INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team HELIX

Informatics and genomics

Grenoble - Rhône-Alpes

THEME BIO

Table of contents

Texte a mettre

1. Team

Research Scientist

Alain Viari [Research Director (DR) Inria]
Marie-France Sagot [Research Director (DR) Inria, HdR]
Eric Tannier [Research Associate (CR) Inria]
Vincent Daubin [Research Associate (CR) Cnrs]
Laurent Duret [Research Director (DR) Cnrs, HdR]
Manolo Gouy [Research Director (DR) Cnrs, HdR]
Daniel Kahn [DR Inra, HdR]
Gabriel Marais [Research Associate (CR) Cnrs]
Guy Perrière [Research Director (DR) Cnrs, HdR]
Franck Picard [Research Associate (CR) Cnrs, started October 2007]
Jean Thioulouse [Research Director (DR) Cnrs, HdR]

Faculty Member

Christian Gautier [Professor, University Claude Bernard, HdR]
Philippe Genoud [Associate Professor, University Joseph Fourier, on sabbatical in Canada]
Laurent Guéguen [Associate Professor, University Claude Bernard]
Jean Lobry [Professor, Universite Claude Bernard, HdR]
Dominique Mouchiroud [Professor, University Claude Bernard, HdR]
Sylvain Mousset [Associate Professor, University Claude Bernard]
Raquel Tavares [Associate Professor, University Claude Bernard]
Eric Coissac [Associate Professor, University Joseph Fourier, HdR]

Technical Staff

Corinne Lachaize [Project technical staff, Swiss Institute of Bioinformatics]
Anne Morgat [Project technical staff, Swiss Institute of Bioinformatics]
Edouard Blondeau [CDD Aventis-Pasteur until August 2008]
Ludovic Cottret [CDD ANR]
Jean-Francois Gout [CDD ANR]
Adrien Maudet [Associate Engineer]
Emmanuel Prestat [CDD ANR]

PhD Student

Sophie Abby [scholarship Ministère de la Recherche, supervisors: Vincent Daubin and Manolo Gouy]
Vicente Acuña [scholarship Conicyt (Chile) and INRIA, supervisors: Marie-France Sagot and Christian Gautier]
Bastien Boussau [scholarship BDI, CNRS, supervisor: Manolo Gouy, defended in November 2008]
Yves-Pol Deniélou [scholarship ENS, supervisors: Alain Viari and Marie-France Sagot]
Marc Deloger [scholarship BDI, CNRS, supervisors: Cristina Vieira (LBBE) and Marie-France Sagot]
Marília Dias Vieira Braga [scholarship AlBan, supervisors: Marie-France Sagot and Eric Tannier]
Claire Guillet [scholarship ENS Lyon, supervisor: Laurent Duret]
Joan Ho-Huu [scholarship Ministère de la Recherche, supervisor: Gabriel Marais]
Claire Lemaitre [scholarship Ministère de la Recherche, supervisors: Marie-France Sagot and Christian Gautier, PhD defended in November 2008, currently postdoc at Université de Bordeaux 2]
Yann Letrillard [scholarship Institut National du Cancer, supervisor: Guy Perrière]
Nuno Mendes [scholarship Portuguese Ministry of Research (FCT), supervisors: Ana Teresa Freitas (IST, Lisbon, Portugal) and Marie-France Sagot]
Anamaria Necsulea [scholarship Ministère de la Recherche, supervisor: Jean Lobry, defended in June 2008]

Alexandra Popa [scholarship Ministère de la Recherche, supervisors: Christian Gautier and Dominique Mouchiroud, started September 2007]
Clément Rezvoy [scholarship Ministère de la Recherche, supervisors: Frédéric Vivien, LIP-ENS Lyon and Daniel Kahn]
Anne-Sophie Sertier [scholarship Ministère de la Recherche, supervisors: Daniel Kahn and Vincent Daubin]
Patrícia Simões [scholarship Portuguese Ministry of Research (FCT), supervisors: Sylvain Charlat (LBBE) and Marie-France Sagot]

Post-Doctoral Fellow

Matteo Brilli [scholarship FP7 project]
Paulo Gustavo Soares da Fonseca [scholarship ANR]
Thomas Bernard [scholarship FP7 project]
Augusto F. Vellozo [scholarship ANR]

Visiting Scientist

Ana Teresa Freitas [Instituto Superior Técnico, Lisbon, Portugal, 4 months in 2008]
Pierre Peterlongo [INRIA Symbiose, various short visits]

Administrative Assistant

Françoise de Coninck [Secretary (SAR) Inria - Grenoble]
Gaelle Tworkowski [Secretary (SAR) Inria - Lyon (since May 2008)]

2. Overall Objectives

2.1. Overall Objectives

This is the last report for the HELIX team which ended in december 2008 and led to the creation of two other teams whose proposals have already been submitted while a third team should be proposed later in 2009. HELIX spanned two different locations, Grenoble and Lyon, and three different scientific cultures, biology, computer science and mathematics. This pluridisciplinary characteristic of HELIX will continue to be present in each of the three teams that should issue from it. The first offspring is IBIS headed by Hidde de Jong in Grenoble that will focus on the modelling, simulation, measurement, and control of bacterial regulatory networks. The second will be BAMBOO headed by Marie-France Sagot at Lyon which, by having evolution as its main research motivation, will naturally cover a broader area of computational biology, from the genomes to the networks of interactions between the main components of a cell. The third team will be headed by Alain Viari and have as main topic computational proteomics.

HELIX conducted research on computational biology by developing new algorithms and applying them to such bioinformatic objects as DNA and protein sequences, phylogenetic trees, and graphs formalising gene interactions or metabolic networks. The three new teams will also put strong emphasis on developing methods for analysing biological data and inferring from it knowledge on the functioning of living organisms.

One of the founding principles of the research conducted in the HELIX team, and likewise in the new ones is that the mathematical basis of any of our approaches should be clearly stated. An important part of the activity of the new teams, like for the old one, will therefore concentrate on the (re)formulation of biological questions into mathematical models suitable for computer analysis. The fundamental problem is how to design such a model in a way that should be simple enough to be practically useful but not so simple as to miss the subtleties of the biological questions addressed. A solution to this problem requires more than a simple remote collaboration between mathematicians and computer scientists on one side and biologists on the other, but calls instead for a real “symbiosis” between the two sets of cultures, symbiosis which, again, will characterise the three teams that issue from HELIX.

The IBIS team will do a report on their own this year already. The current one thus represents only the activities of the remaining of HELIX. We shall also further concentrate mainly on the activities of the members of HELIX who will remain members of an INRIA team-project.

3. Scientific Foundations

3.1. Comparative genomics

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

Participants: Sophie Abby, Bastien Boussau, Vincent Daubin, Marc Deloger, Marília Dias Vieira Braga, Laurent Duret, Christian Gautier, Jean-Francois Gout, Manolo Gouy, Laurent Guéguen, Claire Guillet, Daniel Kahn, Claire Lemaitre, Jean Lobry, Gabriel Marais, Dominique Mouchiroud, Sylvain Mousset, Anamaria Necsulea, Guy Perrière, Alexandra Popa, Marie-France Sagot, Anne-Sophie Sertier, Patrícia Simões, Paulo Gustavo Soares da Fonseca, Eric Tannier, Raquel Tavares.

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 probabilistic modelling techniques, general data analysis and text algorithmic methods, phylogenetic reconstructions, and combinatorics.

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

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 and inferring ancestor genomes. 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.

Another challenge regarding phylogeny concerns the study of co-evolution. Co-evolution refers to the mutual evolutionary influence between two (or more) species. Each party in a co-evolutionary relationship exerts selective pressures on the other, each thereby affecting the other's evolution. HELIX is more particularly interested in studying the co-phylogeny (co-speciation) of host species and their parasites by using molecular data. Recent results in co-phylogenetic methods show that all may lead to inaccurate results, thereby potentially seriously compromising their interpretation with a view to understanding the evolutionary dynamic of parasites in communities. This work involves a collaboration with Sylvain Charlat from the LBBE "Génétique et Évolution des interactions Hôtes-Parasites (GEIHP)" team, as well as other experimentalists from the GEIHP.

