INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Symbiose

SYstèmes et Modèles BIOlogiques, BIOinformatique et SEquences

Rennes - Bretagne-Atlantique

Theme : Computational Biology and Bioinformatics

Activity Report

2009

Table of contents

1. Team **1**

2. Overall Objectives **2**

2.1. A Bioinformatics center 2

2.2. Scientific axes 2

2.3. Highlights of the year 3

3. Scientific Foundations **4**

3.1. A short introduction to bioinformatics 4

3.2. Optimized algorithms on parallel specialized architectures 4

3.3. Modeling sequences and structures 5

3.3.1. Formal Languages and biological sequences 5

3.3.2. Machine Learning : from Pattern Discovery to Grammatical Inference 6

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 Biogenouest 9

5.3. New activities of the bioinformatics platform in 2009 10

5.3.1. National Project BioMAJ (BIOlogie Mise A Jour) 10

5.3.2. National Projects GRISBI and MobyleNet 10

5.4. Activity of transfer from Symbiose to GenOuest 10

5.4.1. Logol 10

5.4.2. ModuleOrganizer 11

5.4.3. CRISPI 11

5.4.4. Tuiuiu 11

5.4.5. FROSTO and A purva 11

5.4.6. PLAST and GASSSTT 11

5.4.7. Bioquali Cytoscape plugin 11

5.4.8. FlowCore: a Bioconductor package for high throughput flow cytometry 12

6. New Results **12**

6.1. Intensive sequence comparison and filtering 12

6.1.1. Intensive Comparison on FPGA 12

6.1.2. Genome mapping and Next generation sequencing 13

6.1.3. Primer design 13

6.1.4. Data Filtration 13

6.1.5. Average-case analysis of indexes 14

6.2. Modeling motifs and structures on sequences 14

6.2.1. Finding modules in sequences 14

6.2.2. Logical grammars 14

6.2.3. RNA and Protein Folding 14

6.2.3.1. RNA Secondary structure 15

6.2.3.2. Sequence/structure alignment 15

6.2.3.3. Protein structure comparison 15

6.2.4. Learning automata and grammars on biological sequences 15

6.2.4.1. Discovery of new protein 15

6.2.4.2. Characterizing protein fold with automata 15

6.2.4.3. Integrating scores on automata 16

6.2.4.4. Integrating motif discovery methods 16

6.2.4.5. Learning context-free grammars (CFG) 16

6.3. Systems biology: analysing data and modeling interactions 16

6.3.1. Algorithms for the analysis of large-scale models 16
6.3.1.1. Curating a Large-scale Regulatory Network by Evaluating its Consistency with Expression Datasets 16
6.3.1.2. Relating inter-patient gene copy numbers variations with gene expression via gene influence networks 17
6.3.1.3. Knowledge based identification of essential signaling from genome-scale siRNA experiments 17
6.3.2. Construction and analysis of signalling and metabolic pathways 17
6.3.2.1. Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation 18
6.3.2.2. In silico investigation of ADAM12 effect on TGF-beta receptors trafficking 18
6.3.2.3. Regulation of fatty acid metabolism 18
6.3.2.4. Metabolic Flexibility of the Mammary Gland in Lactating Dairy Cows 18
6.3.3. Hierarchical models for complex biological systems 19
6.3.3.1. Hybrid stochastic simplifications for multiscale gene networks 19
6.3.3.2. The effect of time parameters in discrete modeling of gene networks 19
6.3.3.3. Multiclock discrete models of biological systems 19
6.4. Miscellaneous 20
6.4.1. Development of Optimization techniques : the knapsack problem 20
6.4.2. Quality of association rules in Data Mining 20
7. Other Grants and Activities **20**
7.1. Regional initiatives 20
7.1.1. Biogenouest 20
7.1.2. Regional cooperation 21
7.2. National initiatives 21
7.2.1. National projects of the GenOuest platform 22
7.2.1.1. BioSide 22
7.2.1.2. MobyleNet 22
7.2.1.3. BioMAJ 22
7.2.1.4. GRISBI 22
7.2.1.5. Sigenae and Genanimal 22
7.2.2. ANR Projects 22
7.2.2.1. Proteus (Fold recognition and inverse folding problem) 22
7.2.2.2. BioWIC: Bioinformatics Workflows for Intensive Computation 23
7.2.2.3. Modulome: Identifying and displaying modules in genomic sequences 23
7.2.2.4. LepidOLF 23
7.2.2.5. Sitcon: Modeling signal transduction induced by a chimeric oncogene 23
7.2.2.6. DyCoNum: Dynamical and Combinatorial studies of Numeration systems 24
7.3. European and international initiatives 24
7.3.1. International programs 24
7.3.1.1. Integrated Project ACGT 24
7.3.1.2. Application of combinatorial optimization (PHC RILA, Bulgaria) 24
7.3.1.3. SECyT-INRIA-CNRS cooperation program (Argentina) 24
7.3.1.4. PHC Sakura and Amadeus 25
7.3.1.5. Freie Universität Berlin 25
7.3.2. Visiting scientists 25
7.3.2.1. Visitors 25
7.3.2.2. Scientific visit exchanges 25
8. Dissemination **25**
8.1. Leadership within scientific community 25
8.1.1. Administrative functions: scientific commitees, journal bords, jury 25

8.1.2. Jury of PhD Theses 26
8.2. Faculty teaching 26
8.3. Conference and workshop committees, invited conferences 26
8.3.1. International invited conferences 26
8.3.2. National invited conferences 26
8.3.3. Conference committees 27
8.3.4. Seventh meeting of the Bioinformatics platform of Biogenouest 27
8.3.5. BioInfoOuest thematic-day conferences 28
8.4. Theses defenses 28
8.4.1. PhD thesis, Van Hoa Nguyen 28
8.4.2. PhD thesis, Jérémie Gruel 28
9. Bibliography **29**

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 until Sept. 2009, 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, senior Research Engineer, Cnrs]
Hugues Leroy [GENOUEST, senior Research Engineer, Inria]
Olivier Sallou [GENOUEST, senior Research Engineer, Unv. Rennes 1]
Annabel Bourdé [GENOUEST, Engineer, until May 2009, Inria contract genopole]
Anthony Bretaudeau [GENOUEST, Engineer, Inria contract genopole]
Delphine Naquin [GENOUEST, Engineer, Inria contract genopole]
Aurélien Roult [GENOUEST, Engineer, Inria contract genopole]
Romaric Sabas [GENOUEST, Engineer, Inria contract genopole]
Ludmila Sarbu [GENOUEST, Engineer, 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, until May 2009 ANR contract Modulome]
Odile Rousselet [Engineer,ANR contract BioWic]

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 until Sept. 2009]

Noël Malod-Dognin [Inria/Region]
Van Hoa Nguyen [Inria/CORDI until Nov. 2009]
Guillaume Rizk [MENRT]

Post-Doctoral Fellow
Nolwenn Le Meur [Inserm/La Ligue]
Sylvain Blachon [Inria, Sitcon ANR project until Jul. 2009]
Guillaume Launay [ANR Proteus until Aug. 2009]

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

Administrative Assistant
Marie-Noëlle Georgeault [Assistant, Inria]

2. Overall Objectives

2.1. A Bioinformatics center

Symbiose is a bioinformatics research project. It focuses on methodological research at the interface between computer science and molecular biology, excluding "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 core facility **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 resource center GenOuest acts as a facility and software tool provider for the analysis of genomic data generated by numerous laboratories (55) of Biogenouest®. The resource center 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

Our research specificities include our interest in **large scale studies** (genomes, proteomes or regulation networks) and **discrete methods** necessary to handle the associated complexity. We have a global concern for high performance computing and two types of modeling tasks, modeling sequences and structures and modeling regulation networks.

- 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 Gate Arrays (FPGA) or fast components such as Flash memories or Graphical Processing Units (GPU).
- Modeling sequences and structures

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: design of grammatical models on biological sequences and machine learning of grammatical models from sequences; efficient filtering and model matching in data banks; protein structure prediction. Corresponding disciplinary fields are language theory, algorithmic on words, machine learning, data analysis and combinatorial optimization.

- System biology
  We address the question of constructing accurate models of biological systems with respect to available data and knowledge. The availability of high-throughput methods in molecular biology has led to a tremendous increase of measurable data along with resulting knowledge repositories, gathered on the web (e.g. KEGG,MetaCyc, RegulonDB). However, both measurements as well as biological networks are prone to incompleteness, heterogeneity, and mutual inconsistency, making it highly non-trivial to draw biologically meaningful conclusions in an automated way. Based on this statement, we develop methods for the analysis of large-scale biological networks which formalize various reasoning modes in order to highlight incomplete regions in a regulatory model and to point at network products that need to be activated or inactivated to globally explain the experimental data. We also consider small-scale biological systems for a fine understanding of conclusions that can be drawn on active pathways from available data, working on deducible properties rather than simulation. Corresponding disciplinary fields are model checking, constraint-based analysis and dynamical systems.

2.3. Highlights of the year

We have designed a major high performance sequence algorithm for bioinformatics: we propose a very efficient alternative to BLAST for comparing large genomic banks, an issue that becomes increasingly important with the development of high throughput Next Generation Sequencing technologies. The software, called PLAST (Parallel Local Alignment Search Tool), is able to run on a large variety of parallel architectures. Speed-up of X5 are typically observed compared to the multithreaded Blast software, reaching X10 while using GPU boards and X30 with FPGA processing units [22], [41].

From the point of view of modelling, we got several results that show the potential of our methods in biology. We have obtained high impact publications on a study on the canalization of gene expression in the Drosophila blastoderm during early development [21], [20]. This emphasizes the interest of developing precise dynamical models of gene regulations. We have also developed an identification method and database for CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) regrouping the analysis of all known archeal or bacterial genomes (more than 1100 species). It offers the best reference to date to these elements implied in microbial defense mechanisms [26].

At the national level, Genouest is the sole Bioinformatics core facility hosted in a computer science laboratory. It received in 2009 a label “IBISA” [1] and confirmed its ISO 9001:2000 certification in March. Our collaborations with Inra proved to be very fruitful. One of our highlight software, Biomaj, developed in collaboration with Inra has been downloaded more than a hundred of times. It aims at the management of large scale data workflows and offers a great help to bioinformatics platforms and laboratories that need to automatically update data banks on a regular basis. A national Inria development action has started at the end of the year to reinforce the project. We have also investigated insect genomics, hosting AphidBase, a comprehensive genome information resource dedicated to annotation of the pea aphid genome that is used by the International Aphid Genomics Consortium (IAGC).

[1] IBISA is the coordination structure for the technological platforms dedicated to life science in France

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 [96]:

- *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

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*[2]). 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 [97], [106]. Then, with the increasing growth of data, faster algorithms have been designed to drastically speed-up the search. The Blast software [108] 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

[2] http://www.genomesonline.org/

software (JJ Codani, [81]. Other works concern dedicated hardware machines. Several research prototypes such as SAMBA [83], BISP [70], HSCAN [82] or BioScan [112], have been proposed, leading today to powerful commercial products: BioXL, DECYPHER and GeneMatcher coming respectively from Compugen ltd. TimeLogic and Paracel [3].

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 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 [96]. 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. Modeling sequences and structures

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 [84]), Watson Crick complementarity (Sticker-system [90]),inversion, transposition, copy, deletion... Language theorists like A. Salomaa and Gh. Paun [99] 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 [114], [104], [105] ;
- 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 [71]. D. Searls proposed a more systematic approach based on logical grammars and a parser, Genlang [76]. 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] http://www.compugen.co.il/, http://www.timelogic.com, http://www.paracel.com

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 [65], [87]. 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 [86]. Most powerful programs in this field use bayesian procedures, Gibbs sampling and Expectation-Maximization [95]. The linguistic representation, which corresponds to our own work, generally rests on regular expressions. Algorithms use combinatorial enumeration in a partially ordered space [64], [101]. Another track explores variations on the search for cliques in a graph [92], [66].

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 [102]. 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 [94].

Cellular interactions' modeling is an old domain in biology, initiated by people interested in the dynamics of enzymes systems [88]. 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 [109].

The **dynamical framework in systems biology** includes simulations and prediction of behaviors. Models can be either qualitative or quantitative, as reviewed in [74], [69], [93]. 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 [78], [111], [110]. The properties of continuous models can be studied with convex analysis, linear and non-linear control techniques [85], [98], [61]. Stochastic models transform reaction rates into probabilities and concentrations into numbers of molecules, allowing to understand how noise influences a system [89]. 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 [91], [103]. Piecewise linear differential models [75], [80] 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 [107], [69]. 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 [79] or kernels [113] 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 [62], depending on the class of models, that can be discrete [100], continuous [93] or hybrid [67]. 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 [63], [68]. Identification methods are limited to a few dozen components. Model correction or parameter regression can cope with up to hundreds of products [68], [62] 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

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 Biogenouest provides us with a lot of collaborations with biology laboratories. We emphasize here three types of applications with major achievements in the project.

- **Whole genome analysis** is made practical through dedicated data structures and reconfigurable architectures. We have for instance implemented very fast Blast comparisons (Plast), 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*.
- **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.
- **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;
- yearly activity of transfer of the platform, in conjunction with Symbiose.

5.2. GenOuest, the Bioinformatics computing center of Biogenouest

Participants: Olivier Collin, Hugues Leroy, Olivier Sallou, Jacques Nicolas, Annabel Bourdé, Anthony Bretaudeau, Delphine Naquin, Aurélien Roult, Romaric Sabas, Ludmila Sarbu.

Main evolutions of the year have been:

- The GenOuest platform has been recognized as an IBiSA platform by IBiSA, the coordination structure for the technological platforms dedicated to life science in France
- The ISO 9001:2000 certification has been maintained (March 2009).
- The partnership with the "service formation continue" of the University of Rennes 1 has permitted to provide training sessions to 70 people.

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 scientific advances in bioinfomatics. The scientific theme of this year's meeting (26 oct.) was "Structural Biology and Bioinformatics". The platform was involved in the international workshop ISYiP "Information Systems for Insect Pests" (16-17 nov.). GenOuest has also hosted the Biograle meeting on large scale Bioinformatics(24-25 nov.).

The platfom is involved in different coordination activities at a national level. O. Collin is a member of the ReNaBi (Reseau National des plates-formes Bio-informatiques) steering board since July 2008 and a member of the scientific committee of Biogenouest. 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 an active member of Grisbi (GRIlles Support pour la BIologie), a group gathering 6 platforms, recently labelled by IBiSA. GenOuest is also involved in another IBiSA project named MobyleNet with other french bioinformatics platforms. GenOuest has established a partnership with NBCR (National Biomedical Computation Resource) in San Diego for the development of additional code of the Opal Toolkit.

5.2.1. Bioinformatics services hosting

The GenOuest bioinformatics platform is hosting bioinformatics services developed by external research teams who require computing power. It sometimes leads to common publications in case of stronger collaborations. We have for instance worked on the annotation of the plant pathogenic fungus Gaeumannomyces graminis var. tritici [5] and the EST database of the Pacific oyster (Crassostrea gigas)[11].

Main hosted services are [4]:

- *AphidBase* AphidBase, formerly a web application for the analysis of Aphids ESTs has been upgraded to a comprehensive genome information resource dedicated to annotation of the pea aphid genome. It is the result of a collaboration with INRA Bio3P, Le Rheu and it is used by the International Aphid Genomics Consortium (IAGC) [18].
- *Autograph*,an integrated web server for multi-species comparative genomic analysis designed for constructing and visualizing synteny maps and for highlighting evolutionary breakpoints. Developed by UMR6061 Rennes.
- *Germonline and Mimas*, a centralized bioinformatic resource and cross-species knowledge base built around the Ensembl genome Browser, providing microarray data relevant for the cell cycle and gametogenesis [13]. Developed by INSERM U625 Rennes and SIB Lausanne.
- *LepidoDB*, a new bioinformatic platform for the annotation and cross-comparisons of lepidopteran genomes
- *MIPDB*, a relational DB of all Major Intrinsic Proteins, and *RASTA*, a tool for the study of toxin-antitoxin compounds in bacteria (developed by UMR 6026 Rennes).
- *M@ia*, a tool dedicated to micro-array data analysis developed by INSERM U620 Rennes.

[4] http://www.genouest.org/spip.php?rubrique166

5.3. New activities of the bioinformatics platform in 2009

Participants: Olivier Collin, Hugues Leroy, François Moreews, Aurélien Roult, Romaric Sabas, Olivier Sallou.

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 bioinformatics 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 and has been downloaded more than 100 times and it used in production on 8 french bioinformatics (INRA Jouy, INRA Toulouse, PBIL Lyon, Strasbourg...) platforms. [5].

A new development cycle has started in November with a dedicated INRIA ADT (Action for technological development) national project. It will add functionalities like peer-to-peer, real time monitoring, databanks versioning and graphical interface for workflows conception.

5.3.2. National Projects GRISBI and MobyleNet

GRISBI aims to develop a grid dedicated to bioinformatics. This action, funded by IBiSA, involves different actors of the french bioinformatics community : IPCB Lyon (C. Blanchet, C. Eloto, A. Michon), INRA Jouy-en-Josas (J.-F. Gibrat), CNRS Roscoff (C. Caron), CBIB Bordeaux (T. Martin), IGBMC Strasbourg (F. Plewniak), INRIA Lille and GenOuest (O. Collin, A. Roult).

MobyleNet aims to develop a bioinformatics application portal allowing the remote execution of jobs among the different partners computer resources. It gathers different bioinformatics platforms, is managed by P. Tuffery at RPBS, Paris, and is funded by IBiSA. GenOuest provides his technical and development expertise.

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 Bretaudeau, Annabel Bourdé, Carito Guziolowski.

Modeling activity concerns sequences and networks in Symbiose.

5.4.1. 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. The Logol Software Suite is a set of software composed of a Logol language interpreter(biological patterns) and pattern search tool, a graphical web-based editor, and a result analyser. Result files contain the matches on the sequence(s) with all required details. Pattern description supports (among others) word complement, overlaps, substitution and distance errors as well as variables

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

usage. 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.

5.4.2. ModuleOrganizer

ModuleOrganizer is a software package proposing a synthetic view of a set of DNA sequences by providing both a segmentation of them into domains and a classification on the basis of these domains. It has been developed in the framework of ANR Modulome, leaded by Symbiose. It indexes the maximal repeats in the sequences and assembles them for create modules. After a classification step relatively to the presence or the absence of modules, the method results in a graphical view of a hierarchical clustering of the segmented sequences.

5.4.3. CRISPI

The CRISPR genomic structures (Clustered Regularly Interspaced Short Palindromic Repeats) form a family of repeats that is largely present in archea and frequent in bacteria. CRISPI is a user-friendly web interface with many graphical tools and facilities allows extracting CRISPR, finding out CRISPR in personal sequences or calculating sequence similarity with spacers. It offers a reference in this domain with more than 1100 species and is updated automatically on a regular basis. It has been developed during ANR Modulome [26], leaded by Symbiose, in collaboration with LME/Ifremer Brest.

5.4.4. Tuiuiu

In the paper [24] we proposed a filter for speeding-up the multiple repeat search. This filter, called Tuiuiu, defines a necessary condition stronger than previous filters and proposes an efficient way to apply it on large set of sequences.

In collaboration with the Genouest platform, the Tuiuiu filter had been housed on the mobyle portal and is thus publicly available. Using a submission form [6] the Tuiuiu tool can be lunched using the Genouest machines.

5.4.5. FROSTO and A purva

Two programs working on protein structures, FROSTO and A purva, are installed on the GenOuest cluster and available to registered users. FROSTO (PhD thesis G. Collet) is a program that finds remote homolgies between proteins, based on INRA program Frost. It aligns a protein sequence with a database of protein structures by an efficient protein threading method with non-local parameters and uses a dedicated solver based on a Lagrangian relaxation approach. A purva (PhD thesis N. Malod- Dognin) is a tool for computing the similarity of two protein structures, by finding the maximum overlap of their contact maps.

5.4.6. PLAST and GASSSTT

PLAST and GASSSTT are freely downloadable codes available on the software web site of Symbiose. PLAST (PhD thesis V. H. Nguyen) is a parallel Blast, the most used sequence comparison software. GASSST (PhD thesis G. Rizk) finds global gapped alignments of short DNA sequences against large DNA banks.

5.4.7. Bioquali Cytoscape plugin

Bioquali is dedicated to computations on qualitative models represented by interaction graph. Nodes of these graphs represents chemical species and arrows are labeled by positive or negative influences.

[6] http://mobyle.genouest.org/cgi-bin/Mobyle/portal.py?form=tuiuiu

The software 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 2009, we focused our attention on the design of a cytoscape plugin to allow a friendly use of Bioquali functionnalities [7]. The BioQuali plugin [16] facilitates *in silico* exploration of large-scale regulatory networks by combining the user-friendly tools of the Cytoscape environment with high-performance automatic reasoning algorithms. As a main feature, the plugin guides further investigation regarding a system by highlighting regions in the network that are not accurately described and merit specific study.

The BioQuali plugin is implemented in Java, based on the Cytoscape API, and using the REST architectural style. By default, the client component uses an unauthenticated HTTP connection to communicate with the GenOuest Web server. This enables fast remote execution of the algorithm underlying the BioQuali plugin on the GenOuest high performance computing facility. Alternatively, the server side component can be downloaded and installed on any standard PC using the Cytoscape plugin management system. The plugin is available to download from the Cytoscape plugin website [8], under the Plugin/Analysis section or via Java Web Start. It is compiled with the latest Cytoscape API (version 2.6) and packaged as a jar file.

5.4.8. FlowCore: a Bioconductor package for high throughput flow cytometry

Flow cytometry is a technology used for high throughput screening, generating large complex data sets often in clinical trials or drug discovery settings. We developed a set of flexible open source computational tools in the R package flowCore to facilitate the analysis of these complex data. A key component of which is having suitable data structures that support the application of similar operations to a collection of samples or a clinical cohort. In addition, our software constitutes a shared and extensible research platform that enables collaboration between bioinformaticians, computer scientists, statisticians, biologists and clinicians. This platform will foster the development of novel analytic methods for flow cytometry. The software has been applied in the analysis of various data sets and its data structures have proven to be highly efficient in capturing and organizing the analytic work flow. Finally, a number of additional Bioconductor packages successfully build on the infrastructure provided by flowCore, open new avenues for flow data analysis [17].

6. New Results

6.1. Intensive sequence comparison and filtering

The first step of the analysis of a new sequence is to compare it with already known sequences. We work on all aspects allowing to speed up this time consuming task: efficient architectures, efficient indexing schemes and efficient filtering of sequences.

6.1.1. Intensive Comparison on FPGA

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

[7] http://www.irisa.fr/symbiose/projects/bioqualiCytoscapePlugin/

[8] http://www.cytoscape.org

We propose a very efficient alternative to BLAST for sequence bank comparison called PLAST (Parallel Local Alignment Search Tool). An Implementation of PLAST has been developed using the SGI RASC 100 architecture. This platform, composed of two large high performance FPGAs (200K logics cells), is linked to an Altix350 (Intel Itanium2 Core2 1.6GHz) through SGI Numalink bus, providing a theoretical bandwidth of 3,2GBytes/s in each direction. The time-consuming part of the PLAST algorithm (sequence comparison) is deported on FPGA, which implements a specific parallel sequence comparison operator. It is architectured as an array of 192 small dedicated processing elements, each one computing a single alignment. Speed-up of X53 has been achieved over software execution on the Altix350, and X19 over the NCBI tBLASTn software [22], [41].

6.1.2. Genome mapping and Next generation sequencing

Participants: Rayan Chikhi, Dominique Lavenier, Guillaume Rizk.

Next generation sequencing technologies (NGS) produce large quantities of genomic data that are useful for a wide range of large-scale applications. This triggers the need for new algorithms able to map accurately and efficiently millions of short sequences on large genomes. We developed GASSST, a Global Alignment Short Sequence Search Tool. It is a new short read aligner which can map reads with gaps and mismatches at very high speed. It uses the standard seed and extend strategy. The novelty of our approach stands in a new filter step which allows us to discard candidates hits before having to execute the computationally expensive extend step (Needleman-Wunsch algorithm). We developed a series of filters of increasing complexity and efficiency capable of quickly eliminating most false-positive candidate hits for a wide range of execution configurations, with few or many gaps, low or high error rates.

Many genetic mutations can be found by re-sequencing an organism and comparing the data with a reference genome. However, next-generation sequencing data may consist of very short DNA fragments (reads), with lengths starting at 30 base pairs. At this read length, it was shown that only 80% of the human genome can be re-sequenced. Recently, sequencers have been able to produce mate-paired reads, ie. pairs of fragments separated by a known distance. We designed an efficient algorithm [9] to determine which part of a genome can be re-sequenced using mate-paired reads. Using this algorithm, we showed that mate-paired reads of 20-25 base pairs suffice to re-sequence 95% of the human genome. We have also started to investigate on de novo eukaryotic genome assembly from mate-paired reads, a NP-hard issue that remains a bottleneck with respect to the use of NGS data.

6.1.3. Primer design

Participants: Pierre Peterlongo, Jacques Nicolas, Raoul Vorc'h.

Discovery of molecular markers for efficient identification of living organisms remains a challenge of high interest faced to the huge amounts of data that will soon become available in all kingdoms of life. The diversity of species can now be observed in details with low cost genomic sequences produced by new generation of sequencers. We developed a method, called c-GAMMA, which formalizes the design of new markers from such data. It is based on a series of filters on forbidden pairs of words, followed by an optimization step on the discriminative power of candidate markers. This method was implemented and tested on a set of microbial genomes (Thermococcales) [42].

6.1.4. Data Filtration

Participants: Pierre Peterlongo, Olivier Sallou, Anthony Bretaudeau.

Multiple alignment between sequences is a NP-Hard problem where useless long computations can be avoided by focusing on the multiple repeats they contain. The key idea is to remove portions of sequences that may not contain researched repeats before the multiple alignment phase. This preliminary work is then seen as a filter. Basically a filter applies a carefully chosen necessary condition on a sequence or a set of sequences. In [24], [50] we proposed a filter for speeding-up the multiple repeat search. called Tuiuiu, improving previous attempts by a stronger necessary condition, and proposed an efficient way to apply it on large set of sequences. It is available on Genouest.

6.1.5. *Average-case analysis of indexes*

Participant: Jérémie Bourdon.

Factor and suffix oracles have been introduced in 1999 in order to provide an economic and efficient solution for storing all the factors and suffixes respectively of a given text. Whereas good estimations exist for the size of the factor/suffix oracle in the worst case, no average-case analysis has been done until now. In [32], we give an estimation of the average size for the factor/suffix oracle of an n-length text when the alphabet size is 2 and under a Bernoulli distribution model with parameter $1/2$. To reach this goal, a new oracle is defined, which shares many of the properties of a factor/suffix oracle but is easier to study and provides an upper bound of the average size we are interested in. Our study introduces tools that could be further used in other average-case analysis on factor/suffix oracles, for instance when the alphabet size is arbitrary.

6.2. Modeling motifs and structures on sequences

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. RNA and protein folding studies issues use general models that need heavy computations. 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 has coordinated 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. For this last year, CRISPI, the most complete database to date on CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) has been built regrouping the analysis of all complete microbial genomes available to date (more than 1100 archeal and bacterial genomes) and is available through the web [26]. CRISPR are formed by a repetitive skeleton including genetic material imported from viruses and plasmid. The theoretical part of this research has been submitted for publication, together with a method and tool -ModuleOrganizer- for the analysis of modules in transpon families.

6.2.2. *Logical grammars*

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

We propose 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, we have refined the Logol implementation and started to apply it on the search of MITE elements, a transposon family largely present in the human genome, in collaboration with GICC Tours (Y. Bigot). We have also studied with this laboratory the ascovirus DpAV4a (family Ascoviridae) [6].

6.2.3. *RNA and Protein Folding*

Participants: Dominique Lavenier, Rumen Andonov, Guillaume Rizk, Guillaume Collet, Guillaume Launay, Noël Malod-Dognin, Israël-César Lerman.

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 secondary structure prediction, sequence/structure alignment (Protein Threading Problem-PTP), and structure/structure comparison (CMO problem). The first problem is polynomial and the two other problems have been proved to be NP-complete.

6.2.3.1. RNA Secondary structure

The importance of the world of non coding RNA has fostered the interest for efficient prediction programs on RNA secondary structures. Associated algorithms have time complexity in O(n^3) that becomes prohibitive for large sequences or large data sets. We have parallelized algorithms on GPU, which provide better performance/price and performance/energy ratios than CPU. The main computation is a dynamic programming algorithm. In addition, by exploiting parallelism at a coarse grain level among several sequences, we were able to provide the GPU with enough independent tasks. Our implementation faced two major GPU-specific issues: computation divergence and complex memory access patterns, which can lead to respectively inefficient use of computational and memory bandwidth resources. We managed to tackle those issues by off-loading part of the divergence to the CPU, and through the careful use of GPU memory spaces : shared, constant and texture memory. This lead to a x17 speedup over the reference sequential CPU code [43]. Ongoing work with a tiled approach allowing a greater reuse of data shows significant improvement for both the GPU and CPU sequential code.

6.2.3.2. Sequence/structure alignment

In [35] We propose a new local alignment method for the protein threading problem (align part of a protein structure onto a protein sequence). Local sequence-sequence alignments are widely used to find functionally important regions in families of proteins. However, as far as we know, no local sequence-struture alignment algorithm has been implemented. We developed five Mixed Integer Programming (MIP) models that can perform local alignements between sequences and structures and compared their performances.

6.2.3.3. Protein structure comparison

Many protein structure comparison methods can be modeled as maximum clique problems in specific k-partite graphs, referred here as alignment graphs. In [58] we proposed a new protein structure comparison method based on internal distances (DAST) which was posed as a maximum clique problem in an alignment graph. We also designed a dedicated algorithm (ACF) for solving such maximum clique problems. ACF is first applied in the context of VAST, a software largely used in the National Center for Biotechnology Information, and then in the context of DAST. The obtained results on real protein alignment instances show that our algorithm is more than 37000 times faster than the original VAST clique solver which is based on Bron & Kerbosch algorithm. We furthermore compare ACF with one of the fastest clique finder, recently conceived by Ostergard. On a popular benchmark (the Skolnick set) we observe that ACF is about 20 times faster in average than the Ostergard's algorithm.

6.2.4. Learning automata and grammars on biological sequences

Participants: Jérémie Bourdon, François Coste [correspondant], Matthias Gallé, Pierre Peterlongo, Rumen Andonov.

We use the inference of automata from samples of (unaligned) sequences as a general learning technique for the characterization of protein families. Automata are graphical models that are more expressive than standard sequence patterns (such as PSSM, Profile HMM, or Prosite Patterns) and enable modelling heterogeneous sequence families. We are also studying how to learn more expressive grammars such as context-free grammars.

6.2.4.1. Discovery of new protein

Last year, a new candidate protein in the family involved in cell apoptosis was discovered thanks to Protomata-Learner. In collaboration with T. Guillaudeux from the team *Microenvironnement et Cancer* (MICA), we have defined more precisely the localization of the complete gene and studied it by comparison with other species collecting *in-silico* evidence that the protein is actually expressed [53].

6.2.4.2. Characterizing protein fold with automata

Protomata-Learner software has been generalized and is able now to characterize a set of protein structures instead of a set of sequences. First experiments on building structural cores for the protein threading program FROSTO are encouraging [51]. Considering the study of newly introduced partial local alignments, we have worked on improving their definition and we have designed a C++ library handling these objects as a first step for the implementation of the alignment algorithms.

6.2.4.3. Integrating scores on automata

Since learned automata are used to predict new members of a family, it is important to associate scores on their transitions. We have worked on the introduction of pseudo-counts based on Dirichlet mixtures in the automata and on the significance of the score on new sequences [60]. Test on classical benchmarks and on particular families of proteins in collaboration with biologists of Genouest are planned.

6.2.4.4. Integrating motif discovery methods

Numerous motif discovery tools are now available for the identification of transcription factors, a crucial task to construct regulatory networks. Combining efficiently their results appears useful for comparing and clustering these motifs in order to reduce redundancies and to identify corresponding transcription factor. We develop a pipeline that produces, compares and clusters a set of motifs and identifies some close motifs in public databases like JASPAR and Transfac. Unlike common comparison methods, where each matrix column is compared independently, we have developed a global approach that helps to reduce false positives. We also proposed an original graph motif model that generalizes the classical position specific pattern matrices. