INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Symbiose

SYstèmes et Modèles BIOlogiques, BIOinformatique et SEquences

Rennes - Bretagne-Atlantique

THEME BIO

Table of contents

1. Team **1**
2. Overall Objectives **2**
2.1. A Bioinformatics center 2
2.2. Scientific axes 2
2.2.1. Optimized algorithms on parallel specialized architectures 2
2.2.2. Modeling sequence/structure relationships 3
2.2.3. Systems biology: network modeling and analysis 3
2.3. Highlights of the year 3
3. Scientific Foundations **3**
3.1. A short introduction to bioinformatics 3
3.2. Optimized algorithms on parallel specialized architectures 4
3.3. Syntactical Analysis of sequences 5
3.3.1. Formal Languages and biological sequences 5
3.3.2. Machine Learning : from Pattern Discovery to Grammatical Inference 5
3.4. Systems biology: network modeling and analysis 6
4. Application Domains **8**
5. Software **8**
5.1. Introduction 8
5.2. GenOuest, the Bioinformatics computing center of Ouest-genopole 8
5.2.1. Bioinformatics services hosting 9
5.2.2. Specific advanced developments 9
5.3. New activities of the bioinformatics platform in 2008 10
5.3.1. National Project BioMAJ (BIOlogie Mise A Jour) 10
5.3.2. National Project BioWorkFlow 10
5.4. Activity of transfer from Symbiose to GenOuest 10
6. New Results **11**
6.1. Optimized Algorithms on Parallel Specialized Architectures 11
6.1.1. Dedicated Architectures for genomic comparisons 11
6.1.2. Optimal indexing for protein similarity search 12
6.1.3. Detection of new microRNA in the A. Pisum genome 13
6.1.4. ACGT OncoSimulator 13
6.1.5. Development of Optimization techniques : the knapsack problem 13
6.2. Modelling sequence/structure relationships 14
6.2.1. Finding modules in sequences 14
6.2.2. Logical grammars 15
6.2.3. Learning automata and grammars on biological sequences 15
6.2.4. Protein structure alignment: algorithms and applications 15
6.3. Systems biology: analysing data and modeling interactions 17
6.3.1. Construction and analysis of signalling and metabolic pathways 17
6.3.2. Modularity, hierarchical models, and robustness of complex biological systems 18
6.3.3. Algorithms for the analysis of large-scale models 19
6.4. Some statistical developpments 20
6.4.1. Modeling synthetic lethality 20
6.4.2. Constraint-based mining of SAGE data 21
6.4.3. Quality of association rules in Data Mining 21
7. Other Grants and Activities **22**
7.1. Regional initiatives 22
7.1.1. OUEST-genopole 22
7.1.2. Regional cooperation 22

7.2. National initiatives 22
7.2.1. National projects of the GenOuest platform 23
7.2.1.1. BioSide 23
7.2.1.2. BioWorkFlow 23
7.2.1.3. BioMAJ 23
7.2.1.4. GRISBI 23
7.2.1.5. Sigenae and Genanimal 24
7.2.2. ANR Projects 24
7.2.2.1. PFTV 2007 24
7.2.2.2. Proteus (Fold recognition and inverse folding problem) 24
7.2.2.3. PARA: Parallelism and Improvement of Application Performances 24
7.2.2.4. Modulome: Identifying and displaying modules in genomic sequences 24
7.2.2.5. Sitcon: Modeling signal transduction induced by a chimeric oncogene 25
7.2.2.6. DyCoNum: Dynamical and Combinatorial studies of Numeration systems 25
7.3. European and international initiatives 25
7.3.1. International programs 25
7.3.1.1. Integrated Project ACGT 25
7.3.1.2. Modularity, Abstraction, Robustness of Network Models in Molecular Biology 25
7.3.1.3. Application of combinatorial optimization (PHC RILA, Bulgaria) 26
7.3.1.4. SECyT-INRIA-CNRS cooperation program (Argentina) 26
7.3.1.5. PHC Sakura and Amadeus 26
7.3.2. Visiting scientists 26
7.3.2.1. Visitors 26
7.3.2.2. Scientific visit exchanges 27
8. Dissemination **27**
8.1. Leadership within scientific community 27
8.1.1. Administrative functions: scientific commitees, journal bords, jury 27
8.1.2. Jury of PhD Theses 27
8.2. Faculty teaching 27
8.3. Conference and workshop committees, invited conferences 28
8.3.1. International invited conferences 28
8.3.2. National invited conferences 28
8.3.3. Conference committees 28
8.3.4. ICGI: 9th International Colloquium on Grammatical Inference 29
8.3.5. Bioinformatics animations 29
8.3.6. BioInfoOuest thematic-day conferences 29
8.4. Theses defenses 30
8.4.1. PhD thesis, Goulven Kerbellec 30
8.4.2. Habilitation thesis, Jacques Nicolas 30
8.4.3. Habilitation thesis, Anne Siegel 30
9. Bibliography **30**

The Symbiose project has been created in 2002. Its general purpose concerns bioinformatics, that is, modeling and analysis of large scale genomic and post-genomic data. Our goal is to assist the molecular biologist for the formulation and discovery of new biological knowledge from the information gained through public data banks and experimental data. This project is thus clearly application-oriented and combines multiple research fields in computer science towards this goal.

1. Team

Research Scientist

Jacques Nicolas [Team Leader, Research director, Inria, HdR]
Jérémie Bourdon [Inria on leave from Univ. Nantes]
François Coste [Resarch scientist, Inria]
Dominique Lavenier [Research director, Cnrs on leave at ENS Cachan Bretagne since Oct. 2008, HdR]
Pierre Peterlongo [post-doc followed by a research scientist position in Oct 2008, Inria]
Anne Siegel [Research scientist, Cnrs, HdR]

Faculty Member

Rumen Andonov [Professor, Univ. Rennes 1, HdR]
Catherine Belleannée [Associate Professor, Univ. Rennes 1]
Michel Le Borgne [Associate Professor, Univ. Rennes 1]
Israël-César Lerman [Emeritus Professor, Univ. Rennes 1, HdR]
Basavanneppa Tallur [Associate Professor, Univ. Rennes 1, HdR]
Raoul Vorc'h [Associate Professor, Univ. Rennes 1]

External Collaborator

Ovidiu Radulescu [Associate Professor, IRMAR, Univ. Rennes 1, HdR]
Nathalie Theret [Research director, INSERM, Rennes, HdR]

Technical Staff

Olivier Collin [GENOUEST, permanent Research Engineer, Cnrs]
Hugues Leroy [GENOUEST, permanent Research Engineer, Inria]
Olivier Sallou [GENOUEST, permanent Research Engineer, Unv. Rennes 1]
Anthony Assi [GENOUEST, Engineer, until september 2008, Inria contract genopole]
Annabel Bourdé [GENOUEST, Engineer,Inria contract genopole]
Olivier Filangi [GENOUEST, Engineer, until june 2008, Inria contract genopole]
Laetitia Guillot [GENOUEST, Engineer, until april 2008, Inria contract genopole]
Sophie Roucan [GENOUEST, Engineer, until april 2008, Inria contract genopole]
Fabrice Legeai [permanent Engineer, INRA, 20% time dedicated to the symbiose project]
François Moreews [permanent Engineer, INRA, 20% time dedicated to the symbiose project]
Alexandre Cornu [Engineer national ANR contract Para]
Julien Jacques [Engineer European project ACGT]
Christine Rousseau [Engineer, ANR contract Modulome]

PhD Student

Pierre Blavy [INRA]
Rayan Chikhi [MENRT/ENS]
Guillaume Collet [MENRT]
Matthias Gallé [Inria/CORDI]
Jérémy Gruel [Inserm/Région]
Serge Guelton [MENRT Univ. de Bretagne Occidentale]
Carito Guziolowski [Conicyt/Ambassade de France/Inria]
Thibaut Henin [ENS/MENRT]
Goulven Kerbellec [Inria until March 2008]
Noël Malod-Dognin [Inria/Region]

Van Hoa Nguyen [Inria/CORDI]
Guillaume Rizk [MENRT]

Post-Doctoral Fellow
Nolwenn Le Meur [Inserm/La Ligue]
Sylvain Blachon [Inria, Sitcon ANR project]
Guillaume Launay [ANR Proteus]

Visiting Scientist
Nicola Yanev [2 months, project RILA and ANR Proteus]
Yavor Vutor [2 months, project RILA and ANR Proteus]

Administrative Assistant
Marie-Noëlle Georgeault [Assistant, Inria]

2. Overall Objectives

2.1. A Bioinformatics center

Bioinformatics has a quite large acceptation. In our framework, it focuses on methodological research at the interface between computer science and molecular biology, excluding all "standard" informatics (“biocomputing”) for routine management of biological data. However, it is hard to achieve in depth research in this domain without participating to biology-oriented developments. In order to favor cooperative studies with biological labs we have decided to create a Bioinformatics Center, with a research team, **Symbiose**, leaning back against a bioinformatics platform **GenOuest** (or the converse...). This report is mainly focused on the research project. Our research specificities include our interest in **large scale studies** (genomes, proteomes or regulation networks) and **discrete methods** necessary to handle the associated complexity. Our methods relate on discrete optimization, analysis of systems of qualitative equations and formal language modeling. Our goal is to push forward their range of applicability by exploring the impact of **specialized machines or algorithms**.
The bioinformatics platform GenOuest acts as a facility and software tool provider for the analysis of genomic data generated by numerous laboratories (55) of OUEST-genopole®. The platform provides at first computing power but also a comprehensive list of software dedicated to sequence analysis. On a national level, the platform is developing an expertise in the field of pattern matching and pattern discovery tool. It received a national RIO label in 2003 and 2006 and is supported by national and regional contracts. The platform is a mediator between computer science and biological labs. This leads to consulting, partnership and transfer actions. This activity is described in the section 5.1.

2.2. Scientific axes

The three *Scientific axes* on which the project focuses derive from our choice of modeling complex biological systems in a discrete framework, while managing efficiency issues. We have a global concern for high performance computing and two types of modeling tasks:

2.2.1. Optimized algorithms on parallel specialized architectures

First and foremost, large scale studies need a fine tuning and management of computational resources. We investigate the practical usage of parallelism to speed up computations in genomics. Topics of interest range from intensive sequence comparisons to pattern or model matching, including structure prediction. We work on the co design of algorithms and hardware architectures tailored to the treatment of such applications. It is based on the study of reconfigurable machines employing Field Programmable Logical Arrays (FPGA) or fast components such as Flash memories or Graphical Processing Units.

2.2.2. Modeling sequence/structure relationships

This track concerns the search for relevant (e. g. functional) spatial or logical structures in macromolecules, either with intent to model specific spatial structures (secondary and tertiary structures, disulfide bounds ...) or general biological mechanisms (transposition ...). In the framework of **language theory and combinatorial optimization**, we address various types of problems: the design of grammatical models on biological sequences; efficient filtering and model matching in data banks; protein structure prediction; and machine learning of grammatical models from sequences. Corresponding disciplinary fields are algorithmic on words, machine learning, data analysis and combinatorial optimization.

2.2.3. Systems biology: network modeling and analysis

The ultimate goal, for the biologist, is to explain how the combination of genetic and metabolic interactions determines the phenotype which is observed at the molecular level, particularly in case of diseases. The scarcity of quantitative data on biological phenomena implies the use of qualitative models. Our approach is based on the definition of graph models of interaction networks and the derivation of discrete or differential models for explaining and predicting (in a broad meaning) the behavior of the observed biological system. A special attention is paid to the diagnosis of **large scale** models described by their interaction graph.

2.3. Highlights of the year

This year is marked by two new "habilitations" (J. Nicolas and A. Siegel), that correspond to the maturation of two important fields of research in the team.

The first one concerns the linguistic analysis of sequences. Important results have been obtained at the lexical level in the framework of the project Modulome: a formal modeling of relevant repeats in genomes has been proposed and applied to two types of non-coding sequences, transposable elements in eukaryotes and CRISPR structures in archaea and bacteria. At the syntactical level, the PhD thesis of Goulven Kerbellec has demonstrated the interest of finite automata for modeling protein active sites and allowed the discovery of a new protein of a family involved in cell apoptosis.

The second one concerns the modeling of dynamic systems and biological networks. We have particularly developed this year a new approach for checking expression data against a regulatory network, using answer set programming (ASP). Our inference approach has been applied with success on a large scale network, the S. cerevisiae transcriptional network (2419 nodes and 4344 interactions), by combining ChIP-chip data and expression profiles. It has been able to isolate inconsistencies and to report predictions for a significant portion of the model (15% of all the interactions in both cases). The corresponding paper in BMC Bioinformatics has become a highly accessed paper (>1800 accesses to date).

The other component of the team, the bioinformatics platform GenOuest, has reached a new step in its organization: it has been certified Iso9001, a quality label that has only been attributed twice for national bioinformatics platforms (the other is the platform of Strasbourg).

3. Scientific Foundations

3.1. A short introduction to bioinformatics

Studying life at macromolecular level (DNA, RNA, protein or metabolites) involves multiple researches in mathematics and informatics [109]:

- *Data and Knowledge management.* Multiple technologies are producing raw data that have to be cleared and assembled into meaningful observations. It is the realm of statistical studies, with sophisticated normalization procedures, most of them being included in routine treatments. Information is produced in a highly distributed way, in each laboratory. Standardization, structuring of data banks, detection of redundancies and inconsistencies, integration of several sources of data and knowledge, extraction of knowledge from texts, all these are very crucial tasks for bioinformatics. High throughput techniques are also a source of algorithmic issues (assembling of fragments, design of probes).

- *Comparative genomics.* Referring to a set of already known sequences is the most important method for studying new sequences, in the search for homologies. The basic issue is the alignment of a set of sequences, where one is looking for a global correspondence between positions of each sequence. A more complex issue consists in aligning structures. More macroscopic studies are also possible, involving more complex operations on genomes such as permutations. Genotyping studies consider Single Nucleotide Polymorphism data, which correspond to mutations observed at given positions in a sequence with respect to a population. Analyzing this type of data and relating them to phenotypic data leads to new research issues. Once sequences have been compared, phylogenies, that is, trees tracing back the evolution of genes, may be built from a set of induced distances.
- *From structural analysis to systems biology.* This large domain aims at extracting biological knowledge from Xome studies, where X varies from genes to metabolites. Biological sequences and networks of components in the cell must verify a number of important constraints with respect to stable and accessible conformations, functional mechanisms and dynamics. These constraints result in the conservation during evolution of "patterns" and types of interactions to be deciphered. Many advanced researches consider now the study of life as a system, abstracted in a network of components governed by interaction laws, mostly qualitative or quantitative for reduced systems.

3.2. Optimized algorithms on parallel specialized architectures

Keywords: *dedicated architectures*, *grids*, *parallel architectures*, *reconfigurable architectures*.

Mixing parallel computing and genomics is both motivated by the large volume of data to handle and by the complexity of certain algorithms. Today, (dec. 2008) more than 800 genomes – including the human genome – are completely sequenced, and there exist a lot more sequencing projects (1000 human genomes, Human Microbiome Project,..., see *Genomes online database*[1]). Huge data bases become necessary whose volume approximatively doubles every year. This exponential growth is not expected to decline in the next few years due to low cost sequencing technologies and new needs such as isolation of important conserved structures in close species or metagenomics for ecological studies.

The problem is to efficiently explore these banks, and extract relevant informations. A routine activity is to perform content-based searches related to unknown DNA or protein sequences: the goal is to detect similar objects in the banks. The basic assumption is that two sequences sharing any similarities (identical characters) allow further investigations on some related functionality.

The first algorithms for comparing genomic sequences, essentially based on dynamic programming techniques, have been developed in the seventies [111], [122]. Then, with the increasing growth of data, faster algorithms have been designed to drastically speed-up the search. The Blast software [124] acts now as a reference to perform rapid searches over large data bases. But, in spite of its short computation time (compared to the first algorithms) a growing number of genomic researches require much lower computation time. Parallelizing the search over large parallel computers is a first solution implemented for instance in the LASSAP software (JJ Codani, [92]. Other works concern dedicated hardware machines. Several research prototypes such as SAMBA [95], BISP [78], HSCAN [94] or BioScan [130], have been proposed, leading today to powerful commercial products: BioXL, DECYPHER and GeneMatcher coming respectively from Compugen ltd. TimeLogic and Paracel [2].

Beyond the standard search process, this huge volume of available (free) data naturally promote new field of investigation requiring much more computing power such as, for example, comparing a set of complete genomes, classifying all the known proteins (decrypton project), establishing specific databases (ProDom), etc. Of course, the solutions discussed above can still be used, even if for 3-4 years, new alternative has appeared with the *grid* technology. Here, a single treatment is distributed over a group of computers geographically scattered and connected by Internet. Today, a few grid projects focusing on genomics applications are under deployment: the bioinformatics working group (WP 10) of the European DataGRID project; the BioGRID

[1] http://www.genomesonline.org/
[2] http://www.compugen.co.il/, http://www.timelogic.com, http://www.paracel.com

subproject from the EuroGRID project; the GenoGRID project deploying an experimental grid for genomics application; the GriPPS (Grid Protein Pattern Scaning) project.

Note that the large amount of genomic data is not the only motivation for parallelizing computations. The complexity of certain algorithms is also another strong motivation, especially for the analysis of structures in sequences [BMW03]. For instance, predicting the 3D structure of a protein from its amino acid sequence is an extremely difficult challenge, both in term of modeling and computation time. The problem is investigated following many ways ranging from *de novo* folding prediction to protein threading techniques [109]. The underlying algorithms are NP-complete and require both combinatorial optimization and parallelization approaches to calculate a solution in a reasonable amount of time.

For the last 2-3 years, GPU boards (Graphical Processing Units) have seen their computational power highly increasing. They now become a real alternative for deporting very time consuming general purpose computation. This activity is referred as GPGPU, standing for General-Purpose computation on GPUs. Many bioinformatics algorithms present interesting features allowing them to provide efficient parallelization. In 2007, we have started investigating the potentiality of this hardware support on several basic bioinformatics algorithms.

3.3. Syntactical Analysis of sequences

Keywords: *Data Analysis*, *Grammatical Inference*, *Logic Grammars*, *Machine Learning*, *Pattern Discovery*, *Pattern Matching*.

3.3.1. Formal Languages and biological sequences

Biological sequences may be abstracted as words on an alphabet of nucleic or amino acids. Structural and functional constraints on families of sequences lead to the formation of true languages whose knowledge would enable to predict the properties of these families. The theory of languages offers an ideal framework for the in depth formal or practical study of such languages:

- Formal: the goal is to define and study the most adapted classes of formal languages for the description of observed natural phenomena: crossing over (splicing systems of Head [97]), Watson Crick complementarity (Sticker-system [103]),inversion, transposition, copy, deletion... Language theorists like A. Salomaa and Gh. Paun [115] have explored standard questions (complexity, decidability) when faced with natural operations on biological sequences. The current agreement is that the necessary expressivity is the class of "mildly context sensitive" languages, well-known in natural language analysis [132], [120], [121] ;
- Practical: the goal is to provide to the biologist the means of formalizing his model using a grammar, which submitted to a parser will then make it possible to extract from public data banks relevant sequences with respect to the model. J. Collado Vides was one of the first interested in this framework for the study of the regulation of genes [79]. D. Searls proposed a more systematic approach based on logical grammars and a parser, Genlang [86]. Genlang still required advanced competences in languages and seems not used any more. We started our own work from this solution, keeping in mind the need for better accessibility of the model to biologists.

3.3.2. Machine Learning : from Pattern Discovery to Grammatical Inference

In practice, building relevant models is hard and frequently requires the assistance of Machine Learning techniques. Machine Learning addresses both theoretical (learnable classes) and practical issues (algorithms and their performances). Recent techniques mix both points of view, like *boosting* techniques (allowing good performances from initial weak learner) or *support vector machines* (applying structural risk minimization principle from statistical learning theory). Statistical tools are everywhere: reinforcement learning, classification, statistical physics, neural networks or hidden Markov models (HMM). HMM contain the mathematical structure of a (hidden) Markov chain with each state associated with a distinct independent and identically distributed (IID) or a stationary random process. Estimation of the parameters following maximum likelihood

or related principles has been extensively studied and good algorithms relying on dynamic programming techniques are now available in bioinformatics. When available, domain knowledge may help to design HMM structure but it is often very simple in practice (Profile HMM) and its discriminative power relies mostly on its parameter choice.

Because of its practical importance in genomic sequence analysis, a high number of pattern discovery methods have been proposed [73], [100]. One can primarily represent a language either within a probabilistic framework, by a distribution on the set of possible words, or within a formal languages framework, by a production system of the set of accepted words. At the frontier, Hidden Markov Models and stochastic automata have very good performances, but there structure is generally fixed and learning is achieved on the parameters of the distribution. Distributional representations are expressed via various modalities: consensus matrices (probability of occurrence of each letter at each position), profiles (adding gaps), weight matrices (quantity of information). A typical algorithmic approach scans for short words in the sequences and produce alignments by dynamic programming around these "anchoring" points [99]. Most powerful programs in this field use bayesian procedures, Gibbs sampling and Expectation-Maximization [108]. The linguistic representation, which corresponds to our own work, generally rests on regular expressions. Algorithms use combinatorial enumeration in a partially ordered space [72], [117]. Another track explores variations on the search for cliques in a graph [105], [74].

There exists a fundamental limitation in most studies: it is primarily the presence at a given position of some class of letters which will lead to the prediction. Purely statistical learning reaches its limit when relation between distant sites -frequent in biology- needs to be taken into account, because many parameters need to be adjusted. The theoretical framework of formal languages, where one can seek to optimize the complexity of the representation (parsimony principle), seems to us more adapted. We are studying this problem in the general framework of Grammatical Inference.

A grammatical inference problem is an optimization problem involving the choice of a) a relevant alphabet and a class of languages; b) a class of representations for the languages and a definition of the hypothesis space; c) a search algorithm using the hypothesis space properties and available bias (domain knowledge) to find the “best” solution in the search space. State of the art in grammatical inference is mostly about learning the class of regular languages (at the same level of complexity than HMM structures) for which positive theoretical results and practical algorithms have been obtained. Some results have also been obtained on (sub-)classes of context-free languages [118]. In the Symbiose project, we are studying more specifically how grammatical inference algorithms may be applied to bioinformatics, focusing on how to introduce biological bias and on how to obtain explicit representations. Our main focus is on the inference of automata from samples of (unaligned) sequences belonging to a structural or functional family of proteins. Automata can be used to get new insights into the family, when classical multiple sequence alignments are insufficient, or to search for new family members in the sequence data banks, with the advantage of a finer level of expressivity than classical sequence patterns permitting to model heterogeneous sequence families.

3.4. Systems biology: network modeling and analysis

Recent advances in functional genomics and in the study of complex diseases, such as cancer, immunodeficiencies, responses to infections, mitochondrial diseases, metabolic syndrome or aging, have shown the necessity of a new way of thinking in biology, which considers pathology and physiology as resulting from interactions between many processes at various scales. Systems biology emerged from this need. This scientific field addresses the study of genes (expression, evolution), protein interactions, biochemical reaction networks, cell populations and tissues in organisms considered as dynamical systems. It aims at studying the biological properties that result from the interaction of many components, investigating processes at different scales and achieving their integration.

Understanding will not arise from simulation alone (virtual cell or organism) but rather from the identification of relevant components for a given behavior and the reconstruction of the mechanisms involved. It concerns standard mathematical and physical tools, some borrowed from out-of-equilibrium thermodynamics and dynamical systems. New tools are also required. As coin by S. Brenner, complementary to bottom-up or

top-down approaches, a middle-out strategy starting from the cell is likely to be efficient in the analysis of biological systems. Ultimately, injecting the systemic vision in the understanding of human physiopathology could lead to novel differential diagnosis and improve medical care [21].

Cellular interactions' modeling is an old domain in biology, initiated by people interested in the dynamics of enzymes systems [101]. Models for transcriptional networks appeared as soon as gene interactions were discovered. The simplest static model consists in an oriented labeled graph, with labels + for activation or - for inhibition. Such graph representations are used to store known interactions in general databases. They are also the framework of Bayesian representations, used to infer gene networks from micro-array data, with the support of literature information [125].

The **dynamical framework in systems biology** includes simulations and prediction of behaviors. Models can be either qualitative or quantitative, as reviewed in [84], [77], [106]. A first approach makes use of continuous models: the concentrations of products are modeled by continuous functions of time, governed by differential equations. This framework allows one to state biological properties of networks, eventually by using simulation software [88], [127], [126]. The properties of continuous models can be studied with convex analysis, linear and non-linear control techniques [98], [114], [68]. Stochastic models transform reaction rates into probabilities and concentrations into numbers of molecules, allowing to understand how noise influences a system [102]. Finally, in discrete models each component is assumed to have a small number of qualitative states, and the regulatory interactions are described by discrete functions [104], [119]. Piecewise linear differential models [85], [91] try to build a bridge between continuous and discrete models.

These methods addresses fine dynamical properties such as the existence of attractors (limit cycles or steady states) and the behavior of these with respect to changes in the parameters [123], [77]. However, they need accurate data on chemical reactions kinetics or qualitative information. These data are scarcely available. Furthermore, these methods are also computationally demanding and their practical use is restricted in practice to a small number of variables.

Model identification addresses a different objective, that is, to build or update a model consistently with respect to a set of data. When large amounts of data are available, Bayesian networks [89] or kernels [131] have to be used. Another efficient approach formalizes a priori knowledge as partially specified models. Fitting models to data is obtained by means of various techniques [69], depending on the class of models, that can be discrete [116], continuous [106] or hybrid [75]. Qualitative reasoning, hybrid system, constraint programming or model-checking allow either to identify a subset of active processes explaining experimental time-series data or to correct the models and infer some parameters from data [70], [76]. Identification methods are limited to a few dozen components. Model correction or parameter regression can cope with up to hundreds of products [76], [69] provided that the biomolecular mechanisms and supplied kinetic data are accurate enough.

Reasoning on models Model-based identification can hardly cope with errors and variability that commonly affect measured expression levels in DNA microarrays. Moreover, time series data are absent in many situations, meaning that they inform more on steady state shifts under perturbations than on the dynamics of the system.
Testing and refining models become central issues in such a situation cumulating incomplete knowledge and partial observations. Our own work addresses these questions using formal methods of constraint resolutions. Our purpose is to study large-scale incomplete networks with efficient qualitative equation solvers. Diagnosis of incoherent parts of the networks use specific consistency rules depending on interactions types. Then, specific dynamical modeling procedures can be applied on these subgraphs to exhibit new biological insights.

Dynamical modeling, signalling and cancer Signalling mechanisms are essential in biological systems and represents a major research topic. At the cellular level, signalling networks allow detection and response to changes of the microenvironment and control various biological processes such as mobility, adhesion, differentiation, proliferation and apoptosis. The conservation during evolution of many signalling pathways and their implication in numerous pathologies such as cancer underlines the importance of these pathways for the life of the cell.

Research on molecular targets for cancer therapy relies to an increasing extent on understanding complex dynamical mechanisms, non-linear in time and space. Systems biology becomes a key appraoch in the understanding of such dynamical behaviours of cells from interaction between their components.

4. Application Domains

4.1. Application Domains

Keywords: *"life sciences"*, *"target discovery"*, *cancer*, *genomics*, *health.*

The main stakes of bioinformatics are to assist biologists in the processes of discovering prognostic, diagnostic and therapeutic targets and the understanding of biological mechanisms. The local context of Ouest-genopole provides us with a lot of collaborations with biology laboratories. We emphasize here three types of applications with major achievements in the project.

- **Targeted gene discovery** is studied with a syntactical approach. Models are built for proteins or promoters and then searched in whole genomes. We have for instance applied this strategy for the discovery of new beta-defensins, a family of anti-microbial peptides, in the human genome or the identification of all olfactive receptors genes in the dog genome.
- **Whole genome analysis** is made practical through dedicated data structures and reconfigurable architectures. We have for instance implemented very fast Blast comparisons on the human genome, built a software for bacterial genome fragmentation, GenoFrag, that helps to study genomes variations via Long Range PCR, and studied the occurrences of various structures like retro-transposons in the genome of *Arabidopsis thaliana* or micro-RNA in *Drosophilia melanogaster*.
- **Cancer** is a privilegied domain for the application of systems biology. Each cancer has its specificities, resulting from different functioning modes of interacting pathways, within different environments and submitted to different genetic alterations. This implies experiments producing large and heterogeneous data sets. Developing modeling tools should have an impact on development of new drugs, on diagnosis and prognosis and on multiple therapy optimisations when a combination of drugs is used.
  We have studied the pathways of TGF-β and NF-κB, that are central to the control of proliferation and apoptosis. We plan to consider also the Notch pathway. This project includes collaborations with INSERM Rennes, Curie Institute and NCBS Bangalore.

5. Software

5.1. Introduction

Most prototypes built during our researches are transferred on the platform GenOuest for further development and integration in a suitable environment for biologists. However, GenOuest has its own activity in relation with the service it has to offer and shares also studies with other french bioinformatics platforms (BioMAJ, BioWorkFlow, Grisbi, etc.). This section contains three parts in accordance with this organization scheme :

- general elements of the activity of the platform;
- new results of the platform projects and in collaboration with Inra;
- yearl activity of transfer of the platform, in conjunction with Symbiose.

5.2. GenOuest, the Bioinformatics computing center of Ouest-genopole

Participants: Olivier Collin, Hugues Leroy, Anthony Assi, Annabel Bourdé, Laetitia Guillot, Olivier Filangi, Sophie Roucan, Dragana Jovanoska, Eddy Farhat, Elie El Soueidy.

Main evolutions of the year have been:

- The GenOuest platform is ISO 9001:2000 certified since May 2008. The platform is now the second bioinformatics platform in France possessing this label.
- The computing capacity of the platform has been improved with the addition of new nodes and with new storage servers
- The GenOuest platform has established a partnership with the "service formation continue" of the University of Rennes 1 and is providing a catalog of short training session ranging from basic use of bioinformatics tools to more elaborate methodological sessions.

Since its creation, the platform organizes an annual meeting including technical conferences on the platform's achievments but also invited speakers that give the opportunity to discover new organizations (other bioinformatics plateforms), new technologies (softwares), or new scientific advances in bioinfomatics.

The platfom is involved in different coordination activities at a national level. H. Leroy is a member of the ReNaBi (Reseau National des plates-formes Bio-informatiques) steering board since 2005. He has been replaced by O. Collin at this position in July 2008. The GenOuest platform is in charge of the BioMAJ project, a joint project with INRA Jouy and INRA Toulouse. The platform is also involved in the BioWorkFlow program, a joint project with 5 other french bioinformatics platforms. GenOuest is also an active member of Grisbi (GRIlles Support pour la BIologie), a group gathering 6 platforms, recently labelled by IBiSA.

5.2.1. Bioinformatics services hosting

The GenOuest bioinformatics platform is hosting bioinformatics services developed by external research teams who require computing power. Among these services are

- *Autograph*, an interactive web server for automating and visualizing comparative genome maps
- *Germonline*, a cross-species knowledge base providing microarray data relevant for the cell cycle and gametogenesis
- *MIPDB*, a relational database of All Major Intrinsic Proteins of Archaea, Bacteria and Eukaryota.
- *M@ia*, a tool dedicated to micro-array data analysis.
- *RASTA*, a tool for the study of toxin-antitoxin compound in bacteria

5.2.2. Specific advanced developments

Web platform Test framework: A bioinformatics Web Platform gives access to specific services for a community of bioinformaticians and biologists. One needs to reduce repetitive consuming test tasks needed to guaranty the quality of the services provided by bioinformatics portal. Based on existing testing frameworks, we have developed a prototype of an automated system and functional test engine with various functionalities (service availability, scheduling, evaluation, load test...).

Generation of primers: The accurate determination of primers is of a key importance for the biologist carrying experiments. The platform has setup a set of tools dedicated to this kind of analysis[3]. The most original one is *iThos*, an improvment of*GenoFrag*, a software dedicated to Whole Genome PCR Scanning (WGPS), a means for analysing bacterial genome plasticity. The program initially seeks all the potential primers on a chromosome and calculates the best distribution of primer pairs, thanks to combinatorial optimization algorithms [71], [67].

Pattern matching and pattern discovery Our goal is to be recognized as an expertise resource center in this bioinformatics domain. Particularly, three algorithms have been developed in-house by the Symbiose research team. STAN is based on a suffix tree data structure. This tool scans complete genomes or user submitted sequences with grammatical patterns [4]. WAPAM inputs are protein patterns expressed by weighted automata in banks (nucleic acids or proteins), complete genomes and personal sequences. In both cases, the input patterns

[3] http://genouest.org/primers.php
[4] http://genouest.org/patterndiscovery.php

can be more complex than the usual regular patterns, such as PROSITE ones, since errors (substitutions and indels) and gaps of any size can be defined. In addition STAN provides string variables and its expressivity will be further improved in 2009 (Project Logol). An additional tool, Protomata-Scan, is available on the platform since September 2007. It is based on an extension of WAPAM named WASCAN which is able to handle the full expressiveness of automata. This program scans data banks with respect to automata generated by the Protomata-Learner program.

As a practical application of these tools, a collaboration has been established with the INRA team BIO3P working on plant pathogenic fungi for the analysis of the way a fungus triggers the regulation of an associated rhizobacteria. Results have been published in [16].

5.3. New activities of the bioinformatics platform in 2008

Participants: Hugues Leroy, Anthony Assi, François Moreews, Olivier Filangi, Olivier Collin.

5.3.1. National Project BioMAJ (BIOlogie Mise A Jour)

Biological knowledge, in proteomics and genomics context is mainly based on transitive bioinformatics analysis consisting in periodic comparison of data newly produced again corpus of known information. This approach needs on one hand accurate bioinformatics softwares, pipelines, interfaces... and on another hand numerous heterogeneous biological banks, which are distributed around the world.

Data integration represents a major challenge and bottleneck in bioinformatics. Parameters of this complexity include heterogeneity, size (several Tera bytes), number of banks, cross-linked sources, multiplicity of dedicated post treatments with respect to various bioinformatics software (blast, SRS, emboss, gcg, ...), variable banks frequency update, ... A first stake consists in automating the heavy process of updating the data banks for the administrator. Another significant stake to resolve is for the "quality" of service, providing to the users a clear vision of the integrity of data (state, exact origin, ...) constitutive of their workspaces.