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

Genomic sequences are characterised 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 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, inversions of whole segments, duplications, 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.

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. Search and inference problems are 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 motifs act in general cooperatively and cooperative sets should in general be inferred together. However, this requires more sophisticated algorithms and statistical approaches.

Most problems in this section are conducted in collaboration with various international groups among which the main is the group of Arlindo Oliveira and Ana Teresa Freitas from the Instituto Superior Técnico, Lisbon, Portugal.

3.2. Functional genomics

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

Participants: Vicente Acuña, Thomas Bernard, Matteo Brilli, Ludovic Cottret, Christian Gautier, Laurent Guéguen, Daniel Kahn, Anne Morgat, Guy Perrière, Franck Picard, Emmanuel Prestat, Marie-France Sagot, Paulo Gustavo Soares da Fonseca, Raquel Tavares, Jean Thioulouse, Alain Viari.

Functional genomics refers to arriving at an understanding of the different features of a genome such as genes, non-coding RNAs etc. This requires in general understanding how such features are related to each other, 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, functional 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), and combinatorics (concerning in particular random graph models). Functional genomics further requires 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 Protéines" (LCP) at the CEA in Grenoble. This activity will be the basis of a team proposal in 2009 by Alain Viari who wishes to spend a sabbatical year in a proteomic lab in between.

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. 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. One of the main strategies for transcriptome analysis is hybridisation 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).

One particular type of array, the so-called CGH (Comparative Genomic Hybridisation) arrays, is more of interest to HELIX. The purpose of array-based CGH is to detect and map chromosomal aberrations, on a genomic scale, in a single experiment. CGH arrays have been used in particular for identifying recurrent chromosomal aberrations that occur in some types of cancer tumors. Work on such arrays is been conducted in collaboration notably with Pascal Roy in the "Santé et Statistique" team of the LBBE.

3.2.2. Modelling and structural analysis of biological networks

It is now commonly accepted that the functioning and development of a living organism is controlled by the networks of interactions between its genes, proteins, and small molecules. Studying such networks and their underlying complexity is therefore crucial. This requires improving the mathematical and algorithmic theory needed to accurately model, and then explore and analyse highly intricate systems. Biological networks may represent protein-protein interactions, the metabolism of an organism, its system of gene expression regulation, or even, mixed networks that contain information coming from various of the previous sources plus from the genome organisation itself.

Before analysing such networks, the information allowing to model them needs to be gathered. A large amount is now available on the molecular basis of cellular processes. Such data are quite heterogeneous. The challenge today is to relate and integrate the various types of data, usually dispersed in the literature and difficult to exploit, so as to answer questions involving the different levels of structural, functional, and spatial organisation of a cell. A major contribution of bioinformatics is therefore the development of databases and knowledge bases allowing 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 in collaboration with the SwissProt group at the Swiss Institute of Bioinformatics (database of metabolic pathways UNIPATHWAY).

The analysis of the networks reconstructed, by homology or *ab-initio*, then involve various graph-algorithmic problems such as searching for common topological or connected subgraph motifs, and more generally comparing valued graphs based on their topology or on their capacities (*i.e.*, comparing paths or flows). In the case of metabolic networks, simple valued graphs are not sufficient to fully capture the characteristics of metabolism, which is better represented by valued hypergraphs. This raises interesting and difficult questions as hypergraphs have been little studied in general, and more specifically in terms of algorithmics, combinatorics and statistics. This work is done in collaboration with Alberto Marchetti-Spaccamela from the University of Rome in Italy, Leen Stougie from the Free University and the CWI at Amsterdam in the Netherlands, the combinatorics group at the University of São Paulo in Brazil, and, for the statistical aspects, with Sophie Schbath at the INRA in Jouy-en-Josas and Stéphane Robin at the AgroTechParis in Paris.

3.2.3. Inter- and intra-chromosomal regulatory networks

For many years, work in the area of gene regulation concentrated on finding the sequences upstream of genes that could correspond to promoter or other regulatory sequences, that is, to sites where protein, RNA or protein/RNA complexes would bind and thereby initiate, stop, up or down-regulate the level of expression of a given gene.

The advent of high-throughput microarray technologies has added one important level of information to this image, as it enabled to measure the co-expression of sets of genes in a given tissue in given conditions. These studies suggest a correlated action of different elements of the gene regulation machinery and their potential interaction. More recently, the importance and extent of regulation at the epigenetic level started to be fully

realised. This refers to heritable changes in gene regulation that occur without a change in the DNA sequence and are therefore not encoded at the genomic level. Epigenetic regulation was nevertheless recognised first at the scale of DNA molecules. This concerns in particular the chromatin structure and the possible chemical modification of some DNA bases.

While these aspects of epigenetic regulation remain still largely unexplored by computational biologists, another level of complexity has in the last two decades emerged in the study of gene regulation. It is indeed now increasingly realised that, besides chromatin structure and DNA modification, the spatial arrangement of the chromosomes of a eukaryotic genome inside a cell is related to gene regulation, and possibly also to other important life processes. During interphase, chromosomes thus occupy distinct territories with preferred radial locations inside the nucleus, that is with preferential locations relative to the nuclear center. These preferred locations are cell type-specific and are conserved in the same cell type accross different primates. In the last decade, intra-chromosomal interaction or simple spatial proximity between genetic elements situated at often distant positions along the genome have started revealing their importance in gene expression.

HELIX has started addressing the issues of helping set up the experiments that will enable to infer full intra- and inter-chromosomal networks and to analyse them. This work is done in close collaboration with an experimental group headed by Ana Pombo, from the MRC at the Imperial College, London.

4. Software

4.1. ade4 and ade4TkGUI

Keywords: *environmetrics*, *graphics*, *multivariate*, *psychometrics*, *spatial*.

Participants: Jean Lobry, Jean Thioulouse [Correspondent].

Analysis of Ecological Data: Exploratory and Euclidean methods in Environmental sciences.
http://pbil.univ-lyon1.fr/ade4
http://pbil.univ-lyon1.fr/ade4TkGUI
Externals: Daniel. Chessel, Stéphane Dray, Anne-Béatrice Dufour, Sébastien Ollier, Sandrine Pavoine,Simon Penel

4.2. Alfacinha

Keywords: *sequence evolution*, *CpG effect*.

Participants: Laurent Guéguen, Leonor Palmeira [Correspondent].

Simulation of sequence evolution with neighbouring-site dependencies.
http://pbil.univ-lyon1.fr/software/alfacinha

4.3. BaobabLuna

Keywords: *reversal distance*.

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

Manipulation of signed permutations in the context of genomic evolution.

4.4. C3P

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

Participants: Anne Morgat, Alain Viari [Correspondent].

Merging two or more graphs representing biological data (e.g. pathways, ...).
http://www.inrialpes.fr/helix/people/viari/cccpart
External: Frédéric Boyer

4.5. DNA array analysis tool

Keywords: *DNA array analysis*.

Participant: Guy Perrière [correspondent].

New resampling strategy for the statistical analysis of DNA array data sets.
http://pulmogene.unibas.ch/articles/optimization
Externals: Michel Bihl, Desmond Higgins.

4.6. Ed'Nimbus

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

Participant: Marie-France Sagot [Correspondent].

Detecting and filtering repeats in sequences prior to multiple alignments.
http://igm.univ-mlv.fr/~peterlon/officiel/ednimbus/index.php
External: Pierre Peterlongo.

4.7. FamFetch

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

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

Set of tools to search for tree patterns in databases of phylogenetic trees.
http://pbil.univ-lyon1.fr/software/famfetch.html
Externals: Jean-François Dufayard, Simon Penel.

4.8. GeM

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

Participants: Christian Gautier [Correspondent], Bruno Spataro.

Database for comparative genomic analysis of complete vertebrate genomes.
http://pbil.univ-lyon1.fr/gem/gem_home.php Externals: Gisèle Bronner, Bruno Spataro.

