor, Université Claude Bernard]
Michel Page [Associate professor, Université Pierre Mendès-France]
Guy Perrière [Research associate, CNRS]
François Rechenmann [Research director, INRIA]
Marie-France Sagot [Research director, INRIA]
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

Stéphane Bruley [Junior technical staff, INRIA]
Antoine Brun [Project technical staff, INRIA]
Pierre-Emmanuel Ciron [Project technical staff, INRIA]
Véronique Dupierris [Project technical staff, INRIA]
Gilles Faucherand [Project technical staff, INRIA]
Agnès Iltis [Project technical staff, INRIA]
Erwan Reguer [Project technical staff, INRIA]
Emma Ribes [Junior technical staff, INRIA]
Timothée Silvestre [Project technical staff, CNRS]
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]
Estelle Nugues [Project technical staff, Rhône-Alpes Genopole]

Post-doctoral fellows

Abdel Aouacheria [Association pour la Recherche sur le Cancer]

Stéphanie Mérienne [INRIA]
Simon Penel [European contract]
Delphine Ropers [INRIA]
Eric Tannier [INRIA]

PhD students

Grégory Batt [scholarship ENS Lyon, supervisors: Hidde de Jong, François Rechenmann]
Frédéric Boyer [scholarship Ministère de la Recherche, supervisors: Laurent Trilling, Alain Viari]
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]
Jean-François Dufayard [scholarship Ministère de la Recherche, supervisors: Manolo Gouy, François Rechenmann]
Adel Khelifi [scholarship Ministère de la Recherche, supervisor: Dominique Mouchiroud]
Christelle Melo de Lima [scholarship Ministère de la Recherche, supervisors: Christian Gautier, Didier Piau, François Rechenmann]
Julien Meunier [scholarship ENS Lyon, supervisor: Laurent Duret]
Vincent Navratil [scholarship INRA, supervisor: Christian Gautier]
Marie Semon [scholarship Ministère de la Recherche, supervisor: Laurent Duret]
Marina Zelwer [scholarship Ministère de la Recherche, supervisors: Maxime Crochemore, Marie-France Sagot]
The HELIX group at Lyon benefits from the human resources of the UMR 5558, in particular as concerns administrative assistance.

2. Overall Objectives

The information necessary to the development and the maintenance of a living organism is contained in its genome, materialized 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 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 3 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 such 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 during evolution. In biology, it is impossible to ignore this historical aspect. Evolution allows us to compare and decipher the meaning of structure, the modification of metabolic pathways, the preservation and variation of signalling systems, and so on.

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 post-genomic 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.

The overall objective of HELIX is to develop methods and computer tools to help biologists represent, access, and analyse the huge amounts of genomic and post-genomic data available today. Six 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 organization of genomic information;
3. Motif search and inference;
4. Modelling of metabolism: molecular components, regulation, and pathways;
5. Modelling and simulation of genetic regulatory networks;
6. Computational proteomics.

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

The HELIX project at the same time bridges two geographical locations and two different bioinformatic cultures. While one group is located in Grenoble and has its origin in computer science, the other group resides in Lyon and has its roots in biology. However, a long tradition of collaboration between the two groups gives 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 two 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.

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;
- PBIL (Pôle BioInformatique Lyonnais) in collaboration with the IBCP (UMR CNRS 5086) federates the methodological developments of the group at Lyon. This platform should soon extend to the whole Rhône-Alpes region under the name of PRABI and thus further integrate all the groups within the HELIX project or associated with it. The amount of computer resources necessary for the functioning of the platform has given rise to a growing collaboration with the IN2P3 computer center.

The two platforms, GENOSTAR and PBIL (PRABI) offer services that are complementary and address the biological community in different ways. The first platform offers a sophisticated modelling tool enabling the biologist to work with powerful methods of analysis, but requires an initial effort to ensure its effective application. PBIL is a Web platform directly accessible to researchers in biology, but does not offer the possibility of creating one's own strategies of analysis.

3. Scientific Foundations

3.1. Evolution of species and of gene families

Participants: Jean-François Dufayard, Laurent Duret, Christian Gautier, Manolo Gouy [Correspondent], Guy Perrière, Eric Tannier, Marie-France Sagot [Correspondent].

Evolution is the main characteristic of living systems. Biological diversity results from the succession of two independent processes: one introducing 'errors' 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 error, where the frequency of occurrence of a very small fraction of the errors will increase in the population until such errors become the 'norm'.

The analysis of these two processes and of their consequences on a genome underlies an important part of the field of molecular computational biology. It will therefore appear in almost all the topics developed within the HELIX project. One can in particular mention:

- *The 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 very 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 fixing, even the process of error itself, depends on these two aspects and leaves 'diagnostic' traces in the genome that are interpretable 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 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 the organism they belong to. Their comparison allows therefore, *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 possible 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 to go from one permutation to another. Sensitive algorithms are required to deal with the problem.

A phylogenetic reconstruction that uses 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 thus 6000 different trees. The management and comparison of such trees is a computational and mathematical problem that requires the expertise of the two groups composing the HELIX project and is thus characteristic of the collaboration existing between them.

The modification of a base from a parental to a descendant genome results from the interaction between a purely physical event (a radiation-induced mutation for instance) and various biological processes. The main example of such an interaction is DNA reparation: most genome modifications will be repaired, but not all with the same efficiency. More complex interactions may exist which favour or inhibit the production of errors (methylation of some bases, variable time of separation of the two strands, *etc.*). The fixation of a mutation may itself be subject to various types of constraints. The best known are those said to be the consequence of natural selection. Such mutations translate the fact that, if an individual presenting a mutation has more descendants than others, the mutation has a higher probability of becoming fixed. Other subtler constraints modulate the effect of natural selection (population size, recombination). The HELIX project is interested in all these phenomena, both in the prokaryotic and eukaryotic worlds.

3.2. Modelling and analysis of the spatial organization of genomic information

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

The Grenoble and Lyon groups of the HELIX project participated very early (over more than 20 years ago) in the discovery of strong genomic heterogeneities having biological and statistical characteristics. In particular, neighbouring genes along a genome often share multiple properties, whose nature is both structural (size and number of introns for instance) and statistical (related to base and codon frequencies). In certain cases, such neighbouring structures have been interpretable in terms of biological processes. For instance, in bacteria, a neighbouring structure 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 generate and maintain them. The most characteristic such structure concerns isochores in vertebrates. Taking into account the structure in 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.*)

During the course of evolution, the spatial organization 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 or to compare the rates of rearrangements observed in different genomic regions. The HELIX project has thus 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 organization is conserved between two species).

Modelling and analysing genomic maps requires expertise in various areas such as knowledge representation, statistics and algorithmics. The knowledge representation system AROM – originally developed in the SHERPA project, precursor of the HELIX project – is used both in Grenoble and in Lyon for modelling, respectively, prokaryotes (GENOSTAR, GEB) and eukaryotes (GEMCORE). 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 partitionning (or segmentation) techniques. Finally, difficult algorithmical problems appear in the definition and computation of distances between maps.

3.3. Motif search and inference

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

The term *motif* is a very general one used for designing 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, for instance, evolutionary trees or graphs, and biochemical or genetic networks (see sections 3.4 and 3.5 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, retrotranposons, *etc.*

Identifying motifs, whether using a model established from previously obtained examples of a conserved structure or proceeding *ab initio*, represents an important area of research in computational biology. Such identification covers two main aspects:

1. Feature identification: this 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 (*i.e.* tandem repeats) or transposable elements (dispersed repeats with a specific structure), regulatory regions etc.
2. Relational identification: the goal in this case is to find the 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 same cellular process, or their physical interaction.

Search and inference problems, whether they concern features or relations, are in fact only the two extreme facets of a continuum of problems that range from seeking for something well-known to trying to identify objects about which very little is known. 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. The information that must be manipulated in this case – cooperative signals, operons, regulons, reaction pathways or molecular assemblies – are, however, 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 inprint that may, in some cases, be identified by comparison inside or between different genomes or organisms. Such inprint is often modelled as 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, section 5.20), gene finding (UTOPIA, section 5.21; EMKOV, section 5.22), satellites identification (SATELLITES, section 5.19), RNA common substructure inference (MIGAL, section 5.11) *etc.*

Correctly inferring such information may then enable to assign one (or more) function(s) to the various biological entities.