BioMAJ is a joint development between three bioinfomatics platforms : INRA Toulouse (David Allouche), INRA Jouy-en-Josas (Christophe Caron) and our platform. BioMAJ is written using state-of-the-art technologies (java, xml, ..) and is based on a parameterizable workflow engine. Post processes are written for the usual formats (gcg, blast, srs, ...) and are easily customisable. BioMAJ has been relased under an opensource licence in April 2008 [5]. It is used regularly on different bioinformatics platfoms in France (INRA Jouy, INRA Toulouse, PBIL Lyon, Strasbourg and of course, on GenOuest)

5.3.2. National Project BioWorkFlow

The bioinformatics community is increasingly interested in ways to quickly build workflows and efficiently execute them on a distributed network of resources. The conception of workflows from distributed resources is often based on the use of web services. The BioWorkFlow action aims at evaluating the state-of-the-art, both in available solutions to provide tools through web services and existing softwares for generating workflows.

This action, funded by ReNaBi, involves many actors of the french bioinformatics community including different platforms : INRA Jouy-en-Josas (C. Caron, M. Wessner, V. Martin), LIPM Toulouse (J. Gouzy, S. Carrere), CNRS Roscoff (E. Corre), RPBS Paris (P. Tuffery), Institut Pasteur (C. Letondal, H. Menager, B. Neron), INRA Evry (F. Samson, M. Hoboeke), LabSTICC Brest (P. Picouet) and GenOuest (O. Collin).

5.4. Activity of transfer from Symbiose to GenOuest

Participants: Olivier Sallou, Michel Le Borgne, Israël-César Lerman, Hugues Leroy, Jacques Nicolas, Anne Siegel, Basavanneppa Tallur, Anthony Assi, Annabel Bourdé, Laetitia Guillot, Grégory Ranchy, Carito Guziolowski, Goulven Kerbellec, Sébastien Tempel, Philippe Veber.

Modeling activity concerns sequences and networks in Symbiose. Both of them have been subject to major developments during 2008.

[5] http://biomaj.genouest.org/

Logol The first software suite aims at offering a platform to search for complex models within both DNA and protein sequences. It is based on previous works made within the team in order to propose an expressive language (Stan and Wapam) that goes beyond pattern matching in biological sequences and study modeling needs of biologists at the level of whole genomes. Project's concerns are expressiveness, efficiency and ergonomy. The modeling language, called Logol allows writing a particular form of Definite Clause Grammars. The platform includes two components, a graphical interface- Logol Designer, [6]- and a parser -Logol Analyser [7] -. The interface provides a drag and drop facility to build interactively Logol grammars from graphical templates. The Logol Designer is written in Java script and licensed under the CeCILL v2 license. The analyser is written in Prolog. It may be run in command-line mode and on a personal computer or via a scheduler web page submitting jobs to the genouest cluster. Coupled with BioMAJ, the tool allows to parse updated versions of public banks or personal sequences.

The second activity of transfer concerns network analysis.

Pyquali is dedicated to computations on qualitative models represented by interaction graph. Nodes of these graphs represents chemical species and arrows are labeled by $\{+, -\}$ representing influence of the variation of a specie on another specie. An efficient representation of qualitative equations as ternary decision diagrams have been developped in this software.

Pyquali offers several functionnalities for the confrontation of networks and observation data: (i) The *internal consistency* of the network corresponds to checking that the whole set of constraints have at least a solution. (ii) *Consistency between a network and datasets* corresponds to checking that a partial set of variations on node can be extended to a whole solution to the set of constraints. (iii) *Diagnosing* an inconsistent network means that if a system does not check the basic rule, we shall identify a subset of interactions and data that bear inconsistencies. (iv) *Predicting* new variations corresponds to identifying the variables that have the same sign in all solutions of the set of constraints.

In 2008, A new version of Bioquali has been implemented. It is now possible to use a "no-variation" label for nodes and an "no-influence" label for edges. These improvements in the formalism of influence graphs paved the way to multiclock discrete dynamical models. Notice that a kind of dynamic influence graph is obtained by replacing the "no-influence" label on edges by a clock. In this kind of model, influences are not instantaneous but can propagate at different speeds depending on the state of the system.

Additionnally, new reasoning rules, that describe post-translational interactions (see sec. 6.3.3), can now be used by the introduction of user defined functions on qualitative values.

Several forms are proposed to use bioquali:

- A python module allows a user-friendly manipulation of equations with a python interface. [8].
- Most useful applications of pyquali are accessible through a web access named Bioquali [9].
- Aa cytoscape plugin has been developped to use and visualize the main features of Bioquali.

6. New Results

6.1. Optimized Algorithms on Parallel Specialized Architectures

6.1.1. Dedicated Architectures for genomic comparisons

Participants: Dominique Lavenier, Alexandre Cornu, Van Hoa Nguyen, Xianchun Ye.

[6] http://genoweb3.irisa.fr:8383/LogolDesigner
[7] http://genoweb3.irisa.fr:8383/org.irisa.genouest.logol.LogolAnalyser
[8] http://genoweb.univ-rennes1.fr/Serveur-GPO/outils/help/datafiles_test/pyquali.tgz
[9] http://genoweb.univ-rennes1.fr/Serveur-GPO/outils_acces.php3?id_syndic=226

With the increasing amount of available complete genomes, the need for inter or intra genome comparison is now a reality and supercomputer manufacturers now propose to include dedicated accelerator boards in their machines. Through the ANR *PARA project* (Parallélisme et Amélioration du Rendement des Applications), in cooperation with the BULL R&D team (Les Clayes sous bois), we have designed a reconfigurable accelerator tightly connected to the host processor.

Even if FPGA components offer consequent processing power, one challenge remains to have these resources fed from the main memory at a very high speed. The PCI-express (PCIe) interface, especially the second generation, achieves this goal by providing an agregated bandwidth of 10 Gbytes/sec.

Since 2006, we have developped the concept of "reconfigurable interface" for simplifying the generation of high performance interface between reconfigurable operators and host processor. Such a device is made of DMA transfer actors which can be assembled to form a dedicated PCIe interface able to connect FPGA operators trough DMA channels. This work has been achieved in 2008 and tested on various configurations. The great advantage of this approach is that a PCIe interface tailored to specific FPGA operators can be designed in a very short time. It remains to generate automatically this interface from a high level specification.

During 2008, we have also developed a reconfigurable genomic operator dedicated to protein processing. It is composed of 256 pico-processors able to process long genomic streams of amino acids for detecting relevant similarities. This architecture (reconfigurable operator and reconfigurable interface) has been implemented on a FPGA board. Large speed-up with an optimized software version – also developed in the context of the PARA project – has been obtained. Depending of the data set, speed-up from 10 to 80 have been measured on the part of the application deported to the FPGA operator.

Simple sequence comparison is still a basic preprocessing task in bioinformatics. One of the best program, BLAST, is based on a powerfull heuristic providing both good sensitivity and fast execution time. However, BLAST is not optimized for comparing large sets of data. First, its primarily goal is to query large genomic banks. Second, even if a multithreading version exist, the BLAST algorithm is fundamentally sequential.

Based on these facts, we have developed a new algorithm, based on the same BLAST heuristics, but more suited for handling large databases in a parallel way [40]. The design of this algorithm explicitely targets parallel structures such as SSE instructions (Streaming SIMD Extensions), manycores architectures, GPU hardware and FPGA platforms. An implementation of this algorithm, named *PLAST (Parallel Local Alignment Search Tool)*, is currently under development (V.H. Nguyen). Performances measured on the pure parallel soft version, including SSE optimizations, running on a 8-core system indicate a speed-up of 6 over the optimized BLAST multithreading version [44].

Experiment conducted with the Institute of Computing Technology from the Chinese Academy of Sciences in Beijing for implementing this algorithm onto the multicore Godson-T chip (64 MIMD processors on a single die) have also demonstrated the validity of our approach [45] : the speed-up is linearly proportional to the number of processors.

GPU implementation also provides interesting speed-up, ranging from 5 to 10 [43] compared to a BLAST execution on a single-core processor. FPGA implementation is under development.

6.1.2. Optimal indexing for protein similarity search

Participants: Dominique Lavenier, Pierre Peterlongo, Van Hoa Nguyen.

Efficient similarity search on genomic data sets requires fast indexes providing, in constant time, all word positions in the sequences together with some additional information for speeding up the whole process. With the tremendous increase of amount of data, storing large indexes in the computer memory becomes a serious bottleneck.

In order to decrease the index size without losing quality, we generate specific substitution tables adapted to a given reduced protein alphabet. Generally, proteins are represented as strings over 20 letters alphabets, leading to an encoding of 5 bits per amino acids. By partitioning the alphabet into 16 or 8 groups, 4-bit or 3-bit character encoding can be considered. Index sizes can then decrease accordingly. The generation of new substitution tables, based on these reduced alphabet, aims to compensate the lost of information induced

by the reduction. We have provided all the statistical tools associated to a given reduced protein alphabet. Experiments have shown an index size reduction of 35% without loss of quality.

This work has been published in BMC Bioinformatics [28]. It is the result of a closed collaboration with the INRIA Sequoia group, Lille, started in 2006 in the context of the INRIA ARC (Coordinated Research Action) program FLASH (http://www.irisa.fr/remix/arc.html)

6.1.3. Detection of new microRNA in the A. Pisum genome

Participants: Dominique Lavenier, Jacques Nicolas, Guillaume Rizk, Fabrice Legeai.

Genome anotation aims to analyse new sequenced genome to locate all potential genes and other pattern of interest such as microRNAs. In 2008, we have started a fruitfull collaboration with the BIO3P laboratory (Biologies des Organismes et des Populations appliquée à la Protection des Plantes), INRA, Rennes, for discovering new microRNAs in the *A. Pisum* genome using an original *in-silico* approach.

MicroRNAs are small non coding RNAs (ncRNAs) that play crucial roles in regulation of gene expression in plants and animals. In insects, microRNAs have been shown to be involved in multiple mechanisms such as embryonic development, tissue differentiation, metamorphosis, caste polyphenism and circadian rhythm.

We have developed a bioinformatics workflow to rapidely screen full genomes with criteria specific to microRNAs. Currently, 139 microARNs have been identified from which 72 are new and *A. Pisum* specific.

Tuning the workflow requires many runs. Unfortunately, the method intensively used secondary structure prediction algorithms (MFOLD), leading to very long computation time (typically 50 hours to process the *A. Pisum* genome). For shortening the process, graphic boards including the last generation of graphical processing units (GPU) have been investigated to parallelize MFOLD. Its parallelisation on one Telsa NVIDIA board provides a speed-up of 10 compared to a sequential execution. Experimentation on a 3 GHz Pentium equiped with 2 Telsa NVIDIA boards has demonstrated that the complete screening of the *Pisum* genome can then be done in 2h30 [55].

6.1.4. ACGT OncoSimulator

Participants: Dominique Lavenier, Julien Jacques.

In the Europeen ACGT project we actively participate to the development of a simulator for modeling cancer tumor development. This activity, referred as "In Silico Oncology" is a complex and multiscale combination of sciences and technologies in order to simulate malignant tumour growth and normal tissue response to therapeutic modalities at all levels of biocomplexity. The aim is to better understand cancer and related phenomena and to optimize therapeutic interventions by performing in silico (on the computer) experiments based on the individual data (clinical, imaging, histopathologic, molecular) of the patient.

We have in charge to parallelize the OncoSimulator code either for providing a better interactivity (shorter response time) or a better accuracy of the model in a reasonable amount of time [41]. Three approaches have been investigated: parallelization on the ACGT GRID, parallelization on clusters of PC, and parallelization on multicores.

The first one, which is relatively effortless, deals with the repartition of the many independent simulation on the available machines of the grid. This approach is used when there are a lot of combinations on different parameters to simulate. It is currently implemented in the ACGT GRID through a web portal.

The second one aims to parallelize a single simulation over a distributed memory system. Its execution time is then reduced and/or the precision is increased due to a higher mesh resolution. However, it requires to split the problem into rather independant tasks. We found that the architecture of the oncosimulator was not suited for this type of parallelization and we didn't go further in this direction.

The third approach is currently under investigation and will be developed during the last year of the ACGT project and combined with the GRID approach.

6.1.5. Development of Optimization techniques : the knapsack problem

Participants: Rumen Andonov, Nicola Yanev.

Besides our involvment in finding efficient architecture-algorithms pairs, we maintain a fundamental research in optimization since large scale bioinformatics also relies on a clever use of advanced optimization techniques (like Dynamic Programing (DP), Branch&Bounds (B&B), Lagrangian Relaxation (LR), diverse Heuristics etc). Applications in bioinformatics are presented in section 6.2.4.

Our paper [15] presents a preprocessing procedure for the 0-1 multidimensional knapsack problem. First, a non-increasing sequence of upper bounds is generated by solving LP-relaxations. Then, a non-decreasing sequence of lower bounds is built using dynamic programming. The comparison of the two sequences allows either to prove that the best feasible solution obtained is optimal, or to fix a subset of variables to their optimal values. In addition, a heuristic solution is obtained. Computational experiments with a set of large-scale instances show the efficiency of our reduction scheme. Particularly, it is shown that our approach allows to reduce the CPU time of the leading commercial software CPLEX of ILOG.

The result in [29] presents a new approach for exactly solving the Unbounded Knapsack Problem (UKP) and proposes a new bound that was proved to dominate the previous bounds on a special class of UKP instances. Integrating bounds within the framework of sparse dynamic programming led to the creation of an efficient and robust hybrid algorithm, called EDUK2. This algorithm takes advantage of the majority of the known properties of UKP, particularly the diverse dominance relations and the important periodicity property. Extensive computational results show that, in all but a very few cases, EDUK2 significantly outperforms both MTU2 and EDUK, the currently available UKP solvers, as well the well-known general purpose mathematical programming optimizer CPLEX of ILOG. These experimental results demonstrate that the class of hard UKP instances needs to be redefined, and the authors offer their insights into the creation of such instances.

6.2. Modelling sequence/structure relationships

Several lines of research are carried out using pattern matching, formal languages and combinatorial analysis techniques in order to identify structural models on sequences. Biologists may either want to design and test hypothetical models or to infer such models from a set of sequences sharing a functional or structural property. The goal is always to get an explicit view of the organization of the sequences and possibly to get new candidates with a similar organization in new sequences or to validate hypothetical mechanisms.

6.2.1. Finding modules in sequences

Participants: Jacques Nicolas [correspondant], François Coste, Dominique Lavenier, Israël-César Lerman, Anne Siegel, Basavanneppa Tallur, Pierre Peterlongo, Christine Rousseau.

J. Nicolas coordinates the national ANR project *Modulome* that aims at modeling the structure of genomes in terms of assembly of «modules» that may be copied and move inside or between genomes. This is supported by three applications on genomic mobile elements in cooperation with URGI/Inra Versailles, LME/Ifremer Brest and LEPG/CNRS Tours.

The theoretical part of this research as been concluded this year, leading to a fine characterization of several important properties with respect to relevant repeats in biological sequences [64]. Mainly, we have introduced concepts characterizing the locality of occurrences of repeats and precisely delimiting repeated entities with respect to their maximal local extension (largest maximal repeats). These concepts have shown to be widely applicable to natural repeated non coding structures in transposons and in imported genetic material.

As for the applied part of the research this year, we have achieved the complete analysis of an intriguing genetic structure in bacteria and archaebacteria, CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats). These are formed by a repetitive skeleton including genetic material imported from viruses and plasmid. CRISPR are likely to represent a fundamental immune system in prokaryotes. We have developped an identification method using very few parameters thanks to the previously described theoretical background. An important database has been built regrouping the analysis of all complete genomes available to date (more than 800 archeal and bacterial genomes), including flexible facilities for comparing structures in different organisms, the location of associated genes (CAS) and the possibility to check a user's sequence against the whole database.

6.2.2. Logical grammars

Participants: Jacques Nicolas [correspondant], Catherine Belleannée, Pierre Peterlongo, Olivier Sallou.

Last year we proposed the specification of a new modelling language, called Logol, intended to express structure-based models for biological sequences, based on a particular form of Definite Clause Grammars. We did deeply revisit String Variable Grammars (SVG) for this purpose in the line of Searls' work. This year was dedicated to Logol implementation and the resulting software is described in section 5.4. Technically, we decided to implement the parsers in the Prolog developing language (Sicstus Prolog), in order to take into account the declarative nature of the models, and to include use of the suffix array data structure (Vmatch), to allow parsing in large biological sequences, that is whole genomes.

At the same time we went on further investigations on the common ancestor concept, a promising means to express evolutionary biological models. On a theorical point of view, this concept is related to the median string and central string problems. We are just starting a collaboration with URGI and UPVD to apply it on the study of LTR retrotransposons.

6.2.3. Learning automata and grammars on biological sequences

Participants: François Coste [correspondant], Goulven Kerbellec, Matthias Gallé, Pierre Peterlongo, Olivier Baldellon.

Our main focus is on the inference of automata from samples of (unaligned) sequences belonging to a structural or functional family of proteins. Automata can be used to get new insights into heterogeneous families, when classical multiple sequence alignments are insufficient, or to search for new family members in the sequence data banks, with the advantage of a finer level of expressivity than classical sequence patterns (such as PSSM, Profile HMM, or Prosite Patterns).

This year, we developped a weighting scheme used to scan more accurately the data banks when searching for new members of the protein family. Formalization of the whole approach has been refined and additional experiments have been achieved for Goulven Kerbellec's PhD. report [11]. The corresponding tool "Protomata-Learner" is available on the Genouest bioinformatics platform [10]. We have introduced a new kind of sequence alignment named "partial local multiple sequence alignment" which is currently found by a greedy procedure. We have formalized the problem of searching for the best partial local multiple sequence alignment as the search of an optimal k-matching under constraints from pairwise local alignments [60].

We are working in collaboration with several biologists to apply this tool to model their sequence families. We are thus looking for new proteins of a family involved in cell apoptosis studied by T. Guillaudeux in the team *Microenvironnement et Cancer* (MICA). In the framework of the Proteus ANR project, we are currently applying Protomata-Learner's ideas in order to build the templates that are needed in protein threading programs (see section 6.1.5).

We are also studying how to learn more expressive grammars such as context-free grammars. It is the subject of the PhD study of Matthias Gallé. Focusing rather on learning the structure than the language, we developed an approach related to the smallest grammar problem, aiming at finding an optimal recursive coding of repeated elements. Such approaches heavily depend on efficient data structures for the lexical analysis of large sequences such as genomes. To detect and score repeats, we use suffix arrays that need to be regularly updated after each rewriting of repeat. We proposed an algorithm to update a suffix array after the substitution in the indexed text of some (possibly all) occurrences of a given word by a new character. Our algorithm uses the specific internal order of suffix arrays in order to update simultaneously groups of entries, and ensures that only entries to be modified are visited. The algorithm was implemented and showed a significant run time speed-up compared to the construction of suffix arrays from scratch at each step [36].

6.2.4. Protein structure alignment: algorithms and applications

Participants: Rumen Andonov [correspondant], Guillaume Collet, Guillaume Launay, Noel Malod-Dognin, Israël-César Lerman, Yavor Vutor.

[10] http://protomata-learner.genouest.org/

Computational problems related to spatial structures are inherently much more complex than those considering only the sequence level. A theoretical basis that could support a rigourous analysis and understanding of structure prediction models is almost non-existent, as the problems are blend of continuous and discrete mathematics. In our group we focus notably on creating efficient algorithms for solving combinatorial optimization problems yielded by sequence/structure alignment (Protein Threading Problem-PTP), and structure/structure comparison. Both applications have been proved to be NP-complete and attract the researchers' interest since more than fifteen years [107], [93], [83].

Sequence/structure alignment

We have been working on PTP since 2001 in close collaboration with two teams: MIG from INRA, Jouy-en-Josas, and the Operation Research Departement from the University of Sofia, Bulgaria whose leader, Prof. N. Yanev is regularly visiting Symbiose. FROST (Fold Recognition Oriented Search Tool) was initially created by MIG, and is currently the only french software participating in the prestigious competition CAFASP (Critical Assessment of Fully Automated Structure Prediction). The optimization algorithms in this tool are now based on the approaches developed by our group and published in a series of results last year [57] and this year [32].

The weak point of FROST (as well as all currently available methods) is that *only* global sequence/structure alignment is allowed (*all* Secondary Structures Elements (SSE) have to be aligned). Designing algorithms for flexible structure alignment (relaxing the former constraint) is our current focus of interest and is the main subject in the PhD study of Guillaume Collet. Some ideas to tackle this problem were presented in [80], based on Integer Programming models for semi-global and local sequence/structure alignment. This axis was reinforced in 2008 by G. Launay, postdoctorant in the framework of the Proteus ANR project [11]. G. Launay is working on improving various scoring functions ([113]) used in the latest version of FROST, with a particular focus on their parallel implementations. The parallel framework was initially developed by Y. Vutov from the Institute for Parallel Processing of the Bulgarian Academy of Sciences (http://parallel.bas.bg/SciComp/) who visited SYMBIOSE for 2 months in 2008 in the framework of Proteus.

The scoring functions developed to solve the single protein fold recognition problem should be somehow transferable to the multiple proteins complex recognition problem. G. Launay is currently assessing the feasability of such methods [20].

Structure/structure comparison

Among the measures for quantifying the similarity between protein 3D structures, contact map overlap (CMO) maximization deserved sustained attention during the past decade. Despite this large involvement, current algorithms possess a modest performance and are not applicable for large scale comparison.

Our WABI'08 conference result offers a clear advance in this respect. We propose a new integer programming model for CMO and an exact B&B algorithm with bounds obtained by a novel Lagrangian relaxation. The efficiency of the approach is demonstrated on a popular small benchmark (Skolnick set, 40 domains). On this set our algorithm significantly outperforms the best existing exact algorithms. Many hard CMO instances have been solved for the first time. In order to validate the accuracy of the pairwise association result, Likelihood Linkage Analysis (LLA) hierarchical classification method (CHAVL program) has been performed on the whole protein set. The adopted similarity index is derived from the lower bound obtained by the Lagrangian relaxation. Right classifications have been obtained for the Skolnick set. To assess furthermore our approach, we constructed a large scale set of 300 protein domains. Using the similarity measure for any of the 44850 pairs, we obtained a classification in excellent agreement with SCOP. Our algorithm provides thus a powerful classification tool for large structure databases [35]. New questions arose that will be the focus of our future research: what is the influence of the similarity index choice on classification stability and what is the influence of the number of iterations ?

Our second center of interest in this research axis relates to improving the tool VAST (Vector Alignment Search Tool). It is the NCBI's structure-structure similarity search service http://www.ncbi.nlm.nih.gov/Structure/VAST/. It compares 3D coordinates of a newly determined pro-

[11] http://migale.jouy.inra.fr/proteus

tein structure to those in the MMDB/PDB database. VAST first aligns SSE and then extends this alignment to amino-acids. The SSE alignment is presented as a maximum clique problem in a particular graph. In [53] we propose a new integer programming model for various maximum weight clique problems and we successfully applied it in the VAST framework.

6.3. Systems biology: analysing data and modeling interactions

The ultimate goal of this axis is to build dynamical systems that model interactions implied in biological processes. It both adresses medium scale modeling with differential equations and large scale modeling using model reduction techniques and logical constraints.

6.3.1. Construction and analysis of signalling and metabolic pathways

Participants: Michel Le Borgne [correspondant], Ovidiu Radulescu, Anne Siegel, François Moreews, Nolwenn Le Meur, Pierre Blavy, Jeremy Gruel.

Effect of ADAM12 in TFGβ signalling pathway.

The transforming growth factor TGF-β governs a variety of cellular process including differentiation, proliferation and apoptosis and is the main contributor to development of hepatic fibrosis followed by cirrhosis, a major risk for the development of hepatocellular carcinoma. Accordingly, the regulation of TGF-β signaling is complex and depends on specific protein context. We are collaborating on this liver injury with the team of N. Theret (INSERM Rennes) since several years. They recently showed that, the disintegrin ADAM12 is expressed during hepatic fibrosis, interacts with TGF-β receptor and modulates it's trafficking among membranes, a crucial point in TGF-β signaling.

We have investigated the potential effect of ADAM12 on different steps of TGF-β receptor trafficking. Using previously established mathematical models and physiological data, we built two differential models implementing two hypothesis on the degradation of receptors. Our main hypothesis was that ADAM12 does not change the structure of receptor traffic but rather acts on rates of transfer between compartments and rates of degradation. We investigated two specific rates compatible with experimental measures in presence of ADAM12 that led to an hypothesis on TGF-βsignaling. This hypothesis was tested against experimentations and was not observed. After this negative result, instead of further trying to fit numerous alternative models, we generated a family of models producing a behavior compatible with the one observed in presence of ADAM12. The compatibility criteria includes qualitative results from experimentations. The family of models is parameterized by a set of rate coefficients. The invariant properties of this set (its shape) gives insight on the effect of ADAM12 on TGF-βreceptors trafficking. Our model predicts that ADAM12 increases TGF-β receptors trafficking rates between the cell surface and the endosomal membrane. It also appears that ADAM12 induces a modification into TGF-β signal shape favoring a permanent response by removing the transient component observed under physiological conditions [51].

NF-κB signalling system

The transcription factor NF-κB is also involved in a wide diversity of domains (immune and inflammatory responses, cell survival and apoptosis, neuro-degenerative diseases, cancer and development). NF-κB is sequestered in the cytoplasm by inactivating proteins named IκB. NF-κB bound to IκB molecules is transported to the nucleus to activate its target genes. Five members of the NF-κB family (including p65) and seven members of the IκB family are known in mammals [90]. This generates a large combinatorial complexity of dimers, affinities and transcriptional capabilities. Moreover, the gene coding for IκBα, is transcriptionally activated by NF-κB. This negative feed-back loop can give rise to oscillations of the activity of NF-κB [112]. Phosphorylation of IκBα upon signalling is provided by a kinases complex associated to a regulating protein.

In collaboration with A. Lilienbaum, (CNRS EA300, U. Paris 7) we proposed a rather detailed dynamical model of this complex NF-κB pathway, which takes into account transcription, translation and degradation of different NF-κB units. The three dimers of NF-κB are characterized by different affinities for DNA sites, and associate differentially to three IκB species generating thus 9 species with different abundances and characteristics upon signalling and degradation. Many transcription factors have been considered in the

model and kinase activation/inactivation module was borrowed from previously existant models. The complex resulting model allows us to cover a large area of situations, including non-conservative scenarios such as those used by our collaborators in Bangalore (over-expression of $p65$ unit). In a recent paper [30] we have used this complex model and the hierarchy of reduced models derived from it as a benchmark to validate model reduction and robustness techniques.

Fatty acid metabolism.

In collaboration with laboratories of INRA and supported by a co-tutored PhD thesis (ASC Inra program, P. Blavy), we continued investigations on the regulation of fatty acids metabolism in hepatic cells.

P. Blavy proposed a modeling strategy to determine from a large set of reactions the smallest set compatible with a compendium of required properties and experimental data. We applied this strategy to define a model of hepatic fatty acid metabolism integrating both metabolic pathways and the contribution of some major genetic regulators. We considered the fatty acid metabolism in hepatic cells of mice during a 72 hours fed-to-fast transition and we tested the key role of oxidation and elongation in fatty acid metabolism. The modeling approach suggests that a small set of reactions are enough to properly fit the model to both mutant and wild type fatty acid data.

We concluded that a minimal model in fasted PPAR-/- mice must include synthesis of C22:6ω3 from C18:3ω3. Fasting experiment data suggest that C22:6ω3 production in PPAR-/- mice should decrease; but our analysis pointed out a constant accumulation rate. This encouraged us to measure other activity levels (D6D) in order to refine the model with new regulations. Our analysis also suggested that PPARα has a strong effect on oxidation, independent from the nutritional status. Since the model fits with a constant rate, the role of PPARα in the regulation of oxidation during fasting is negligible with respect to its constitutive function as a regulator of FA oxidation enzymes. These results were discussed in [46].

6.3.2. Modularity, hierarchical models, and robustness of complex biological systems

Participants: Ovidiu Radulescu [correspondant], Vincent Noël.

Model reduction techniques

Cellular processes such as metabolism, differentiation, signalling, etc., can be modeled as large networks of biochemical reactions. In order to understand the functioning of these systems, there is a strong need for general model reduction techniques allowing to simplify models without loosing their main properties. In systems biology we also need to compare models or to couple them as parts of larger models. In these situations reduction to a common level of complexity is needed [21].

We propose a systematic treatment of model reduction of multiscale biochemical networks. First, we considered linear kinetic models and proposed a reduction algorithm which is based on a generalized theory of the limiting step that we have developed in [18]. Second, for non-linear systems we developed an algorithm based on dominant solutions of quasi-stationarity equations. For oscillating systems, quasi-stationarity and averaging are combined to eliminate time scales much faster and much slower than the period of the oscillations. In all cases, we obtain robust simplifications and also identify the critical parameters of the model.

Our approach produces hierarchies of models that are useful when one needs to zoom in and out several levels of complexity ("middle-out" strategy). It allows also critical parameter identification, an important issue in systems biology with potential applications to biological control and therapeutics. Our approach also deals naturally with the presence of multiple time scales, which is a general property of systems biology models. Similar approches have been developed for stochastic models, based on piecewise deterministic approximations. Piecewise deterministic processes (PDP) are hybrid processes consisting of discrete and continuously valued components, and of a sequence of random events. An event consists in a change (jump) of the discrete component, and in a jump of the state and/or of a change of the differential equation governing the dynamics of the continuous component. Starting from the full stochastic Markov jump process we obtain the PDP approximation as a weak limit.

Robustness studies : cancer and development systems

Biological robustness, defined as stability against external perturbations and internal variability, represents a common feature of living systems. The fittest organisms are those resisting to diseases, imperfections or damages of regulatory mechanisms, and that can function reliably in various conditions. Furthermore, abnormal, pathological behavior such as cell proliferation in carcinoma are also very robust. Understanding robustness as an organizing principle is thus an important issue in biology and represents one of the major (and still unsolved) objectives of the new field of systems biology.

We have studied two types of biological robustness. The first type is Waddington's canalization. Using experimental arguments and modeling, we show that the gap gene network in Drosophila's development has canalization properties (joint work with the team of J.Reinitz in Stony Brook). Another form of robustness is the reduced variability of biological processes with respect to variability in pathways details and interaction strengths. This was numerically illustrated by von Dassow for several models. We claim that von Dassow's robustness is a case of variability reduction via projection. Robustness of large size, multiscale models, can be understood as resulting either from mapping a huge number of parameters into synthetic, critical parameters, and/or by dominating the effects of pathways and of submodels by main pathways and dominant submodels. A natural theoretical framework to handle high-dimensional phenomena is provided in statistical mechanics by Gromov/Talagrand concentration of measure on metric spaces. We have applied it for the first time to molecular biology (joint work with A.Gorban, U.Leicester and A.Zinovyev, Institut Curie) [30].

6.3.3. Algorithms for the analysis of large-scale models

Participants: Michel Le Borgne, Jacques Nicolas, Ovidiu Radulescu, Anne Siegel [correspondant], Carito Guziolowski, Thibaut Henin, Sylvain Blachon, Annabel Bourdé.

Post-translational interactions: dedicated modeling and identification In 2008, we have investigated refined post-translational interactions models and proposed neat improvments in the diagnosis of such models described by influence graphs (see the Bioquali application described in Sec. 5.4).

1. On previous works, we demonstrated the feasibility of large-scale network diagnosis with the comparison of the whole E. Coli transciptional network with a set of stress nutritional data. The process resulted in a small number of sub-networks to be refined by a few post-translational interactions in order to obtain a consistent network [96]. This year was dedicated to the *prediction* of unknown values in the network. This task heavily depends on the design of precise interaction rules. We have introduced a rule for complex formation increasing the prediction rate from 1% to 25% in E.coli network. We validated them with a set of real measurements, obtaining an agreement of 80%. This result, comparable to other methods working on *E. coli* data [81], [82], [87], is remarkable since we used only a transcriptional model without including metabolic regulations [37].
2. In collaboration with T. Baumuratauva (Irmar, Rennes), we applied the diagnosis process to a EWS/FLI1 regulatory network coupled with transcriptome time series data. The network was built by G. Stoll and partners of the Sitcon ANR project at Institut Curie. The issue is to check whether the level of mRNA of a certain gene correlate with the behavior of EWS/FLI1. Our analysis suggested the necessity to explicitely introduce genes, the corresponding mRNAs and active proteins in the network. This improved network allows describing the various influences received by proteins and correctly explains why the levels of some proteins are opposite to the levels of the corresponding mRNAs. The next step of the study will be to apply constraints solvers on the model, as explained in the next paragraph.

Towards efficient diagnosis procedures

The study of EWS/FLI1 network as well as yeast network have shown that reduction procedures on decision diagrams implemented in our software application (Bioquali) [128] are not sufficient to handle the complexity of the network. M. Le Borgne and C. Guziolowski, in collaboration with P. Veber at Postdam University have designed efficient algorithms to treat the networks.

- In collaboration with T.Schaub's lab (Potsdam University), we have studied the applicability of Answer Set Programming (ASP) for checking consistency between a whole organism regulatory

model and experimental data. The conclusion drawn from this preliminary work is twofold: first ASP solvers can solve such large sets of qualitative constraints corresponding to several thousands of genes and regulations [129]. Second, ASP taken as a language, provides a simple and reasonably efficient way of developing and testing some extensions of our framework (in particular, stronger notions of consistency between models and data). This approach is complementary to decision diagrams that allow to study the full set of solutions to the system of constraints (percentage of solutions satisfying a given condition) but are limited to at most 100 nodes.

We have illustrated the efficiency of this approach with a contribution to *sign inference* in transcriptional regulation network. Based on the analysis of artificial datasets, we have proved that about 30% of the signs of the *E. Coli* network are expected to be inferred from about 20 DNA-chips. We have then proposed to deal with errors by combining the predictions on several different datasets. With this approach, we were able to infer correctly about 10% of the signs of *E. Coli* network [31]. We had to combine decisions diagrams and constraint solvers to achieve these performance.

- Given a family of influence graph and a set of qualitative observations, we are interested in the computation of so called hard components which are the invariants of the family. Our preceding brute force approach computed polynomial representing the conjunction of local constraints on each node of the network. Polynomials are represented with a BDD like structure (QDD). It took 6H to compute hard components for the graph used in the study of the Ewing tumor.

  We have designed a new algorithm which avoids to compute the conjunction of local constraints before quantifier elimination. This algorithm uses the structure of the dependance graph between local constraints to perform progressive quantifier elimination interweaved with conjunctions. A special covering tree drives the computations. The complexity of this algorithm is not easy to establish since it depends on fine parameters of the dependance graph structure but it is fairly efficient for set of loosely coupled constraints. The compatibility test run in a few seconds on the Ewing network. This algorithm is published as a new Python module in Bioquali (see Sec. 5.4).