4.9. GenoStar

Keywords: *bioinformatics environment*.

Participants: Anne Morgat, Alain Viari [Correspondent].

Integrated bioinformatics environment with data and knowledge management.
http://www-helix.inrialpes.fr/article121.html
Externals: François Rechenmann,Danielle Ziébelin

4.10. Herbs

Keywords: *annotation support*.

Participants: Anne Morgat, Alain Viari [Correspondent].

Reannotation of complete bacterial proteomes by consistency analyses.
http://www-helix.inrialpes.fr/article542.html Externals: Swiss Institute of Bioinformatics.

4.11. Hogenom and Hovergen

Keywords: *databases*, *genomes*.

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

Databases of homologous genes between fully-sequenced genomes.

http://pbil.univ-lyon1.fr/databases/hogenom.html
External: Simon Penel.

4.12. Hoppsigen

Keywords: *database*, *pseudogenes*.

Participant: Dominique Mouchiroud [Correspondent].

Nucleic database of homologous processed pseudogenes.
http://pbil.univ-lyon1.fr/databases/hoppsigen.html

4.13. HoSeqI

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

Participants: Manolo Gouy, Guy Perrière [correspondent].

Automated homologous sequence identification in gene family databases.
http://pbil.univ-lyon1.fr/software/HoSeqI
External: Anne-Muriel Arigon.

4.14. Identitag

Keywords: *SAGE*, *database*.

Participants: Laurent Duret, Dominique Mouchiroud.

Relational database for SAGE tag identification and interspecies comparison.
http://pbil.univ-lyon1.fr/software/identitag/
Externals: Céline Keime, Francesca Damiola, Olivier Gandrillon.

4.15. LalnView

Keywords: *local alignment*, *visualizer*.

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

Visualising local alignments between two proteic or nucleic sequences.
http://www-helix.inrialpes.fr/article124.html

4.16. MareyMap

Keywords: *recombination rate estimator*.

Participants: Laurent Guéguen [Correspondent], Gabriel Marais.

R package for meiotic recombination rate estimation.
http://pbil.univ-lyon1.fr/software/mareymap/
Externals: Delphine Charif.

4.17. Migal

Keywords: *RNA*, *tree comparison*.

Participant: Marie-France Sagot [Correspondent].

RNA structure comparison.
http://www-igm.univ-mlv.fr/~allali/logiciels/index..en.php
External: Julien Allali.

4.18. MotusWEB

Keywords: *coloured motif search and inference*.

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

Searching and inferring coloured motifs in metabolic networks.
http://pbil.univ-lyon1.fr/software/motus/
Externals: Odile Rogier, Fabien Jourdan.

4.19. Motus

Keywords: *coloured motif search and inference*.

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

Searching and inferring coloured motifs in undirected graphs.
http://genome.crg.es/~vlacroix/motus/
Externals: Odile Rogier, Fabien Jourdan.

4.20. Njplot

Keywords: *phylogenetic tree drawing*.

Participant: Manolo Gouy [correspondent].

Drawing phylogenetic trees. Updates allowing unresolved trees to be processed.
http://pbil.univ-lyon1.fr/software/njplot.html

4.21. OBIWarehouse - formerly MicrOBI

Keywords: *database*.

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

Integrated and synchronized heterogeneous public data on micro-organisms.
http://www.grenoble.prabi.fr/obiwarehouse

4.22. Oriloc

Keywords: *replication origin and terminus*.

Participant: Jean Lobry [Correspondent].

Prediction of putative origin and terminus of replication in prokaryotes.
http://pbil.univ-lyon1.fr/software/oriloc.html

4.23. PepLine

Keywords: *proteomic data analysis*.

Participant: Alain Viari [Correspondent].

Pipeline for the high-throughput analysis of proteomic data.
http://www.grenoble.prabi.fr/protehome/software/pepline
External: Jérôme Garin.

4.24. PhyloJava

Keywords: *phylogenetic reconstruction*.

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

Server for grid-powered phylogenetic reconstruction.
http://pbil.univ-lyon1.fr/software/phylojava/phylojava.html. PHYLOJAVA was developed also with

External: Timothée Sylvestre.

4.25. ProDom

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

Participants: Lauranne Duquenne, Daniel Kahn [correspondent].

Protein domain families automatically generated from protein databases.
http://prodom.prabi.fr/prodom/current/html/home.php
External: Aurélie Laugraud.

4.26. Priam

Keywords: *enzyme gene detection*, *metabolic inference*.

Participants: Thomas Bernard, Daniel Kahn [correspondent].

Sequence profiles generated from the ENZYME database.
http://priam.prabi.fr/
External: Clotilde Renard.

4.27. PSbR

Keywords: *perfect sorting by reversals*.

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

Testing for the evolution and conservation of common clusters of genes.
External: Yoan Diekmann.

4.28. Remote Acnuc Access

Keywords: *access to molecular databases*.

Participant: Manolo Gouy [correspondent].

Network protocol for remote access to PRABI molecular databases.
http://pbil.univ-lyon1.fr/databases/acnuc/remote_acnuc.html

4.29. Repseek

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

Participants: Eric Coissac [Correspondent], Alain Viari.

Finding approximate repeats in large DNA sequences.
External: Guillaume Achaz.

4.30. RFDD

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

Participant: Guy Perrière [correspondent].

Analysis of restriction fragments obtained by *in silico* digestion..
http://pbil.univ-lyon1.fr/software/RFDD/
External: Hélène Simonnet.

4.31. Sarment

Keywords: *sequence partitioning*.

Participant: Laurent Guéguen [correspondent].

HMM sequence partitioning and Maximal Predictive Partitioning.
http://pbil.univ-lyon1.fr/software/sarment/

4.32. SeaView

Keywords: *editor of multiple sequence alignments*.

Participant: Manolo Gouy [correspondent].

Editing multiple sequence alignments.
http://pbil.univ-lyon1.fr/software/seaview.html

4.33. SeqinR

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

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

R package for the analysis and management of biological sequences.
http://cran.univ-lyon1.fr/src/contrib/Descriptions/seqinr.html
Externals: Delphine Charif.

4.34. Smile and Riso

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

Participant: Marie-France Sagot [Correspondent].

Motif inference algorithms taking as input a set of biological sequences.
http://algos.inesc-id.pt/~asmc/software/riso.html
Externals: Laurent Marsan, Alexandra Carvalho.

4.35. SymBioCyc

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

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

Database of metabolic data dedicated to endosymbiotic organisms.
http://pbil.univ-lyon1.fr/software/symbiocyc.

4.36. UniPathway

Keywords: *database*, *metabolism*.

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

Database of manually curated pathways developped with the Swiss-Prot group..
http://pbil.univ-lyon1.fr/software/symbiocyc

5. New Results

5.1. Comparative genomics

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

The evolutionary process may create new gene families by assembling new genes from copies of pieces of various older genes, rapidly building new functions from a novel collection of already reliable parts. This process was termed "modular evolution" and many proteins are thus organised as a succession of modules called *domains*. Such protein modular evolution is the topic of the PhD of Anne-Sophie Sertier. HELIX is also involved in the development of InterPro (http://www.ebi.ac.uk/interpro), an integrated resource for protein families, domains and functional sites, which integrates different protein signature databases, including PRODOM, the database of protein domains established and maintained by Lauranne Duquenne with Daniel Kahn.

The resolution of the combinatorial assortments of protein sequences into domains is a prerequisite for protein sequence interpretation. However the recognition and clustering of homologous domains from sequence databases typically scales quadratically with respect to their size which grows exponentially, making it essential to parallelise the process. This has been instrumental in the construction of the PRODOM database of protein domain families. This was challenging because of the dependencies between program iterations, their extremely heterogeneous run times and communication bottlenecks that could arise because of the large size of the data [46]. This work is part of the PhD of Clément Rezvoy and was done in collaboration with Frédéric Vivien from the GRAAL team-project at the ENS, Lyon.