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

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

Advanced experimental techniques in genomics have produced huge amounts of data on the molecular basis of cellular processes. These data are quite heterogeneous, including among other things the genomic sequences of organisms as well as information on the organization 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 evolution with time 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 that 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 organization of a cell. An example question is: what is the location of the genes of organism x coding for enzymes that catalyse the biochemical reactions of metabolic pathway y?

The genomic data gathered over the past few decades are usually dispersed over 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, such as EcoCyc, KEGG (http://www.genome.ad.jp/kegg/) for metabolic reactions, and RegulonDB (http://www.cifn.unam.mx/Computational_Genomics/regulondb/) for the structure and regulation of operons. However, the integration of the data in the different bases existing today is strongly hampered by their lack of *interoperability*, that is, the possibility to relate the data in the different bases. This is partially due to technical problems, arising from the choice of platform and programming language. However, a more fundamental problem is the incompatibility of the models underlying the data and knowledge bases, a situation often aggravated by the absence of an explicit formulation of the models.

In HELIX, the interoperability problem is countered by giving priority to the development of explicit, formal models of the molecular components of the cell and their organization. This approach has been put to work in the system GENOEXPERTBACTERIA (GEB), formerly called Panoramix. Object-oriented models of three aspects of cellular processes in bacteria have been developed for GEB:

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

GEB provides a rich source of information to analyse the chromosomal organization 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 putative metabolic pathways connecting a set of given compounds without any other knowledge than a set of reactions (and the chemical compounds they involve). This approach allows the prediction of unknown, possibly unrealistic, pathways that should be further assessed. A second use of the information accessible through GEB, currently at a more embryonic stage, concerns the inference of evolutionary conserved network motifs from the comparison of metabolic networks in different organisms.

3.5. Modelling and simulation of genetic regulatory networks

Participants: Grégory Batt, 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, we need to elucidate the networks of interactions involved in gene regulation, metabolism, signal transduction, and other cellular and intercellular processes.

Genetic regulatory networks control the spatiotemporal expression of genes in an organism, and thus underlie complex processes like cell differentiation and development. Genetic regulatory networks consist of genes, proteins, metabolites, and other small molecules, as well as their mutual interactions. Their study 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, an intuitive understanding of their dynamics is difficult to obtain and may lead to erroneous conclusions. *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 equation models* and *stochastic models*. 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 differential equations and stochastic 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 our research 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, we have developed a method for the *qualitative simulation* of genetic regulatory networks that has been implemented in the computer tool GENETIC NETWORK ANALYZER (GNA) (section 5.7). 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).

3.6. Computational proteomics

Participants: Estelle Nugues, Erwan Reguer, 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. However, while the term genome can be used to denote the complete set of genes of a living organism, the definition of *proteome* is far less straightforward. As a matter of fact, a same polypeptide may lead to several proteins which differ from each other by post-translational modifications. In this context, the term proteomics is understood to be the identification of the set of proteins which are expressed in a cell at a given time under given conditions.

Recent progress in *mass spectrometry (MS)* has resulted in efficient techniques allowing for the large-scale analysis of proteomes. Two main approaches towards protein identification by means of mass spectrometry can be distinguished. Both of these require a preliminary step in which the protein under study is digested by a specific enzyme, usually trypsine.

1. The *MS approach* consists in weighing, using time-of-flight (TOF) measurements, each of the trypsin ions obtained in the preliminary step. The results can be plotted in a mass spectrum, representing a fingerprint of the protein being studied.

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 in turn by mass spectrometry. From these measurements, so-called *peptide sequence tags (PSTs)* are generated. The PSTs are short peptide sequences (3 to 5 amino acids), flanked by two polypeptides of known, measured mass. They can be used to identify the protein by comparing the pattern with the sequences stored in protein databases.

The second approach lies at the heart of a collaboration between HELIX and the Laboratoire Chimie des Protéines at the CEA in Grenoble.

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, *i.e.* the determination of the amino acid sequence of each peptide, 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.

The aim of the collaboration between HELIX and the CEA is to develop such computer tools. In particular, efficient algorithms have been produced 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 the modules TAGGOR and PEPMAP, combined in the proteomics software pipeline PEPLINE (section 5.14).

5. Software

5.1. Bibi

Participant: Guy Perrière [Correspondent].

BIBI (http://pbil.univ-lyon1.fr/bibi) was developed in order to simplify sequence analysis for the purpose of bacterial identification. This program combines similarity search tools in the sequence databases and phylogeny display programs. It implements a chaining of two well-known tools: BLAST and CLUSTAL W.

5.2. Box

Participants: Antoine Brun, 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. For more information, see http://www-helix.inrialpes.fr/article397.html.

5.3. FactorTree

Participant: Marie-France Sagot [Correspondent].

FACTORTREE (http://www.inrialpes.fr/helix/people/sagot/programs/factortree.html) 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 space economy 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. The code for FACTORTREE (in C++) is freely available to academics and non profit organizations upon request to Julien Allali (allali@univ-mlv.fr) or Marie-France Sagot (Marie-France.Sagot@inria.fr).

5.4. 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 new tools to search for tree patterns in databases of phylogenetic trees.

5.5. GenoExpertBacteria (GEB)

Participants: Frédéric Boyer, Christophe Bruley, Stéphane Bruley, 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. For more information, see http://www-geb.inrialpes.fr/.

5.6. GenoStar

Participants: Pierre-Emmanuel Ciron, Véronique Dupierris, 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 the data which are manipulated by the application modules, and all the results they produce, are explicitely represented in an entity-relationship model: AROM. Within a module, the methods are organized into strategies, the execution of which addresses complex analysis problems.

The second version (1.2) of GENOSTAR is made up of three application modules: GENOANNOT, GENOLINK et 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 to understand 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 indeed 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.

5.7. Genetic Network Analyzer (GNA)

Participants: Grégory Batt, 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.

5.8. 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. It 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.

5.9. 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 from fully sequenced genomes, structured under the ACNUC sequence database management system. It allows to select sets of homologous genes among, respectively, general or vertebrate species, and to visualize 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.

5.10. ISee

Participants: Philippe Genoud [Correspondent], Stéphanie Merriene, Anne Morgat, Alain Viari, François Rechenmann, Danielle Ziébelin.

ISEE (IN SILICO BIOLOGY E-LEARNING ENVIRONMENT) explains the principles of the main bioinformatic algorithms and illustrates their use on real data. In its present state, the environment is structured into three main chapters: sequence comparison (dotplots and Needleman-Wunsch dynamic programming algorithm), statistical analysis of DNA sequences for the identification of coding regions, basic pattern-matching algorithms including the use of regular expressions. This small set of algorithms is combined into a bacterial gene finding strategy. A preliminary chapter introduces the genetic code and the translation process. ISEE can be adapted and extended to be used at different levels. Other courses can be written along the same principles, relying on the resources of the environment. In addition, new algorithmic modules can be developed and integrated. For more information, see http://www-helix.inrialpes.fr/article124.html.

5.11. Migal

Participants: Julien Allali, Marie-France Sagot [Correspondent].

MIGAL is an algorithm that compares two RNA structures. The C++ code of the MIGAL prototype is freely available to academics and non-profit organizations upon request to Julien Allali (allali@univ-mlv.fr) or Marie-France Sagot (Marie-France.Sagot@inria.fr).

5.12. 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 GLIMMER2 outputs to discriminate between codon positions.

5.13. PBIL

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

PBIL (http://pbil.univ-lyon1.fr/) is a set of web interfaces to access databases of gene families.

5.14. PepLine

Participants: Estelle Nugues, Erwan Reguer, 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 locating the gene coding for the protein. For more information, see http://www-helix.inrialpes.fr/article228.html.

