Project-Team: HELIX

Informatics and genomics
Theme bio
Rhone-Alpes

January 31, 2007

Contents

debug : moreinfo

Texte a mettre

1 Team

debug : Module named "en-tete" for the project HELIX, section composition

Head of the team

Alain Viari [Profession=Chercheur] [Category=INRIA] [Research Director (DR) Inria]

Project assistant

Françoise de Coninck [Profession=Assistant] [Category=INRIA] [Secretary (SAR) Inria]

Research scientists (Inria)

Hidde de Jong [Profession=Chercheur] [Category=INRIA] [Research Director (DR) Inria] [HDR=habilite]

François Rechenmann [Profession=Chercheur] [Category=INRIA] [Research Director (DR) Inria] [HDR=habilite]

Delphine Ropers [Profession=Chercheur] [Category=INRIA] [Research Associate (CR) Inria (since Sept. 2006)]

Marie-France Sagot [Profession=Chercheur] [Category=INRIA] [Research Director (DR) Inria] [HDR=habilite]

Eric Tannier [Profession=Chercheur] [Category=INRIA] [Research Associate (CR) Inria]

Alain Viari [Profession=Chercheur] [Category=INRIA] [Research Director (DR) Inria]

Research scientists (external)

Sandrine Charles [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Universite Claude Bernard] [HDR=habilite]

Vincent Daubin [Profession=Chercheur] [Category=CNRS] [Research Associate (CR) Cnrs]

Laurent Duret [Profession=Chercheur] [Category=CNRS] [Research Director (DR) Cnrs] [HDR=habilite]

Christian Gautier [Profession=Enseignant] [Category=UnivFr] [Professor, Universite Claude Bernard] [HDR=habilite]

Johannes Geiselmann [Profession=Enseignant] [Category=UnivFr] [Professor, Université Joseph Fourier, since October 2006] [HDR=habilite]

Philippe Genoud [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Universite Joseph Fourier]

Manolo Gouy [Profession=Chercheur] [Category=CNRS] [Research Director (DR) Cnrs] [HDR=habilite]

Laurent Guéguen [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Universite Claude Bernard]

Daniel Kahn [Profession=Chercheur] [Category=AutreEtablissementPublic] [DR Inra] [HDR=habilite]

Jean Lobry [Profession=Enseignant] [Category=UnivFr] [Professor, Universite Claude Bernard] [HDR=habilite]

Gabriel Marais [Profession=Chercheur] [Category=CNRS] [Research Associate (CR) Cnrs]

Dominique Mouchiroud [Profession=Enseignant] [Category=UnivFr] [Professor, Universite Claude Bernard] [HDR=habilite]

Sylvain Mousset [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Universite Claude Bernard]

Michel Page [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Universite Mendes France]

Guy Perrière [Profession=Chercheur] [Category=CNRS] [Research Director (DR) Cnrs] [HDR=habilite]

Raquel Tavares [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Universite Claude Bernard]

Jean Thioulouse [Profession=Chercheur] [Category=CNRS] [Research Director (DR) Cnrs] [HDR=habilite]

Danielle Ziébelin [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Universite Joseph Fourier] [HDR=habilite]

External members

Eric Coissac [Profession=Enseignant] [Category=UnivFr] [Associate Professor, Université Joseph Fourier] [HDR=habilite]

Corinne Lachaize [Profession=Technique] [Category=UnivEtrangere] [Project technical staff, Swiss Institute of Bioinformatics]

Anne Morgat [Profession=Technique] [Category=UnivEtrangere] [Project technical staff, Swiss Institute of Bioinformatics]

Visitors

José Luis Aguirre [Profession=Enseignant] [Category=UnivEtrangere] [Professor, Technologico de Monterrey, Mexico, 1 year]

Fabio Martinez [Profession=Visiteur] [Category=UnivEtrangere] [Federal University of Mato Grosso do Sul, Brazil, 3 weeks]

Alberto Marchetti-Spaccamela [Profession=Visiteur] [Category=UnivEtrangere] [University of Rome, Italy, 1 week]

Christelle Melo de Lima [Profession=Visiteur] [Category=UnivFr] [University of Lyon II, 1 year]

Mark Musters [Profession=Visiteur] [Category=UnivEtrangere] [Eindhoven University of Technology, Netherlands, 3 months]

Vincent Navratil [Profession=Visiteur] [Category=AutreEtablissementPublic] [INRA, 2 years]

Alair Pereira do Lago [Profession=Visiteur] [Category=UnivEtrangere] [University of São Paulo, Brazil, 2 weeks]

Pierre Peterlongo [Profession=Visiteur] [Category=UnivFr] [University of Marne-la-Vallée, various short visits]

Paulo Gustavo Soares da Fonseca [Profession=Visiteur] [Category=UnivEtrangere] [Federal University of Pernambuco, Recife, Brazil, 2 years]

Leen Stougie [Profession=Visiteur] [Category=UnivEtrangere] [Eindhoven University of Technology, Netherlands, 1 week]

Technical staff

Bruno Besson [Profession=Technique] [Category=INRIA] [CDD European project Hygeia, since November 2006]

Ludovic Cottret [Profession=Technique] [Category=AutreAffiliation] [CDD ANR]

Stephane Delmotte [Profession=Technique] [Category=CNRS] [Technical Staff, CNRS]

Estelle Dumas [Profession=Technique] [Category=INRIA] [CDD Graduate Engineer INRIA, since October 2006]

Jean-Francois Gout [Profession=Technique] [Category=AutreAffiliation] [CDD ANR, started Oct. 2006]

Sophie Huet [Profession=Technique] [Category=AutreAffiliation] [CDD Fondation Rhône-Alpes Futur, RNG, since October 2006]

Vincent Lombard [Profession=Technique] [Category=AutreAffiliation] [CDD ANR]

Emmanuel Prestat [Profession=Technique] [Category=AutreAffiliation] [CDD ANR]

Bruno Spataro [Profession=Technique] [Category=CNRS] [Technical Staff, CNRS]

Post-doctoral fellows

Frédéric Boyer [Profession=Postdoc] [Category=INRIA] [scholarship Sanofi Pasteur, until Sept. 2006]

Cinzia Pizzi [Profession=Postdoc] [Category=INRIA] [scholarship Inria, starting Dec. 1, 2006]

Adrien Richard [Profession=Postdoc] [Category=INRIA] [scholarship INRIA, with POPART and HELIX, since October 2006]

Patricia Thébault [Profession=Postdoc] [Category=INRIA] [scholarship Inria]

PHD Students

Sophie Abby [Profession=PhD] [Category=UnivFr] [scholarship Ministère de la Recherche, supervisors: Vincent Daubin and Manolo Gouy, started Oct. 2006]

Vicente Acuña [Profession=PhD] [Category=UnivFr] [scholarship Conicyt (Chile) and INRIA, supervisors: Marie-France Sagot and Christian Gautier]

Anne-Muriel Arigon [Profession=PhD] [Category=UnivFr] [scholarship Ministère de la Recherche, supervisors: Manolo Gouy and Guy Perrière]

Elise Billoir [Profession=PhD] [Category=UnivFr] [scholarship Ministère de la Recherche, supervisors: Sandrine Charles]

Bastien Boussau [Profession=PhD] [Category=CNRS] [scholarship BDI, CNRS, supervisor: Manolo Gouy]

Yves-Pol Deniélou [Profession=PhD] [Category=CNRS] [scholarship ENS, supervisors: Alain Viari and Marie-France Sagot, started Oct. 2006]

Marc Deloger [Profession=PhD] [Category=CNRS] [scholarship BDI, CNRS, supervisors: Cristina Vieira (LBBE) and Marie-France Sagot, started Oct. 2006]

Marília Dias Vieira Braga [Profession=PhD] [Category=UnivFr] [scholarship AlBan, supervisors: Marie-France Sagot and Eric Tannier]

Samuel Druhle [Profession=PhD] [Category=INRIA] [scholarship ENS Cachan, supervisor: Hidde de Jong]

Claire Guillet [Profession=PhD] [Category=UnivFr] [scholarship ENS Lyon, supervisor: Laurent Duret, started Oct. 2006]

Vincent Lacroix [Profession=PhD] [Category=CNRS] [scholarship BDI, CNRS, supervisor: Marie-France Sagot]

Claire Lemaitre [Profession=PhD] [Category=UnivFr] [scholarship Ministère de la Recherche, supervisors: Marie-France Sagot and Christian Gautier]

Yann Letrillard [Profession=PhD] [Category=UnivFr] [scholarship Institut National du Cancer, supervisor: Guy Perrière]

Nuno Mendes [Profession=PhD] [Category=UnivEtrangere] [scholarship Portuguese Ministry of Research (FCT), supervisors: Ana Teresa Freitas (IST, Lisbon, Portugal) and Marie-France Sagot, started November 2006]

Pedro Monteiro [Profession=PhD] [Category=UnivEtrangere] [scholarship Portuguese Ministry of Research (FCT), Supervisors: Ana Teresa Freitas (IST Lisbon, Portugal), Hidde de Jong and Radu Mateescu, started November 2006]

Anamaria Necsulea [Profession=PhD] [Category=UnivFr] [scholarship Ministère de la Recherche, supervisor: Jean Lobry]

Leonor Palmeira [Profession=PhD] [Category=UnivFr] [scholarship Ministère de la Recherche, supervisors: Laurent Guéguen and Jean Lobry]

Former PHD Students

Grégory Batt [Profession=PhD] [Category=UnivFr] [scholarship ENS Lyon, supervisor: Hidde de Jong, PhD defense in February 2006, currently postdoc at Boston University]

Stéphane Descorps-Declère [Profession=PhD] [Category=UnivFr] [CIFRE convention, GENOME express, supervisors: Alain Viari, Pierre Netter, Université Paris 6, PhD defended in July 2006, currently postdoc at the University of Orsay]

Master Students

Juliet Ansel [Profession=Stagiaire] [Category=UnivFr] [Master M2 Pro, University Rouen and University Claude Bernard, supervisors: Marie-France Sagot with Ludovic Cottret and Vincent Lacroix]

Nicolas Blayet [Profession=Stagiaire] [Category=UnivFr] [Master M2 Research, University Claude Bernard, supervisor: Guy Perrière]

Hugo Devillen [Profession=Stagiaire] [Category=UnivFr] [Master M2 Research, University Claude Bernard, supervisors: Jean Lobry (with Frédéric Menu, LBBE)]

Anne-Sophie Sertier [Profession=Stagiaire] [Category=UnivFr] [Master M2 Research, University Claude Bernard, supervisors: Daniel Kahn and Vincent Daubin]

Clément Rezvoy [Profession=Stagiaire] [Category=UnivFr] [Master M2 Research, University Claude Bernard, supervisors: Daniel Kahn (with Frédéric Vivien, LIP-ENS Lyon]

Christian Gautier also co-supervises a PhD student, Caroline Truntzer. C. Truntzer is a member of the "Biostatistics-Health" team at the LBBE. Her co-supervisor is Pascal Roy. Another of the PhD students Christian Gautier co-supervised (with René Écochard, also from the "Biostatistics-Health" team of the LBBE), Maud Tournoud, defended her PhD in October 2006. Maud remains as postdoc in the laboratory of her PhD. In the same way, Marie-France Sagot co-supervised with Maxime Crochemore a PhD student from the University of Marne-la-Vallée, Pierre Peterlongo, who defended his PhD in September 2006. Pierre is now postdoc in the SYMBIOSE project of the INRIA.

The HELIX group in Lyon benefits from the administrative assistance of the "Laboratoire de biométrie et de biologie évolutive" (UMR 5558).

2 Overall Objectives

debug : Module named "presentation" for the project HELIX, section presentation

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

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

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

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

1. Comparative genomics;
   (a) Computational analysis of the evolution of species and gene families;
   (b) Modelling and analysis of the spatial organisation and dynamics of genomes;
   (c) Motif search and inference;
   (d) Knowledge representation for genomics;
2. Functional genomics
   (a) Computational proteomics and transcriptomics;
   (b) Modelling of metabolism: molecular components, regulation, and pathways;
   (c) Modelling and simulation of genetic regulatory networks;

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

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

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

- GenoStar is a bioinformatics platform for exploratory genomics which integrates methods and tools for modelling genomic data and knowledge developed both within and outside the project (Section 5.11).

- PRABI is a Web server resource providing software which may be downloaded or used through facilities available on the Web. The HELIX group is one of the major participants in the development and maintenance of this platform, which is recognized at the national level as one of the RIO and Genopole platforms. The facilities offered by the PRABI cover such areas as genomics, structural biology, proteomics, health, and ecology. The director of the PRABI (C. Gautier) is a member of HELIX.