  This algorithm can also be generalized to other kind of projection than quantifier elimination. In particular, we also get an optimisation algorithm which can be used for statistical inference of biological influence graphs. This algorithm gives exact results on hard components and moreover establishes a link between algebraic inference and statistical inference.

Reasoning on networks with constraint solvers.

The necessity to assist biologists in various aspects of the management of complex qualitative models of interactions requires a common theoretical and practical framework tailored to the multiple types of relevant queries in this context. Besides consistency checking, A. Siegel showed in her Habilitation thesis the variety of tasks to be addressed, including explanations, predictions on the behavior of the system, planning and control. Logical representations, automated reasoning and constraint solving are well suited fields of research in this respect and this has oriented us to Answer Set Programming.

Automated reasoning on metabolic pathways is the subject of a collaboration with the university of Potsdam and of a co-tutored PhD thesis (T. Henin). We are also collaborating with UMR 6026 (Molecular and Cellular Interaction) on the modeling of the regulations of bacteria genes implied in oxydative stress. We have worked this year on the design of a new action language, called T, for the specification of biological transport networks. It includes membranes and consumes, produces and needs propositions and is compatible with the C action language defined by M. Gelfond and V. Lifschitz.

6.4. Some statistical developpments

We have a small research activity that mainly involves data analysis techniques.

6.4.1. Modeling synthetic lethality

Participant: Nolwenn Le Meur.

Synthetic lethality defines a genetic interaction where the combination of mutations in two or more genes leads to cell death. The implications of synthetic lethal screens have been discussed in the context of drug development as synthetic lethal pairs could be used to selectively kill cancer cells, but leave normal cells relatively unharmed. We propose statistical and computational tools that can be used to find relationships between synthetic lethality and cellular organizational units. In Saccharomyces cerevisiae, we identified multi-protein complexes and pairs of multi-protein complexes that share an unusually high number of synthetic genetic interactions. We found that synthetic lethality can arise from subunits of an essential multi-protein complex or between pairs of multi-protein complexes. In addition, using multi-protein complexes allowed us to take into account the pleiotropic nature of the gene products. Modeling synthetic lethality using current estimates of the yeast interactome is an efficient approach to disentangle some of the complex molecular interactions that drive a cell [23]. This project was done in collaboration with Dr. Robert Gentleman, head of the computational biology group at the Fred Hutchinson Research Cancer Center and was funded through a Human Frontiers in Science program research Grant RGP0022/2005.

6.4.2. Constraint-based mining of SAGE data

Participant: Sylvain Blachon.

Current analyses of co-expressed genes are mainly based on global approaches such as clustering. An alternative is to employ local methods and search for patterns - sets of genes displaying specific expression properties in a set of situations. In [19], we implement and test a flexible constraint-based framework using domain knowledge from various sources (literature databases, biological ontologies and experiment annotations). It enables the effective mining and representation of meaningful over-expression patterns representing intrinsic associations among genes and biological situations. In collaboration with Insa Lyon team, S. Blachon contributed to the development of a web application called SQUAT. This tool gives access to both raw SAGE data and patterns mined from these data, for three species (human, mouse, chicken). This database allows to make simple queries as well as much more complex queries like: “what are the genes that are frequently co-over-expressed with my gene of interest in given biological situations?”. Connections with external web databases enrich biological interpretations, and enable sophisticated queries [26].

6.4.3. Quality of association rules in Data Mining

Participant: Israël-César Lerman.

Likelihood Linkage Analysis (LLA) hierarchical classification is a general methodology with respect to the mathematical, logical or statistical structure of the data description. Two principles sustain this approach. First, set theoretical and relational representation of the description is set up. Second, probabilistic similarity measure between combinatorial and statistical descriptive structures is built with respect to an independence (no relation) statistical hypothesis. A comprehensive presentation of this approach has been given in [47].

In the field of Data Mining, one fundamental objective consists in building asymmetrical association rule measures. The interest of a rule A -> B may be evaluated with the LLA approach using an implication index (measure) that evaluates in a certain way the propensity of B, knowing A. However, a local definition becomes non discriminant for large data bases. In these conditions we have proposed a discriminant extension of the probabilistic indices obtained with respect to a set of potential interest rules. This work has been performed in collaboration with J. Azé (LRI, University of Paris-Sud) [110]. LLA has also strongly influenced another approach, the "Statistical Implicative Analysis" [66]. This method directly uses asymmetrical similarities and build an oriented ascendant binary hierarchical classification. New extensive analysis including formal logical and statistical aspects of this original construction is provided in [25]. The formal and mathematical aspects have been more deeply and more precisely reconsidered in [38]. A future and extensive analysis of these last aspects will lead to a future paper in collaboration with P. Kuntz (LINA Laboratory, Site Polytech’Nantes).

Two types of criteria have been studied theoretically and experimentally in the framework of hierarchical clustering under spatial contiguity constraint, the Ward Criterion and the LLA criterion. We show that the latter is the most robust with respect to the "inversion phenomenon" in the cluster dissimilarities. Accelerated algorithm using the nearest neighbors aggregation principle leads to a linear behavior of the computing

complexity. New and more accurate statement is given in [33]. This work is a result of collaboration with K. Bachar (CREDO Laboratory, ESSC Angers).

7. Other Grants and Activities

7.1. Regional initiatives

7.1.1. OUEST-genopole

OUEST-genopole, the eighth national genopole, funded in January 2002 acts as a strategic project for higher education and research in life sciences, bioinformatics, and for the economic development in the fields of *marine sciences*, *agriculture and food processing* and *human health*. It is a network, federated through a GIS structure (Scientific Interest Groupment), of the various academic organisms involved in these fields (About 55 laboratories from Inra, Inserm, Ifremer, Inria, CNRS, Universities of Rennes, Nantes, Brest and Angers) in western France (Region Bretagne and Pays de la Loire). A network of technological platforms is proposed to all members.

OUEST-genopole is headed by M. Renard (Inra Le Rheu). J. Nicolas and J. Bourdon in charge of the bioinformatics research field and O. Collin in charge of the bioinformatics platform, participates in the monthly meetings of the OUEST-genopole committee.

7.1.2. Regional cooperation

The Symbiose project has collaborations with many laboratories, mostly biological, in western France. Collaborations are detailed in the section devoted to new results. Among the most advanced, let us mention:

- Agrocampus-Inra Rennes - Laboratoire de Génétique Animale: Analysis of gene regulation involved in the lipid metabolism (P. Blavy, O. Radulescu, A. Siegel).
- Inserm U456 (Détoxication et réparation tissulaire). Study of gene regulations in TGFβ signalling in liver cancer (J. Gruel, M. Le Borgne, O. Radulescu).
- Biologie des Organismes et des Populations appliquées à la Protection des Plantes BIO3P, INRA, Rennes. Analysis of micro RNA (D. Lavenier, J. Nicolas)
- Centre de Recherche et d'Etudes de l'Ouest en Gestion des Entreprises de l'ESSCA: Ascendant hierarchical classification (I.-C. Lerman)
- Ecole Nationale Supérieure des Télécommunications de Bretagne (ENSTB), Brest. Automatic design of reconfigurable accelerator (D. Lavenier)
- INSERM SERAIC, Rennes, cell cycle robustness (O. Radulescu).
- Laboratoire d'informatique, Université de Nantes : Oriented classification (I.-C. Lerman).
- MIcroenvironnement et CAncer, UPRES UA 3889 (Thierry Guillaudeux) : Tumor Necrosis Factor (F. Coste, G. Kerbellec)
- UMR-CNRS 6026 Structure et Dynamique des Macromolécules, Rennes (C. Delamarche): Major Intrincsic Proteins (MIP) (F. Coste, G. Kerbellec, G. Ranchy)

7.2. National initiatives

The Symbiose project is involved in the following national collaboration programs (detailed hereafter):

- Inra projects Genanimal, Sigenae, BioWorkFlow and BioMAJ .
- ANR contracts Proteus, Modulome, PARA, SITCON, DyCoNum.

Inside these collaboration programs or beside, the main teams that we cooperate with are:

- Curie Institute, biology, medecine and bioinformatics, Paris (O. Radulescu, A. Siegel, M. Le Borgne)
- ABI, Paris VI University (Joël Pothier): structural cores (F. Coste, R. Andonov).
- Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS), Clermont-Ferrand, Analysis of associative rules (I.-C. Lerman).
- CNRS, biology, Paris : Alain Lilienbaum. (O. Radulescu)
- CEA, Saclay: Laboratoire d'Intégration des Systèmes et des Technologies. Reconfigurable Architectures Reconfigurable computing (D. Lavenier).
- Ecole Polytechnique (T. Simonson) : protein structures, ANR Proteus (R. Andonov, G. Launay).
- Institut de Biologie et de Chimie des Protéines (G. Déleage) (R. Andonov, F. Coste)
- IML, Marseille and LIRMM, Montpellier: substitutive dynamical systems (A. Siegel).
- LIFL, Lille, Genomic String Indexing (D. Lavenier, P. Peterlongo).
- MIG, Inra, Jouy en Josas (J.-F. Gibrat, A. Marin): Protein threading,ANR Proteus (R. Andonov, F. Coste, D. Lavenier)
- LEPG Tours (Y. Bigot) : ANR Modulome (J. Nicolas, C. Rousseau, F. Coste).
- URGI Versailles (H. Quesneville): ANR Modulome (J. Nicolas, C. Rousseau).

7.2.1. National projects of the GenOuest platform

Participants: Olivier Collin, Hugues Leroy, François Moreews, Jacques Nicolas, Anthony Assi, Olivier Filangi.

7.2.1.1. BioSide

BioSide is a collaboration with ENSTB-Brest (P. Picouet, S. Bigaret, P. Tanguy) and Station Biologique Roscoff (FR2424 - X. Bailly, E. Corre, G. Le Corguille). It is an environment where bioinformatic programs are thoroughly described through a set of metadata (including the semantics of their parameters and the execution policy) thus providing access to source programs. The intuitive BioSide interface allows the design, the execution and the storage of workflows (scenarios). Metadata is used both to provide high level help to the final user and to guarantee dynamic extensibility of BioSide. A standalone version is already available for phylogenetic programs. A server version will be soon available.

7.2.1.2. BioWorkFlow

BioWorkFlow is a collaboration with INRA Jouy (C. Caron, V. Martin, J.F. Gibrat), INRA Toulouse (J. Gouzy, S. Carrère), Institut Pasteur (C. Letondal), RPBS (P. Tuffery), ENST-Bretagne (P.Picouet), Station Biologique Roscoff (E.Corre), URGI (F. Samson), Laboratoire Statistique et Génome Evry (M. Hoebeke). This project, funded by ReNaBi, initiated a reflexion concerning the setup of workflows at a national level. At first, the different solutions available for bioinformatics have been analysed and a first prototype of a trusted network of web services will be implemented.

7.2.1.3. BioMAJ

BioMAJ is a joint collaboration between INRA Toulouse (David Allouche), INRA Jouy (Christophe Caron) and IRISA for databanks management. See section 5.3.1.

7.2.1.4. GRISBI

The GRISBI Platform is a joined initiative between six French bioinformatics platforms: PRABI Lyon, MIGALE Jouy-en-Josas, Genouest Rennes, CBIB Bordeaux, BIPS Strasbourg, CIB Lille. This effort tends to set up a grid infrastructure devoted to Bioinformatics at the national level. The goal is to address challenging bioinformatics applications dealing with large scale systems : comparative genomics and genome annotation, protein function prediction, molecular interaction like protein-protein or DNA-protein... This will be reached by sharing and by mutualizing the existing resources of the six platforms with grid software components and coordination tools: computing and storage hardware resources, but also database and software resources. The GRISBI project has been labelled by IBiSA in october 2008.

7.2.1.5. Sigenae and Genanimal

The SIGENAE program (Analysis of Breeding Animals' Genome), coordinated by Inra Toulouse, is the Inra national program of animal genomics. It aims at identifying the expressed part of genomes, developing the map-making of entire genomes and studying genetic diversity of breeding animals (pig, chicken, trout, cow). A privileged international partner is the american ARS (Agricultural Research Service) which develops a comparable project. The transcriptome of trout, chicken and pig are studied in Rennes.

Symbiose collaborates to this program via an Inra engineer, F. Moreews, contributing to the Sigenae information system. We are with UMR Agrocampus-Inra 598 on the modeling of fatty acids metabolism (see Sec. 6.3) and contract eQTL. QTL (Quantitative Trait Loci) are biomarkers of genomic regions responsible of a substantial part of variations deserved on a given character. The aim of the project eQTL is to relate QTL regions obtained by linkage analysis and regions obtained by transcriptomic studies, responsible of the regulation of a set of genes.

7.2.2. ANR Projects

7.2.2.1. PFTV 2007

Participants: Olivier Collin, Anthony Assi, Hugue Leroy.

Funding was given by the ANR in 2007 for the improvement of the computing infrastructure of the GenOuest platform. It allowed to add new computers with high level of memory to perform expensive calculations. The storage infrastructure was also extended with the addition of new storage blades for the Panasas servers. A few machines are also dedicated to hosted databases of GenOuest: GermOnline (INSERM), Autograph (CNRS) and M@IA (CNRS, University Rennes 1).

7.2.2.2. Proteus (Fold recognition and inverse folding problem)

Participants: Rumen Andonov, Guillaume Collet, Noël Malod-Dognin, François Coste.

The project PROTEUS (ANR-06-CIS6-008) started in January 2007 and involves also BIOS at Ecole Polytechnique (coordinator T. Simonson), MIG at INRA Jouy-en-Josas, the Physics Lab. at Ecole Normale Supérieure of Lyon and ABI at UPMC, Paris 6. The standard but difficult «fold recognition» problem requires identifying the 3D structure among a library of possible structures. A complementary approach turns the problem around, and poses the «inverse folding problem»: to enumerate all the amino acid sequences compatible with a given 3D structure. On the one hand, we will predict the fold of all bacterial proteins of unknown structure (300.000 proteins). On the other hand, we will solve the inverse folding problem for 1300 folds, out of 2300 known today (SCOP database), using the emerging technique of directed evolution, which mimics the natural evolutionary process. Reports of the project are available on its web site [12].

7.2.2.3. PARA: Parallelism and Improvement of Application Performances

Participants: Dominique Lavenier, Alexandre Cornu, Julien Jacques.

The aim of this ANR project is to study and develop optimization methods to better exploit all parallelism aspects coming from modern computers. In this project, the Symbiose team is involved in the optimization of intensive comparison algorithms, and their implementations on a reconfigurable accelerator. Scientific results are detailed in section 6.1.1.

7.2.2.4. Modulome: Identifying and displaying modules in genomic sequences

Participants: Catherine Belleannée, François Coste, Dominique Lavenier, Jacques Nicolas, Pierre Peterlongo, Christine Rousseau.

This ANR project, Modulome [13], aims at providing methods for the identification, visualization and formal modeling of the structure of genomes in terms of an assembly of nucleotides "modules" that are repeated along a genome or between several genomes. Three other teams of Biologists and bioinformaticians are involved in this project: LDGE (Dynamique du Génome et Evolution), Institut Jacques Monod, Paris; LEPG (Etude des Parasites Génétiques), Tours; LM2E (Microbiologie des Environnements Extrêmes), Brest See details in section 6.2.1.

[12] http://migale.jouy.inra.fr/proteus
[13] http://www.irisa.fr/symbiose/projets/Modulome/

7.2.2.5. Sitcon: Modeling signal transduction induced by a chimeric oncogene

Participants: Carito Guziolowski, Ovidiu Radulescu, Michel Le Borgne, Anne Siegel, Sylvain Blachon.

This ANR project belongs to the "Biologie Systémique" program. The Ewing inducible cellular model, developed by one of the biologist partners of the project, is characterized by a malignant genomic translocation and appearance of a chimeric gene EWS/FLI-1 whose activity leads to the uncontrolled cell growth. The goals of the projects are to reconstruct the corresponding interaction network, including signal transduction pathways and from a detailed model of functioning, to propose new validation experiments. See details in section 6.3.3

7.2.2.6. DyCoNum: Dynamical and Combinatorial studies of Numeration systems

Participant: Anne Siegel.

The "Jeunes chercheurs" program funded a project named DyCoNum aiming to consider by a transversal approach digital expansions in several number systems. This project focuses on integer base expansions, non-standard systems with integer base (signed digit expansions), beta-expansions and substitutive numeration systems, (generalized) continued fractions. This program involves W. Steiner and C. Frougny (LIAFA, Paris 7) and B. Adamczewski (Institut Camille Jordan, Lyon 1).

7.3. European and international initiatives

The main international teams we cooperate with are the following

- Argentina, Universidad Nacional de Córdoba Grammatical inference (F. Coste, M. Gallé)
- Bulgaria, IPP (Yavor Vutov) Protein structures (R. Andonov)
- Bulgaria, Sofia University, Protein structures (R. Andonov)
- China, Institute of Computing Technology, Beijing. Parallelization of bioinformatics algorithms onto multicore processors (D. Lavenier)
- Greece, Institute of Communication and Computer Systems, National Technical Univ. of Athens. Oncosimulator. (D. Lavenier, A. Assi, J. Jacques)
- Germany, Postdam university, Prof. T. Schaub's lab. Logic programming and boolean constraint solving. (J. Nicolas, T. Henin; C. Guziolowski).
- US, Stony Brook University, Drosophila developmental biology: J. Reinitz (O. Radulescu).
- Russia, St. Petersburg, mathematics: Sergei Vakulenko, modeling : V. Gursky, bioinformatics: M. Samsonova. (O. Radulescu).
- India, NCBS Bangalore, systems biology: Upi Bhalla, biophysics : M. Rao. (O. Radulescu).
- UK, Department of Mathematics, University of Leicester: A. Gorban (O. Radulescu, P. Blavy)

7.3.1. International programs

7.3.1.1. Integrated Project ACGT

Participants: Dominique Lavenier, Israël-César Lerman, Basavanneppa Tallur, Julien Jacques, Anthony Assi, Jacques Nicolas.