Despite a large agreement between ribosomal RNA and concatenated protein phylogenies, the phylogenetic tree of the bacterial domain remains uncertain in its deepest nodes. For instance, the position of the hyperthermophilic Aquificales is debated, as their commonly observed position close to Thermotogales may proceed from horizontal gene transfers, long branch attraction or compositional biases, and may not represent vertical descent. Indeed, another view, based on the analysis of rare genomic changes, places Aquificales close to epsilon-Proteobacteria. To get a whole genome view of the Aquifex relationships, all trees containing sequences from Aquifex in the HOGENOM database were surveyed. The bioinformatic study conducted by Bastien Bousseau as part pf his PhD revealed that Aquifex is most often found as a neighbour to Thermotogales [9]. Moreover, informational genes, which appeared to be less often transferred to the Aquifex lineage than non-informational genes, most often placed Aquificales close to Thermotogales.

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

5.1.2.1. Spatial organisation

There was good progress in the study of the impact of recombination on the emergence and evolution of genomic structures, notably isochores (which correspond to long regions of DNA with a relatively homogeneous base composition – homogeneous rate of G+C – and well delimited frontiers). Various mammalian genomes have been analysed and correlations have been identified between the isochores enriched in G+C and such biological properties as intron length and number, gene compactness and density, distribution of repeated sequences etc. Those correlations were not detected in some of the fishes analysed (namely, medaka and danio) or appear inverted. Six sequences of medaka from Japan have been recovered and compared. This enabled to demonstrate that the process of biased gene conversion (BCG) is not present in this species. A mathematical model to measure the impact of BCG on a population of infinite size was also built and simulations of such an impact were performed on populations of finite size. The conclusions of these preliminary models are that the BCG process could enable to keep in a population lethal recessive alleles.

In the context of his PhD, co-supervised between Grenoble and Lyon, Yves-Pol Deniélou has substantially extended the "common connected components" (CCC) approach initially developed by Frédéric Boyer during his PhD in HELIX to address the problem of looking for conserved gene locations across several bacterial species while allowing for some permutations in the gene order. Yves-Pol thus introduced a notion of quorum in the definition of conserved blocks: when looking for syntenies (blocks of similar genes whose relative location is conserved across species) between more than 2 genomes, some of the conserved genes may be missing in a few species. The resolution of this problem has required a complete re-design of the initial algorithm that could not be extended in a trivial way [50]. It was also extended to deal with any type of data, including of mixed origin, that may be modelled as a graph.

5.1.2.2. Dynamics

A methodological framework was established to estimate the chromosomal organisation of lost species whose descendants are known, that is, to reconstruct the ancestral genomes of vertebrates. This framework allowed to propose ancestral genomes for the boreoeutherians and for the amniotes, and takes into account duplicated genomes despite the difficulty in identifying the syntenies in this case [47].

The inference of ancestral genomes has motivated the computation of medians which consist, given the genomes of three or more species, to recreate the chromosomal organisation of their ancestors together with the evolutionary events that could explain their genomic differences. We thus progressed in the algorithmic study of medians by establishing the theoretical complexity of variants of the problem for the case where the genomes are multi-chromosomal such as in vertebrates, and possibly duplicated [48]. The computation of genomic medians intensively uses the computation of rearrangement distances. The complexity of the latter influences therefore greatly the efficiency of the methods for computing ancestral configurations. We made progress on the computation of rearrangement distances that preserve groups of genes co-localised between two genomes for a variant of the reversal distance called DCJ [44].

A method for detecting breakpoints in mammalian genomes was published this year as part of the PhD of Claire Lemaitre (defended in November 2008) [27]. Thanks to this method, we now have at our disposal much more precisely delimited breakpoint regions than were possible with previous approaches. This has allowed to study the correlations between such regions on the human genome and various other genomic structures, such as the organisation of replication domains, isochores and gene density along the genome.

This method was used also to study the evolution of the human chromosomes X and Y (the so-called sexual chromosomes). This enabled to identify inverted duplications at the borders of reversals on the Y chromosome, thereby reinforcing the hypothesis that the Y chromosome has diverged from the X by successive reversals that were responsible for the progressive stop of the recombination between these two chromosomes. This result is part of a larger-scale work on the analysis of all optimal scenarii of rearrangements between two chromosomes (or genomes) which uses a method of sorting by reversals developed as part of the PhD of Marília Braga (to be defended Jan. 30, 2009) [10].

This method which allows, for relatively small reversal distances (up to 20 approximately), to compute the reversal scenarii and represent them in a compact form, has been extended to take into account important biological constraints when the ancestral genomes are known or when the studied organisms are bacteria. In the first case, the preservation of groups of genes co-localised between two genomes considers not only these two but also all intermediate ancestors of the last common ancestor to the two genomes. In the second case, reversals that are symmetric to the start of replication only are considered (in bacteria; eukaryotes have more than one replication start). The method is currently been applied to study the evolution of the Rickettsias and Mycobacteria.

5.1.3. Motif search and inference

The work on filtering sequences previous to a multiple alignment or to the detection of long repeats in whole chromosomes or genomes that was in preparation last year was submitted with results that allow to identify, for instance, families of ALUs (degenerate repeats of approx. 300 bp) in a whole human chromosome in reasonable time with a very good selectiveness. This work will be extended into a multiple aligner and into a more general repeats detector in a collaboration with Pierre Peterlongo from the Symbiose team-project at the

INRIA Bretagne Atlantique, and with a long-term Italian collaborator, Nadia Pisanti, from the University of Pisa, Italy.

A collaboration was set up on the topic of RNA motifs detection with the group of Eric Westhof at the IBMC at Strasbourg to analyse a dataset of microRNAs (abbrev. into miRNAs) that were sequenced in the lab. The primary goal of that project was to predict the most likely origin of each of the miRNAs while also searching for novel candidates. It was an opportunity to test several criteria for miRNA precursor identification. This is now being extended to cover the detection of both the miRNAs and their targets after an extensive review of the algorithms currently available for either problem, and of their limits. This review has been submitted to *Nucleic Acids Research*. The new method developed by Nuno Mendes in his PhD, co-directed between HELIX and Ana Teresa Freitas from the Instituto Superior Técnico of Lisbon, will be applied to the bacterium *Buchnera*'s host, the aphid, in a collaboration with the BF2I lab at the INSA-Lyon, two of which members (Hubert Charles and Yvan Rahbé) will be members of BAMBOO.

HELIX is also involved in another RNA-motivated project which was funded by the ANR in 2006. and is being conducted in collaboration with Julien Allali, ex-PhD student with a member of HELIX and since 2006 Associate Professor at the LABRI, University of Bordeaux. Julien had during his PhD introduced a new data structure, called MIGAL for "Multiple Graph Layers", composed of various graphs linked together by relations of abstraction/refinement. The new structure has proved useful for representing information that can be described at different levels of abstraction, each level corresponding to a graph. An algorithm was proposed for comparing two MIGALS [6]. MIGAL was compared to other available software on benchmark datasets that we also helped to set up together with two other collaborators, Claude Thermes and Yves d'Aubenton from the Centre de Génétique Moléculaire (CGM) at Gif-sur-Yvette. The comparison was presented at the French Bioinformatics Conference this year and will now be written into a paper.

As part of his PhD (defended in March 2008), Paulo G. Fonseca was concerned with the identification of transcription regulation modules, *i.e.* groups of co-regulated genes and their regulators. One important distinction of this work in relation to what is available in the literature is that he proposes to identify modules that are evolutionarily conserved. A paper on one part of the method being developed was accepted this year [20] and another is in preparation. A stand-alone Java application with graphical user interface is also being developed as part of the work.

5.2. Functional genomics

5.2.1. Computational proteomics and transcriptomics

PepLine is a fully automated software which maps MS/MS fragmentation spectra of trypsic peptides to genomic DNA sequences [19]. The approach is based on Peptide Sequence Tags (PSTs) obtained from partial interpretation of QTOF MS/MS spectra. PSTs are then mapped on the six-frame translations of genomic sequences giving hits. Hits are clustered to detect potential coding regions. Our work aimed at optimising the algorithms of each component to allow the whole pipeline to proceed in a fully automated manner using raw nucleic acid sequences (*i.e.*, genomes that have not been "reduced" to a database of ORFs or putative exons sequences). Our results demonstrate that PepLine can compete with protein database searching softwares and is fast enough to potentially tackle large data sets and/or high size genomes.

The techniques used in molecular biology are evolving at an incredible pace and current microarray systems are generating hundreds of thousands of point per signal. Our objective is to develop non parametric methods to study such signals, wavelets being a model of predilection. Indeed, they allow to identify not only sudden but also much subtler changes in the signal. This work is being conducted in collaboration with the group of Anestis Antoniadis at the IMAG, Grenoble, and with Marie-Noelle Prioleau and Thierry Grange at the Institut Jacques Monod at Paris. In 2008, a paper was also submitted on mixed linear models for the joint segmentation of multivariate Gaussian processes.

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

The method for identifying sets of precursors mentioned in the previous report was fully formalised and extended to finding all mininal sets of precursors in a metabolic network. A paper was accepted at an international conference [45]. The work was done by Ludovic Cottret and is part of the PhD of Vicente Acuña (to be defended end of 2009) in collaboration with Alberto Marchetti-Spaccamela (University of Rome), Leen Stougie (Eindhoven University of Technology), Fábio Martinez and his student, Paulo Vieira Milreu from the Federal University of Campo Grande do Sul, Brazil, the two latter in the context of a STIC-AmSud project accepted in 2007 and involving teams from Brazil and Chile. It is the first work on this topic that provides an exact algorithm for identifying such minimal sets without eluding the difficult question of how to deal with the cycles in the hypergraphs that are used to model metabolic networks. As part of his work, Ludovic Cottret has also extensively used the algorithm to analyse the network of the endocytobiote *Buchnera aphidicola*, in particular to infer the precursors for the amino acids that are central to the metabolism of the bacterium and that the latter provides to its host, the aphid. It will be possible to check some of the discoveries made once the genome of the aphid is fully annotated and its metabolic network are available. HELIX has been intensively participing in both these processes as part of the postdoc of Augusto Vellozo and a paper is in preparation presenting the reconstructed network (soon available in a public database, APHIDCYC, that mirrors the SIMBIOCYC database of symbiont networks that Ludovic Cottret has established).

With the aim of defining the specificities of the metabolic network that were selected during the evolution of different live styles, HELIX has also been realising a topological comparison of the networks of 35 bacteria that live in wildly different conditions, from free living to various kinds of symbiotic relationships with their host (parasitism, mutualism, commensalism). We first computed the reactions and metabolites that are common to all 35 bacteria and found that there was only one single reaction and 31 metabolites in the intersection. Even taking possible annotation problems into account, such very low numbers are surprising and one could thus question the notion of a "minimal metabolism", in particular when one considers bacteria with a reduced metabolism that live in an obligate symbiosis with their host (either would die without the other). We have then tried to determine what properties of the networks could correlate with the life style of each bacterium. Various measures have been used for that. The results obtained will be written into a paper in early 2009.

Unicellular organisms that separate their somatic and germinal function into different nuclei, the paramecium presents characteristics that are at the same time simple and complex, making it a very interesting object of study. Three whole genome duplications have thus been detected and the question is what was the impact of such duplications on the metabolic network of the paramecium. As part of the postdoc of Amélie Véron, we studied how duplicate enzymatic genes are distributed along the metabolic network, whether there is a special pattern observed for those genes that have kept their duplicates. It seems that the answer is positive, namely, reactions catalysed by enzymatic genes that have retained their duplicates form statistically significant aggregates in the network. The results are being further analysed.

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, the complexity of the problem had never been fully analysed. This work was done by Vicente Acuña in collaboration with Alberto Marchetti-Spaccamela (University of Rome), Leen Stougie (Eindhoven University of Technology) and Vincent Lacroix (Center for Regulatory Genomics, Spain). A first paper indicated as submitted last year was accepted [5].

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 standardised representation of metabolic data in the UniProtKB/Swiss-Prot database. These metabolic data are explicitely represented and stored into a relational database (UniPathwayDB). In 2008, almost 300 more pathways were added to the database in relation to the end of 2007, representing around 300 more distinct biochemical reactions. The database is available through a web site hosted at the INRIA Rhône-Alpes (http://www.grenoble.prabi.fr/obiwarehouse/unipathway).

6. Contracts and Grants with Industry

6.1. Sanofi Pasteur

Participants: Edouard Blondeau, Alain Viari.

In 2008, HELIX concluded a contractual relation with Sanofi Pasteur (the vaccine division of the Sanofi Aventis group) located near Lyon. This collaboration was a follow-up of a previous contract on the (re)annotation and comparative analysis of pathogenic bacteria of interest to Sanofi Pasteur. It extended the analysis to expression data provided by Sanofi Pasteur and aimed at understanding some global regulation processes involved in pathogenicity.

7. Other Grants and Activities

7.1. National projects

Project name	Caractérisation et modélisation de la "fonction symbiotique" de *Buchnera aphidicola* chez le puceron du pois *Acyrthosiphon pisum*
Coordinator	H. Charles (INSA-INRIA Lyon)
HELIX participants	L. Cottret, V. Lacroix, M-F. Sagot
Type	AgroBI INRA (2006-2008)

Project name	GENOMICRO
Coordinator	L. Duret
HELIX participants	L. Duret, V. Daubin G. Marais, S. Mousset, E. Tannier, J. Lobry
Type	ANR jeunes chercheurs (2006-2008)

Project name	MetaGenoReg
Coordinator	D. Kahn
HELIX participants	M. Brilli, D. Kahn [correspondent]
Type	ANR BIOSYS (2006-2009)

Project name	ECONOMIC-RMQS
Coordinator	L. Ranjard
HELIX participants	J. Thioulouse
Type	ANR RMQS (2006-2009)

Project name	REGLIS
Coordinator	M-F. Sagot
HELIX participants	V. Acuña, M.D.V. Braga, L. Cottret, M. Deloger, C. Gautier, L. Guéguen, V. Lacroix, C. Lemaître, L. Palmeira, E. Prestat, M-F. Sagot, P.G.S. Fonseca, E. Tannier, P. Thébault, A. Viari
Type	ANR Blanc (2006-2008)

Project name	PATHOS-RMQS
Coordinator	S. Nazaret
HELIX participants	J. Thioulouse
Type	ANR RMQS (2007-2010)

Project name	MIRI
Coordinator	M-F. Sagot
HELIX participants	C. Gautier, V. Acuña, P. Simoẽs, A. Vellozo
Type	ANR Blanc (2008-2012)

Project name	CoGeBi
Coordinator	L. Duret
HELIX participants	C. Gautier, A. Popa, L. Duret, E. Tannier
Type	ANR Génomique Animale (2008-2011)

Project name	NeMo
Coordinator	S. Robin
HELIX participants	F. Picard, M-F. Sagot
Type	ANR Blanc (2008-2011)

Project name	PhylAriane
Coordinator	V. Berry (CNRS, LIRMM, Montpellier)
HELIX participants	V. Daubin, E. Tannier
Type	ANR (2008-2011)

Project name	Ecogenome
Coordinator	X. Nesme
HELIX participants	L. Guéguen, V. Daubin
Type	ANR (2008-2011)

7.2. European projects

Project name	IMPACT
Coordinator	S. Hunter
HELIX participants	T. Bernard, D. Kahn [correspondent]
Type	FP7 capacities: Scientific Data Repositories (2008-2010)

Project name	An integrated experimental-computational approach to modeling cellular networks and its application to analysing the LDB-based transcription complex
Coordinator	R. Sharan (Israel) and M.-F. Sagot
HELIX participants	Various members of Helix
Type	France-Israel project (2007-2008)

Project name	Evolutionary recurrent sequences in proteins: identification of protein domain families and their relationships in complete proteomes
Coordinators	M. Linial and D. Kahn
HELIX participants	D. Kahn, A.S. Sertier
Type	France-Israel project (2007-2008)

Project name	Reconstruction de génomes ancestraux
Coordinator	E. Tannier
HELIX participants	E. Tannier
Type	ECO-NET (2008-2009)

Project name	EMBRACE
Coordinator	G. Cameron
HELIX participants	L. Duquenne, D. Kahn [correspondent], A. Laugraud
Type	FP6 Network of excellence (2005-2010)

Project name	Chromonet
Coordinator	M-F. Sagot
HELIX participants	V. Acuña, L. Cottret, M. Deloger, C. Gautier, V. Lacroix, C. Lemaître, E. Prestat, M-F. Sagot, P.G.S Fonseca, E. Tannier, A. Véron, A. Viari
Type	ARC INRIA (2007-2008)

Project name	MetNet4SysBio
Coordinator	H. Charles
HELIX participants	V. Acuña, C. Gautier, D. Kahn, C. Pizzi, M-F. Sagot
Type	ANR Biosys (2007-2010)

7.3. International projects

Project name	Patagonia
Coordinator	M-F. Sagot
HELIX participants	V. Acuña, L. Cottret, M. Deloger, C. Gautier, V. Lacroix, C. Lemaître, E. Prestat, M-F. Sagot, P.G.S Fonseca, E. Tannier, A. Viari
Type	Stic AmSud INRIA (2007-2009)

8. Dissemination

8.1. Talks

Speaker	Title and Location	Date
B. Boussau	Parallel adaptations to high temperatures in the archean eon, Society of Molecular Biology and Evolution, Barcelona	June 7
B. Boussau	Parallel adaptations to high temperatures in the archean eon, Bayesian Phylogeny, Rényi Institute, Budapest	June 26
M.D.V. Braga	Exploring the solution space of sorting by reversals when analyzing genome rearrangements, Atelier de BioInformatique, Paris	May 13
M.D.V. Braga	Exploring the solution space of sorting by reversals when analyzing genome rearrangements, Faculty of Technology, University of Bielefeld, Germany	July 14
L. Duret	Analyse comparative des introns et autres régions non-codantes dans les génomes eucaryotes : mais où est donc passée la sélection naturelle ?, IBMC, Strasbourg	January 11
L. Duret	Evolution des séquences dans les points-chauds de recombinaison: le talon d'Achille de notre génome, LEGS, Gif-sur-Yvette	January 18
L. Duret	The impact of recombination hotspots on sequence evolution: the Achille's heel of our genome, Université de Genève	February 12
L. Duret	Recombination hotspots: the Achille's heel of mammalian genomes, St Radboud University, Nijmegen, The Netherlands	April 15
L. Duret	The mystery of intron splicing, JOBIM, Lille	June 30
L. Duret	The mystery of intron splicing, RECOMB Comparative Genomics, Paris	October 13-15
L. Duret	The mystery of intron splicing, Center for Integrative Genomics, Lausanne	October 20
P.G.S. Fonseca	Efficient p-value computation for high order Markov motifs, Valparaiso, Chili	March 27
M. Gouy	Adaptations parallèles aux hautes températures pendant l'archéen, Centre de Recherche Pétrographique et Géochimique, Nancy	March 27
D. Kahn	Metabolic Control Theory and the analysis of biological regulation, Lille Spring School on Modelling Complex Biological Systems in the Context of Genomics, Villeneuve d'Ascq	April 7-11
D. Kahn	Organisation of the EMBRACE Advanced Protein Domain Analysis Training Workshop, Lyon	May 18-20
D. Kahn	Phylogenomics of modular proteins: relative importance of domain shuffling vs. domain innovation, SMBE 2008, Barcelona, Spain	June 5-8
D. Kahn	Calcul intensif en bio-informatique et génomique, IBM Campus Day High Performance Computing : nouvelles perspectives pour la Recherche, Lyon	November 13
D. Kahn	Parallel large scale inference of protein domain families, 14th Intl Conference on Parallel and Distributed Systems, Melbourne.	December 8-10

Speaker	Title and Location	Date
C. Lemaitre	Precise detection of rearrangement breakpoints in mammalian chromosomes, Rencontres Alphy, Lyon, France	January 2008
C. Lemaitre	A method to detect precisely rearrangement breakpoints in mammalian genomes, Dynamics of genomes, workshop à Valparaiso, Chili	March 2008
C. Lemaitre	Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes, IPG 2008, Lyon	November 21
G. Marais	The evolution of dioecy and sex chromosomes in Silene, International workshop “Silene : from populations to genes”, Monte Verita, Ascona, Switzerland	May 2008
G. Marais	Mutation rate and genome reduction in endosymbiotic and free-living bacteria, SMBE, Barcelona, Spain	June 2008
G. Marais	Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes, GDR 1928 “Génomique des populations”, Sètes	October 2008
A. Necsulea	Is there evidence for a replication-related genome organization in human?, GDRE, Barcelona, Spain	October 23
L. Palmeira.	Sequence effects on nucleosome positioning and chromatin structure, Seminar at the Laboratoire Arago, Banyuls	June 2008
E. Prestat	Modeling pharmacogenomics data with Bayesian Networks, Workshop on Modeling of Genetic Regulatory and Metabolic Networks, Valparaiso, Chili	March 28
E. Prestat	Using bayesian networks in cancer research, International conference on system science in health care (ICSSHC 2008), ENS-Lyon	September 6
E. Tannier	Reconstruction de génomes ancestraux: les deux parcimonies, Laboratoire LIAFA, Paris	January 22
E. Tannier	Predicting the past of chromosomes, Université Libre de Bruxelles	April 21
E. Tannier	Predicting the past of chromosomes, Conférence MIEP, Montpellier	May 10
E. Tannier	Prédictions de synténies dans le génome ancestral des amniotes, Jobim, Lille	June 30
E. Tannier	Cassures évolutives et organisation des génomes de mammifères, Séminaire Cytogénomique, Toulouse	October 9
E. Tannier	Cassures évolutives et réplications dans les génomes de mammifères, GTGC, satellite JOBIM, Lille	July 3

8.2. Editorial and reviewing activities

Laurent Duret

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

Manolo Gouy

Type	Journal or Conference
Editorial Board	*Molecular Biology and Evolution*

Daniel Kahn

Type	Journal or Conference
Editorial Board	*Biology Direct*
Faculty Member	Faculty of 1000

Marie-France Sagot

Type	Journal or conference
Member Steering Committee	European Conference on Computational Biology (ECCB) International Symposium on Bioinformatics Research and Applications (ISBRA)
Editor-in-Chief	*IEEE/ACM Transactions on Computational Biology and Bioinformatics*, starting in Jan. 2009
Editorial Board	*Journal of Discrete Algorithms* *Lecture Notes in BioInformatics* *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, until Dec 2008 (see above) *BMC Algorithms for Molecular Biology* *BMC Bioinformatics*
Member Program Committee	ECCB (steering), ISMB, PSW, RECOMB Satellite Conf. on Regulatory Genomics, RECOMB Satellite Conf. on Comp. Genomics, APBC, IEEE BIBM, ISBRA

Jean Thioulouse

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

8.3. Administrative activities

L. Duret is member of the scientific committee of the “programme fédérateur INRA de biologie intégrative animale, végétale et microbienne (agroBI)”, of the section 22 and of the interdisciplinary commission 43 of the CoNRS.

C. Gautier was until September 2008 chair of the section 29 of the CoNRS, and is director of the PRABI.

M. Gouy is member of the “Comité National des Universités”, section 67 (Ecology & Evolution), of the selection committee of the CNRS ATIP Biodiversity, and of the Scientific Advisory Board of the Swiss Institute of Bioinformatics.

D. Kahn is member of the Scientific Committee of the French National Sequencing Centre (Génoscope, Evry) and represents the INRIA at the Scientific Board of Lyon 1 University.

D. Mouchiroud is director of the LBBE (UCBL, UMR 5558) and is member of the section 29 of the CoNRS.

G. Perrière is President of the "Société Française de Bioinformatique".

Marie-France Sagot is a member of the section 01 and of the interdisciplinary commission 43 of the CoNRS, as well as of the scientific committee of the course "Informatique en Biologie" of the Institut Pasteur in Paris. She participated in the reviewing process of candidates for a research position and of projects for the University of Uppsala in Sweden and for the Swedish Research Council.

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

8.4. Teaching

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

Various members of the project have developed over the years courses in biometry, bioinformatics and evolutionary biology at all levels of the University as well as at the "École Normale Supérieure" (ENS) of Lyon and the INSA ("Institut National de Sciences Appliquées").

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). It leads to Master diplomas in the scientific and medical fields.

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 (37h), Manolo Gouy (26h), Daniel Kahn (6h), Guy Perrière (37h), Franck Picard (8h), Marie-France Sagot (24h), Eric Tannier (16h).

9. Bibliography

Year Publications

Doctoral Dissertations and Habilitation Theses

[1] E. BILLOIR. *Modélisation dynamique et inférence bayésienne pour l'analyse des données en écotoxicologie*, Ph. D. Thesis, Université Claude Bernard Lyon 1, 2008.

[2] B. BOUSSAU. *Evolution Profonde et Phylogénie*, Ph. D. Thesis, Université Claude Bernard Lyon 1, 2008.

[3] C. LEMAITRE. *Réarrangements chromosomiques dans les génomes de mammifères : caractérisation des points de cassure*, Ph. D. Thesis, Université Claude Bernard Lyon 1, 2008.

[4] A. NECSULEA. *Etude des patrons d'évolution asymétrique dans les séquences d'ADN*, Ph. D. Thesis, Université Claude Bernard Lyon 1, 2008.

Articles in International Peer-Reviewed Journal

[5] V. ACUÑA, F. CHIERICHETTI, V. LACROIX, A. MARCHETTI-SPACCAMELA, M.-F. SAGOT, L. STOUGIE. *Modes and cuts in metabolic networks: Complexity and algorithms*, in "Biosystems", vol. in press, 2008.

[6] J. ALLALI, M.-F. SAGOT. *A multiple layer model to compare RNA secondary structures*, in "Software Practice and Experience", vol. 38, 2008, p. 775–792.

[7] A.-M. ARIGON, G. PERRIÈRE, M. GOUY. *Automatic identification of large collections of protein-coding or rRNA sequences*, in "Biochimie", L1, vol. 90, n^{o} 4, 2008, p. 609-614.

[8] B.-A. BENAYOUN, S. CABURET, A. DIPIETROMARIA, M. BAILLY-BECHET, F. BATISTA, M. FELLOUS, D. VAIMAN, R.-A. VEITIA. *The identification and characterization of a FOXL2 response element provides insights into the pathogenesis of mutant alleles*, in "Human Molecular Genetics", PL1, it, vol. 17, n^{o} 20, 2008, p. 3118-3127.

[9] B. BOUSSAU, L. GUEGUEN, M. GOUY. *Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria*, in "BMC Evolutionary Biology", L1, vol. 8, n^{o} 272, 2008.

[10] M. BRAGA, M.-F. SAGOT, C. SCORNAVACCA, E. TANNIER. *Exploring the Solution Space of Sorting by Reversals, with Experiments and an Application to Evolution*, in "Transactions on computational biology and bioinformatics", L1, Web of science, vol. 5 - nÂ°3, 2008, p. 348-356.

[11] C. BROCHIER-ARMANET, B. BOUSSAU, S. GRIBALDO, P. FORTERRE. *Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota*, in "Nature Reviews Microbiology", L1, vol. 6, 2008, p. 245-252.

[12] J.-C. CADORET, F. MEISCH, V. HASSAN-ZADEH, I. LUYTEN, C. GUILLET, L. DURET, Q. QUESNEVILLE, M.-N. PRIOLEAU. *Genome-wide studies highlight indirect links between human replication origins and gene regulation*, in "Pnas", L1, vol. 105-41, 2008, p. 15837-15842.

[13] C. CHAUVE, E. TANNIER. *A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes*, in "PLoS Computational Biology", in press, 2008.

[14] Y. COUTÉ, K. KINDBEITER, S. BELIN, R. DIECKMANN, L. DURET, L. BEZIN, J.-C. SANCHEZ, J.-J. DIAZ. *ISG20L2, a novel vertebrate nucleolar exoribonuclease involved in ribosome biogenesis*, in "Molecular and cellular proteomics", L1, vol. 7, n^{o} 3, 2008, p. 546-559.

[15] J.-J. DAUDIN, F. PICARD, S. ROBIN. *A mixture model for random graphs*, in "Statistics and Computing", PL1, vol. 18, n^{o} 2, 2008, p. 173-183.

[16] L. DURET, P. ARNDT. *The Impact of Recombination on Nucleotide Substitutions in the Human Genome*, in "PloS Genetics", L1, de, vol. 4, n^{o} 5, 2008, p. 1-19.

[17] L. DURET, J. COHEN, C. JUBIN, P. DESSEN, J.-F. GOÛT, S. MOUSSET, J.-M. AURY, O. JAILLON, B. NOËL, O. ARNAIZ, M. BÉTERMIER, P. WINCKER, E. MEYER, L. SPERLING. *Analysis of sequence*

variability in the macronuclear DNA of Paramecium tetraurelia: A somatic view of the germline, in "Genome Research", L1, vol. 18, n^{o} 4, 2008, p. 585-596.

[18] J. DUTHEIL, B. BOUSSAU. *Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs*, in "BMC Evolutionary Biology", vol. 8:255, 2008.

[19] M. FERRO, M. TARDIF, E. REGUER, R. CAHUZAC, T. BRULEY, E. NUGUES, M. VIGOUROUX, Y. VANDENBROUCK, J. GARIN, A. VIARI. *PepLine: a software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences.*, in "J Proteome Res.", vol. 7, n^{o} 5, 2008, p. 1873–1883.

[20] P. G. S. FONSECA, C. GAUTIER, K. GUIMARÃES, M.-F. SAGOT. *Efficient representation and p-value computation for high order Markov motifs*, in "Bioinformatics", vol. 24, 2008, p. i160–i166.

[21] M. GOUY, M. CHAUSSIDON. *Ancient bacteria liked it hot*, in "Nature", L1, vol. 451, 2008, p. 635-636.

[22] M. GOUY, S. DELMOTTE. *Remote access to ACNUC nucleotide and protein sequence databases at PBIL*, in "Biochimie", L1, vol. 90, 2008, p. 555-562.

[23] T. HANDORF, N. CHRISTIAN, O. EBENHÖH, D. KAHN. *An environmental perspective on metabolism*, in "Journal of Theoretical Biology", L1, de, vol. 252, 2008, p. 530-537.

[24] O. JAILLON, K. BOUHOUCHE, J.-F. GOÛT, J.-M. AURY, B. NOËL, B. SAUDEMONT, M. NOWACKI, V. SERRANO, B. PORCEL, B. SÉGURENS, A. LE MOUAL, G. LEPÈRE, V. SCHAECHTER, M. BÉTERMIER, J. COHEN, P. WINCKER, L. SPERLING, L. DURET, E. MEYER. *Translational control of intron splicing in eukaryotes*, in "Nature", L1, vol. 451, n^{o} 7176, 2008, p. 359-362.

[25] V. LACROIX, L. COTTRET, P. THÉBAULT, M.-F. SAGOT. *An introduction to metabolic networks and their structural analysis*, in "IEEE-ACM Transactions in Computational Biology and Bioinformatics", vol. in press, 2008.

[26] C. LEMAITRE, M.-F. SAGOT. *A small trip in the untranquil world of genomes a survey on the detection and analysis of genome rearrangement breakpoints*, in "Theoretical Computer Science", vol. 395, n^{o} 2-3, 2008, p. 171-192.

[27] C. LEMAITRE, E. TANNIER, C. GAUTIER, M.-F. SAGOT. *Precise detection of rearrangement breakpoints in mammalian chromosomes*, in "BMC Bioinformatics", L1, vol. 9, 2008, p. 286-322.