3 Scientific Foundations

3.1 Comparative genomics

debug : Module named "compgen" for the project HELIX, section fondements, topic Comparative genomics

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

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

debug : moreinfo

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

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

3.1.1 Computational analysis of the evolution of species and gene families

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

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

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

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

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

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

3.1.3 Motif search and inference

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

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

Search and inference problems, whether they concern features or relations, are in fact the extremes of a continuum of problems that range from seeking for something well-known to trying to identify unknown objects. The main difficulty lies in the fact that features and the relations holding between them should in general be inferred together. However, the information that must be manipulated in this case (cooperative signals, operons, regulons, reaction pathways or molecular assemblies) is more complex than the initial genome data and thus requires a higher degree of abstraction, and more sophisticated algorithms or statistical approaches. Various search and inference methods have already been developed by HELIX. These include methods for DNA and protein sequence motifs inference, gene finding, satellites and repeats identification and RNA common substructure inference. More recent work concerns the definition of motifs in graphs representing, for instance, metabolic pathways. In the last year, work has started also on mixing information from various often quite heterogeneous sources to infer motifs. These include for now sequence information with information on gene expression coming from microarray experiments and information provided by the signals evolution imprints into genomes. The final objective is to be able to automatically infer whole cellular modules, that is in fact small or, in the longer term, larger-scale biological networks. This new topic provides a strong link between comparative and functional genomics, molecular biology seen at the linear level of a genome and networks.

3.2 Functional genomics

debug : Module named "funcgen" for the project HELIX, section fondements, topic Functional genomics

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

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

debug : moreinfo

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

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

3.2.1 Computational proteomics and transcriptomics

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

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

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

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

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

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

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

3.2.3 Modelling and simulation of genetic regulatory networks

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

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

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

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

Topic x undefined

4 Application Domains

4.1 Panorama

debug : Module named “panorama” for the project HELIX, section domaine, topic

Keywords: medicine, agriculture.

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

Topic x undefined

5 Software

5.1 AROM

debug : Module named "AROM" for the project HELIX, section logiciels, topic

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

Keywords: knowledge representation.

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

Topic x undefined

5.2 C3P

debug : Module named "C3P" for the project HELIX, section logiciels, topic

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

Keywords: graph merging, multigraph common connected component.

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

Topic x undefined

5.3 DNA array analysis tool

debug : Module named "DNA array analysis tool" for the project HELIX, section logiciels, topic

Participants: Guy Perrière [correspondent].

Keywords: DNA array analysis.

In collaboration with the groups of Michel Bihl (University of Basel) and Desmond Higgins (University of Dublin), we have developed a new resampling strategy for the statistical analysis of DNA array data sets [64]. This strategy can be applied with any supervised clustering method used for the analysis of gene expression data. The corresponding software is available as R scripts at `http://pulmogene.unibas.ch/articles/optimization`.

Topic x undefined

5.4 Ed'Nimbus

debug : Module named "Ed'Nimbus" for the project HELIX, section logiciels, topic

Participants: Marie-France Sagot [Correspondent].

Keywords: filter for sequence alignment and repeat identification.

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

Topic x undefined

5.5 Evolution by Reversals Manipulation Tool

debug : Module named "Evolution by Reversals Manipulation Tool" for the project HELIX, section logiciels, topic

Participants: Marilia Dias Vieira Braga, Marie-France Sagot, Eric Tannier [correspondent].

Keywords: sorting by reversals.

This algorithm was implemented by Celine Scornavacca and Marília Braga, based on up-to-date results on the problem of sorting permutations by reversals. In particular, it implements an algorithm designed by members of HELIX that presents in a compact way the whole set of solutions of this problem.

Topic x undefined

5.6 FamFetch

debug : Module named "FamFetch" for the project HELIX, section logiciels, topic

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

Keywords: tree pattern search, phylogenetic trees, database.

FamFetch is a set of tools to search for tree patterns in databases of phylogenetic trees. FamFetch is available for download at (`http://pbil.univ-lyon1.fr/software/famfetch.`

html). It was developed with Jean-François Dufayard who did his PhD in HELIX. He is now IR at the LIRMM, Montpellier.

Topic x undefined

5.7 GeM

debug : Module named "GeM" for the project HELIX, section logiciels, topic

Participants: Christian Gautier [Correspondent], Vincent Navratil, Bruno Spataro.

Keywords: database, comparative genomics, vertebrates.

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

Topic x undefined

5.8 Genepi (Genepi)

debug : Module named "Genepi" for the project HELIX, section logiciels, topic

Participants: Stéphane Descorps-Declere, François Rechenmann, Alain Viari, Danielle Ziébelin,.

Keywords: annotation system, blackboard architecture.

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

Topic x undefined

5.9 GenoExpertBacteria (GEB)

debug : Module named "GenoExpertBacteria (GEB)" for the project HELIX, section logiciels, topic

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

Keywords: data analysis environment, knowledge base, graphical interface.

GenoExpertBacteria is an environment for the analysis of genomic and metabolic data in bacteria. It integrates a knowledge base and a graphical user interface facilitating the exploration and analysis of the available data. GEB has now been integrated (under the name "PathwayExplorer") into the Iogma bioinformatics environment developed and distributed by the Genostar company (see 7.1). For more information, see: http://www-helix.inrialpes.fr/article141.html

Topic x undefined

5.10 Genetic Network Analyzer (GNA)

debug : Module named "GNA" for the project HELIX, section logiciels, topic

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

Keywords: genetic network qualitative modelling.

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

Topic x undefined

5.11 GenoStar

debug : Module named "GenoStar" for the project HELIX, section logiciels, topic

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

Keywords: bioinformatics environment.

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

Topic x undefined

5.12 Herbs

debug : Module named “Herbs” for the project HELIX, section logiciels, topic

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

Keywords: annotation support.

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

Topic x undefined

5.13 Hogenom and Hovergen

debug : Module named “HOGENOM and HOVERGEN” for the project HELIX, section logiciels, topic

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

Keywords: databases, genomes.

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

Topic x undefined

5.14 Hoppsigen

debug : Module named “HOPPSIGEN” for the project HELIX, section logiciels, topic

Participant: Dominique Mouchiroud [Correspondent].

Keywords: database, pseudogenes.

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

Topic x undefined

5.15 HoSeqI

debug : Module named “HoSeqI” for the project HELIX, section logiciels, topic

Participants: Anne-Muriel Arigon, Manolo Gouy, Guy Perrière [correspondent].

Keywords: sequence identification, gene family database.

HoSeq1 (Automated homologous sequence identification in gene family databases) is a web service available at `http://pbil.univ-lyon1.fr/software/HoSeqI` The user can position a protein or a DNA sequence relatively to a database of families of homologous sequences and identify the family to which the sequence belongs, as well as its position within the multiple alignment and the evolutionary tree of this family.

Topic x undefined

5.16 Identitag

debug : Module named “Identitag” for the project HELIX, section logiciels, topic

Participants: Laurent Duret, Dominique Mouchiroud.

Keywords: database, SAGE.

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

Topic x undefined

5.17 ISee

debug : Module named “ISee” for the project HELIX, section logiciels, topic

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

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

Topic x undefined

5.18 LalnView

debug : Module named "LalnView" for the project HELIX, section logiciels, topic

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

Keywords: local alignment, visualizer.

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

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

Topic x undefined

5.19 Mentalign

debug : Module named "Mentalign" for the project HELIX, section logiciels, topic

Participants: Manolo Gouy [Correspondent], Guy Perrière.

Keywords: multiple alignment, phylogenetic trees.

Mentalign is an incremental algorithm for performing a multiple alignment and building the phylogenetic tree of members of a same gene family. When a new sequence is added to a pre-aligned family, the alignment and the tree are modified rather than fully recomputed. For more information please contact: mgouy@biomserv.univ-lyon1.fr Mentalign was developed with Jean-François Dufayard who did his PhD in HELIX. He is now IR at the LIRMM, Montpellier.

Topic x undefined

5.20 MicrOBI

debug : Module named "MicrOBI" for the project HELIX, section logiciels, topic

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

Keywords: database.

MicrOBI is a relational database devoted to microorganisms, integrating and synchronizing heterogeneous data from various public sources: genome data (EBI genome files), proteome data (Swiss-Prot and HAMAP), metabolic data (Enzyme and KEGG) and functional classification (GeneOntology). It has been implemented using PosgreSQL and ZOPE and uses trigger mechanisms for automatic updates and data consistency checks. It acts as a data source for GEB (Section 5.9), but can also be used as a stand-alone database. Since 2006, it is part of a new opensource project called "OBISchema". For more information please contact: Eric.Coissac@inrialpes.fr

Topic x undefined

5.21 Motus

debug : Module named "Motus" for the project HELIX, section logiciels, topic

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

Keywords: reaction motif search and inference.

Motus is an algorithm for searching and inferring reaction motifs in metabolic networks. A metabolic network is modelled as an undirected graph. Nodes (reactions) are labelled with EC numbers (code of 4 numbers expressing the chemistry of a reaction). A reaction motif is defined as a set of (partial) EC numbers. An occurrence of the motif in the network is a set of nodes (reactions) which is connected and labelled by EC numbers that are similar to the ones of the motif. Motus is available at: `http://pbil.univ-lyon1.fr/software/motus/`. The algorithm was developed by Vincent Lacroix, and the web interface by Ludovic Cottret and Odile Rogier from the PRABI.

Topic x undefined

5.22 Migal

debug : Module named "Migal" for the project HELIX, section logiciels, topic

Participants: Marie-France Sagot [Correspondent].

Keywords: RNA, tree comparison.

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

Topic x undefined

5.23 Njplot

debug : Module named "Njplot" for the project HELIX, section logiciels, topic

Participants: Manolo Gouy [correspondent].

Keywords: phylogenetic tree drawing.

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

Topic x undefined

5.24 Oriloc

debug : Module named "Oriloc" for the project HELIX, section logiciels, topic

Participant: Jean Lobry [Correspondent].

Keywords: replication origin and terminus.

Oriloc is a program to predict the putative origin and terminus of replication in prokaryotic genomes. The program works with unannotated sequences and therefore uses Glimmer2 outputs to discriminate between codon positions. For more information see : http://pbil.univ-lyon1.fr/software/oriloc.html

Topic x undefined

5.25 PepLine

debug : Module named "PepLine" for the project HELIX, section logiciels, topic

Participant: Alain Viari [Correspondent].

Keywords: proteomic data analysis.

PepLine is a software pipeline supporting the high-throughput analysis of proteomic data, in particular the identification of proteins from MS/MS spectra. At present, PepLine consists of two components: Taggor and PMMatch. Taggor generates so-called PSTs (Peptide Sequence Tags) from MS/MS data, while PMMatch maps the PSTs to sequences in protein databanks, or to the complete translated genome of an organism, thus helping to locate the gene coding for the protein. PepLine was developed with Estelle Nugues and Erwan Reguer through a collaboration with the Laboratoire de Chimie des Protéines headed by J. Garin at the CEA of Grenoble. For more information, see : http://www-helix.inrialpes.fr/article228.html.

Topic x undefined

5.26 PhyloJava

debug : Module named "PhyloJava" for the project HELIX, section logiciels, topic

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

Keywords: phylogenetic reconstruction.

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

Topic x undefined

5.27 ProDom

debug : Module named "ProDom" for the project HELIX, section logiciels, topic

Participants: Daniel Kahn [correspondent].

Keywords: protein domain families, database.

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

Topic x undefined

5.28 PSbR

debug : Module named "PSbR" for the project HELIX, section logiciels, topic

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

Keywords: perfect sorting by reversals.

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

Topic x undefined

5.29 Remote Acnuc Access

debug : Module named "Remote Acnuc Access" for the project HELIX, section logiciels, topic

Participants: Manolo Gouy [correspondent].

Keywords: access to molecular databases.

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

Topic x undefined

5.30 Repseek

debug : Module named "Repseek" for the project HELIX, section logiciels, topic

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

Keywords: approximate repeat detection, DNA sequences.

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

Topic x undefined

5.31 RFDD

debug : Module named "RFDD" for the project HELIX, section logiciels, topic

Participants: Guy Perrière [correspondent].

Keywords: creation and update database, transcriptomic technique.