The project [14] aims at delivering the cancer research community an integrated CIT environment enabled by a powerful GRID infrastructure. Our contribution concerns parallelism (Grid development, tumor growth simulation) and data mining (integration of CHAVL in a R environment) (see Sec. 6.1.4).

7.3.1.2. Modularity, Abstraction, Robustness of Network Models in Molecular Biology

Participants: Ovidiu Radulescu, Pierre Blavy.

[14] http://eu-acgt.org/home.html

Principal investigators of this 2007-2008 Egide project are A.N. Gorban (Leicester, UK), O. Radulescu (Univ. Rennes, France), INRA Génétique animale (Rennes, Toulouse) and Institut Curie (Paris). It involves mathematics, molecular biology and bio-informatics. Its purpose is to develop mathematical techniques for the analysis of large-scale non-linear differential equations systems describing molecular biology networks. We address reduction tecniques and robustness isues (see Sec. 6.3.2), in the context of a simplified model for fatty acids metabolism including metabolic, genetic and hormonal regulations.

7.3.1.3. Application of combinatorial optimization (PHC RILA, Bulgaria)

Participants: Rumen Andonov, Nicola Yanev.

This program is managed by the French Ministry of Foreign Affairs [15]. The project focusses on the application of combinatorial optimization techniques in the domain of protein structure comparison/prediction. This domain is rich in NP-hard problems and the goal of the project is to propose and to analyse new mathematical models for solving these problems. Publications [35] have been partially supported by this program.

7.3.1.4. SECyT-INRIA-CNRS cooperation program (Argentina)

Participants: François Coste [correspondant], Matthias Gallé.

With G. Infante-Lopez, head of the *GPLN*, Universidad Nacional de Córdoba, we started a new project entitled "linguistic modeling of genomic sequences by grammatical inference" thanks to the international cooperation program *SECYT (Argentine)/CNRS-INRIA (France)*. It aims at studying how formalisms and grammatical inference methods developped for natural language processing can be adapted for genomic sequences. This includes the co-tutored PhD thesis of M. Gallé on learning context-free grammars (see section 6.2.3). M. Gallé spent 3 months in Argentina while F. Coste and G. Infante-Lopez spent 2 weeks in the other partner team. We have also worked on the definition of a new approach for unsupervised learning of derivation trees.

7.3.1.5. PHC Sakura and Amadeus

Anne Siegel is also implied in programs supported by PHC Sakura (Japan) and PHC Amadeus (Austria) dedicated to the study of the dynamical properties of expansions in non integer basis and their relations with fractal theory and discrete geometry.

7.3.2. Visiting scientists

International cooperation have allowed to fund the following visits.

7.3.2.1. Visitors

The following scientists visited the Symbiose project.

- prof B. Morgenstern (Univ. Göttingen, Mikrobiologie und Genetik) 3 days in june 2008.
- Professor G. Infante-Lopez (Univ. Nacional de Cordoba, Argentina) 2 weeks in oct. 2008.
- Professor S. Vakulenko (Institute of Print, St.Petersbourg, Russie) 1 month.
- M. V. (D. di Mecanica et Aeronautica, Univ. Roma, Italy) 1 week.
- Professor J. Adams (Cavendish Laboratory, Univ. of Cambridge, UK) 1 month.
- Post-Doc T. Maliekal (NCBS Bangalore, Inde) 1 week.
- Associated Professor L. Jun (Canton university, China) 4 days.
- Prof. N. Yanev from the Univ. of Sofia 2 months.
- Y. Vutov from the Department of Scientific Computations, Institute for Parallel Processing of the Bulgarian Academy of Sciences (http://parallel.bas.bg/SciComp/) 2 months

[15] http://www.egide.asso.fr/uk/programmes/

7.3.2.2. Scientific visit exchanges

- Bangalore, India, June 2008 [O. Radulescu]
- Cordoba university, Argentina, 2 weeks visiting funded by a SECyT-INRIA-CNRS cooperation program [F. Coste]
- Cordoba, Argentina, 3 months visiting (co-tutored PhD thesis) [M. Gallé]
- Kanasawa university and Keio university, 2 weeks visiting, PHC SAKURA program [A. Siegel]
- Postdam university, Germany. (co-tutored PhD thesis) 2 months [T. Henin].
- Sofia, Bulgaria, 2 weeks in Sofia [R. Andonov]
- Santiago de Chile, Chile, Center for Mathematical Modelling, 1 week visiting [A. Siegel]
- Sao Carlos, Brasil, Feb. 2008 [O. Radulescu]

8. Dissemination

8.1. Leadership within scientific community

8.1.1. Administrative functions: scientific commitees, journal bords, jury

- Editorial Board of *Mathématiques et Sciences Humaines, Mathematics and Social Sciences* [I.-C. Lerman]
- Member of the Scientific & Technical Council of the European Morpheus project [D. Lavenier]
- Member of the Evaluation Comitee of Inria [A. Siegel]
- Scientific Advisory Board of Ouest-genopole [O. Collin, J. Nicolas].
- Scientific Advisory Board of Bioinfapa, INRA (Bioinformatics for Animal genomics) [J. Nicolas].
- Scientific Advisory Board of Institut Génétique et développement, INSERM [J. Nicolas].

8.1.2. Jury of PhD Theses

- Referee, PhD Thesis of V. Derrien, *Heuristiques pour la résolution du problème d'alignement multiple de séquences*, march 2008 [R. Andonov]
- Referee, PhD Thesis of Y. Zhou, Univ Paris-Sud XI and Graduate School of Wuhan Univ. , *Application of RNA Bioinformatics in decoding RNA structure and regulation*, dec. 2008 [J. Nicolas]
- Referee, PhD Thesis of K. Maria Popovici, *Environnement de programmation Multi Niveau pour Architectures Hétérogènes MPSoC*, INP de Grenoble, feb. 2008 [D. Lavenier]
- Referee, PhD Thesis of S. Syed Yazdani, *Coordinated Concurrent Shared Memory Accesses on a Reconfigurable Multimedia Accelerator*, Université de Bretagne Occidentale, nov. 2008 [D. Lavenier]
- Jury, Habilitation thesis of A. Siegel, *Analyse de systèmes dynamiques par discrétisation. Exemples d'applications en théorie des nombres et en biologie moléculaire*, Université de Rennes 1, dec. 2008 [J. Nicolas].
- Jury, PhD Thesis of G. Kerbellec, *Apprentissage d'automates modélisant des familles de séquences protéiques*, June 2008 [F. Coste].
- Jury, Habilitation Thesis of V. Fleury, April 2008 [O. Radulescu]

8.2. Faculty teaching

Members of the Symbiose project are actively involved in the bioinformatics teaching program proposed by the University of Rennes 1. Rumen Andonov is in charge with the Master Research Degree in Computer Science (http://www.irisa.fr/master/) (he shares this responsibility with P. Sebillot). The members of Symbiose are also actively involved in the 4th and 5th year bioinformatics master degrees, with biologist colleagues from the Life Science department *Vie-Agro-Santé*. The originality of this 2 year training program lies in recruiting both biologists and computer scientists.

Besides usual teachings of the faculty members, the Symbiose project is involved in many programs:

1. ENSAI, Rennes (D. Lavenier, M. Gallé)
2. Formation permanente en bioinformatique, Université de Rennes 1 (N. Le Meur)
3. Magistere Informatique et Télécommunications, ENS Cachan, Ker Lann (D. Lavenier)
4. Master 1 & 2 Modeling biological systems. (R. Andonov, O. Radulescu, A. Siegel)
5. Master 2 Mathematics. (B. Tallur)
6. Master 2 Computer Science / Engineer diplom, IFSIC. (F. Coste, H. Leroy)
7. IRISA formation program [PERL] (F. Legeai)
8. ERASMUS Program Université de Rennes 1/ Université de Sofia, Bulgarie (R. Andonov)

8.3. Conference and workshop committees, invited conferences

8.3.1. International invited conferences

- Sao Carlos, Brésil, Feb. 2008, *What Drosophila gap genes patterns teach us about redundancy, robustness and epigenesis* [O. Radulescu]
- Journées CMM-INRIA, Santiago du Chili, *from questions in systemic biology to constraints resolution* [A. Siegel]

8.3.2. National invited conferences

- Troisièmes Journées Thématiques Apprentissage Artificiel et Fouille des Données, Apr. 2008, LIPN, Université Paris-Nord, Invited conference [I.-C. Lerman]
- CEA Saclay, *Architectures matérielles spécialisées pour la bioinformatique*, Feb. 2008 [D. Lavenier]
- Gen2Bio, the biotech symposium organised by OUEST-genopole, *Protomata-Learner : la rencontre entre alignement multiple et découverte de motifs* [F. Coste]
- INSERM workshop on gene regulation and noise, *Hierarchical modeling of the NFkB signaling module*, May 2008 [O. Radulescu]
- Atelier Substitutions et automorphismes de groupes libres. Porq'roll 2008 *Propriétés topologiques des fractals de Rauzy* [A. Siegel].
- Université de Rennes I, Séminaire de systèmes dynamiques, Feb. 2008, *Propriétés topologiques des fractals de Rauzy* [A. Siegel]
- Station biologique de Roscoff, Oct. 2008, *Aider le biologiste par une automatisation des raisonnements sur des systèmes* [A. Siegel]

8.3.3. Conference committees

- Biograle 2008, Lyon : BIOlogie à GRAnde echelLE (O. Collin, Program Committee)
- CAp 2008, Conférence Apprentissage (F. Coste, P.C.)
- EGC 2008: Extraction et Gestion des Connaissances (I.-C. Lerman, P.C.)
- ERSA: International Conference on Engineering of Reconfigurable Systems and Algorithms (D. Lavenier, P.C.).
- FPL: International Conference on Field Programmable Logic and Applications (D. Lavenier, P.C.).
- HiCOMB: Workshop on High Performance Computational Biology (D. Lavenier, P.C.)
- HPPC: Workshop on Highly Parallel Processing on a Chip (D. Lavenier, P.C.)
- ICGI: 9th International Colloquium on Grammatical Inference, (F. Coste, chair, see below)

- JOBIM, Lille, (A. Siegel, P.C.)
- RCeducation: Workshop on Reconfigurable Computing Education (D. Lavenier, P.C.)
- ReConfig: International Conference on Reconfigurable Computing and FPGAs ReConFig (D. Lavenier, P.C.)
- SPL: Southern Conference on Programmable Logic (D. Lavenier, P.C.)
- Sympa: Symposiun en Architecture de Machine (D. Lavenier, P.C.)
- Workshop Systems Biology of Signalling in Cancer, London (O. Radulescu, main organizer)

8.3.4. ICGI: 9th International Colloquium on Grammatical Inference

The ninth edition of ICGI, the major biennial international conference on grammatical inference, was organized in Saint Malo on September 22-24, 2008, by F. Coste (Conference Chair) helped by A. Clark, London Univ., and L. Miclet, Enssat / Univ. de Rennes 1 (Programme Committee co-chairs), M. Gallé and M.-N. Georgeault from Symbiose, and E. Lebret (Organization committee). The event was sponsored by PASCAL2 (Network of Excellence on Pattern Analysis, Statistical Modeling and Computational Learning), INRIA, Région Bretagne and University of Rennes 1.

The previous edition of ICGI were held in Tokyo (Japan) in 2006. ICGI'08 attracted 50 participants, mainly from France, Japan, Usa and Spain. The program proposed the presentation of 21 accepted papers (acceptation rate 57%) and 2 invited talks (F. Yvon,LIMSI, Paris and Y. Sakakibara (Keio University,Department of Biosciences and Informatics, Japan). The proceedings of ICGI-2008 are published by Springer-Verlag as volume 5278 in Lecture Notes in Articifial Intelligence subseries [59] and video-recorded (freely available on the VIDEOLECTURES.net website [16]).

8.3.5. Bioinformatics animations

Sixth meeting of the Bioinformatics platform of OUEST-genopole The meeting held at Irisa, Rennes on Oct. 21, 2008. It was organized by Olivier Collin. ([17]). Invited speakers included Philippe DESSEN (Directeur de Recherche - CNRS FRE2939) and Fabien PETEL (Chef de Projet "Carte d'identité des tumeurs®" - Ligue Nationale Contre le Cancer).

8.3.6. BioInfoOuest thematic-day conferences

The Symbiose project regularly organizes thematic-day conferences on bioinformatics subjects[18]. The public of this thematic-day is made of computer scientists as well as biologists. Usually, this public gathers 50 persons (with 50 % of biologists) coming from all western France.

- *Metabolic networks*. This conference day was organized by P. Peterlongo. Invited speakers were Gilles Curien (Laboratoire de Physiologie Cellulaire Végétale, CEA Grenoble), Vincent Lacroix (Genome Bioinformatics Research Group - CRG PRBB, Barcelona - Spain) and Vincent Schachter (CNRS-Genoscope-Université d'Evry).
- *WorkFlows*. This conference day was organized by D. Lavenier and F. Moreews. Invited speakers were Marc Wessner (Inra, Jouy), Franck Tanoh (myGrid Team, Manchester, UK) and Philippe Picouet (LUSSI, ENST Bretagne).
- *micro-RNA*. This conference day was organized by A. Siegel. Invited speakers were Daniel Gautheret (Univ. Paris XI), Stéphanie Jaubert and Denis Tagu (INRA, Rennes) and Stephan Ossowski (Max Planck Institute for Developmental Biology, Tübingen).

As a complement to these thematic days, the **team seminar** is held on a bi-weekly basis. 19 talks were given in this framework during the year 2008, including 4 international speakers.

[16] http://videolectures.net/icgi08_saint_malo/
[17] http://genoweb1.irisa.fr/Serveur-GPO/article.php3?id_article=622
[18] http://www.irisa.fr/sci-events/seminars/bioinfo

8.4. Theses defenses

8.4.1. PhD thesis, Goulven Kerbellec

Learning automata modeling families of protein sequences [11]. This thesis shows a new approach out of discovering protein families signatures. Given a sample of (unaligned) sequences belonging to a structural or functional family of proteins, this approach infers non-deterministic automata characterizing the family. A new kind of multiple alignment called PLMA is introduced in order to emphasize the partial and local significant similarities. Various experiments on protein families including MIP and TNF have been presented.

Keywords: Bioinformatics, grammatical inference, protein family, automata, multiple alignment.

The thesis was supervised by Francois Coste. The jury included L. Miclet (Univ. de Rennes 1), B. Morgenstern (Univ. Gottingen), H. Touzet (Univ. Lille), R. Andonov (Univ. Rennes 1), F. Coste (INRIA) and J. Pothier (Paris VI).

8.4.2. Habilitation thesis, Jacques Nicolas

Syntax, Reasoning and Genomes[12] The core of my research work consider sets of sequences representing macromolecules of life as formal languages and develop necessary concepts for their linguistic analysis.

This starts with lexical studies on the genomic vocabulary. Meaningful units are based on repeated words but natural repeats in biological sequences can be very complex as it appears clearly in mobile genetic elements. We have proposed a formalization and practical algorithms and tools for their study.

Syntactical analysis is the second stage and needs to design dedicated modeling languages. We have proposed Stan and Logol, languages based on string variables that has been applied on the search of proteins and genetic elements (human defensins, promoters, olfactive receptor genes).

Finally, when no model is available, it must be inferred from a sample of sequences. I have supervised various studies on context-free or regular inference and applied them on promoter and protein sequences.

8.4.3. Habilitation thesis, Anne Siegel

Analysing dynamical system with discrete approaches. Application in number theory and molecular biology. [13] This work presents theoretical and practical contributions to the theory of symbolic codings of dynamical systems. Available knowledge on studied systems is introduced via algorithmic approaches and discretization methods.

The first type of system we adress consist in toral autormorphisms and translations. Inspired by the one-dimensional case (beta-numeration and sturmian sequences), we exhibit a toral fundamental domain and a partition of this domain in which the dynamics is coded by simple symbolic systems. We used a decidable approach to describe the fractal boundary of the domain in the Pisot case and its tiling properties and we derived criteria to decide whether it generates a good coding. From their topological properties we deduced applications in number theory (characterization of rational numbers with a finite or purely periodic expansion in a non-integer basis) and discrete geometry (condition for the iterative generation of discrete planes).

The second class of question we address in the framework of large-scale dynamical systems in molecular biology is how to produce information on a system that is only partially known and observed. We introduced a formalism to interpret observations in molecular biology, with a dynamical background. The main purpose was to guide model correction and, in the future, experiment design. Qualitative aspects are replaced by properties of shift equilibria, so that analyzing data is reduced to writing and then solving constraints on discrete sets.

9. Bibliography

Major publications by the team in recent years

[1] R. ANDONOV, S. BALEV, N. YANEV. *Protein Threading Problem: From Mathematical Models to Parallel Implementations*, in "INFORMS Journal on Computing", vol. 16, n^{o} 4, 2004, p. 393–405.

[2] N. BEN ZACOUR, M. GAUTIER, R. ANDONOV, D. LAVENIER, P. VEBER, A. SOROKIN, Y. LELOIR. *GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification*, in "Nucleic Acid Research", vol. 32, n^{o} 1, 2004.

[3] P. DURAND, F. MAHÉ, A.-S. VALIN, J. NICOLAS. *Browsing repeats in genomes: Pygram and an application to non-coding region analysis*, in "BMC Bioinformatics", vol. 7, 2006, 477, http://www.biomedcentral.com/content/pdf/1471-2105-7-477.pdf.

[4] S. GUYETANT, M. GIRAUD, L. L'HOURS, S. DERRIEN, S. RUBINI, D. LAVENIER, F. RAIMBAULT. *Cluster of re-configurable nodes for scanning large genomic banks*, in "Parallel Computing", vol. 31, n^{o} 1, 2005.

[5] C. GUZIOLOWSKI, J. GRUEL, O. RADULESCU, A. SIEGEL. *Curating a large-scale regulatory network by evaluating its consistency with expression datasets*, in "CIBB 2008 : computational Intelligence Methods for Bioinformatics and Biostatistics, Salerno Italie", 2008, http://hal.inria.fr/inria-00330589/en/.

[6] J. NICOLAS, P. DURAND, G. RANCHY, S. TEMPEL, A.-S. VALIN. *Suffix-Tree ANalyser (STAN): looking for nucleotidic and peptidic patterns in chromosomes.*, in "Bioinformatics", oct. 2005, http://dx.doi.org/10.1093/bioinformatics/bti710.

[7] P. QUIGNON, M. GIRAUD, M. RIMBAULT, P. LAVIGNE, S. TACHER, E. MORIN, E. RETOUT, A.-S. VALIN, K. LINDBLAD-TOH, J. NICOLAS, F. GALIBERT. *The dog and rat olfactory receptor repertoires*, in "Genome Biology", vol. 6, n^{o} 10, 2005, R83.

[8] S. TEMPEL, J. NICOLAS, A. EL AMRANI, I. COUÉE. *Model-based Identification of Helitrons Results in a New Classification of Their Families in Arabidopsis thaliana*, in "Gene", vol. 403, 2007, http://hal.inria.fr/inria-00180376/en/.

[9] P. VEBER, C. GUZIOLOWSKI, M. LE BORGNE, O. RADULESCU, A. SIEGEL. *Inferring the role of transcription factors in regulatory networks*, in "BMC Bioinformatics", vol. 9, 2008, http://hal.inria.fr/inria-00330578/en/.

[10] N. YANEV, R. ANDONOV, P. VEBER, S. BALEV. *Lagrangian Approaches for a class of Matching Problems in Computational Biology*, in "Computers & Mathematics with Applications", vol. 55, 2008, p. 1054-1067, http://hal.inria.fr/inria-00336582/en/.

Year Publications

Doctoral Dissertations and Habilitation Theses

[11] G. KERBELLEC. *Apprentissage d'automates modélisant des familles de séquences protéiques*, Ph. D. Thesis, Université Rennes 1, http://tel.archives-ouvertes.fr/tel-00327938/en/.

[12] J. NICOLAS. *Syntaxe, raisonnement et génomes*, HDR, Université Rennes 1, http://tel.archives-ouvertes.fr/tel-00355156/en/.

[13] A. SIEGEL. *Analyse de systèmes dynamiques par discrétisation. Exemples d'applications en théorie des nombres et en biologie moléculaire*, HDR, Université Rennes 1, http://tel.archives-ouvertes.fr/tel-00358996/en/.

Articles in International Peer-Reviewed Journal

[14] S. AKIYAMA, G. BARAT, V. BERTHE, A. SIEGEL. *Boundary of central tiles associated with Pisot beta-numeration and purely periodic expansions*, in "Monatshefte für Mathematik", vol. 155, 2008, p. 377-419, http://hal.inria.fr/inria-00330567/en/.

[15] S. BALEV, N. YANEV, A. FRÉVILLE, R. ANDONOV. *A dynamic programming based reduction procedure for the multidimensional 0-1 knapsack problem*, in "European Journal of Operational Research", vol. 186, 2008, p. 63-76, http://hal.inria.fr/inria-00184771/en/.

[16] M. BARRET, P. FREY-KLETT, M. BOUTIN, A.-Y. GUILLERM-ERCKELBOUDT, F. MARTIN, L. GUILLOT, A. SARNIGUET. *The plant pathogenic fungus Gaeumannomyces graminis var. tritici improves bacterial growth and triggers early gene regulations in the biocontrol strain Pseudomonas fluorescens Pf29Arp*, in "New Phytologist", vol. 181, n^{o} 2, 2009, p. 435-447, http://hal.inria.fr/inria-00359115/en/.

[17] O. FILANGI, Y. BEAUSSE, A. ASSI, L. LEGRAND, J.-M. LARRE, V. MARTIN, O. COLLIN, C. CARON, H. LEROY, D. ALLOUCHE. *BioMAJ: a flexible framework for databanks synchronization and processing*, in "Bioinformatics", vol. 24, 2008, p. 1823-1825, http://hal.inria.fr/inria-00327502/en/.

[18] A. GORBAN, O. RADULESCU. *Dynamic and static limitation in multiscale reaction networks, revisited*, in "Chemical Engineering Science", vol. 34, 2008, p. 103-173, http://hal.inria.fr/inria-00184976/en/.

[19] J. KLÉMA, S. BLACHON, A. SOULET, B. CRÉMILLEUX, O. GANDRILLON. *Constraint-based knowledge discovery from SAGE data*, in "In Silico Biology", vol. 8, 2008, p. 157-75, http://hal.inria.fr/inria-00351641/en/.

[20] G. LAUNAY, T. SIMONSON. *Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations*, in "BMC Bioinformatics", 2008, http://hal.inria.fr/inria-00338202/en/.

[21] C. LAVELLE, H. BERRY, G. BESLON, F. GINELLI, J.-L. GIAVITTO, Z. KAPOULA, A. LE BIVIC, N. PEYRIERAS, O. RADULESCU, A. SIX, V. THOMAS-VASLIN, P. BOURGINE. *From Molecules to Organisms: Towards Multiscale Integrated Models of Biological Systems*, in "Theoretical Biology Insights", vol. 1, 2008, p. 13-22, http://hal.inria.fr/inria-00331281/en/.

[22] D. LAVENIER, X. JIANG, S. YAU. *Coding Region Prediction Based on a Universal DNA Sequence Representation Method*, in "Journal of Computational Biology", vol. 15, 2008, p. 1237-1256, http://hal.inria.fr/inria-00347594/en/.

[23] N. LE MEUR, R. GENTLEMAN. *Modeling synthetic lethality*, in "Genome Biology", vol. 9, 2008, R135, http://hal.inria.fr/inria-00337291/en/.

[24] F. LEGEAI, E. PAUX, N. GUILHOT, A.-F. ADAM-BLONDON, M. ALAUX, J. SALSE, P. SOURDILLE, P. LEROY, C. FEUILLET. *Physical mapping in large genomes: accelerating anchoring of BAC contigs to genetic maps through in silico analysis.*, in "Funct Integr Genomics.", vol. 8, 2008, p. 29-32, http://hal.inria.fr/inria-00329085/en/.

[25] I.-C. LERMAN. *Analyse logique, combinatoire et statistique de la construction d'une hiérarchie implicative ; niveaux et noeuds significatifs*, in "Revue Mathématiques et sciences Humaines Mathematics and Social Sciences", 2009, http://hal.inria.fr/inria-00323840/en/.

[26] J. LEYRITZ, S. SCHICKLIN, S. BLACHON, C. KEIME, C. ROBARDET, J.-F. BOULICAUT, J. BESSON, R. PENSA, O. GANDRILLON. *SQUAT: A web tool to mine human, murine and avian SAGE data*, in "BMC Bioinformatics", vol. 9, 2008, 378, http://hal.inria.fr/inria-00351639/en/.

[27] M. MOROLDO, S. PAILLARD, R. MARCONI, F. LEGEAI, A. CANAGUIER, C. CRUAUD, V. DE BERARDINIS, C. GUICHARD, V. BRUNAUD, I. LE CLAINCHE, S. SCALABRIN, R. TESTOLINI, G. DI GASPERO, M. MORGANTE, A.-F. ADAM-BLONDON. *A physical map of the heterozygous grapevine 'Cabernet Sauvignon' allows mapping candidate genes for disease resistance*, in "BMC Plant Biology 2008, 8:66doi:10.1186/1471-2229-8-66", vol. 8, 2008, http://hal.inria.fr/inria-00329169/en/.

[28] P. PETERLONGO, L. NOÉ, D. LAVENIER, V. H. NGUYEN, G. KUCHEROV, M. GIRAUD. *Optimal neighborhood indexing for protein similarity search*, in "BMC Bioinformatics", vol. 9, 2008, http://hal.inria.fr/inria-00340510/en/.

[29] V. POIRRIEZ, N. YANEV, R. ANDONOV. *A Hybrid Algorithm for the Unbounded Knapsack Problem*, in "Discrete Optimization", vol. 6, 2009, p. 110-124, http://hal.inria.fr/inria-00335065/en/.

[30] O. RADULESCU, A. GORBAN, A. ZINOVYEV, A. LILIENBAUM. *Robust simplifications of multiscale biochemical networks*, in "BMC Systems Biology", vol. 2:86, 2008, http://hal.inria.fr/inria-00331212/en/.

[31] P. VEBER, C. GUZIOLOWSKI, M. LE BORGNE, O. RADULESCU, A. SIEGEL. *Inferring the role of transcription factors in regulatory networks*, in "BMC Bioinformatics", vol. 9, 2008, http://hal.inria.fr/inria-00330578/en/.

[32] N. YANEV, R. ANDONOV, P. VEBER, S. BALEV. *Lagrangian Approaches for a class of Matching Problems in Computational Biology*, in "Computers & Mathematics with Applications", vol. 55, n^{o} 5, 2008, p. 1054-1067, http://hal.inria.fr/inria-00336582/en/.

Articles in National Peer-Reviewed Journal

[33] I.-C. LERMAN, K. BACHAR. *Comparaison de deux critères en classification ascendante hiérarchique sous contrainte de contiguïté, Application en imagerie numérique*, in "Journal de la Société Française de Statistique & Revue de Statistique Appliquée", vol. 149, 2008, p. 45-74, http://hal.inria.fr/inria-00322083/en/.

Articles in Non Peer-Reviewed Journal

[34] D. LAVENIER. *RDISK : une architecture reconfigurable pour l'exploration des banques génomiques*, in "Techniques de l'Ingenieur", vol. November, 2008, http://hal.inria.fr/inria-00347596/en/.

International Peer-Reviewed Conference/Proceedings

[35] R. ANDONOV, N. YANEV, N. MALOD-DOGNIN. *An Efficient Lagrangian Relaxation for the Contact Map Overlap Problem*, in "WABI '08: Proceedings of the 8th international workshop on Algorithms in Bioinformatics", 2008, p. 162–173, http://hal.inria.fr/inria-00327135/en/.

[36] M. GALLÉ, P. PETERLONGO, F. COSTE. *In-place Update of Suffix Array while Recoding Words*, in "Prague Stringology Conference 2008", J. HOLUB, J. ZDÁREK (editors), 2008, p. 54–67, http://hal.inria.fr/inria-00327582/en/.

[37] C. GUZIOLOWSKI, J. GRUEL, O. RADULESCU, A. SIEGEL. *Curating a large-scale regulatory network by evaluating its consistency with expression datasets*, in "CIBB 2008 : computational Intelligence Methods for Bioinformatics and Biostatistics", 2008, http://hal.inria.fr/inria-00330589/en/.

[38] P. KUNTZ, I.-C. LERMAN. *Directed binary hierarchies and directed ultrametrics*, in "First joint meeting of the French and Italian Classification Societies", U. DE NAPLES (editor), 2008, p. 337-340, http://hal.inria.fr/inria-00322110/en/.

[39] B. LASNE, O. RADULESCU, S. LEROUGE. *Timescales and instabilities of shear thinning solutions of wormlike micelles*, in "International Congress on Rheology 2008", 2008, http://hal.inria.fr/inria-00331217/en/.

[40] D. LAVENIER. *Ordered Index Seed Algorithm for Intensive DNA Sequence Comparison*, in "HiCOMB 2008 : Seventh IEEE International Workshop on High Performance Computational Biology", online proceeding : http://www.hicomb.org/HiCOMB2008/, 04 2008, http://hal.archives-ouvertes.fr/hal-00322696/en/.

[41] D. LAVENIER, J. JACQUES. *Parallelizing the ACGT OncoSimulator*, in "3rd International Advanced Research Workshop on In Silico Oncology", online proceedings: http://www.3rd-iarwiso.iccs.ntua.gr/, 09 2008, http://hal.archives-ouvertes.fr/hal-00322703/en/.

[42] F. LEGEAI, J.-P. GAUTHIER, D. TAGU. *Apollo: a tool for the manual annotation of the pea aphid genome*, in "Pea Aphid Genome Annotation Workshop", 2008, http://hal.inria.fr/inria-00329196/en/.

[43] V. H. NGUYEN, D. LAVENIER. *Parallélisation de la recherche de similarités entre séquences protéiques sur GPU*, in "RenPar'18 : Rencontres francophones du Parallélisme", 2008, http://hal.inria.fr/inria-00321456/en/.

[44] V. H. NGUYEN, D. LAVENIER. *Speeding up Subset Seed Algorithm for Intensive Protein Sequence Comparison*, in "6th IEEE International Conference on research, innovation & vision for the future", 2008, http://hal.inria.fr/inria-00321457/en/.

[45] X. YE, V. H. NGUYEN, D. LAVENIER, D. FAN. *Efficient Parallelization of a Protein Sequence Comparison Algorithm on Manycore Architecture*, in "Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies", online proceedings: http://www.cs.otago.ac.nz/pdcat08/, 12 2008, http://hal.archives-ouvertes.fr/hal-00322732/en/.

National Peer-Reviewed Conference/Proceedings

[46] P. BLAVY, F. GONDRET, H. GUILLOU, S. LAGARRIGUE, P. MARTIN, O. RADULESCU, A. SIEGEL, J. VAN MILGEN. *A minimal and dynamical model for fatty acid metabolism in mouse liver*, in "JOBIM : Journées Ouvertes en Biologie Informatique et Mathématiques", 2008, http://hal.inria.fr/inria-00330590/en/.

[47] I.-C. LERMAN. *Analyse de la Vraisemblance des Liens Relationnels*, in "3èmes Journées Thématiques Apprentissage Artificiel et Fouille des Données", 2008, http://hal.inria.fr/inria-00325174/en/.

[48] V. H. NGUYEN, D. LAVENIER. *Comparaison intensive des séquences protéiques : indexation des banques et usage des instructions SIMD des microprocesseurs*, in "Journées Ouvertes Biologie Informatique Mathématiques - JOBIM", 2008, http://hal.inria.fr/inria-00325020/en/.

[49] M. WESSNER, M. SENGER, F. SAMSON, P. PICOUET, F. MOREEWS, H. MÉNAGER, V. MARTIN, S. LETORT, C. LETONDAL, M. HOEBEKE, J. GOUZY, J.-F. GIBRAT, E. CORRE, O. COLLIN, S. CARRERE, C. CARON, P. TUFFÉRY, B. NÉRON. *BioWorkFlow: Web Services toolkit and workflow applications evaluation to deploy a confidence network*, in "Jobim (Journées Ouvertes Biologie Informatique Mathématiques)", 2008, http://hal.inria.fr/inria-00327528/en/.

Workshops without Proceedings

[50] C. DÉSERT, P. BLAVY, F. MOREEWS, M. DUCLOS, P. LEROY, M. DOUAIRE, C. DIOT. *Transcriptome Profiling Of Feeding-To-Fasting Transition In Chicken Liver Using A Chicken 20K Oligo Microarray*, in "International Plant & Animal Genomes XVI Conference", 2008, http://hal.inria.fr/inria-00329926/en/.

[51] J. GRUEL, N. LE MEUR, M. LE BORGNE, N. THÉRET. *In silico investigation of ADAM12 effect on TGF-beta receptors trafficking*, in "IPG 2008", 2008, http://hal.inria.fr/inria-00337265/en/.

[52] C. KLOPP, F. MOREEWS, M. AUBRY, S. LAGARRIGUE. *Contribution To The Annotations Of The Chicken 20K Oligo Microarray Of ARK-Genomics*, in "International Plant & Animal Genomes XVI Conference", 2008, http://hal.inria.fr/inria-00329915/en/.

[53] N. MALOD-DOGNIN, R. ANDONOV, N. YANEV, J.-F. GIBRAT. *Modèle de PLNE pour la recherche de cliques de poids maximal*, in "ROADEF 2008", 2008, http://hal.inria.fr/inria-00327118/en/.

[54] F. MOREEWS. *GEO EXPORTER : a tools to automate publishing of microarray data to Gene Expression Omnibus*, in "ISMB 2008", ISCB (editor), 2008, http://hal.inria.fr/inria-00329898/en/.

[55] G. RIZK, D. LAVENIER. *GPU Accelerated RNA-RNA Interaction Algorithm*, in "EMBnet", 2008, http://hal.inria.fr/inria-00323566/en/.