[28] P. LIO, M. BRILLI, R. FANI. *Topological Metrics in Blast Data Mining: Plasmid and Nitrogen-Fixing Proteins Case Studies*, in "BIRD", L1, uk, it, vol. 13, 2008, p. 207-220.

[29] G. MARAIS, M. NICOLAS, R. BERGERO, P. CHAMBRIER, E. KEJNOVSKY, F. MONÉGER, R. HOBZA, A. WIDMER, D. CHARLESWORTH. *Evidence for Degeneration of the Y Chromosome in the Dioecious Plant Silene latifolia*, in "Current biology", L1, vol. 18, 2008, p. 545-549.

[30] S. MORISSET, F. PEYRON, J. LOBRY, J. GARWEG, J. FERRANDIZ, K. MUSSET, J. GOMEZ-MARIN, A. DE LA TORRE, M. DEMAR, B. CARME, C. MERCIER, J.-F. GARIN, M. CESBRON-DELAUW. *Serotyping of*

Toxoplasma gondii: striking homogeneous pattern between symptomatic and asymptomatic infections within Europe and South America, in "Microbes and Infection", L1, ch, co, vol. 10, n^{o} 7, 2008, p. 742-747.

[31] N. MUGNIER, L. GUEGUEN, C. VIEIRA, C. BIÉMONT. *The heterochromatic copies of the LTR retrotransposons as a record of the genomic events that have shaped the Drosophila melanogaster genome*, in "Gene", L1, vol. 411, n^{o} 1-2, 2008, p. 87-93.

[32] V. NAVRATIL, S. PENEL, S. DELMOTTE, D. MOUCHIROUD, C. GAUTIER, A. AOUACHERIA. *DigiPINS: A database for vertebrate exonic single nucleotide polymorphisms and its application to cancer association studies*, in "Biochimie", L1, vol. 90, 2008, p. 563-569.

[33] G. PERRIÈRE. *Bioinformatics in the complete genome sequence era*, in "Biochimie", L1, vol. 90, n^{o} 4, 2008, p. 553-554.

[34] P. PETERLONGO, J. ALLALI, M.-F. SAGOT. *Indexing gapped-factors using a tree*, in "International Journal of Foundations of Computer Science", vol. 19, 2008, p. 71–88.

[35] P. PETERLONGO, N. PISANTI, F. BOYER, A. P. DO LAGO, M.-F. SAGOT. *Lossless filter for multiple repetitions*, in "Journal of Discrete Algorithms", in press, vol. 6, 2008, p. 124–135.

[36] F. PICARD, J.-J. DAUDIN, M. KOSKAS, S. SCHBATH, S. ROBIN. *Assessing the exceptionality of network motifs*, in "Journal of Computational Biology", PL1, vol. 15, n^{o} 1, 2008, p. 1-20.

[37] P. REMIGI, A. FAYE, A. KANE, M. DERUAZ, J. THIOULOUSE, M. CISSOKO, Y. PRIN, A. GALIANA, B. DREYFUS, R. DUPONNOIS. *The exotic legume tree species Acacia holosericea alters microbial soil functionalities and the structure of the arbuscular mycorrhizal community*, in "Applied and Environmental Microbiology", L1, sn, vol. 74, n^{o} 5, 2008, p. 1485-1493.

[38] C. SANTOS, F. TAVARES, J. THIOULOUSE, P. NORMAND. *A phylogenomic analysis of bacterial helix-turn-helix transcription factors*, in "FEMS Microbiology Reviews", in press, 2008.

[39] S. SCHBATH, V. LACROIX, M.-F. SAGOT. *Assessing the exceptionality of coloured motifs in networks*, in "EURASIP Journal on Bioinformatics and Systems Biology", vol. in press, 2008.

[40] A. STUDER, S. PENEL, L. DURET, M. ROBINSON-RECHAVI. *Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes*, in "Genome Research", L1, CH, vol. 18, n^{o} doi:10.1101/gr.076992.108, 2008, p. 1393-1402.

[41] N. VACHIERY, G. MAGANGA, T. LEFRANÇOIS, Y. KANDASSAMY, F. STACHURSKI, H. ADAKAL, C. FERRAZ, A. MORGAT, A. BENSAID, E. COISSAC, F. BOYER, J. DEMAILLE, A. VIARI, D. MARTINEZ, R. FRUTOS. *Differential strain-specific diagnosis of the heartwater agent: Ehrlichia ruminantium.*, in "Infect Genet Evol.", vol. 8, n^{o} 4, 2008, p. 459–466.

[42] C. VALIENTE MORO, J. THIOULOUSE, C. CHAUVE, P. NORMAND, L. ZENNER. *Bacterial taxa associated with the hematophagous mite, Dermanyssus gallinae detected by 16S rDNA PCR amplification and TTGE fingerprint*, in "Research in Microbiology", vol. in press, 2008.

International Peer-Reviewed Conference/Proceedings

[43] S. ADI, M. BRAGA, C. FERNANDES, C. FERREIRA, F. MARTINEZ, M.-F. SAGOT, M. STEFANES, C. TJANDRAATMADJA, Y. WAKABAYASHI. *Repetition-free longest common subsequence*, in "IV Latin-American Algorithms, Graphs and Optimization Symposium, Puerto Varas, CHILE", L1, vol. 30, Electronic Notes in Discrete Mathematics, 2008, p. 243-248.

[44] S. BÉRARD, A. CHATEAU, C. CHAUVE, C. PAUL, E. TANNIER. *Perfect DCJ Rearrangement*, in "RECOMB-CG 2008", S. VIALETTE, C. NELSON (editors), Lecture Notes in Bioinformatics, L1, ca, vol. 5267, Springer-Verlag, 2008.

[45] L. COTTRET, P.-V. MILREU, V. ACUÑA, A. MARCHETTI-SPACCAMELA, F. VIDUANI-MARTINEZ, M.-F. SAGOT, L. STOUGIE. *Enumerating Precursor Sets of Target Metabolites in a Metabolic Network*, in "WABI 2008", J. LAGERGREN, K. CRANDALL (editors), L1, It, Br, Nl, vol. LNBI 5251, Springer-Verlag Berlin Heidelberg 2008, 2008, p. 233-244.

[46] D. KAHN, C. REZVOY, F. VIVIEN. *Parallel large scale inference of protein domain families*, in "Proceedings of the 14th IEEE International Conference on Parallel and Distributed Systems (ICPADS'08)", 2008.

[47] R. LENNE, C. SOLNON, T. STÜTZLE, E. TANNIER, M. BIRATTARI. *Reactive Stochastic Local Search Algorithms for the Genomic Median Problem*, in "EvoCOP 2008, Berlin", J. VAN HEMERT, C. COTTA (editors), Lecture Notes in Computer Science, L1, vol. 4972, Springer Verlag, 2008, p. 266-276.

[48] E. TANNIER, C. ZHENG, D. SANKOFF. *Multichromosomal Genome Median and Halving Problems*, in "Proceedings of WABI'08", J. LAGERGREN, K. CRANDALL (editors), Lecture notes in Bioinformatics, L1, ca, vol. 5251, Springer-Verlag, 2008, p. 1-13.

National Peer-Reviewed Conference/Proceedings

[49] F. BOYER, C. CHAUVE, E. TANNIER. *Prédiction de synténies dans le génome ancestral des amniotes*, in "Actes des Journée Ouvertes de Biologie, Informatique et Mathématiques, JOBIM 2008", J. VAN HELDEN, Y. MOREAU (editors), L1, ca, 2008.

[50] Y.-P. DENIÉLOU, F. BOYER, M.-F. SAGOT, A. VIARI. *Recovering isofunctional genes: a synteny-based approach*, in "Actes des Journée Ouvertes de Biologie, Informatique et Mathématiques, JOBIM 2008", J. VAN HELDEN, Y. MOREAU (editors), L1, ca, 2008.

Scientific Books (or Scientific Book chapters)

[51] E. TANNIER. *Sorting Signed Permutations by Reversal (Reversal Sequence) 2004; Tannier, Sagot*, in "Encyclopedia of algorithms", M.-Y. KAO (editor), L1, Springer, 2008, p. 860-863.