In collaboration with Helene Simonnet (Centre de Genetique Moleculaire et Cellulaire, UMR CNRS 5534, Lyon), we have developed a web service for bioinformatic analysis of RFDD (Restriction Fragment Differential Display) results. RFDD is a transcriptomic technique derived from AFLP (Analysis of Fragment Length Polymorphism). This service uses a relational database of restriction fragments obtained by *in silico* digestion of all human and rat transcribed sequences present in the RefSeq database. We have developed software for the creation and update of this database and for providing web access to it (http://pbil.univ-lyon1.fr/software/RFDD/). A first biological result has been derived from this tool by analysis of transcriptome data from rat under hypoxic conditions [23].

Topic x undefined

5.32 SeaView

debug : Module named "SeaView" for the project HELIX, section logiciels, topic

Participants: Manolo Gouy [correspondent].

Keywords: editor of multiple sequence alignments.

This program for editing multiple sequence alignments has been updated by interfacing it with the muscle sequence alignment algorithm, in addition to the clustalw algorithm. Available at: http://pbil.univ-lyon1.fr/software/seaview.html.

Topic x undefined

5.33 SeqinR

debug : Module named "SeqinR" for the project HELIX, section logiciels, topic

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

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

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

Topic x undefined

5.34 Smile and Riso

debug : Module named "SmileRiso" for the project HELIX, section logiciels, topic

Participants: Marie-France Sagot [Correspondent].

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

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

Topic x undefined

5.35 SymBioCyc

debug : Module named "SymBioCyc" for the project HELIX, section logiciels, topic

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

Keywords: database, metabolism, endosymbiont.

SymBioCyc is a database of metabolic data dedicated to endosymbiotic organisms (that is, to organisms that live within the body or cells of another organism). SymBioCyc follows the same representation model as BioCyc (`http://www.biocyc.org/`) but the data it contains has been expertised both automatically and manually by Ludovic Cottret. It will be made available to the academic research community in early 2007.

Topic x undefined

5.36 WellReader

debug : Module named "WellReader" for the project HELIX, section logiciels, topic

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

Keywords: gene expression data analysis.

WellReader is a program for the analysis of gene expression data obtained by means of fluorescent and luminescent reporter genes. WellReader reads data files in a XML format or in a format produced by microplate readers, and allows the user to detect outliers, perform background corrections and spline fits, and compute promoter activities and protein concentrations. WellReader has been written in Matlab and will be made available to the academic research community in early 2007.

Topic x undefined

5.37 Other software developed in HELIX

debug : Module named “others” for the project HELIX, section logiciels, topic

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

HELIX has contributed to the development of software by other members of the PRABI (Section 2). This is in particular the case for:

- ROSO (INSA, N. Raymond), which supports the efficient design of eukaryotic DNA chips;
- RTKdb (CGMC, université Claude Bernard, J. Grassot), which is a database dedicated to the tyrosine kinase receptors. RTKdb uses the FamFetch environment (Section 5.6);
- BIBI (LBBE, J.-P. Flandrois), which is a powerful tool for identifying pathogenic bacteria from genomic sequences.

Several other programs have resulted from the activities of HELIX members, but are no longer being actively developed. This concerns the following programs (with the contact person between brackets): Acnuc (Manolo Gouy), Alice (Marie-France Sagot), Combi (Marie-France Sagot), CoSamp (Marie-France Sagot), DomainProteix (Alain Viari), Druid (Marie-France Sagot), EMKov (Alain Viari), FactorTree (Marie-France Sagot), GeM (Bruno Spataro), Jadis (Dominique Mouchiroud), Mtdp (Alain Viari), Satellites (Marie-France Sagot), and SeaView (Manolo Gouy), Utopia (Marie-France Sagot).

Topic x undefined

6 New Results

6.1 Comparative genomics

debug : Module named “rescompgen” for the project HELIX, section resultats, topic

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

In several contexts such as 1. species classification, 2. confrontation of a new sequence to a database, 3. update of homologous (*i.e.* descending from a same ancestor) gene family sequence databases, the classification of a new sequence into a collection is needed. This classification allows the identification of which family the sequence belongs to and contributes to the assessment of its evolutionary relationships. Today, massive sequencing techniques are routinely used and the number of new available sequences grows up quickly. Furthermore, the identification task requires the chaining of different programs (for similarity search, alignment and phylogenetic tree computation) that are sometimes complex to handle. Some results have also to be manually checked. Doing these tasks sequentially makes the work of sequence identification tedious and time-consuming. Automated bioinformatic methods are thus necessary to carry out these operations in an accurate and fast way. As part of her PhD, Anne-Muriel Arigon has developed a method that allows to automatically assign sequences to homologous gene families from a set of databases [6]. After identification of the most similar gene family to the query sequence, this sequence is added to the whole alignment, and the phylogenetic tree of the family is rebuilt. The phylogenetic position of the query sequence in its gene family can then be easily identified.

Recent heuristic advances have made maximum likelihood phylogenetic tree estimation tractable for hundreds of sequences. Noticeably, these algorithms are currently limited to reversible models of evolution. The reversible property is a technical one that leads to more tractable models but is clearly not verified by evolutionary processes. As part of his PhD, Bastien Boussau has shown that by reorganising the way likelihood is computed, one can efficiently compute the likelihood of a tree from any of its nodes with a nonreversible model of sequence evolution, and hence benefit from cutting-edge heuristics. This computational trick can be used with reversible models of evolution without any extra cost. Bastien then introduced nhPhyML [12], an adaptation of the nonhomogeneous nonstationary model of Galtier and Gouy (1998; *Mol. Biol. Evol.* 15:871-879) to the structure of PhyML (2003; *Syst Biol.* 52(5):696-704), as well as an approximation of the model in which the set of equilibrium frequencies is limited. This new version shows good results both in terms of exploration of the space of tree topologies and ancestral G+C content estimation. Moreover, the approach argues for a hypothesis that members of the HELIX team already defended in the past: the last common ancestor of all of today's living organisms did not live at high temperatures. nhPhyML was applied to the slowly evolving sites of rRNA sequences. The result is that the model and a wider taxonomic sampling still do not plead in favour of a hyperthermophilic last universal common ancestor.

Beside research in phylogeny, Helix also conducts activities in e-learning in the context of the ISee platform (see Section 5.17). In 2006, A. Chamontin has finished the development of an ISee course dedicated to the computation of phylogenetic trees from a set of genomic sequences. In the first part of the course, the user can follow the execution of the UPGMA algorithm step by step. In a second part, he/she has to make the right choices for the algorithm to progress correctly. For instance, at each iteration, he/she has to choose the next term of the distance matrix and to recompute the remaining terms. The course also provides three smaller programs which illustrate the limits of the method. They essentially explain why the distance between two sequences, in terms of differences, may differ from the evolutionary distance between the organisms. This course will be used by the CCSTI at Grenoble for its "École de l'ADN" targeted to graduate students.

The arrival of Daniel Kahn in HELIX has brought to the team an internationally recognised expertise in the modular evolution of protein sequences. This expertise is currently stored in a database, ProDom, that has been transferred to the PRABI and is maintained and regularly updated by Daniel Kahn and members of the PRABI (see Section 5.27). The analysis of evolutionary scenarios of protein domain families from ProDom (2005; *Nucleic Acids Res.* 33(Database issue):212-215) showed that only a small minority of domain families is truly ancestral. Far from being static, the protein domain repertoire undergoes a continuous innovation process. Therefore the tremendous diversity of modular proteins results from both the combinatorial assortment of protein domains and an ongoing process of protein domain innovation.

In mammals, females carry 2 X chromosomes and males only one. To avoid a higher gene expression in females, a mechanism inactivates one of the X chromosomes in each cell. This inactivation is managed by a non coding gene, called Xist. Laurent Duret has shown [31] that Xist evolves from a coding gene since the separation between eutherians (mammals that have a placenta) and marsupials (mammals, such as the kangooroo, in which the female typically has a pouch where its youngs are reared through early infancy). This result is an important step towards understanding sex determination in mammals. It confirms that, although sex is one of the most universal properties among eukaryotes, the way it is determined is submitted to a very quick evolution.

Statistical analysis of the global composition of genomes and its link with environmental or metabolic characteristics has been the focus of considerable interest. As part of Leonor Palmeira's PhD, the hypothesis that pyrimidine dinucleotides (dinucleotides composed of Cs and Ts) are avoided in light-exposed genomes as the result of a selective pressure due to high ultraviolet (UV) exposure was investigated. The main damage to DNA produced by UV radiation is known to be the formation of pyrimidine dimers as the product of a photochemical reaction between adjacent pyrimidines. All available complete prokaryotic genomes and the model organism *Prochlorococcus marinus* were statistically analysed and it was found [50] that pyrimidine dinucleotides are not systematically avoided. This suggests that prokaryotes must have sufficiently effective protection and repair systems for UV exposure to not affect their dinucleotide composition.

The evolution of nucleic sequences is usually modelled by point substitutions and under the hypothesis that sites evolve independently of each other. This hypothesis is mainly kept for mathematical purposes and has no biological foundation, as it is now clear that molecular substitution mechanisms frequently involve adjacent bases. The most typical example is the highly frequent spontaneous chemical transformation of CpG dinucleotides observed on some sequences (the " CpG" notation is used to distinguish a cytosine C followed by guanine G from a cytosine base paired to a guanine). Berard, Gouere and Piau (2006; private communication) have shown that in some special cases which include the CpG transformation, neighbour-dependent models are solvable: equilibrium distributions can be determined. Still as part of Leonor Palmeira's PhD, the system was solved for a number of specific biological models. Under the assumption of stationarity, it was further shown that it is easy to compute the substitution rates acting on any given sequence.

The relationship between codon usage in prokaryotes and their ability to grow at extreme temperatures has been given much attention over the past years. Previous studies have suggested that the difference in synonymous codon usage between (hyper)thermophiles and mesophiles is a consequence of a selective pressure linked to growth temperature. A hyperthermophile is an organism that thrives in extremely hot environments while a mesophile grows best in moderate temperature. As part of her PhD, Anamaria Necsulea performed an updated analysis [46]. The conclusion reached is that the difference in synonymous codon usage between (hyper)thermophilic and non-thermophilic species cannot be clearly attributed to a selective pressure linked to growth at high temperatures. Strong efforts are currently under way to determine the genome of major eukaryotic human parasites. Anamaria Necsulea has started a bioinfomatics analysis of some of these genomes, mainly in the *Leishmania* gender [49]. She has shown a new and surprising usage of synonymous codons (codons that refer to the same amino acid) in these organisms. Several biological interpretations are possible: this may either be explained by a better adaptation to the translational process, or be the result of mutational bias. New genomes are being sequenced and this gives hope that it will be possible to discriminate between the two hypotheses in a near future.

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

Spatial organisation

Genomes are organised as a succession of region having different functional roles: introns, exons, intergenic regions, etc. Each of these regions has different statistical properties that are required by their functional role. However, other regional organisations exist at broader scales, isochores being the most studied ones. Isochores occur mainly in mammal and bird genomes. An isochore is a large region (more than 300kb) with a relative homogeneity in base frequencies, particularly in C+G. An analysis of the isochore organisation of a genome needs therefore to separate structures that occur at different scales. Before the PhD work of Christelle Melodelima [47] [71], either the local structure was ignored or the analysis was restricted to exons. C. Melodelima proposed an HMM approach that simultaneously builds a modelling of the local organisation and allows (through a bayesian approach) a model selection that leads to the segmentation of the genome in its isochores. Moreover, this original approach has lead to new biological results, for instance, on the relationships between the organisation in isochores and the sequences coming just before and after each gene (5'UTR and 3'UTR).

Hidden Markov Models are one possible way of segmenting a biological sequence. Another that was developed in HELIX by Laurent Gueguen is maximal predictive partioning (MPP) (2001; *LNCS* vol. 2066, pages 32-44). MPP is a method that, given a set of models (for instance, related to sequence composition), builds the best partition of a sequence into k segments. MPP scores the adequacy of a model to a segment. During the partitioning operation, several models are compared in order to optimally segment a sequence of letters into homogeneous parts. Such a model may, for instance, be Markovian. Using dynamic programming, MPP computes the k-partitions for successive values of k in time linear with k and the length of the sequence. Hence, an MPP gives a multi-scale representation of the sequence. Another way to compute a multi-scale segmentation of sequences is to use a hierarchical process. The segments are in this case recursively divided.

A package of Python modules, called SARMENT was developed for easy building and manipulation of sequence segmentations (2005; *Bioinformatics* 21(16):3427-3428). The first aim of SARMENT is to provide an efficient implementation of the HMM segmentation algorithms (Viterbi and Forward-Backward) and of the MPP method. The second aim of SARMENT is to allow easy manipulation of the models that are used in HMMs and in MPPs. An algorithm for computing the probability of a given segmentation has also been submitted for publication (submitted paper).

Gene order is not random with regard to gene expression in mammals: coexpressed genes, and in particular housekeeping genes (*i.e.* genes that are transcribed at a relatively constant level), are clustered along chromosomes more often than expected by chance. To understand the origin of these clusters, and to quantify the impact of this phenomenon on genome organisation, Laurent Duret and his former PhD student, Marie Semon, analysed clusters of coexpressed genes in the human and mouse genomes. They showed that neighbouring genes experience continuous concerted expression changes during evolution, which leads to the formation of coexpressed gene clusters [56]. The pattern of expression within these clusters evolves more slowly than the genomic average. Moreover, by studying gene order evolution, it was shown that some clusters are maintained by natural selection and, therefore, have a functional meaning. However, it was also demonstrated that some coexpressed gene clusters are the result of neutral coevolution effects, as illustrated by the clustering of genes escaping inactivation on the X chromosome.

Dynamics

Genomes undergo large scale changes through evolution, called rearrangements. As part of the PhD of Claire Lemaitre, we are interested in the specific sequences where a rearrangement took place, more precisely we seek to identify characteristics in these breakpoint regions specific to rearrangements [45]. The detection of such breakpoints is made by a comparative genome analysis between two species using the annotated orthologous genes. The breakpoint region is then refined by the alignment of intergenic regions. This method has been applied on the human and mouse genomes and it allows to analyse precisely the sequences around the breakpoints, and to compare them with other sequences in the genome. The aim is to find sequence characteristics linked with genome dynamics to understand the molecular mechanisms underlying the process leading to rearrangements.

The mammal chromosomes X and Y have evolved from an identical autosome pair. This process is at the origin of sexual differentiation - the female XX and the male XY pairs. Due to the recombination mechanism (recombination is the genetic transmission process intra and inter chromosomes), female organisation favours X chromosome conservation. On the other hand, the male XY pair evolution causes Y chromosome degeneration, as this chromosome loses gradually the capacity of recombining with its X partner.

Current theories show that the rearrangement process followed by these two chromosomes was mainly composed by a few big reversals, which have happened in an ordered way, from the end to the beginning of the Y chromosome. Nevertheless these theories still present controversial aspects. As part of the PhD of Marilia Braga, we are trying to elucidate the question by reconstructing the rearrangement process with the available information we get in public databases. In addition, we are developing a new algorithmic model for the genome rearrangement problem, which might be better adapted to these ordered big reversal events.

In comparative genomics, algorithms that sort permutations by reversals are often used to propose evolutive scenarios of large scale genomic mutations between species. One of the main problems of such methods is that they give one solution while the number of optimal solutions is huge, with no criteria to discriminate among them. One previous study by Bergeron and colleagues (2006; *IEEE/ACM TCBB*, in press) has tried to give some structure to the set of optimal solutions, in order to be able to give more presentable results than only one solution or a complete list of all solutions. The structure is a set of partially ordered sets (posets), of which all linear extensions are solutions. However, no algorithm existed so far to compute this set of posets except through the enumeration of all solutions, which takes too much time even for small permutations. With an Italian master student, Celine Scornavacca, and with Marilia Braga, we devised such an algorithm, which gives all the posets and counts the number of solutions, with a better theoretical and practical complexity than a complete enumeration (paper in preparation). Several biological examples are provided where the result is more relevant than a unique optimal solution or the list of all solutions, the latter being often impossible to compute.

Another way to deal with the huge number of solutions provided by rearrangement reconstitution algorithms is to add some biological constraints, such as favouring small inversions or inversions that do not cut clusters of co-localised genes. This approach is called “perfect” sorting. Together with a German master student, Yoan Diekmann, Eric Tannier and Marie-France Sagot devised an algorithm that is able to test if there is one solution that respects the contraints of gene clusters, and gives it if this is the case [26]. This was tested on gene order data of several species, and some statistics were provided for the cases where there exists one solution or not. The algorithm was then extended by giving some solutions that minimize the number of gene clusters that have to be broken by a rearrangement scenario.

The reconstructions of evolution scenarios of genomic rearrangements are sometimes very different in computational biology than they are in cytogenetics, where different kinds of data are analysed. Cytogenetics is the study of the structure of chromosome material. With the PhD students Marilia Braga and Claire Lemaitre, and in collaboration with Thomas Faraut from the INRA Toulouse, we started a deep analysis of the differences in the methods in the two domains. This work is done in collaboration with Bernard Dutrillaux and Florence Richard from the Museum National d'Histoire Naturelle of Paris, who are cytogeneticists who have been doing research on rearrangement scenarios for many years and have what is probably the richest collection of expertly assessed cytogenetic data in the world. We intend to group the methods, data and results of both research domains, in order to propose better algorithms, and a unifying theory of the modes of speciation.

Rearrangements and accelerated mutation rates are also observed in the most simple organisms such as bacteria when subjected to environmental changes. In the context of a collaboration with the group of Roger Frutos (CIRAD Montpellier), we have performed the complete annotation of two strains of *Ehrlichia ruminatium*, an obligatory pathogen and causative agent of heartwater, a major tick-borne disease of livestock in Africa and Caribbean. The most specific feature of these genomes is their exceptionally large intergenic regions (acually the largest amongst bacteria) and the presence of long-period tandem repeats associated to expansion/contraction of these intergenic regions. Following the publication of the complete genome [35], we have performed a comparative genomic analysis of these strains as well as additional species of *Ehrlichia*. This has revealed the presence of an active and specific mechanism of genomic plasticity, probably following the exposure to a diverse environment (different hosts), which could explain the limited field-efficiency of vaccines against *E. ruminantium* [36]. Most of these studies have been conducted by using the GenoStar platform and have therefore represented the first real-size test bed for this platform.

Gene duplication has different outcomes: pseudogenization (death of one of the two copies), gene amplification (both copies remain the same), sub-functionalization (both copies are required to perform the ancestral function) and neo-functionalization (one copy acquires a new function). Asymmetric evolution (one copy evolves faster than the other) is usually seen as a signature of neo-functionalization. However, it has been proposed that sub-functionalization could also generate asymmetric evolution among duplicate genes when they experience different local recombination rates. Gabriel Marais and Raquel Tavares together with an L2 student, Yves Clément, tested this idea with about 100 pairs of young duplicates from the *Drosophila melanogaster* genome [17]. They found that dispersed pairs tend to evolve more asymmetrically than tandem ones. Among dispersed copies, the low recombination copy tends to be the fast-evolving one. They also tested the possibility that all this was explained by a confounding factor (expression level) but found no evidence for it. In conclusion, their results do support the idea that asymmetric evolution among duplicates is enhanced by restricted recombination. However, further work is needed to clearly distinguish between sub-functionalization and neo-functionalization for the asymmetrically-evolving duplicate pairs that they found.

The duplication of entire genomes has long been recognized as having great potential for evolutionary novelties, but the mechanisms underlying their resolution through gene loss are poorly understood. In collaboration with the groups of Jean Cohen (CGM, Gif), Eric Meyer (ENS, Paris) and the Genoscope (P. Wincker, Evry), Laurent Duret, Vincent Daubin and Jean-François Gout from HELIX are involved in the analysis of the genome of the unicellular eukaryote *Paramecium tetraurelia*. In this ciliate, most of the nearly 40,000 genes arose through at least three successive whole-genome duplications. Phylogenetic analysis indicates that the most recent duplication coincides with an explosion of speciation events that gave rise to the *P. aurelia* complex of 15 sibling species. We observed that gene loss occurs over a long timescale, not as an initial massive event. Genes from the same metabolic pathway or protein complex have common patterns of gene loss, and highly expressed genes are over-retained after all duplications. The conclusion of this analysis is that many genes are maintained after whole-genome duplication not because of functional innovation but because of constraints on the number of copies of a given gene present in a cell or nucleus [7] [18].

6.1.3 Motif search and inference

One large part of the algorithmic studies in the HELIX project concerns the search for regularities at many levels of living systems. Regularities may be seen as motifs in DNA sequences, RNA structures or protein structures, as well as motifs in metabolic networks.

Concerning motifs in DNA sequences, Pierre Peterlongo, a PhD student co-supervised by HELIX and the University of Marne-la-Vallee who defended in September 2006, designed two algorithms, called Nimbus (2005; *LNCS*, vol. 3772, pages 179-190) and Ed'Nimbus [74], for filtering sequences prior to finding repetitions occurring more than twice in a sequence, or in more than two sequences. Nimbus and Ed'Nimbus use gapped seeds that are indexed with a new data structure, that can be either a bi-factor [73] or array [74]. Experimental results show that the filter can be very efficient. This work is being done in collaboration with Nadia Pisanti from the University of Pisa, Italy, and with Alair Pereira do Lago from the University of São Paulo, Brazil.

In the same vein, Frédéric Boyer and Eric Coissac, in collaboration with Guillaume Achaz at the Université Paris VI, have developed a new, space efficient implementation of the Karp-Miller-Rosenberg algorithm to look for exact repeats in very large DNA sequences (such as complete human chromosomes) [2]. This implementation forms the basis of the Repseek program that looks for approximate repeats, *i.e* allowing for deletion or substitution scores. In the statistical framewotk of extreme distributions, the parameters of the score distribution, as a function of the DNA GC content, have been empirically determined.

Concerning RNA secondary structures, Julien Allali, ex-PhD student of Marie-France Sagot and now Associate Professor at the LABRI, University of Bordeaux, introduced a new data structure, called MiGaL for "Multiple Graph Layers", that is composed of various graphs linked together by relations of abstraction/refinement. The new structure is useful for representing information that can be described at different levels of abstraction, each level corresponding to a graph. We proposed an algorithm for comparing two MiGaLs [63] [3]. MiGaLs represent a very natural model for comparing RNA secondary structures that may be seen at different levels of detail, going from the sequence of nucleotides, single or paired with another to participate in a helix, to the network of multiple loops that is believed to represent the most conserved part of RNAs having similar function.

As part of the PhD of Nuno Mendes in co-supervision with Ana Teresa Freitas from the Instituto Superior Tecnico of Lisbon, Portugal, work has just started on the development of new algorithms and models for predicting small functional RNA motifs. In particular, several models for an RNA sequence will be studied, allowing a flexible and general approach to RNA interference problems (problems of interference of RNAs with regulation of gene expression). This work will concern more specially microRNAs (denoted by miRNAs). Such RNAs are 20 to 24 nucleotides in length, single stranded and predominantly derived from intergenetic regions. Due to the difficulty of systematically detecting miRNAs by existing experimental techniques, researchers increasingly turn to computational methods to identify new miRNAs. Important computational tools have been used but they are limited in practice since they are based on comparative sequence analyses that can fail when sequence conservation is too low (that is, leads to poor alignments) or too high (that is, leads to lack of sequence covariation – variation of bases at a given position that is correlated with variation at another position). Some of them perform simultaneous multiple sequence alignment and folding that is computationally expensive. A very recent approach proposes a probabilistic algorithm that uses covariance models for motif description to predict miRNAs motifs in unaligned sequences. However, the approach is not able to find the best solution and cannot be used to identify miRNAs motifs that are only present in a subset of the input sequences. To push forward future developments in this field, new algorithms to identify miRNAs motifs unrelated to previously known ones have to be developed.

Finally, as part of the long-term visit among us of Paulo G. S. Fonseca (Paulo is a PhD student of Katia Guimaraes at the Federal University of Pernambuco, Brazil who came to

France with a “sandwich” scholarship and is being funded for his second year with us by HELIX), we are working on the problem of integrating various sources of information (sequence motifs, gene expression profiles and evolution) to infer genetic network modules.

6.1.4 Knowledge representation for genomics

Genome annotation can be viewed as an incremental, cooperative, data-driven, knowledge-based process that involves multiple methods to predict gene locations and structures. This process may have to be executed more than once and may be subjected to several revisions as the biological (new data) or methodological (new methods) knowledge evolves. In this context, although many annotation platforms already exist, there is still a strong need for computer systems which take care, not only of the primary annotation, but also of the update and advance of the associated knowledge. We propose to adopt a blackboard architecture when designing such a system. In his PhD, defended in July, S. Descorps-Declère has developed a prototype, called GENEPI, which validates these conceptual and technical options [25]. Specific adaptations to the classical blackboard architecture have been required, such as the description of the activation patterns of the knowledge sources by using an extended set of Allen’s temporal relations.

Recent works around the AROM platform (see 5.1), in collaboration with Jérôme Gensel (LSR-IMAG) and Cécile Capponi (LIF Marseille), concerns 1. the evolution of the AROM knowledge representation meta-model (to integrate whole-part relationships on a basis close to the one proposed in the UML-2 specification), and 2. the integration into the AROM2 platform of an Algebraic Modelling Language (AML) which allows the writing of equations involving variables of classes and associations. These equations are part of an AROM model and can be used to infer variables values.

Finally, Helix has welcomed Dr. José Luis Aguirre, Professor at the Technologico de Monterrey (Mexico) for a one year visit, starting in August 2006. Dr. Aguirre’s research interests are directed toward multi-agent systems and their application to information search and exchange. One of the expected outcomes of this visit is to study a multi-agent system for facilitating the access to heterogeneous and distributed biological data according to a user’s specific profiles known by the system.

Topic x undefined

6.2 Functional genomics

debug : Module named “resfuncgen” for the project HELIX, section resultats, topic

6.2.1 Computational proteomics and transcriptomics

This year has seen the thorough rewriting of the PEPLINE software, in collaboration with users at the Laboratoire de Chimie des Protéines (LCP, CEA Grenoble), in order to improve both the accuracy and the overall performance of the pipeline. The PST generation algorithm has been modified in order to incorporate a *de novo* step and a new rank-based scoring scheme. This resulted in a large improvement of the accuracy (70 % of correctly predicted PSTs). The performance of the chromosome mapping step has also been improved; the complete analysis of the chromosomes of *A. thaliana* can now be performed in a few minutes. These new algorithms have been added into the GenoProteo module of GenoStar by Jérémie Turbet and Marianne Tardif at the LCP with the help of the Genostar development team. This provides the end user with an intuitive and easy to use interface to PepLine embedded in the GenoStar environment. A paper describing the approach as well as some applications to the chloroplastic membrane of *A. thaliana* is in preparation.

Transposons, also called transposable elements, are sequences of DNA that can move around to different positions within the genome of a single cell, a process called transposition. Transposons that are still active are transcribed and we know that such transcription is regulated but genome-scale studies of their profile of expression has rarely been attempted. As part of the Master of Florence Cavalli (currently doing a PhD at the EBI and the University of Cambridge), we started addressing this question, using the genome of *Drosophila melagonaster* as our model. This work was done in collaboration with Cristina Vieira, from the team of "Genome and Populations" at the LBBE. We used for that data from ESTs (Expressed Sequence Tags – these are short sub-strings of a transcribed protein-coding or non-protein-coding nucleotide sequence originally intended as a way to identify gene transcripts) available in various public databases. The difficulty of the problem in the case of transposon is in correctly assigning each EST to its transposon in the genome sequence. Indeed, while the sequencing of ESTs is not error-free, the transposons in a same family are almost exact copies of one another, much more than genes that are duplicates, and the families are in general bigger than gene families. This kind of problem is similar to the one faced when assembling a genome from its sequenced fragments. The initial results obtained by Florence Cavalli seem to indicate a different profile of expression of transposons in the X chromosome, and a correlation between the number of transposons in a family to which an EST maps (and that one may therefore assume is expressed) and the number of copies in that family. The second result in particular is in contradiction with previous ones. Both will continue being investigated during the PhD of Marc Deloger that started in October of this year in co-supervision with Cristina Vieira.

The main goal of pharmacogenomics is to predict the effect drugs may have based on the genomic information of the patient. Using microarray technologies, this may help improve both diagnosis and the posology policy to be adopted. A few research structures are now ready to simultaneously screen the transcriptome and the genome of patients in order to reveal possible correlations with drug effects, and to consider adapting the medical protocols with this new information. However, it is not easy to extract knowledge from the amount of data this entails. To try to address this issue, we are developing new methods that use bayesian networks. Bayesian networks permit to represent causal relations (through a DAG), and then to estimate state probabilities. This method has already been applied to microarray data (2000; *J. Comp. Biol.* 7:601-620). Our contribution will be to append clinical information to the network and to specify a model that takes pharmacogenomics constraints into account. This should enable to do more accurate predictions. The pharmacogenomics structure of Lyon will keep all the data treated in several projects (each corresponding to a different cancer disease), with several samples for a given tumor at different stages of the disease. This gives us the possibility to model the time dimension using dynamic bayesian networks (2003; *Brief Bioinform.* 4:228-235). It will also be interesting to compare the information on the different diseases (by comparing the corresponding bayesian networks) which can reflect some cancer mechanisms. This represents an original approach, which could then be applied to other biological systems. This work will be done by Emmanuel Prestat and Christian Gautier.

Molecular data on biodiversity both in health and ecology represent new challlenges for data analysis. In particular, the large number of probes present in this case on the DNA chips invalidate all discriminant analyses. HELIX has concentrated its effort on addressing this issue. Two main results that have been obtained. The first, by Jean Thioulouse, is that the association between DNA chips devoted to biodiversity analysis and environment data cannot be made by a classical maximization of correlation (canonical correlation analysis); however the use of co-inertia analysis (CIA) that maximizes covariance have proven its efficiency on several studies of soil microbial biodiversity [28] [29] [40] [54] [55]. The second result was obtained by Caroline Truntzer, a PhD student of the "Biostatistics-Health" team of the LBBE co-supervised by Christian Gautier. Based on both simulated and "real" results, C. Truntzer made a comparison of different multivariate analyses that had been performed to discriminate between clinical states by using human pangenomic chips. The two works used the R software, more particularly the ADE4 package developed in HELIX.

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

Topological motifs have been extensively studied in the context of genetic and protein interaction networks but they seem to be not adapted to capture the functional information of metabolic networks. Therefore, as part of the PhD of Vincent Lacroix, we have defined a new type of motif (called coloured motifs and for which the topology of the subgraph is not given, only the labels of the nodes are known). We have worked on the problem of searching for all the occurrences of such motifs in a graph. We now have an exact algorithm for solving this problem as well as a proof that this problem is NP-complete [42].

To define ways of assessing the over and under-representation of such motifs in metabolic networks we have then collaborated with Sophie Schbath (INRA), Stephane Robin (InaPG) and their group. From this initial goal, we worked in two directions. The first concerns the conception of realistic random graph models (which model well the distribution of the degrees of the nodes, as well as the modularity of metabolic networks). The main achievement of this part is the extension of the Erdos-Renyi random graph model to a mixture model (ERMG) for which general properties such as clustering coefficient have been studied [67] [70]. The second direction is to search for an analytic formula (to avoid simulations) for the expectation and the variance of the number of occurrences of a motif in a Erdos-Renyi random graph. This work has been successfully applied to topological motifs and we are now working on coloured of motifs (work in preparation).

As part of the postdoc of Patricia Thebault, the relation between metabolic motifs in general, and coloured motifs in particular on one hand, and gene expression on the other is also being investigated using various statistical approaches. The organism chosen for such study is *Saccharomyces cerivisiae* with gene expression data taken from the *Saccharomyces* Genomes Database (SGD) and gene regulation data from Yeastract (2006; *Nucleic Acids Res.* 34:446-451, `http://db.yeastgenome.org/cgi-bin/expression/expressionConnection.pl`) that is maintained by our Portuguese collaborators, Ana Teresa Freitas and Arlindo Oliveira from the

Instituto Superior Tecnico in Lisbon. This work is one aspect of a more general work on the links between metabolic and genetic netwoorks. The main aim is to be able to provide a framework for the modelling of the relations between the genotype and phenotype. This work should also lead to proposing new models of evolutionary and functional modularity in biological networks.

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

The tools that are available to draw, and to manipulate the drawings of metabolism are usually restricted to metabolic pathways. This limitation becomes problematic when studying processes that span several pathways. In collaboration with Fabien Jourdan (INRA Toulouse), Romain Bourqui and David Auber (LABRI, University of Bordeaux), Vincent Lacroix and Ludovic Cottret are participating in the development of a method which enables to draw the entire metabolic network while also taking into account its structuration into pathways [65] [11].

Anne Morgat, from the Swiss-Prot group at the Swiss Institute for Bioinformatics, has continued her work on the Unipathway project in the framework of the BioSapiens NOE and the UniProt grants. The project aims at providing a standardized representation of metabolic data in the UniProtKB/Swiss-Prot database. These metabolic data are explicitely represented and stored into a relational database (UniPathwayDB). They are hierarchically decomposed into super-pathways, pathways, linear sub-pathways and reactions (steps). The development of UniPathwayDB (using postgreSQL) was performed through a collaboration with Eric Coissac at the Université Joseph Fourier. The database is populated with manually expertised metabolic data (from the Swiss-Prot group) and public data (UniProtKB/Swiss-Prot, complete proteomes (UniProtKB/Swiss-Prot and UniProtKB/TrEMBL), GenomeReview complete genomes, Enzyme). By the end of year 2006, more than 260 pathways were manually curated, representing about 450 distinct biochemical reactions. This covers more than 30 000 Swiss-Prot entries (about 70% of the total number of entries related to metabolism). The database will be made available in early 2007 through a web site hosted at the INRIA Rhône-Alpes. The server, as well as one full time engineer (Sophie Huet) who was hired in october 2006, have been provided by the PRABI (Génopole Rhône-Alpes) to this purpose.

6.2.3 Modelling and simulation of genetic regulatory networks

The group of Hidde de Jong has continued their efforts on the application of the qualitative simulation tool Genetic Network Analyzer (GNA) (section 5.10)) to the modeling of actual genetic regulatory networks. In particular, we study the nutritional stress response in the bacterium *Escherichia coli* in collaboration with experimental biologists in the laboratory of Johannes Geiselmann (Université Joseph Fourier, Grenoble, on leave in HELIX since October 2006). The original model developed by Delphine Ropers, published in a special issue of *BioSystems* [53], has been extended with additional genes and proteins in order to account for observed discrepancies between the model predictions and published data. Moreover, Delphine Ropers has compared, by means of a Monte-Carlo simulation study, the detailed nonlinear differential equation model of the stress response network with the reduced piecewise-linear differential equation model used in GNA. The project EC-MOAN, funded in the framework of the FP6 NEST programme of the European Commission (2006-2009), and the project MetaGenoReg, funded by the ANR in the framework of the BioSys programme (2006-2009), will allow us to maintain and extend these modeling activities.

The *E. coli* stress response model has given rise to predictions that cannot be tested by currently available experimental data. This has motivated an experimental programme carried out in the laboratory of Johannes Geiselmann, using fluorescent and luminescent gene reporter systems to obtain precise measurements with a high sampling density. Several members of HELIX have contributed to the design of the experiments, while Bruno Besson has (re)developed the program WellReader for the analysis of the gene reporter measurements (section 5.36)). The systematic comparison of the experimental results and the model predictions is currently under way. Other experiments are being carried out in collaboration with Irina Mihalcescu of the Laboratoire de Spectométrie Physique (Université Joseph Fourier, Grenoble).

In addition to HELIX, various other groups are using GNA in their modeling projects. In a number of cases, we have been actively involved in the formulation of the biological problem and the actual application of the tool. The current version 5.6 of GNA has been deposited at the APP and is distributed by the company Genostar. It has also been integrated in the Iogma platform for exploratory genomics developed by the company Genostar. The European project Cobios (FP6 NEST), which is due to start early 2007, will provide additional support to achieve this integration, notably by providing modules for the formulation of simulation models and facilitating the exchange of models with other modeling and simulation tools.

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

First of all, instead of generating the entire state transition graph, it is often sufficient to compute the steady states of the system and to analyze the neighbouring states in order to determine the stability of the steady states. Based on the mathematical characterisation of equilibria of piecewise-linear differential equation models and their stability, carried out in collaboration with Jean-Luc Gouzé (INRIA Sophia-Antipolis) and Tewfik Sari (Université Haute Alsace, Mulhouse) [13], Michel Page and Hidde de Jong have developed an attractor search module for GNA. This module transforms the search of steady states into a SAT problem and exploits existing, efficient SAT solvers to find all steady states of networks of more than thousand genes. The work on the attractor search module has been submitted for publication.

A second solution for the upscaling problems consists in the use of model-checking techniques for the automated verification of properties of state transition graphs. In the framework of his PhD thesis, Grégory Batt has pursued this approach in collaboration with Radu Mateescu and his colleagues of the VASY project. This has resulted in another version of GNA, currently only available as a prototype for internal use, which connects the simulation tool to state-of-the-art model checkers. In order to achieve this, a refined simulation method has been developed that exploits the concept of discrete abstraction developed in the hybrid systems community. The work initiated by Grégory Batt is now being carried on in several directions. Estelle Dumas recently joined the HELIX and VASY projects, on an INRIA associate engineer contract, to develop a user-friendly web interface between GNA and the model checker CADP. In the framework of his PhD thesis, Pedro Monteiro has started to study appropriate temporal logics and high-level specification languages for helping the user to formulate biological properties the model has to satisfy. Adrien Richard, in collaboration with Gregor Goessler (POP-ART), is currently investigating the use of modular approaches to verify larger networks.

The above-mentioned work has focused on the analysis of models obtained through literature study and human expertise. The PhD of Samuel Drulhe, supervised by Hidde de Jong and Giancarlo Ferrari-Trecate (University of Pavía) within the framework of the European project HYGEIA, takes a different direction. It concerns the development of methods for the identification of piecewise-linear differential equation models of genetic regulatory networks from gene expression data, adapting existing methods for the identification of hybrid systems. A paper summarizing the first results on simulated data has been presented at the major annual hybrid systems conference [27], while a longer version of the method has been submitted for a journal publication. Shortly, the application of the method to gene reporter data on the *E. coli* nutritional stress response will be undertaken.

7 Contracts and Grants with Industry

7.1 Genostar

debug : Module named “Genostar” for the project HELIX, section contrats

Participants: François Rechenmann.

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

7.2 sanofi pasteur

debug : Module named "sanofi pasteur" for the project HELIX, section contrats

Participants: Frédéric Boyer, Alain Viari.

In September 2004, HELIX started a two-year contractual relation with the company Sanofi Pasteur located in Lyon. The collaboration concerns the in-depth (re)annotation of pathogenic bacteria of interest to Sanofi Pasteur. The (re)annotation of several strains was completed in 2006 and followed, in 2007, by their comparative analysis in order to pinpoint genomic or metabolic specificities. This work has been done by F. Boyer in the context of his post-doc between Lyon and Grenoble.

8 Other Grants and Activities

8.1 National projects

debug : Module named "national" for the project HELIX, section international

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

Project name	Evolutionary dynamics of global gene regulatory networks in *Escherichia coli*
Coordinator	J. Geiselmann
HELIX participants	H. de Jong, M. Page, D. Ropers
Type	inter-EPST Microbiologie (2004-2006)
Web page	Not available

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

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

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

Project name	Integrated Biological Networks (IBN)
Coordinator	M.-F. Sagot
HELIX participants	V. Acuña, M. D. V. Braga, L. Canet, L. Cottret, M. Deloger, H. de Jong, C. Gautier, L. Gueguen, V. Lacroix, C. Lemaitre, L. Palmeira, E. Prestat, M.-F. Sagot, P. G. S. Fonseca, E. Tannier, P. Thébault, A. Viari
Type	ARC Inria (2005-2006)
Web page	`http://biomserv.univ-lyon1.fr/baobab/rubrique.php3?id_rubrique=63`

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

Project name	ModelPhylo
Coordinator	N. Galtier
HELIX participants	M. Gouy, B. Boussau
Type	ACI IMPBio (2004-2006)
Web page	`http://pari-stic.labri.fr/IMPBio/MODELPHYLO_2004.pdf`

Project name	PBIL
Coordinators	M. Gouy and G. Deleage
HELIX participants	M. Gouy
Type	Plan Pluri Formations
Web page	N/A

Project name	PBIL-Extension
Coordinator	C. Combet
HELIX participants	M. Gouy, G. Perrière, J. Thioulouse, L. Duret, S. Penel, D. Kahn, V. Lombard
Type	ACI IMPBio (2004-2006)
Web page	`http://impbio.lirmm.fr/PROJETS_ACCEPTES/paper85.html`

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

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

8.2 European projects

debug : Module named "europe" for the project HELIX, section international

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

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

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

8.3 International projects

debug : Module named "international" for the project HELIX, section international

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

Project name	ArcoIris
Coordinator	M.-F. Sagot and Y. Wakabayashi
HELIX participants	V. Acuña, M. D. V. Braga, L. Canet, L. Cottret, M. Deloger, C. Gautier, L. Guéguen, V. Lacroix, C. Lemaitre, L. Palmeira, E. Prestat, M.-F. Sagot, P. G. S. Fonseca, E. Tannier, P. Thébault, A. Viari
Type	Associated Team INRIA-USP (2005-2007)
Web page	`http://biomserv.univ-lyon1.fr/baobab/rubrique.php3?id_rubrique=65`

Project name	BemTeVi
Coordinators	C. E. Ferreira and M.-F. Sagot
HELIX participants	V. Acuña, M. D. V. Braga, L. Canet, L. Cottret, M. Deloger, C. Gautier, L. Guéguen, V. Lacroix, C. Lemaitre, L. Palmeira, E. Prestat, M.-F. Sagot, P. G. S. Fonseca, E. Tannier, P. Thébault, A. Viari
Type	Project FAPESP, Brazil (2005-2006)
Web page	`http://biomserv.univ-lyon1.fr/baobab/rubrique.php3?id_rubrique=65`

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

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

‘

9 Dissemination

9.1 Talks

debug : Module named “talks” for the project HELIX, section diffusion

Bastien Boussau

Title	Event and location	Date
Evolution of thermophily	Meeting "LUCA, ten years after", Fondation Les Treilles, Tourtour, France	Sept. 2006

Eric Coissac

Title	Event and location	Date
Unipathway: a metabolic door to UniProtKB/Swiss-Prot	Third Indo-French Bioinformatics Meeting (IFBM 2006)	Bangalore (India), June 2006
Du gène à l'environnement	École CNRS Microbiologie moléculaire, Carry Le Rouet	Oct. 2006
Unipathway: une porte d'entrée vers le métabolisme dans Swiss-Prot	Séminaire de l'unité "Génétique moléculaire de la levure", Institut Pasteur	Dec. 2006

Laurent Duret

Title	Event and location	Date
Homology-dependent methylation of repetitive DNA in mammals: what can we learn from comparative genomics?	Reunion Alphy, Lyon	Jan. 2006
Deciphering substitution pattern variations along mammalian genomes: control of transposable elements and other stories	VIB, Ghent, Belgium	Feb. 2006
Analysis of excision polymorphism in the macronucleus of *Paramecium tetraurelia*	"International Paramecium Genomics Meeting", Dourdan	May 2006
The impact of recombination on the evolution of mammalian genomes	University of Bern, Switzerland	Jun. 2006
The Xist RNA Gene Evolved in Eutherians by Pseudogenization of a Protein-Coding Gene	Second International Conference on X-inactivation", Paris	Sept. 2006
The impact of whole genome duplications: insights from *Paramecium tetraurelia*	"Otto Warburg International Summer School and Workshop 2006 on Evolutionary Genomics", Berlin, Germany	Sept. 2006
The impact of whole genome duplications: insights from *Paramecium tetraurelia*	Reunion GTGC, Nantes	Oct. 2006
The impact of whole genome duplications: insights from *Paramecium tetraurelia*	University of Lausanne, Switzerland	Dec. 2006

Samuel Drulhe

Title	Event and location	Date
Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks	Hybrid Systems: Computation and Control (HSCC) 2006, Santa Barbara, USA	Mar. 2006

Christian Gautier

Title	Event and location	Date
Some bioinformatics tools for studying microbial diversity	Meeting COST 853, Working group "Bioinformatics and information dissemination", Zurich, Switzerland	Mar. 2006

Manolo Gouy

Title	Event and location	Date
Contribution of comparative genomics to the analysis of mitochondrial evolution in protists	University of Milano, Italy	Mar. 2006
Efficient likelihood computation with non-reversible evolutionary models	University of Ottawa, Canada	May 2006

Hidde de Jong

Title	Event and location	Date
Qualitative modeling and simulation of the carbon starvation response in *Escherichia coli*	TIGEM seminar, Naples	Feb. 2006
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Annual Conference of the Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM), Jena, Germany	Mar. 2006
Qualitative modeling and simulation of bacterial stress responses	4th Workshop on Statistical methods for molecular biology high throughput data, Toulouse	Mar. 2006
Modeling, analysis, and simulation of genetic regulatory networks (with Jean-Luc Gouzé)	CEA-EDF-INRIA School on Nonsmooth Dynamical Systems: Analysis, Control, Simulation, and Applications, Rocquencourt	May 2006
Qualitative modeling of the carbon starvation response in *E. coli*	Workshop Towards Molecular Systems Biology, Bielefeld University, Centre for Interdisciplinary Research (ZIF), Germany	Jun. 2006
Introduction to the modelling of genetic regulatory networks, Qualitative analysis of piecewise-linear models of genetic regulatory networks & Application: qualitative analysis of the carbon starvation response in *E. coli*	EPSRC (Engineering and Physical Research Council) Graduate Course on Networks, Bristol, UK	Jul. 2006
Qualitative modeling and simulation of genetic regulatory networks: From piecewise-affine differential equations to reporter gene data (and back)	Keynote lecture 20th International Workshop on Qualitative Reasoning, Dartmouth, USA	Jul. 2006
Qualitative modeling and simulation of genetic regulatory networks	BioIT Management and Modeling, EuroBio06, Paris	Oct. 2006
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Colloque SFG Genetics meets Systems Biology. Theory and Practice	Nov. 2006
An overview of methods for the modelling and simulation of genetic regulatory networks & Qualitative modelling and simulation of bacterial regulatory networks	BioInfoSummer symposium, Canberra, Australia	Dec. 2006

Daniel Kahn

Title	Event and location	Date
Modularité et plasticité des systèmes régulateurs chez les bactéries: l'exemple de la régulation de la fixation de l'azote	Laboratoire de Biométrie et Biologie Evolutive, University Claude Bernard	Jan. 2006
L'évolution des familles de protéines analysée par réseau Bayésien	Génoscope, Evry	Feb. 2006
Modularité des protéines et identification des enzymes codés dans les génomes	Swiss Institute of Bioinformatics, Geneva	May 2006
Approaches for reconstructing metabolism and its evolution	Workshop on Molecular Systems Biology, Bielefeld	Jun. 2006
Protein domain families and protein innovation	Celebrating the 20th Anniversary of Swiss-Prot, Fortaleza	Aug. 2006
Comment modéliser l'interaction entre régulations métaboliques et régulations géniques ?	RIAMS 2006, Lyon	Nov. 2006

Anne Morgat

Title	Event and location	Date
UniPathway: a metabolic door to UniProtKB/Swiss-Prot	"In silico analysis of proteins", Fortaleza, Brazil	Aug. 2006
UniPathway project	BioSapiens Third AGM, Barcelona	Mar. 2006

Leonor Palmeira

Title	Event and location	Date
Models of DNA evolution with neighbor-dependent substitutions	"Otto Warburg International Summer School and Workshop 2006 on Evolutionary Genomics", Berlin, Germany	Sep. 2006
Theoretical approaches for the study of dinucleotide content in genomes	Theoretical Approaches for the Genome, Annecy, France	Nov. 2006

Guy Perrière

Title	Event and location	Date
Automated homologous sequence identification	Annual Meeting of the Society for Molecular Biology and Evolution, Tempe	May 2006
Molecular sequence databases in the genomic age	3rd Meeting of the Global U8 Consortium, University of the Havre	Jun. 2006

Delphine Ropers

Title	Event and location	Date
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Systems Biology Workshop, Paris	Feb. 2006
Qualitative simulation of the carbon starvation response in *Escherichia coli*	Seminar of the SAFE Consortium (European Association for Food Safety), Brussels, Belgium	Mar. 2006

Marie-France Sagot

Title	Event and location	Date
Computational biology: Negative thoughts and (maybe) some positive actions	XV Congress of the FESPB	Jul. 2006

Eric Tannier

Title	Event and location	Date
Conservation and Rearrangements in Genomes	INRA, Toulouse	Mar. 2006
Conservation and Rearrangements in Genomes	University of Rome, Italy	Oct. 2006

9.2 Organization of conferences, workshops and meetings

debug : Module named "animation" for the project HELIX, section diffusion

Hidde de Jong

Type	Location	Date
Second International Symposium on Positive Systems (POSTA 06)	Université Joseph Fourier	Aug. 2006

Marie-France Sagot

Type	Location	Date
Biological Networks III: Modularity and Genome Evolution	University of Bologna Residential Center, Bertinoro, Italy	Jun. 2006