5.15. PhyloJava

Participants: Manolo Gouy [Correspondent], Guy Perrière, Timothée Sylvestre.

PhyloJava is a Graphic User Interface (GUI) devoted to editing and calculating phylogenetic trees. It is implemented in Java to allow port to a wide-range of systems and is based on a client-server architecture. PhyloJava offers the possibility to visualize protein or DNA sequence alignments and to select/deselect sites or sequences. An advantage of PhyloJava is that one can add at any time new phylogenetic methods. Another main functionality is the possibility to parallelize jobs using a grid environment.

5.16. Phylo_win

Participants: Manolo Gouy [Correspondent], Guy Perrière.

Phylo_win is a graphical colour interface for molecular phylogenetic inference that was developed by Nicolas Galtier during his PhD under the supervision of Manolo Gouy. It uses neighbour-joining, parsimony and maximum likelihood methods and may apply bootstrap to any of them. Many distances can be used including Jukes & Cantor, Kimura, Tajima & Nei, HKY, Galtier & Gouy, LogDet for nucleotidic sequences, Poisson correction for protein sequences, Ka and Ks for codon sequences. Species and sites to include in the analysis are selected by mouse. Reconstructed trees can be drawn, edited, printed, stored and evaluated according to numerous criteria.

5.17. Roso

Participants: Laurent Duret, Guy Perrière [Correspondent], Jean-François Dufayard.

ROSO (http://pbil.univ-lyon1.fr/roso/Home.php) is a software to design optimized oligonucleotide probes for microarrays.

5.18. RTKdb

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

The RTKdatabase (http://pbil.univ-lyon1.fr/RTKdb/) is the web-based interface of RTKDB which is a database containing all the protein sequences of receptor tyrosine kinases (RTK), organized into families. It allows the user to select sets of homologous genes from different species (only common species for the moment) and to visualize multiple alignments and phylogenetic trees. Thus RTKDB is particularly useful for comparative genomics, phylogeny, and molecular evolution studies. RTKDB is the PhD work of Julien Grassot, from the 'Center for Molecular and Cellular Genetics' with which the Lyon group of HELIX collaborates.

5.19. Satellites

Participant: Marie-France Sagot [Correspondent].

SATELLITES (http://www.inrialpes.fr/helix/people/sagot/programs/satellites.html) 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 organizations by sending an email to Marie-France.Sagot@inria.fr.

5.20. Smile

Participant: Marie-France Sagot [Correspondent].

SMILE (http://www.inrialpes.fr/helix/people/sagot/programs/smile.html) is a motif inference algorithm that takes as input a set of DNA (RNA) or protein sequences. The code (in C) can be freely obtained by academics and non-profit research organizations by simply sending a mail to marsan@univ-mlv.fr or to Marie-France.Sagot@inria.fr.

5.21. Utopia

Participant: Marie-France Sagot [Correspondent].

UTOPIA (http://www.inrialpes.fr/helix/people/sagot/programs/utopia.html) 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. The current version (in C++) together with scripts for post-processing may be freely recovered by academics and non-profit research organizations by simply sending a mail to Marie-France.Sagot@inria.fr.

5.22. Other software developed in HELIX

Participants: François Rechenmann [Correspondent], Manolo Gouy [Correspondent].

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 (A. Viari), DRUID (Marie-France Sagot), EMKOV (A. Viari), GEM (Bruno Spataro), JADIS (Dominique Mouchiroud), MTDP (A. Viari), SEAVIEW (Manolo Gouy).

6. New Results

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

Participants: Jean-François Dufayard, Laurent Duret, Christian Gautier, Manolo Gouy [Correspondent], Guy Perrière, Marie-France Sagot [Correspondent].

Alexandra Calteau, PhD student, is currently working on the analysis of the evolutionary origin of thermophilic bacteria. A comparative genomic approach is followed that compares phylogenetic trees of many bacterial genes to distinguish between different hypotheses: single *vs.* multiple origins; horizontal transfer of one *vs.* several genes.

Jean-François Dufayard, PhD student, is working on an incremental algorithm that is able to construct very large multiple sequence alignments (involving several thousands of sequences). This work has much progressed in 2003 and is expected to be submitted for publication next year.

The development of databases of homologous protein genes is continued, with new releases of HOVERGEN (vertebrate genes) and with the creation of HOGENOM (families of homologous genes from all completely sequenced genomes, 117 species in the last release). The retrieval software associated to these databases has been extensively updated, both in terms of functionality and user interface, mainly through the work of Simon Penel. Work towards automation of the update procedures of these family databases is progressing.

A new bioinformatics tool, BIBI, for Bioinformatics Bacterial Identification Tool (see section 5.1), has been developed and made available for web access. It allows users, mostly from the medical field, but also microbiologists such as mycologists, to quickly identify the species of origin of a sample using a short DNA sequence fragment.

Timothée Silvestre has been developing a client-server software system entitled PHYLO_JAVA (see section 5.22), devoted to phylogenetic analyses, which allows access to up-to-date tree-building algorithms through the internet and with a user interface offering easy access. This application is currently being used by *beta* testers.

We have also developed new interfaces to access databases of gene families [47]. The main new features are: display of multiple alignments and phylogenetic trees; complex queries according to features comprising display of multiple alignments and phylogenetic trees; complex queries according to taxonomic criteria (phylogenetic substraction).

Finally, work remains in progress as concerns computing a recombination distance between two evolutionary trees. Two main paths of investigation are currently being followed. One explores a parametric approach of the problem, while the other explores possible relations between computing the recombination distance and finding maximal common subtrees of two trees. This work is carried out in a collaboration between Marie-France Sagot, Estela Maris Rodrigues and Yoshiko Wakabayashi at the University of São Paulo, Brazil.

6.2. Modelling and analysis of the spatial organization of genomic information

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

Isochores have been studied for a long time; various members of the HELIX project, Laurent Duret, Dominique Mouchiroud, and their students, have made very significant progress towards understanding this phenomenon. A first important result is that in mammals, G+C-isochores are not at evolutionary equilibrium: we have previously shown that they tend to disappear, at least in primates and in artiodactyls. A second result is that isochores are likely created by a phenomenon called biased gene conversion (BGC) that tends to increase the G+C content of sequences located in regions of high recombination rate [28]. This phenomenom is also probably at the origin of the relationship between recombination and G+C content in *Drosophila* and the nematode [27].

Work is also in progress as concerns the analysis of nucleic acid sequences using different approaches based on a probabilistic, in particular Markovian, modelling of the sequences. This concerns methods for detecting a possible structuring of DNA sequences given a very general probabilistic criterion, and the study of the influence of context-dependent mutations on current approaches for phylogenetic reconstruction. In addition, a statistical test is being sought to estimate the validity of a sequence partitioning. This work is being conducted by Laurent Guéguen with a master's student. He also collaborates with Christian Mazza from the LAPCS, University of Lyon I.

Christelle Melo de Lima, PhD student with Christian Gautier, Didier Piau (LAPCS, University of Lyon I) and François Rechenmann, is studying, in collaboration with Laurent Guéguen, the behaviour and possible uses of Hidden Markov models with macro states. One of the applications of such models is the prediction of genes.

Finally, Eric Tannier has started working on various problems related to genomic rearrangements: inferring segments conserved under rearrangements for functional or other reasons, calculating a transposition distance between two genomes and, finally, evaluating breakpoint re-use. In the latter case, it can be shown that the number of such re-uses is strongly dependent on the type of rearrangements considered. In particular, the number decreases significantly if transpositions are also possible. Restricted versions of the transposition distance between genomes seem to present special dual properties that are currently been explored for analysing the complexity of the problem. These properties have already led to a new algorithm for calculating the inversion distance between two genomes that answers a long standing question in algorithmics and complexity analysis. This work has just been submitted.

6.3. Motif search and inference

Participants: Stéphane Declère, Christian Gautier, Laurent Guéguen, Christelle Melo de Lima, Guy Perrière, Marie-France Sagot [Correspondent], Alain Viari.

Inferring 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 this, many deep problems remain. In our work, we have tried to make progress in different directions simultaneously. This includes work on biological data to answer precise questions and get a feeling for the problems that remain still beyond us, in terms of either the biological and mathematical models, or the algorithmic complexity of the problems as we would like to formulate them.

The work with biological data has mainly concerned joint work with Anne Morgat, Hidde de Jong, and biological collaborators at the Université Joseph Fourier and the ENS Ulm. This refers to the inference of motifs in bacteria like *E. coli* and *Synechocystis* PCC 6803 related to the promoter and regulatory sequences. These are typical examples of 'structured motifs', that is, motifs composed of different parts (3 to 4 in this case) separated by non-random distances. The HELIX project remains one of the rare research groups that are trying to tackle the inference of such types of motifs. The existing algorithm to achieve this, SMILE, has already been widely distributed in an informal way and is constantly undergoing improvements. These include taking into account correlations believed to exist between different parts of a structured motif (typically as related to the existence of palindromic parts in a motif); changing the underlying data structure used to make the algorithm more efficient and thus applicable to a wider variety of potential motifs; introducing the possibility of insertions and deletions, *etc.* The algorithmical part of this work is being conducted in collaboration with Maxime Crochemore and various co-supervised students at the Université de Marne-la-Vallée, Ana Teresa Freitas and students at the Instituto Superior Técnico of Lisbon, Laurent Marsan from the Université de Versailles, Nadia Pisanti from the Università di Pisa, and Kátia Guimarães and students from the Universidade do Pernanbuco, Brazil, and finally, Costas Iliopoulos and students from King's College, London, UK. The collaboration with Lisbon has also led to an accepted publication this year [41]. Another publication, on various extensions of SMILE, is in preparation with Nadia Pisanti and Laurent Marsan.

Motifs have continued to be studied from a more 'purely' theoretical point of view as well, in particular following our previous work concerning the notion of a basis of motifs. A basis of motifs is a subset of the motifs satisfying certain input constraints that allow, given a certain algebraic operation on motifs, to generate all the motifs not in the basis. Intuitively, a basis will be interesting from both the computational and biological point of view if it is 'small enough' and thus enable very succinctly and compactly to characterize what is commonly conserved among a set of sequences. We showed [49][51] that all currently existing definitions of motifs lead to a basis having an exponential size in some of the input constraints. We have therefore continued our exploration of a definition for motifs that would be satisfying from all points of view considered. A first attempt in that direction has led to a publication this year [42]. This work is being conducted in collaboration with M. Crochemore, Nadia Pisanti and Roberto Grossi from the Università di Pisa, and Costas Iloupoulos and a student from King's College, London.

A special form of motifs may have an indirect usefulness for two other ancient areas of research in computational biology, that concern both gene finding and repeats identification. These are motifs that are employed for filtering purposes. Filtering techniques have long been developed 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. Up to now, however, either only a single such factor has been considered, or several factors, but without taking into account their relative location. In all cases, only pairwise filtering techniques have been developed. We are currently working on filters for multiple sequence comparison. Such filters take this multiplicity into account directly in the counting of common factors. It also considers the maximum number of common exact factors as well as their relative locations. Approximate common factors are also under study. This work is being carried out jointly with the PhD students Pierre Peterlongo and Julien Allali, under co-supervision of Maxime Crochemore.

Finally, work is progressing fast on the inference of RNA common sub-structures from either RNA sequences or full structures that were experimentally determined or predicted. Concerning the latter, an algorithm, MIGAL, will soon be made publicly available. It calculates all sub-structures common to two RNA structures by using a biologically more appropriate distance than simple edit distance between trees, and performs the comparison at different, increasingly more constrained levels of granularity. A publication is currently in preparation. This is joint work with Julien Allali, a PhD student at the Université of Marne-la-Vallée. Work is also in progress concerning the inference of RNA common sub-structures from RNA sequences, also called RNA structural motifs. This uses a data structure called an affix tree that is being adapted for the purposes of identifying motifs composed of small conserved sequence segments located in the close vicinity of a biological palindrome. Eventually, the two problems will be considered together. This is work done in collaboration with Stéphane Vialette from the Université d'Orsay, Christine Gaspin, Thomas Schiex and members of their group at the INRA of Toulouse, and Vincent Moulton and a student at Uppsala University in Sweden.

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

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

GEB has been implemented as an application module of GENOSTAR. All data which are manipulated by the application modules, and all the results they produce, are explicitly represented in the entity-relationship model AROM. The GEB module shares the common services offered by GENOSTAR (data manager, cartographic viewer, query and navigation tools, *etc.*). Hence, the heterogeneous (genomic, proteic, metabolic) data represented in the GEB knowledge base are viewed as a large information network. For each object, its properties as well as its relationships with the other objects of the knowledge base are accessible. Hence, the GEB module allows navigation through the information network and graphical visualization using specialized tools:

1. a cartographic viewer for the genetic elements;
2. a pathway viewer for the metabolic pathways;
3. a biochemical reaction viewer;
4. a low MW compound viewer.

GEB can be run as a stand-alone application or as a GENOSTAR module. In the latter case, GEB can exchange data and results with the other modules of the GENOSTAR environment.

All available completely-sequenced genomes can be browsed in GEB. For several bacterial model organisms, we work on data consistency checks in collaboration with biological experts at the École Normale Supérieure (Paris), Institut Pasteur (Paris), Université Joseph Fourier (Grenoble), and INSA (Lyon). The first version has been released in September 2003. GEB is freely accessible to the academic community for research and teaching purposes (http://www-helix.inrialpes.fr/article141.html).

In the framework of the thesis of Frédéric Boyer, a new approach towards an *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. We then 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 stated formally as finding a composition of partial injections which maximizes the image size. Our approach can be split into two successive problems:

1. define a unique mapping between atoms on each side of a reaction;
2. compute, on the basis of these mappings for the individual reactions, all paths ensuring a minimum transfer of atoms between given source and sink compounds.

The first problem is classical in the area of computational chemistry: it can be expressed as a *Maximum Common Subgraph* problem, that is, the problem of finding an isomorphism between two graphs by deleting the minimum number of edges. We have proven that the second problem, called the *Maximal Partial Injections Composition* problem, is PSPACE-HARD. In order to deal with this latter problem, we have designed an algorithm to construct an automaton that accepts exactly all words corresponding to maximal compositions of partial injections from a source to a sink compound. This algorithm has been successfully applied to the reconstruction of standard metabolic pathways with acceptable run-times.

6.5. Modelling and simulation of genetic regulatory networks

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

A substantial part of the activities has concerned the dissemination of the method and the tool for the qualitative simulation of genetic regulatory networks developed in the HELIX project (section 3.5). The mathematical and computational basis of the method has been accepted fpr the *Bulletin of Mathematical Biology* and presented at the annual conference *Hybrid Systems: Computation and Control* [45]. The computer tool GENETIC NETWORK ANALYZER (GNA) (section 5.7)) has been brought to the attention of the bioinformatics community by means of a paper in the journal *Bioinformatics* [20]. Over the past few years, more than 70 groups have asked a copy of the current version 5.0 or of a previous version of the program. An application of the method and the tool, concerning the modelling and simulation of the initiation of sporulation in *Bacillus subtilis* has also been accepted for publication in the *Bulletin of Mathematical Biology*.

In parallel, work has continued on extensions of the qualitative simulation method and of the computer tool:

- The PhD research of Grégory Batt focuses on the *validation of qualitative models of genetic regulatory networks* by means of gene expression data. An approach based on model-checking techniques has been chosen and explored in a pilot study, presented at the annual *International Workshop on Qualitative Reasoning* [40]. The work is supported by the inter-EPST project ‘Validation de réseaux de régulation génique : Régulation globale de la transcription chez *Escherichia coli* et *Synechocystis* PCC 6803’ (2002-2004).
- In collaboration with Richard Casey and Jean-Luc Gouzé of the project Comore at INRIA Sophia-Antipolis, we have started to work on the mathematical characterization of attractors of piecewise-linear differential equation models of genetic regulatory networks, the class of models employed by the qualitative simulation method. The research is supported by the ACI IMPBio ‘BacAttract. Analyse théorique et expérimentale d’attracteurs de réseaux de régulation génique : régulation globale de la transcription chez *Escherichia coli* et *Synechocystis* PCC 6803’ (2003-2006), and the ARC INRIA ‘GDyn : Analyse dynamique de réseaux de régulation génique’ (2002-2004). The latter project has a larger, more exploratory scope.
- A new user interface for GNA has been developed by Michel Page and Hidde de Jong, completing work initiated by Céline Hernandez during her stay as an associate engineer in the HELIX project (2000-2002). A *beta* version of the new simulator has been distributed to selected partners and will be made available to the research community in 2004.

In the framework of the above-mentioned projects, we have worked on the application of the qualitative simulation method and tool in collaboration with experimental biologists, in particular the groups of Johannes Geiselmann (Université Joseph Fourier, Grenoble) and Jean Houmard (École Normale Supérieure, Paris). The applications center around regulatory processes in prokaryotes, such as the global regulation of transcription in *E. coli*, the transduction of external signals by cyclic nucleotides in *Synechocystis*, and the response of *E. coli* to infection by the bacteriophage Mu. The work on these experimental systems, amounting to the construction of qualitative models and their validation by means of experimental data, has received an additional stimulus from the arrival of Delphine Ropers (INRIA post-doc) and a master's student. It also involves other HELIX members, most notably Anne Morgat and Marie-France Sagot, for issues concerning biological knowledge bases and sequence analysis.

6.6. Computational proteomics

Participants: Estelle Nugues, Erwan Reguer, Alain Viari [Correspondent].

Work on computational proteomics has focused on the development of PepLine, a computer tool for the computer-supported interpretation of mass spectrometry data (section 5.14). It has been developed by Estelle Nugues, Erwan Reguer, and Alain Viari, in collaboration with members of the Laboratoire Chimie des Protéines of the CEA Grenoble.

PepLine consists of the programs PepMap and Taggor. PepMap, developed in 2002, maps peptide sequence tags (PSTs) on protein sequences or on the six translated phases of a genomic sequence. A PST is a peptide sequence of 3 to 5 amino acids long, flanked by two masses corresponding to the left and right-adjacent stretches of amino acids. By means of clustering algorithms, PepMap groups together PSTs that have been mapped to the same protein or to the same gene. Taggor, developed in 2003, provides input to PepMap. More specifically, it generates a list of PSTs from MS/MS data.

In order to evaluate the ability of PepMap to efficiently identify proteins from MS/MS data, two validation studies have been carried out. The data used for the first study, concerning MS/MS data on *Arabidopsis thaliana* proteins, have been provided by the Laboratoire Chimie des Protéines. The performance of PepLine on this data set has been compared with that of Mascot, the most commonly-used program for protein identification. Mascot identified 40 proteins, whereas PepLine arrived at 65 proteins (32 proteins in common with Mascot). Many of the 33 proteins identified by PepLine alone are homologs of the proteins identified by both. In a second study, PepLine and Mascot were compared on a database of *Arabidopsis thaliana* proteins, characterized by a high level of internal consistency between genomic and proteomic data, and containing valuable gene-protein and protein-gene information.

The PST approach embedded in PepLine has demonstrated its usefulness for genome annotation in a test on the localization of intron/extron boundaries. Using the tool, it has been possible to infer the presence of a gene included in the Swiss-Prot database, but not included in the TAIR database.

7. Contracts and Grants with Industry

7.1. Introduction

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

GenoStar is an integrated bioinformatics environment, which has been developed by a consortium of four academic and industrial partners: INRIA, Institut Pasteur, Hybrigenics, and GENOME express. The consortium has been funded by the French Ministry for Research ('Programme Génomique' and 'Programme Gen-Homme'). In addition, GenoStar was an 'action de développement' of the INRIA and as such received financial support in the period 2000-2002.

7.2. PepMap

Participants: Estelle Nugues, Erwan Reguer, Alain Viari [Correspondent].

The PepMap project unites INRIA, CEA, and GENOME express in the development of a tool for the mapping of peptide sequence tags, obtained from MS/MS experiments, to sequences in protein databanks or to the complete translated genome of an organism. This allows to locate the gene coding for the protein (see sections 3.6 and 5.14).

7.3. Sanofi

Participant: Alain Viari [Correspondent].

In September 2002, HELIX started a contractual relation with the company Sanofi Synthélabo in Toulouse. The collaboration concerns the analysis of proteomic data. In particular, the project aims at connecting mass spectrometry data with other biological data available from public or private sources.

8. Other Grants and Activities

8.1. National projects

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

Project name	**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	**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	**GEB**
Coordinators	A. Morgat
HELIX participants	A. Morgat, A. Viari, S. Declère
Type	inter-EPST Bioinformatique (2001-2003)
Web page	http://www-helix.inrialpes.fr/article.php3?id_article=141

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)
Web page	http://bacillus.inrialpes.fr/gna/voorstellen/interepst02/validationproject.html

8.2. Projects funded by international organisms or including international teams

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	

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)
Web page	http://www.inrialpes.fr/helix/people/sagot/projects/epst/2002/epst2002_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	**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	**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	**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/

9. Dissemination

9.1. Talks

Grégory Batt

Title	Event and location	Date
Qualitative analysis of genetic regulatory networks: A model-checking approach	IJCAI-03 Workshop on Model Checking and Artificial Intelligence (MoChArt), Acapulco (Mexico)	10/08/03

Qualitative analysis of genetic regulatory networks: A model-checking approach	Seventeenth International Workshop on Qualitative Reasoning (QR-03), Brasilia (Brazil)	22/08/03
Formal validation of models of genetic regulatory networks	ECCB Satellite Meeting on Modeling and Simulation of Biological Regulatory Processes, Paris	27/09/03

Frédéric Boyer

Title	Event and location	Date
Ab initio reconstruction of metabolic pathways	ECCB 2003	29/9/03

Laurent Duret

Title	Event and location	Date
Databases and tools for large scale phylogenomic analyses	Swiss Institute of Bioinformatics (SIB), Geneva (Switzerland)	15/04/03
Evolution of isochores in mammalian genomes	Charlesworth lab, Edinburgh (UK)	07/05/03
Databases and tools for large scale phylogenomic analyses	INRA, Toulouse	28/05/03
Automatic search for orthologous or paralogous genes in sequence databases	Newport Beach (USA)	26-29/06/03
Evolution of isochores in mammalian genomes	Computational Biology Research Center (CBU), Bergen, Norway	12/09/03

Christian Gautier

Title	Event and location	Date
Genomic mapping and molecular processes	Brazilian Symposium on Mathematical and Computational Biology, Rio de Janeiro, Brazil	26/11/03

Manolo Gouy

Title	Event and location	Date
Contribution à l'analyse phylogénétique des séquences génomiques complètes de procaryotes	University of Lausanne (Switzerland)	27/11/03

Hidde de Jong

Title	Event and location	Date
Qualitative simulation of genetic regulatory networks	Workshop on Dynamical Stochastic Modeling in Biology, University of Copenhagen	08/01/03
Simulation qualitative de réseaux de régulation génique	Séminaire MAGMA, Institut de Mathématiques de Luminy	06/02/03
Qualitative simulation of genetic regulatory networks	Seminar Vlaams interuniversitair Instituut voor Biotechnologie, University of Ghent	24/03/03
Hybrid modeling and simulation of genetic regulatory networks: A qualitative approach	International Workshop on Hybrid Systems: Computation and Control (HSCC'03), Prague	03/04/03
Qualitative modeling and simulation of genetic regulatory networks	Séminaire Apprentissage et Cognition, laboratoire Leibniz, IMAG, Grenoble	07/04/03
Qualitative simulation of genetic regulatory networks	Séminaire Laboratoire des Interactions Plantes-Microorganismes, INRA/CNRS, Toulouse	11/04/03
Modélisation et simulation qualitative de réseaux de régulation génique	Séminaire thématique de Bioinformatique de la Génopole Ouest, IRISA, Rennes (with J. Geiselmann)	12/06/03
Qualitative modeling and simulation of genetic regulatory networks	European Symposium on Intelligent Technologies, Hybrid Systems, and their Implementation on Smart Adaptive Systems, Oulu, Finland	11/07/03
Modeling and simulation of genetic regulatory networks	International Symposium on Positive Systems (POSTA 2003), University of Rome La Sapienza	30/08/03
Modeling and simulation of genetic regulatory networks	Troisième École thématique CNRS-INRA de Biologie végétale, Roscoff	09/10/03

Vers des cellules virtuelles ?	Troisième École thématique CNRS-INRA de Biologie végétale, Roscoff (with F. Rechenmann)	10/10/03
Qualitative modeling and simulation of genetic regulatory networks	Séminaire AC Graphes et Algorithmique pour la Bioinformatique, LIRMM, Montpellier	24/10/03
Qualitative modeling and simulation of prokaryote regulatory networks	Journées thématiques en Bioinformatique, Bordeaux	11/12/03

Jean Lobry

Title	Event and location	Date
De l'usage du code et des acides-aminés chez les bactéries thermophiles	Université de Lausanne, Suisse	19/02/03
Le projet seqinR pour l'analyse des séquences biologiques	Université de Lausanne, Suisse	02/07/03

Estelle Nugues

Title	Event and location	Date
Présentation des activités, avancées et résultats des plates-formes bioinformatiques du génopole Rhône-Alpes (avec C. Bruley)	Journées 'Bioinformatique des génopoles'	22/10/03

Guy Perrière

Title	Event and location	Date
Homologous genes databases for bacteria: application to phylogenomic studies	Microbial Genomes Conference, New Orleans (USA)	31/01/03
Data retrieval and handling tools for the PBIL gene family databases	European Conference on Computational Biology (ECCB) 2003, Paris	29/09/03
Banques de données de séquences biologiques	Séminaire IN'Tech, Bioinformatics: from genomic and post-genomic data to biological knowledge, Lyon	23/10/03

François Rechenmann

Title	Event and location	Date
Présentation et démonstration de GenoStar	Rhône-Alpes Genopole, École Normale Supérieure, Lyon (with A. Viari)	21/01/03
Integrated computer environments for exploratory genomics	Seminar Minatec (Micro et nano-technologies), Palais des Congrès de Grenoble	22/09/03
La bioinformatique : Modélisation et analyse des données génomiques et post-génomiques	Journée Réflexion sur les logiciels : valorisation et diffusion de logiciels, les grands challenges applicatifs, Société de Mathématiques Appliquées et Industrielles, Rocquencourt	09/10/03
Vers des cellules virtuelles ?	Troisième École thématique CNRS-INRA de Biologie végétale, Roscoff (with H. de Jong)	10/10/03

Erwan Reguer

Title	Event and location	Date
A software pipeline dedicated to automatic MS/MS data analysis	ECCB 2003	27/9/03

Marie-France Sagot

Title	Event and location	Date
Some combinatorial algorithms for molecular biology	Seminar of the Department of Computer Science, University of Helsinki, Finland	08/05/03
Tiny steps towards addressing one part of the DNA rearrangement puzzle	Third Haifa Workshop on Interdisciplinary Applications of Graph Theory, Combinatorics and Computing, University of Haifa, Israel	29/05/03

Motifs and regulation	Seminar of the Department of Computer Science, Tel Aviv University, Israel	01/06/03
Tiny steps towards addressing a small part of the DNA rearrangement puzzle	Bertinoro Computational Biology Meeting, University of Bologna Residential Center Bertinoro (Forlé), Italy	07/06/03
Bases of repetead motifs with don’t cares	Workshop on Combinatorics, Algorithms, and Applications, Ubatuba, Brazil	01/09/03
Tiny steps for one part of the DNA rearrangement puzzle	EWM, CIRM, Luminy, Marseille	04/11/03
Biological motif inference	Seminar of the Linnaeus Center for Bioinformatics, Uppsala University, Sweden	11/11/03

9.2. Organization of conferences, workshops and meetings

Hidde de Jong

Type	Location	Date
Organization of ECCB satellite meeting on the Modeling and Simulation of Biological Regulatory Processes (with D. Thieffry)	ENS Paris	01/10/03

Guy Perrière

Type	Location	Date
ECCB 2003 Satellite Workshop: Sequence Databases and Ontologies	Pasteur Institute, Paris	

François Rechenmann

Type	Location	Date
Organization of "Atelier GenoStar" (with A. Viari)	INRIA, Grenoble	24-25/03/03
Participation at Fête de la Science, demonstration of ISee, GEB, and GenoStar	Grenoble	17-19/10/03
Organization of Séminaire IN’Tech "Des données génomiques et post-génomiques aux connaissances biologiques", in association with Rhône-Alpes Génopole	Université Claude Bernard, Lyon	23/10/03

Marie-France Sagot

Type	Location	Date
European Conference on Computational Biology (ECCB) 2003	Centre de Conférences, Parc de la Villette, Paris	27-30/09/03
14th series of the Seminar Algorithmics and Biology: Algorithmics and Combinatorics in Biology	CNRS Amphitheater, La Doua, Lyon	-2-4/04/03

9.3. Editorial and reviewing activities

Laurent Duret

Type	Journal or conference
Member 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
Member Program Committee	International Workshop on Qualitative Reasoning (QR)

Guy Perrière

Type	Journal or conference

François Rechenmann

Type	Journal or conference
Editorial Board	*Bioinformatics*, OUP

Marie-France Sagot

Type	Journal or conference
Member Steering Committee	French national conference on Bioinformatics, Jobim (until mid-2003)
Member Steering Committee	European Conference on Computational Biology (ECCB)
Editorial Board	*Journal of Discrete Algorithms*, Elsevier
Editorial Board	*Research in Microbiology*, Elsevier
Editorial Board	*Lecture Notes in BioInformatics*, Springer Verlag
Editorial Board	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*, IEEE and ACM Press
Member Program Committee	RECOMB, CPM, ECCB, SPIRE

Alain Viari

Type	Journal or conference
Member Steering Committee	French national conference on Bioinformatics, Jobim (until mid-2003)

9.4. Administrative activities

Christian Gautier was ‘chargé de mission pour la bioinformatique’ at the CNRS and responsible for the Bioinformatics inter EPST Program.

Manolo Gouy is a member of the Conseil National des Universités, section 67 and of the scientific committee of the Institut Français de la Biodiversité.

Hidde de Jong is a member of the ‘Commission de Spécialistes’ (section 27) of the Université de Provence (Marseille).

Jean Lobry is a member of the scientific board of the University Claude Bernard Lyon I and a member of the executive board of its Electronic Publication Ressource Facility.

Guy Perrière is a member of the section 67 of the ‘Commission de Spécialistes de l’Enseignement Supérieur’ of the UCBL. Since Septembre 2001, he is second vice-president of this commission. He is also member of the managing committee of the IFR 41 (Institut des Sciences et Méthodes de l’Écologie et de l’Évolution) of the UCBL.

François Rechenmann was a member of the ‘Comité de Coordination des Sciences du Vivant’ (CCSV) and of the scientific committee of the Bioinformatics inter-EPST Program. He is also a member of the technical committee Bioinformatics of GenoPlante. He was coordinator for bioinformatics in the Réseau National des Génopoles and responsible of the bioinformatics platform for the Rhône-Alpes Génopole. He is the scientific head of the Genostar consortium. He is member of the ‘Commission de Spécialistes’ (section 27) of the UFR Informatique et Mathématiques Appliquées (Université Joseph Fourier). He was president of the recruiting committee for CR2 positions at the INRIA Rhône-Alpes in 2003.

Marie-France Sagot is a member of the scientific committee of the ACI Cryptologie of the Ministry of Research, 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 nominated member (substitute) of the ‘Commission d’Évaluation’ of the INRIA. She participated in the recruiting committee for CR2 positions at the INRIA Rennes in 2003.

Alain Viari is a member of the ‘Comité de Coordination des STIC’ (CCSTIC). He was a nominated member (collège A) of the Section 65 (cellular biology) of the Conseil National des Universités.

9.5. Teaching activities

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

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

As part of the new LMD system being set up at all Universities in France, members of the project have created a complete interdisciplinary module of the LMD offering training in biology, mathematics and computer science. The module is called 'Mathématique et Informatique du Vivant'. It leads to master's diplomas in the scientific and medical fields.

A second important educational activity of the project concerns not disconnecting biology from the teaching of mathematics to biologists. To this purpose, various members of the project work in the context of an INCA ('Initiative Campus Action') project together with other Universities in the Rhône-Alpes region to maintain a web site (http://nte-serveur.univ-lyon1.fr/nte/mathsv/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 cover biostatistics, biomathematics and bioinformatics while the second concern general and population genetics, and molecular evolution.

Finally, members of the project have participated in, or sometimes organized numerous courses or teaching modules including at the international level (such as, for instance, the creation and support of a master's course in Ho-Chi-Minh, Vietnam).

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

Laurent Duret

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

Manolo Gouy

Subject	Year	Location	Hours
Molecular phylogeny	3 to 5	UCBL, ENS Lyon, INSA Lyon	33

Hidde de Jong

Subject	Year	Location	Hours
Modelling and simulation of genetic regulatory networks, with G. Batt	4	Université Joseph Fourier, Grenoble	13
Modelling and simulation of genetic regulatory networks	4	INSA, Lyon	14

Guy Perrière

Subject	Year	Location	Hours
Biodiversity of model organisms	3	ENS Lyon	4
Horizontal gene transfer	4	INSA Lyon	8
Plasticity of bacterial genomes	4	UCBL	4

Introduction to bioinformatics	5	UCBL	11
Motif inference	4	InaPG, Paris	3
Motif inference and genome rearrangements	5	INSA Lyon	20

François Rechenmann

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

Marie-France Sagot

Subject	Year	Location	Hours
Algorithmics for biology	5	Université de Marne-la-Vallée	8
Motif inference	5	Université de Marne-la-Vallée	10
Algorithmical complexity and NP-completeness	5	Pasteur Institute, Paris	3
Genome rearrangements	5	Pasteur Institute, Paris	3
Motif inference	5+	Lipari School on Algorithmics for Data Mining and Pattern Discovery, Italy	3

10. Bibliography

Major publications by the team in recent years

[1] L. Duret, G. Perrière, M. Gouy. *HOVERGEN: Database and software for comparative analysis of homologous vertebrate genes.* S. Letovsky, editor, in « Bioinformatics Databases and Systems », Kluwer Academic Publishers, Boston, 1999, pages 13-29.

[2] N. Galtier, N. Tourasse, M. Gouy. *A nonhyperthermophilic common ancestor to extant life forms.* in « Science », volume 282, 1999, pages 220-221.

[3] J. Lobry. *Asymmetric substitution patterns in the two DNA strands of bacteria.* in « Molecular Biology and Evolution », number 5, volume 13, 1996, pages 660-665.

[4] L. Marsan, M.-F. Sagot. *Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification.* in « Journal of Computational Biology », volume 7, 2000, pages 345-362.

[5] G. Matassi, P. Sharp, C. Gautier. *Chromosomal location effects on gene sequence evolution in mammals.* in « Current Biology », number 15, volume 9, 1999, pages 786-791.

[6] C. Médigue, F. Rechenmann, A. Danchin, A. Viari. *Imagene : an integrated computer environment for sequence annotation and analysis.* in « Bioinformatics », number 15, 1999, pages 2-15.

[7] C. MÉDIGUE, M. ROSE, A. VIARI, A. DANCHIN. *Detecting and analyzing DNA sequencing errors: toward a higher quality of the Bacillus subtilis genome sequence.* in « Genome Research », number 11, volume 9, 1999, pages 1116-1127.

[8] G. PERRIÈRE, L. DURET, M. GOUY. *HOBACGEN: Database system for comparative genomics in bacteria.* in « Genome Research », volume 10, 2000, pages 379-385.

[9] D. PROUX, F. RECHENMANN, L. JULLIARD. *A pragmatic information extraction strategy for gathering data on genetic interactions.* in « Proceedings of the 8th International Conference on Intelligent Systems in Molecular Biology (ISMB 2000) », AAAI Press, pages 279-285, 2000.

Articles in referred journals and book chapters

[10] G. ACHAZ, E. COISSAC, P. NETTER, E. ROCHA. *Associations between inverted repeats and the structural evolution of bacterial genomes.* in « Genetics. », number 4, volume 164, 2003, pages 1279-1289.

[11] O. ASSOSSOU, F. BESSON, J. ROUAULT, F. PERSAT, C. BRISSON, L. DURET, J. FERRANDIZ, M. MAYENCON, F. PEYRON, S. PICOT. *Subcellular localization of 14-3-3 proteins in Toxoplasma gondii tachyzoites and evidence for a lipid raft-associated form.* in « Fems. Microbiol. Lett. », volume 224, 2003, pages 161-168.

[12] A. BAUDIN-BAILLIEU, E. FERNANDEZ-BELLOT, F. REINE, E. COISSAC, C. CULLIN. *Conservation of the prion properties of Ure2p through evolution.* in « Mol. Biol. Cell. », number 8, volume 14, 2003, pages 3449-3458.

[13] I. BIECHE, A. LAURENT, I. LAURENDEAU, L. DURET, Y. GIOVANGRANDI, J. FRENDO, M. OLIVI, J. FAUSSER, D. EVAIN-BRION, M. VIDAUD. *Placenta-specific INSL4 expression is mediated by a human endogenous retrovirus element.* in « Biol. Reprod. », number 4, volume 68, 2003, pages 1422-1429.

[14] P. BLAYO, P. ROUZÉ, M.-F. SAGOT. *Orphan gene finding. An exon assembly approach.* in « Theor. Comput. Sci. », volume 290, 2003, pages 1407-1431.

[15] F. BOYER, A. VIARI. *Ab initio reconstruction of metabolic pathways.* in « Bioinformatics », number suppl. 2, volume 19, 2003, pages 26-34.

[16] A. CULHANE, G. PERRIÈRE, D. HIGGINS. *Cross platform comparison and visualisation of gene expression data using co-inertia analysis.* in « BMC Bioinformatics », volume 4, 2003, pages 59.

[17] V. DAUBIN, E. LERAT, G. PERRIÈRE. *The source of laterally transferred genes in bacterial genomes.* in « Genome Biol. », volume 4, 2003, pages 57.

[18] V. DAUBIN, G. PERRIÈRE. *G+C3 structuring along the genome: a common feature in prokaryotes.* in « Mol. Biol. Evol. », number 4, volume 20, 2003, pages 471-483.

[19] H. DE JONG. *Simulation qualitative.* L. TRAVÉ-MASSUYÈS, P. DAGUE, editors, in « Modèles et raisonnements qualitatifs », Hermès,, Paris, 2003, pages 269-329.

[20] H. de Jong, J. Geiselmann, C. Hernandez, M. Page. *Genetic Network Analyzer: Qualitative simulation of genetic regulatory networks.* in « Bioinformatics », number 3, volume 19, 2003, pages 336-344.

[21] H. de Jong, J. Geiselmann, D. Thieffry. *Qualitative modeling and simulation of developmental regulatory networks.* S. Kumar, P. Bentley, editors, in « On Growth, Form, and Computers », Academic Press, London, 2003, pages 109-143.

[22] G. Devulder, G. Perrière, F. Baty, J.-P. Flandrois. *BIBI, a Bioinformatics Bacterial Identification Tool.* in « J. Clin. Microbiol. », number 4, volume 41, 2003, pages 1785-1787.

[23] P. Durand, C. Médigue, A. Morgat, Y. Vandenbrouck, A. Viari, F. Rechenmann. *Integration of data and methods for genome analysis.* in « Current Opinion in Drug Discovery and Development », number 3, volume 6, 2003, pages 346-352.

[24] J. Grassot, G. Mouchiroud, G. Perrière. *RTKdb: database of Receptor Tyrosine Kinase.* in « Nucleic Acids Res. », number 1, volume 31, 2003, pages 353-358.

[25] J. Lobry, D. Chessel. *Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria.* in « J. Appl. Genet. », number 2, volume 44, 2003, pages 235-261.

[26] J. Lobry, J. Louarn. *Polarisation of prokaryotic chromosomes.* in « Curr. Opin. Microbiol. », number 2, volume 6, 2003, pages 101-108.

[27] G. Marais, D. Mouchiroud, L. Duret. *Neutral effect of recombination on base composition in Drosophila.* in « Genet. Res. », number 2, volume 81, 2003, pages 79-87..

[28] J. Meunier, L. Duret. *Recombination drives the evolution of GC-content in the human genome.* in « Mol. Biol. Evol. », 2003, in press.

[29] C. Mougel, J. Thioulouse, G. Perrière, X. Nesme. *A mathematical method for determining genome divergence and species delineation using AFLP.* in « Int. J. Syst. Evol. Microbiol. », number Pt 2, volume 52, 2003, pages 573-586.

[30] G. Perrière, C. Combet, S. Penel, C. Blanchet, J. Thioulouse, C. Geourjon, J. Grassot, C. Charavay, M. Gouy, L. Duret, G. Deléage. *Integrated databanks access and sequence/structure analysis services at the PBIL.* in « Nucleic Acids Res. », number 13, volume 31, 2003, pages 3393-3399.

[31] G. Perrière, J. Thioulouse. *Use of correspondence discriminant analysis to predict the subcellular location of bacterial proteins.* in « Comput. Methods Programs Biomed. », number 2, volume 70, 2003, pages 99-105.

[32] N. Pisanti, M.-F. Sagot. *Network expression inference.* J. Berstel, D. Perrin, editors, in « Applied Combinatorics on Words », Cambridge University Press, 2003, in press.

[33] E. Reguer, E. Mouton, E. Nugues, R. Cahuzac, T. Vermat, M. Ferro, J. Garin, A. Viari. *Outils bioinformatiques pour la protéomique à haut débit : Identification automatique de protéines par MS/MS.* in

« Spectra Analysis », number 233, volume 32, 2003, pages 25-28.

[34] N. Reymond, H. Charles, L. Duret, F. Calevro, G. Beslon, J.-M. Fayard. *Research of optimized oligonucleotide probes for microarrays.* in « Bioinformatics », 2003, in press.

[35] C. Rizzon, E. Martin, G. Marais, L. Duret, L. Segalat, C. Biemont. *Patterns of selection against transposons inferred from the distribution of Tc1, Tc3, and Tc5 insertions in the mut-7 line of the nematode Caenorhabditis elegans.* in « Genetics », 2003, in press.

[36] S. Robin, J.-J. Daudin, H. Richard, M.-F. Sagot, S. Schbath. *Occurrence probability of structured motifs in random sequences.* in « J. Comp. Biol. », volume 9, 2003, pages 761-773.

[37] M.-F. Sagot, Y. Wakabayashi. *Pattern inference under many guises.* B. Reed, C. L. Sales, editors, in « Recent advances in algorithms and combinatorics », Springer-Verlag, 2003, pages 245-288.

[38] D. Schneider, A. Latifi, H. de Jong, J. Geiselmann. *Dynamics and simulation of the genetic regulatory networks in bacteria.* in « Recent Research Developments in Genetics », volume 3, 2003, pages 55-83.

Publications in Conferences and Workshops

[39] G. Batt, H. de Jong, J. Geiselmann, M. Page. *Qualitative analysis of genetic regulatory networks: A model-checking approach.* in « Working Notes of IJCAI Workshop on Model Checking and Artificial Intelligence », pages 51-58, 2003.

[40] G. Batt, H. de Jong, J. Geiselmann, M. Page. *Qualitative analysis of genetic regulatory networks: A model-checking approach.* in « Working Notes of Seventeenth International Workshop on Qualitative Reasoning (QR-03) », B. Bredeweg, P. Salles, editors, pages 31-38, 2003.

[41] A. Carvalho, A. Freitas, A. Oliveira, M.-F. Sagot. *A parallel algorithm for the extraction of structured motifs.* in « Proc. ACM Symposium on Applied Computing (SAC'04) », ACM Press, W. Jones, M. Palakal, J. Chen, editors, 2003, in press.

[42] M. Crochemore, C. Iliopoulos, M. Mohamed, M.-F. Sagot. *Longest repeated motif with k don't cares.* in « Latin America Theoretical INformatics (LATIN'04) », series Lecture Notes in Computer Science, Springer-Verlag, M. Farach-Colton, editor, 2003, in press.

[43] H. de Jong. *Modeling and simulation of genetic regulatory networks.* in « Positive Systems (POSTA'03) », series Lectures Notes in Control and Information Sciences, volume 294, Springer-Verlag, L. Benvenuti, A. D. Santis, L. Farina, editors, pages 111-118, 2003.

[44] H. de Jong, J. Geiselmann, C. Hernandez, M. Page. *Qualitative simulation of the initiation of sporulation in Bacillus subtilis.* in « Proc. of the 5th ESMTB Conference on Mathematical Modeling & Computing in Biology and Medecine », V. Capasso, editor, pages 23-28, 2003.

[45] H. de Jong, J.-L. Gouzé, C. Hernandez, M. Page, T. Sari, J. Geiselmann. *Hybrid modeling and simulation of genetic regulatory networks: A qualitative approach.* in « Hybrid Systems: Computation and

Control (HSCC'03) », series Lecture Notes in Computer Science, volume 2623, Springer-Verlag, A. Pnueli, O. Maler, editors, pages 267-282, 2003.

[46] D. Eveillard, D. Ropers, H. de Jong, C. Branlant, A. Bockmayr. *Multiscale modeling of alternative splicing regulation.* in « Computational Methods in Systems Biology (CMSB'03) », series Lecture Notes in Computer Science, volume 2602, Springer-Verlag, C. Priami, editor, pages 75-87, 2003.

[47] G. Perrière, J.-F. Dufayard, S. Penel, J. Grassot, L. Duret, M. Gouy. *Data Retrieval and Handling Tools for the PBIL Gene Family Databases.* in « Proc. European Conference on Computational Biology », M.-F. S. C. Christophe, editor, 2003, http://www-helix.inrialpes.fr/IMG/pdf/perriere.pdf, (short paper).

[48] H. Philippe, D. Casane, S. Gribaldo, P. Lopez, J. Meunier. *Heterotachy and functional shift in protein evolution.* International Union of Biochemistry and Molecular Biology, 2003, in press.

[49] N. Pisanti, M. Crochemore, R. Grossi, M.-F. Sagot. *A basis of tiling motifs for generating repeated patterns and its complexity for higher quorum.* in « 28th International Symposium on Mathematical Foundations of Computer Science (MFCS'03) », series Lecture Notes in Computer Science, volume 2747, Springer-Verlag, B. Rovan, P. Vojtas, editors, pages 622-632, 2003.

[50] E. Reguer, E. Nugues, R. Cahuzac, M. Ferro, T. Vermat, E. Mouton, J. Garin. *A software pipeline dedicated to automatic MS/MS data analysis.* in « Proc. European Conference on Computational Biology », M.-F. S. C. Christophe, editor, 2003, http://www-helix.inrialpes.fr/IMG/pdf/reguer.pdf, (short paper).

Internal Reports

[51] N. Pisanti, M. Crochemore, R. Grossi, M.-F. Sagot. *Bases of motifs for generating repeated patterns with don't cares.* Technical report, number TR-03-02, University of Pisa, Department of Computer Science, 2003, ftp://ftp.di.unipi.it/pub/techreports/TR-03-02.ps.Z.

[52] I. Vatcheva, O. Bernard, H. de Jong, N. Mars. *Experiment selection for the discrimination of semi-quantitative models of dynamical systems.* Technical report, number 4940, INRIA, 2003, http://www.inria.fr/rrrt/rr-4940.html.