[56] A. SIEGEL. *Topological properties of central tiles and boundary graphs*, in "Number Theory and Ergodic Theory", 2008, http://hal.inria.fr/inria-00330588/en/.

Scientific Books (or Scientific Book chapters)

[57] R. ANDONOV, G. COLLET, J.-F. GIBRAT, A. MARIN, V. POIRRIEZ, N. YANEV. *Recent advances in solving the protein threading problem*, in "Grid Computing for Bioinformatics and Computational Biology", E.-G. TALBI, A. Y. ZOMAYA (editors), Wiley, 2008, http://hal.inria.fr/inria-00180695/en/.

Books or Proceedings Editing

[58] D. CAUCAL, A. SIEGEL (editors). *Special Issue: Journées Montoises d'informatique théorique (Rennes, 2006)*, vol. RAIRO - Theoretical Informatics and Applications 42(3), 2008, http://hal.inria.fr/inria-00330569/en/.

[59] A. CLARK, F. COSTE, L. MICLET (editors). *Grammatical Inference: Algorithms and Applications 9th International Colloquium, ICGI 2008 Saint-Malo, France, September 22-24, 2008 Proceedings*, vol. Lecture Notes in Artificial Intelligence 5278, 2008, http://hal.inria.fr/inria-00338208/en/.

Research Reports

[60] O. BALDELLON. *Alignements multiples locaux et partiels de séquences protéiques à partir de paires de fragments alignables*, Technical report, 2008, http://hal.inria.fr/inria-00327591/en/.

[61] A. CORNU, F. DUSSAUGEY, D. LAVENIER. *Parallel Reconfigurable Operator for Genomic Sequence Comparison: Architecture and Performance Analyses*, Technical report, 2008, http://hal.inria.fr/inria-00348482/en/.

[62] F. LEGEAI. *bioperl : Bibliothèque perl pour la modélisation et le traitement de données issues de la biologie*, Technical report, 2008, http://hal.inria.fr/inria-00331279/en/.

[63] V. H. NGUYEN, D. LAVENIER. *Fine-grained parallelization of similarity search between protein sequences*, Technical report, 2008, http://hal.inria.fr/inria-00275250/en/.

[64] J. NICOLAS, C. ROUSSEAU, A. SIEGEL, P. PETERLONGO, F. COSTE, P. DURAND, S. TEMPEL, A.-S. VALIN, F. MAHÉ. *Modeling local repeats on genomic sequences*, Technical report, 2008, http://hal.inria.fr/inria-00353690/en/.

Other Publications

[65] M. DAROUICH, S. GUYETANT, D. LAVENIER. *Architecture flexible pour la stéréovision embarquée*, 2008, http://hal.inria.fr/inria-00329460/en/.

References in notes

[66] R. GRAS, E. SUZUKI, F. GUILLET, F. SPAGNOLO (editors). *Statistical Implicative Analysis, Studies in Computational Intelligence*, Springer, 2008.

[67] R. ANDONOV, D. LAVENIER, N. YANEV, P. VEBER. *Dynamic programming for LR-PCR segmention of bacterium genomes*, in "HiComb 2004: Third IEEE International Workshop on High Performance Computational Biology, Santa Fe, New Mexico, USA", 2004.

[68] J. ANGELI, J. J. FERRELL, E. SONTAG. *Detection of multi-stability, bifurcations, and hysteresis in a large class of biological positive-feedback systems*, in "PNAS", 2004, p. 1822-1827.

[69] M. BANSAL, V. BELCASTRO, A. AMBESI-IMPIOMBATO, D. DI BERNARDO. *How to infer gene networks from expression profiles*, in "Mol Syst Biol.", vol. 3, n^{o} 78, 2007.

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

[71] N. BEN ZACOUR, M. GAUTIER, R. ANDONOV, D. LAVENIER, P. VEBER, A. SOROKIN, Y. LE LOIR. *GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification*, in "Nucleic Acid Research", vol. 32, n^{o} 1, 2004.

[72] A. BRAZMA, I. JONASSEN, I. EIDHAMMER, D. GILBERT. *Efficient discovery of conserved patterns using a pattern graph.*, in "Cabios", n^{o} 13, 1997, p. 509-522.

[73] A. BRAZMA, I. JONASSEN, I. EIDHAMMER, D. GILBERT. *Approaches to the Automatic Discovery of Patterns in Biosequences*, in "Journal of Computational Biology", vol. 5, n^{o} 2, 1998, p. 277-304.

[74] J. BUHLER, M. TAMPA. *Findind motifs using random projections*, in "Proceedings of RECOMB01, Montreal, Canada", ACM Press, 2001, p. 69-76.

[75] L. CALZONE, N. CHABRIER-RIVIER, F. FAGES, S. SOLIMAN. *A Machine Learning Approach to Biochemical reaction Rules Discovery*, in "Proceedings of Foundations of Systems Biology in Engineering'05, Santa-Barbara", 2005.

[76] N. CHABRIER-RIVIER, M. CHIAVERINI, V. DANOS, F. FAGES, V. SCHÄCHTER. *Modeling and querying biomolecular interaction networks*, in "Theor. Comp. Sci.", vol. 325, n^{o} 1, 2004, p. 25-44.

[77] M. CHAVES, R. ALBERT, E. SONTAG. *Robustness and fragility of Boolean models for genetic regulatory networks*, in "J. Theor. Biol.", vol. 235, 2005, p. 431-449.

[78] E. CHOW, T. HUNKAPILLER, J. PETERSON. *Biological Information Signal Processor*, in "ASAP", 1991, p. 144-160.

[79] J. COLLADO-VIDES. *A Transformational-Grammar Approach to the Study of The Regulation of Gene Expression*, in "J. Theor. Biol.", vol. 13, n^{o} 6, 1989, p. 403-425.

[80] G. COLLET, N. YANEV, A. MARIN, R. ANDONOV, J.-F. GIBRAT. *A flexible model for protein fold recognition*, in "Journées Ouvertes Biologie, Informatique et Mathématiques, Bordeaux France", 2006, http://hal.inria.fr/inria-00185460/en/.

[81] M. COVERT, E. KNIGHT, J. REED, M. HERRGARD, B. PALSSON. *Integrating high-throughput and computational data elucidates bacterial networks*, in "Nature", vol. 429, n^{o} 6987, 2004, p. 92-6.

[82] M. COVERT, B. PALSSON. *Transcriptional regulation in constraints-based metabolic models of Escherichia coli*, in "J biol chem", vol. 277, n^{o} 31, 2002, p. 28058-28064.

[83] P. CRESCENZI, V. KANN. *Approximation on the web: A compendium of NP optimization problems*, in "Randomization and Approximation Techniques in Computer Science", Lecture Notes in Computer Science 1269/1997, 1997, p. 111-118.

[84] H. DE JONG. *Modeling and simulation of genetic regulatory Systems: A literature review*, in "Journal of Computational Biology", vol. 9, n^{o} 1, 2002, p. 69-105.

[85] H. DE JONG, J.-L. GOUZÉ, C. HERNANDEZ, M. PAGE, T. SARI, J. GEISELMANN. *Qualitative simulation of genetic regulatory networks using piecewise-linear models.*, in "Bulletin of Mathematical Biology", vol. 66, 2004, p. 301–340.

[86] S. DONG, D. SEARLS. *Gene structure prediction by linguistic methods*, in "Genomics", vol. 23, 1994, p. 540-551.

[87] J. EDWARDS, B. PALSSON. *The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities*, in "Proc Natl Acad Sci U S A", vol. 97, n^{o} 10, May 2000, p. 5528-33.

[88] R. EISENTHAL, A. CORNISH-BOWDEN. *Prospects for antiparasitic drugs: the case of Trypanasoma brucei, the causative agent of African sleeping sickness*, in "J. Biol. Chem", vol. 272, 1998, p. 5500-5505.

[89] N. FRIEDMAN, D. KOLLER. *Being Bayesian about Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks*, in "Machine Learning", vol. 50, 2003, p. 95-126.

[90] S. GHOSH, M. KARIN. *Missing pieces in the NF-κB puzzle*, in "Cell", vol. 109, 2002, p. S81-96.

[91] R. GHOSHN, C. ANDOMLIN. *Symbolic Reachable Set Computation of Piecewise Affine Hybrid Automata and its Application to Biological Modelling: Delta-Notch Protein Signalling*, in "Systems Biology", vol. 1, n^{o} 1, 2004, p. 170-183.

[92] E. GLEMET, J. CODANI. *LASSAP: a LArge Scale Sequence compArison Package,*, in "Cabios", vol. 13, n^{o} 2, 1997, p. 137-143.

[93] D. GOLDMAN, S. ISRAIL, C. PAPADIMITRIU.. *Algorithmic Aspects of Protein Structure Similarity*, in "FOCS '99: Proceedings of the 40th Annual Symposium on Foundations of Computer Science, Washington, DC, USA", IEEE Computer Society, 1999, p. 512–521.

[94] P. GUERDOUX-JAMET, D. LAVENIER. *Systolic Filter for fast DNA Similarity Search*, in "ASAP'95, International Conference on Application Specific Array Processors, Strasbourg, France", 1995.

[95] P. GUERDOUX-JAMET, D. LAVENIER. *SAMBA: Hardware Accelerator for Biological Sequence Comparison*, in "CABIOS", vol. 13, n^{o} 6, 1997, p. 609-615.

[96] C. GUZIOLOWSKI, P. VEBER, M. LE BORGNE, O. RADULESCU, A. SIEGEL. *Checking Consistency Between Expression Data and Large Scale Regulatory Networks: A Case Study*, in "The Journal of Biological Physics and Chemistry", vol. 7, 2007, p. 37-43, http://hal.inria.fr/inria-00178914/en/.

[97] T. HEAD. *Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviours*, in "Bull. Math. Biology", vol. 49, 1987, p. 737-759.

[98] R. HEINRICH, S. SCHUSTER. *The Regulation of Cellular Systems*, Chapman and Hall, New York, 1996..

[99] J. HENIKOFF, S. HENIKOFF. *BLOCKs database and its applications*, in "Methods Enzymol.", vol. 266, 1996, p. 88-105.

[100] J. HUDAK, M. MCCLURE. *A comparative analysis of computational motif-detection methods*, in "Pacific Symposium of Biocomputing PSB 1999", 1999, p. 138-139.

[101] N. JAMSHIDI, S. JEREMY, J. EDWARD, T. FAHLAND, G. CHURCH, B. PALSSON. *Dynamic simulation of the human red blood cell metabolic network.*, in "Bioinformatics", vol. 17, 2001, p. 286-287.

[102] M. KAERN, T. A. ELSTON, W. J. BLAKE, J. J. COLLINS. *Stochasticity in gene expression: from theories to phenotypes*, in "Nature Rev.Genet.", vol. 6, 2005, p. 451-464.

[103] L. KARI, G. PAUN, G. ROZENBERG, A. SALOMAA, S. YU. *DNA computing, Sticker systems and universality*, in "Acta Informatica", vol. 35, 1998, p. 401-420.

[104] S. KAUFFMAN. *The origin of order, self-organisation and selection in evolution*, Oxford University Press, Oxford, U.K., 1993.

[105] V. KEICH, A. PEVZNER. *Findind motifs in the twilight zone*, in "Proceedings of RECOMB02, Washington, USA", ACM Press, 2002, p. 195-203.

[106] R. KING, S. GARRETT, G. COGHILL. *On the use of qualitative reasoning to simulate and identify metabolic pathways*, in "Bioinformatics", vol. 21, n^{o} 9, 2005, p. 2017-2026.

[107] R. LATHROP. *The protein threading problem with sequence amino acid interaction preferences is NP-complete*, in "Protein Engineering", vol. 255, 1994, p. 1059-1068.

[108] C. E. LAWRENCE, S. ALTSCHUL, M. S. BOGUSKI, J. S. LIU, A. F. NEUWALD, J. C. WOOTTON. *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.*, in "Science", vol. 262, 1993, p. 208-214.

[109] T. LENGAUER. *Bioinformatics. From genoms to Drugs*, Wiley-VCH, 2002.

[110] I.-C. LERMAN, J. AZÉ. *A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link*, in "Quality Measures in Data Mining. Studies in Computational Intelligence", FABRICE. GUILLET, H. HAMILTON (editors), Springer, 2007, p. 207-236, http://hal.inria.fr/inria-00180117/en/.

[111] S. NEEDLEMAN, C. WUNSCH. *A general method applicable to the search of similarities in the amino acid sequences of two protein,*, in "J. Mol. Biol.", vol. 48, 1970, p. 443-453.

[112] D. E. NELSON, AL.. *Oscillations in NF-κB Signaling Control the Dynamics of Gene Expression*, in "Science", vol. 306, 2004, p. 704-708.

[113] A. PANJKOVICH, F. MELO, M. MARTI-RENOM. *Evolutionnary potentials: structure specific knowledge-based potentials exploiting the evolutionary record of sequence homologs*, in "Genome Biology", vol. 9, 2008, R68.

[114] J. PAPIN, J. STELLING, N. PRICE, S. KLAMT, S. SCHUSTER, B. PALSSON. *Comparison of network-based pathway analysis methods*, in "Trends in Biotechnology", vol. 22, 2004, p. 400-405.

[115] G. PAUN, G. ROZENBERG, A. SALOMAA. *DNA Computing. New Computing Paradigms*, Springer-Verlag, 1998.

[116] P. REISER, R. KING, D. KELL, S. MUGGLETON, C. BRYANT, S. OLIVER. *Developing a Logical Model of Yeast Metabolism*, in "Electronic Transaction in Artificial Intellingence", vol. 5, 2001, p. 223-244.

[117] M.-F. SAGOT, A. VIARI. *A Double Combinatorial Approach to Discovering Patterns in Biological Sequences*, in "Proceedings of the7th Annual Symposium on Combinatorial Pattern Matching, Laguna Beach, CA", D. S. HIRSCHBERG, E. W. MYERS (editors), 1075, Springer-Verlag, Berlin, 1996, p. 186-208.

[118] Y. SAKAKIBARA. *Recent advances of grammatical inference*, in "Theoretical Computer Science", vol. 185, 1997, p. 15-45.

[119] L. SANCHEZ, D. THIEFFRY. *A logical analysis of the Drosophila gap-gene system*, in "J. Theor. Biol.", vol. 211, n^o 115-141, 2001.

[120] D. B. SEARLS. *String Variable Grammar: A Logic Grammar Formalism for the Biological Language of DNA*, in "Journal of Logic Programming", vol. 24, n^o 1/2, 1995, p. 73-102.

[121] D. SEARLS. *Formal language theory and biological macromolecules*, in "Theoretical Computer Science", vol. 47, 1999, p. 117-140.

[122] T. SMITH, M. WATERMAN. *Identification of common molecular subsequences*, in "J. Mol. Biol.", n^o 147, 198, p. 195-197.

[123] E. SNOUSSI. *Necessary conditions for multistationnarity and stable periodicity*, in "J. Biol. Syst.", vol. 6, 1998, p. 1-23.

[124] D. STATES, W. GISH, S. ALTSCHUL. *Basic local alignment search tool,*, in "J. Mol. Biol.", vol. 215, 1990, p. 403-410.

[125] Y. TAMADA, S. KIM, H. BANNAI, S. IMOTO, K. TASHIRO, S. KUHARA, S. MIYANO. *Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection*, in "Proceedings of the ECCB'03 conference", 2003.

[126] M. TOMITA, K. HASHIMOTO, K. TAKAHASHI, T. SHIMUZU, Y. MATSUZAKI, F. MIYOSHI, K. SAITO, S. TANIDA, K. YUGI, J. VENTER, J. HUTCHINSON. *E-CELL:software environment of whole-cell simulation*, in "Bioinformatics", vol. 15, 1999, p. 72-84.

[127] J. J. TYSON, C. CHEN, B. NOVÁK. *Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell*, in "Curr. Opinion Cell Biol.", vol. 15, 2003, p. 221-231.

[128] P. VEBER, M. LE BORGNE, A. SIEGEL, S. LAGARRIGUE, O. RADULESCU. *Complex Qualitative Models in Biology: A new approach*, in "Complexus", Doi: 10.1159/000093686, vol. 2, n^o 3-4, 2006, p. 140 – 151, http://content.karger.com/ProdukteDB/produkte.asp?Aktion=JournalHome&ProduktNr=227088.

[129] P. VEBER, C. GUZIOLOWSKI, M. LE BORGNE, O. RADULESCU, A. SIEGEL. *Inferring the role of transcription factors in regulatory networks*, accepted for publication in BMC Bioinformatics, Research Report, INRIA, 2007, http://hal.inria.fr/inria-00185038/en/.

[130] C. WHITE, R. SINGH, P. REINTJES, J. LAMPE, B. ERICKSON, W. DETTLOFF, V. CHI, S. ALTSCHUL. *BioSCAN: A VLSI-Based System for Biosequence Analysis,*, in "IEEE Int. Conf on Computer Design: VLSI in Computer and Processors", 1991, p. 504-509.

[131] Y. YAMANISHI, J.-P. VERT, M. KANEHISA. *Protein network inference from multiple genomic data: a supervised approach*, in "Bioinformatics", vol. 20, 2004, p. i363 - i370.

[132] T. YOKOMORI, S. KOBAYASHI. *DNA Evolutionary Linguistics and RNA Structure Modeling : A Computational Approach*, in "Proc.of 1st International IEEE Symposium on Intelligence in Neural and Biological Systems", 1995, p. 38-45.