Eric Tannier

Type	Location	Date
Workshop on Comparative Genomics	Lyon	Jan. 2006
Workshop on Comparative Genomics	Nantes	Oct. 2006

Alain Viari

Type	Location	Date
Seconde Journée Nationale Bioinformatique pour la Protéomique	Réseau National des Génopoles, INRIA Rhône-Alpes	Jun. 2006

9.3 Editorial and reviewing activities

debug : Module named "editorial" for the project HELIX, section diffusion

Laurent Duret

Type	Journal or conference
Member Steering Committee	French national conference on Bioinformatics, Jobim
Editorial Board	*Systematic biology*

Manolo Gouy

Type	Location
Area Chair for the Evolution & Phylogeny Section of ISMB	Fortaleza, Brazil
Editorial Board	*Molecular Biology and Evolution*

Daniel Kahn

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

Hidde de Jong

Type	Journal or conference
Editorial Board	*ACM/IEEE Transactions on Computational Biology and Bioinformatics*
Editorial Board	*BioSystems* (guest editor of special issue on Dynamical Modelling of Biological Regulatory Networks)
Editorial Board	*Technique et Science Informatiques* (guest editor of special issue on Modélisation et simulation pour la post-génomique)
Program Committee	CompBioNets 06, QR 06, GENSIPS 06, JOBIM 06, IPG 06, RIAMS 06, AIME 07
Scientific Committee	Working group VICANNE (Modélisation dynamique et simulation des systèmes biologiques)
Coordinator (with S. Robin)	Working group on Transcriptome, protéome, modélisation, inférence et analyse des réseaux biologiques of GDR CNRS 3003 Bioinformatique moléculaire

Delphine Ropers

Type	Journal or conference
Member Program Committee	JOBIM 07

Marie-France Sagot

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

Eric Tannier

Type	Location
Member Program Committee of the Recomb-Comparative Genomics Conference	Montreal, Canada

9.4 Administrative activities

debug : Module named "administrative" for the project HELIX, section diffusion

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

C. Gautier is director of the LBBE (UCBL, UMR 5558), deputy director of the IFR of Biology of the UCBL, chair of the section 29 of the CoNRS, and director of the PRABI.

M. Gouy is member of the "Comite National des Universités", section 67 (Ecology & Evolution), the selection committee of the CNRS ATIP Biodiversity, the Scientific Advisory Board of the Swiss Institute of Bioinformatics, the Scientific Board of the "Institut Francais de la Biodiversité", and the committee of the "Centre Intégratif de Génomique (CIG)", University of Lausanne.

Hidde de Jong is a member of the International Relations working group of the Conseil d'Orientation Scientifique et Technologique (COST) of INRIA as well as Europe correspondent of INRIA Rhône-Alpes. He has participated in the reviewing process of projects or candidates for a research position at Delft University of Technology (the Netherlands) and the Australian National University (Australia).

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

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

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

Marie-France Sagot is a member of the section 44 of the CoNRS, and of the scientific committees of the courses "Informatique en Biologie" of the Institut Pasteur in Paris and Computational Biology of the University of Chile in Santiago, Chile. She is Director of the PhD Program on Computational Biology, Instituto Gulbenkian de Ciencia, Lisbon, Portugal. She participated in the recruiting committee for CR2 positions at the INRIA Rhone-Alpes. She also participated in the reviewing process of projects or candidates for a research position for the Universities of Toronto and Vancouver (Canada), the "Fund for Scientific Research" (FWO) (Flanders, Belgium), the EPSRC (UK), the BBSRC (UK), the University of Haifa (Israel) and the Technion (Haifa, Israel).

Alain Viari is a member of the "Commission de spécialistes" section 65 at Université de Paris 6 and of the scientific advisory board of the MIA (Mathematics and Applied Mathematics) at the INRA. He is president of the "Comité des Emplois Scientifiques" at the INRIA Rhône-Alpes and member of the COST-AE (INRIA). He is also co-responsible of the Bioinformatics program of the "Haut Conseil pour la coopération scientifique et technologique entre la France et Israel" and member of the scientific committee of the ANR "Biologie des systèmes".

9.5 Teaching

debug : Module named "enseignement" for the project HELIX, section diffusion

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

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

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

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

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

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

Laurent Duret

Subject	Year	Location	Hours
Bioinformatique	3 to 5	INSA Lyon	26
Bioinformatique	3 to 5	ENS Lyon	7
Bioinformatique	3 to 5	UCBL	7
Bioinformatique	3 to 5	Otto Warburg International Summer School, Germany	6
Bioinformatique	3 to 5	International Postgraduate Course in Genomics, Espagne	20

Hidde de Jong

Subject	Year	Location	Hours
Modelling and simulation of genetic regulatory networks (with D. Ropers)	5	UCBL	8
Modelling and simulation of genetic regulatory networks	4	INSA, Lyon	14
Modelling and simulation of genetic regulatory networks	4	University of Cuernavaca, Morelos, Mexico	2
Modelling and simulation of genetic regulatory networks	5	Instituto Gulbenkian de Cienca, Lisbon, Portugal	14

Manolo Gouy

Subject	Year	Location	Hours
Molecular phylogeny	5	UCBL, ENS Lyon, INSA Lyon	6
Molecular phylogeny	5	UCBL	9
Molecular phylogeny	4	INSA, Lyon	9

Daniel Kahn

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

Guy Perrière

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

Delphine Ropers

Subject	Year	Location	Hours
Modelling and simulation of genetic regulatory networks (with H. de Jong)	5	UCBL	8
Modelling and simulation of genetic regulatory networks	4	UJF, Grenoble	16

Marie-France Sagot

Subject	Year	Location	Hours
Algorithmics for biology	4	UCBL	4
Algorithmics for biology	5	INSA Lyon	8

Eric Tannier

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

10 Bibliography

Books and Monographs

[1] J.-P. Dumas, J.-L. Roch, E. Tannier, S. Varette, *Théorie des codes*, *sciences sup*, Dunod, 2007.

Articles in referred journals and book chapters

[2] G. Achaz, F. Boyer, E. Rocha, A. Viari, E. Coissac, "Repseek, a tool to retrieve approximate repeats from large DNA sequences.", *Bioinformatics*, 2006, in press.

[3] J. Allali, M.-F. Sagot, "A multiple layer model to compare RNA secondary structures", submitted.

[4] A. Aouacheria, C. Geourjon, N. Aghajari, V. Navratil, G. Deleage, C. Lethias, J. Y. Exposito, "Insights into Early Extracellular Matrix Evolution: Spongin Short Chain Collagen-related Proteins are Homologous to Basement Membrane Type IV Collagens and Form a Novel Family Widely Distributed in Invertebrates", *Mol Biol Evol*, 2006.

[5] A. Aouacheria, V. Navratil, A. Barthelaix, D. Mouchiroud, C. Gautier, "Bioinformatic screening of human ESTs for differentially expressed genes in normal and tumor tissues", *BMC Genomics 7*, 2006, p. 94.

[6] A.-M. Arigon, G. Perriere, M. Gouy, "HoSeqI: automated homologous sequence identification in gene family databases", *Bioinformatics 22(14)*, 2006, p. 1786–1787.

[7] J. Aury, O. Jaillon, L. Duret, B. Noel, C. Jubin, B. Porcel, B. Segurens, V. Daubin, V. Anthouard, N. Aiach, O. Arnaiz, A. Billaut, J. Beisson, I. Blanc, K. Bouhouche, F. Camara, S. Duharcourt, R. Guigo, D. Gogendeau, M. Katinka, A. Keller, R. Kissmehl, C. Klotz, F. Koll, A. L. Mouel, G. Lepere, S. Malinsky, M. Nowacki, J. Nowak, H. Plattner, J. Poulain, F. Ruiz, V. Serrano, M. Zagulski, P. Dessen, M. Betermier, J. Weissenbach, C. Scarpelli, V. Schachter, L. Sperling, E. Meyer, J. Cohen, P. Wincker, "Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*", *Nature 444*, 7116, 2006, p. 171–178.

[8] G. Batt, R. Casey, H. de Jong, J. Geiselmann, J.-L. Gouzé, M. Page, D. Ropers, T. Sari, D. Schneider, "Analyse qualitative de la dynamique de réseaux de régulation génique par des modèles linéaires par morceaux", *Technique et Science Informatiques*, 2006.

[9] G. Batt, R. Casey, H. de Jong, J. Geiselmann, J.-L. Gouzé, M. Page, D. Ropers, T. Sari, D. Schneider, "Qualitative analysis of the dynamics of genetic regulatory networks using piecewise-linear models", *in : Mathematical and Computational Methods in Biology*, E. Pecou, S. Martinez, and A. Maass (editors), Editions Hermann, Paris, 2006, p. 206–239.

[10] E. Billoir, A. Péry, S. Charles, "Integration of both lethal and sublethal effects of toxic compounds in population dynamics of *Daphnia magna*: a combination of DEBtox models and matrix population models", *Ecological Modelling*, 2006, in press.

[11] R. Bourqui, L. Cottret, V. Lacroix, D. Auber, P. Mary, M.-F. Sagot, F. Jourdan, "Metabolic network visualization using a constraint planar graph drawing algorithm", *BMC Bioinformatics*, 2006, submitted.

[12] B. Boussau, M. Gouy, "Efficient Likelihood Computations with Nonreversible Models of Evolution", *Systematic Biology 55(5)*, 2006, p. 756–768.

[13] R. Casey, H. de Jong, J.-L. Gouzé, "Piecewise-linear models of genetic regulatory networks: Equilibria and their stability", *Journal of Mathematical Biology 52*, 1, 2006, p. 27–56.

[14] S. Charles, H. Persat, J.-P. Mallet, "Population dynamics of grayling: Modelling temperature and discharge effects", *Mathematical Modelling of Natural Phenomena*, 2006, in press.

[15] J. Charlet, B. Cremilleux, M. D. Carvalho, C. Garbay, J. Jabre, D. Janas, A. Labarre-Vila, V. Luengo, V. Rialle, D. Ziébelin, "Traitement de l'information en médecine et ENMG", *Revue de Neurophysiologie clinique*, 2006.

[16] A. Chaumot, N. Milioni, A. Abdoli, D. Pont, S. Charles, "First step of a modeling approach to evaluate spatial heterogeneity in a fish (Cottus gobio) population dynamics", *Ecological Modelling 197(3-4)*, 2006, p. 263–273.

[17] Y. Clément, R. Tavares, G. Marais, "Does lack of recombination enhance asymmetric evolution among duplicate genes? Insights from the *Drosophila melanogaster* genome", *Gene*, 2006, in press.

[18] H. G. S. Consortium, "Insights into social insects from the genome of the honeybee *Apis mellifera*", *Nature 443*, 7114, 2006, p. 931–949.

[19] M. Crochemore, C. Iliopoulos, M. Mohamed, M.-F. Sagot, "Longest Repeats with a Block of k Don't Cares", *Theor Comput Sci*, 2006, in press.

[20] H. de Jong, C. Chaouiya, D. Thieffry, "Dynamical modeling of biological regulatory networks", *BioSystems 84*, 2, 2006, p. 77–80.

[21] H. de Jong, D. Ropers, "Qualitative approaches towards the analysis of genetic regulatory networks", *in : System Modeling in Cellular Biology: From Concepts to Nuts and Bolts*, Z. Szallasi, V. Periwal, and J. Stelling (editors), MIT Press, Cambridge, MA, 2006, p. 125–148.

[22] H. de Jong, D. Ropers, "Strategies for dealing with incomplete information in the modeling of molecular interaction networks", *Briefings in Bioinformatics 7*, 2006, p. 354–363.

[23] E. de Laplanche, K. Gouget, G. Cléris, F. Dragounoff, J. Demont, A. Morales, L. Bezin, C. Godinot, G. Perrière, D. Mouchiroud, H. Simonnet, "Physiological oxygenation status is required for fully differentiated phenotype in kidney cortex proximal tubules", *Am. J. Physiol. Renal Physiol. 291*, 2006, p. 750–760.

[24] G. Decelier, Y. Letrillard, S. Charles, C. Biémont, "TESD: A population genomics simulation environment for transposable element dynamics", *Bioinformatics*, 2006, in press.

[25] S. Descorps-Declere, D. Ziebelin, F. Rechenmann, A. Viari, "Genepi: a blackboard framework for genome annotation", *BMC Bioinformatics 7*, 2006, p. 450–462.

[26] Y. Diekmann, M.-F. Sagot, E. Tannier, "Evolution under reversals: parsimony and conservation of common intervals", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2006, in press.

[27] S. Drulhe, G. Ferrari-Trecate, H. de Jong, A. Viari, "Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks", *in : Hybrid Systems: Computation and Control (HSCC 2006)*, J. Hespanha and A. Tiwara (editors), *Lecture Notes in Computer Science*, *3927*, Springer-Verlag, Berlin, 2006, p. 184–199.

[28] R. Duponnois, K. Assikbetse, H. Ramanankierana, M. K. M, J. Thioulouse, M. Lepage, "Litter-forager termite mounds enhance the ectomycorrhizal symbiosis between *Acacia holosericea* A. Cunn. Ex G. Don and *Scleroderma dictyosporum* isolates", *FEMS Microbiol Ecol 56(2)*, 2006, p. 292–303.

[29] R. Duponnois, M. Kisa, K. Assigbetse, Y. Prin, J. Thioulouse, M. Issartel, P. Moulin, M. Lepage, "Fluorescent pseudomonads occuring in Macrotermes subhyalinus mound structures decrease Cd toxicity and improve its accumulation in sorghum plants", *Sci Total Environ 370(2-3)*, 2006, p. 391–400.

[30] P. Durand, L. Labarre, A. Meil, J. Divol, Y. Vandenbrouck, A. Viari, J. Wojcik, "GenoLink: a graph-based querying and browsing system for investigating the function of genes and proteins", *BMC Bioinformatics 7*, 2006, p. 21–31.

[31] L. Duret, C. Chureau, S. Samain, J. Weissenbach, P. Avner, "The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene", *Science 312*, 5780, 2006, p. 1653–1655.

[32] L. Duret, A. Eyre-Walker, N. Galtier, "A new perspective on isochore evolution", *Gene*, 2006, in Press.

[33] L. Duret, "The GC Content of Primates and Rodents Genomes Is Not at Equilibrium: A Reply to Antezana", *J Mol Evol 62*, 6, 2006, p. 803–806.

[34] O. Ebenhoh, T. Handorf, D. Kahn, "Evolutionary changes of metabolic networks and their biosynthetic capacities", *Systems Biology 153(5)*, 2006, p. 354–358.

[35] R. Frutos, A. Viari, C. Ferraz, A. Bensaid, A. Morgat, F. Boyer, E. Coissac, N. Vachiery, J. Demaille, D. Martinez, "Comparative Genomics of Three Strains of *Ehrlichia ruminantium*: A Review", *Ann N Y Acad Sci. 1081*, 2006, p. 417–433.

[36] R. Frutos, A. Viari, C. Ferraz, A. Morgat, S. Eychenie, Y. Kandassamy, I. Chantal, A. Bensaid, E. Coissac, N. Vachiery, J. Demaille, D. Martinez, "Comparative genomic analysis of three strains of *Ehrlichia ruminantium* reveals an active process of genome size plasticity", *J Bacteriol. 188*, 2006, p. 2533–2542.

[37] J. Grassot, M. Gouy, G. Perriere, G. Mouchiroud, "Origin and Molecular Evolution of Receptor Tyrosine Kinases with Immunoglobulin-Like Domains", *Mol. Biol. Evol. 23(6)*, 2006, p. 1232–1241.

[38] H. Hallay, N. Locker, L. Ayadi, D. Ropers, E. Guittet, C. Branlant, "Biochemical and NMR study on the competition between proteins SC35, SRp40, and heterogeneous nuclear ribonucleoprotein A1 at the HIV-1 Tat Exon 2 splicing site", *The Journal of Biological Chemistry 281*, 2006, p. 37159–37174.

[39] T. Handorf, O. Ebenhoh, D. Kahn, R. Heinrich, "Hierarchy of metabolic compounds based on their synthesising capacity", *Systems Biology 153(5)*, 2006, p. 359–363.

[40] J. Y. L. Hesran, N. Fievet, J. Thioulouse, P. Personne, B. Maubert, S. M'bidias, D. Etye'ale, M. Cot, P. Deloron, "Development of cellular immune responses to *Plasmodium falciparum* blood stage antigens from birth to 36 months of age in Cameroon", *Acta Trop 98(3)*, 2006, p. 261–269.

[41] A. Khelifi, J. Meunier, L. Duret, D. Mouchiroud, "GC Content Evolution of the Human and Mouse Genomes: Insights from the Study of Processed Pseudogenes in Regions of Different Recombination Rates", *J Mol Evol 62*, 6, 2006, p. 745–752.

[42] V. Lacroix, C. G. Fernandes, M.-F. Sagot, "Motif Search in Graphs: Application to Metabolic Networks", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2006, in press.

[43] T. Lefebure, C. J. Douady, M. Gouy, J. Gibert, "Relationship between morphological taxonomy and molecular divergence within *Crustacea*: Proposal of a molecular threshold to help species delimitation", *Molecular Phylogenetics and Evolution 40(2)*, 2006, p. 435–447.

[44] T. Lefebure, C. J. Douady, M. Gouy, P. Trontelj, J. Briolay, J. Gibert, "Phylogeography of a subterranean amphipod reveals cryptic diversity and dynamic evolution in extreme environments", *Mol. Ecol. 15(7)*, 2006, p. 1797–1806.

[45] C. Lemaitre, M.-F. Sagot, "A Small Trip in the Untranquil World of Genomes", *Theoretical Computer Science*, 2006, accepted with minor revisions.

[46] J. R. Lobry, A. Necsulea, "Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes", *Gene*, 2006, in press.

[47] C. Melodelima, L. Guéguen, D. Piau, C. Gautier, "A computational prediction of isochores based on hidden Markov models", *Gene*, 2006, in press.

[48] V. A. na, G. Didier, A. Maass, "Centered codes", *Theor. Comput. Sci.*, 2006.

[49] A. Necsulea, J. R. Lobry, "Revisiting the directional mutation pressure theory: The analysis of a particular genomic structure in Leishmania major", *Gene*, 2006, in press.

[50] L. Palmeira, L. Guéguen, J. R. Lobry, "UV-targeted dinucleotides are not depleted in light-exposed Prokaryotic genomes", *Mol Biol Evol 23(11)*, 2006, p. 2214–2219.

[51] F. Rechenmann, "Dessine-moi un génome", *La Recherche 402*, 2006, p. 82–83.

[52] E. M. Rodrigues, M.-F. Sagot, Y. Wakabayashi, "The Maximum Agreement Forest Problem: approximation algorithms and computational experiments", *Theor Comput Sci*, 2006, in press.

[53] D. Ropers, H. de Jong, M. Page, D. Schneider, J. Geiselmann, "Qualitative simulation of the carbon starvation response in *Escherichia coli*", *BioSystems 84*, 2, 2006, p. 124–152.

[54] H. Sanguin, B. Remenant, A. Dechesne, J. Thioulouse, T. M. Vogel, X. Nesme, Y. Moenne-Loccoz, G. L. Grundmann, "Potential of a 16S rRNA-based taxonomic microarray for analyzing the rhizosphere effects of maize on *Agrobacterium* spp. and bacterial communities", *Appl Environ Microbiol 72(6)*, 2006, p. 4302–4312.

[55] A. Sanon, P. Martin, J. Thioulouse, C. Plenchette, R. Spichiger, M. Lepage, R. Duponnois, "Displacement of an herbaceous plant species community by mycorrhizal and non-mycorrhizal *Gmelina arborea*, an exotic tree, grown in a microcosm experiment", *Mycorrhiza 16(2)*, 2006, p. 125–132.

[56] M. Sémon, L. Duret, "Evolutionary origin and maintenance of coexpressed gene clusters in mammals", *Mol Biol Evol 23*, 9, 2006, p. 1715–1723.

[57] M. Sémon, J. R. Lobry, L. Duret, "No evidence for tissue-specific adaptation of synonymous codon usage in humans", *Mol Biol Evol 23*, 3, 2006, p. 523–529.

[58] P. Taberlet, E. Coissac, F. Pompanon, L. Gielly, C. Miquel, A. Valentini, T. Vermat, G. Corthier, C. Brochmann, E. Willerslev, "Power and limitations of the chloroplast trnl (uaa) intron for plant dna barcoding", *Nucleic Acids Res in press*, 2006.

[59] P. Thébault, S. de Givry, T. Schiex, C. Gaspin, "Combining constraint network processing and pattern matching to describe and locate structured motifs in genomic sequences", *Bioinformatics 22(17)*, 2006, p. 2074–2080.

[60] I. Vatcheva, O. Bernard, H. de Jong, N. Mars, "Experiment selection for the discrimination of semi-quantitative models of dynamical systems", *Artificial Intelligence 170*, 4, 2006, p. 472–506.

[61] E. Vautrin, S. Charles, S. Genieys, F. Vavre, "Evolution and invasion dynamics of multiple infections with *Wolbachia* investigated using matrix based models", *J of Theor Bio*, 2006.

Publications in Conferences and Workshops

[62] S. Abby, M. Gouy, V. Daubin, "Des milliers d'arbres de gènes pour reconstruire l'histoire du vivant", *in: Journées Ouvertes : Biologie Informatique et Mathématiques (JOBIM)*, 2006.

[63] J. Allali, Y. d'Aubenton Carafa, C. Thermes, M.-F. Sagot, "MiGaL: an efficient tool for RNA secondary structures comparison", *in: Journées Ouvertes : Biologie Informatique et Mathématiques (JOBIM)*, 2006. short talk.

[64] F. Baty, M. P. Bihl, A. C. Culhane, M. Brutsche, G. Perrière, "Optimization of Between Group Analysis of gene expression disease class prediction", *in: Proceedings of the 2nd International Symposium on Mathematical and Computational Biology*, R. P. Mondaini, R. D. ao (editors), World Scientific Publishing, p. 351–366, 2005.

[65] R. Bourqui, D. Auber, V. Lacroix, F. Jourdan, "Metabolic network visualization using a constraint planar graph drawing", *in: 10th conf. on Information Visualization*, p. 489–496, 2006.

[66] A. Chaumot, S. Charles, "Pollution, stochasticity and spatial heterogeneity in the dynamics of an age-structured population of brown trout living in a river network", *in: Population-level Ecotoxicological Risk Assessment: Case Studies*, R. Akcakaya (editor), 2006.

[67] J.-J. Daudin, V. Lacroix, F. Picard, S. Robin, M.-F. Sagot, "Uncovering structure in biological networks", *in: Journées Ouvertes : Biologie Informatique et Mathématiques (JOBIM)*, 2006.

[68] H. de Jong, "Qualitative modeling and simulation of genetic regulatory networks: From piecewise-affine differential equations to reporter gene data (and back)", *in: Proceedings of the Twentieth International Workshop on Qualitative Reasoning, QR-06*, C. Bailey-Kellogg, B. Kuipers (editors), Dartmouth,, 2006.

[69] J. Gensel, C. Capponi, P. Genoud, D. Ziébelin, "Vers une intégration des relations Partie-Tout en AROM", *in: Langages et Modèles à Objets (LMO'06)*, 2006.

[70] M. Mariadrissou, J.-J. Daudin, V. Lacroix, V. Miele, F. Picard, S. Robin, M.-F. Sagot, "Uncovering structure in biological networks", *in: RIAMS'06*, 2006. papers accepted to this conference will be submitted *a posteriori* to *Journal of Biological Physics and Chemistry.*

[71] C. Melodelima, L. Guéguen, C. Gautier, D. Piau, "A Markovian Approach for the Analysis of the Gene Structure", *in: PSC'06*, 2006.

[72] B. Moisuc, P. Genoud, D. Ziébelin, J. Gensel, C. Capponi, “Apports de la modélisation algébrique pour la représentation de connaissances par objets : illustration en AROM”, *in: Langages et Modèles à Objets (LMO'06)*, 2006.

[73] P. Peterlongo, J. Allali, M.-F. Sagot, “The Gapped-Factor Tree”, *in: PSC'06*, 2006.

[74] P. Peterlongo, N. Pisanti, A. P. do Lago, M.-F. Sagot, “Ed'Nimbus: A Lossless Filter for Long Multiple Repetitions with Edit Distance”, *in: Journées Ouvertes : Biologie Informatique et Mathématiques (JOBIM)*, 2006. short talk.

[75] N. Pisanti, A. Carvalho, L. Marsan, M.-F. Sagot, “RISOTTO: Fast extraction of motifs with mismatches”, *in: LATIN'06*, J. R. Correa, A. Hevia, M. Kiwi (editors), *Lecture Notes in Computer Science*, *3887*, Springer-Verlag, p. 757–768, 2006.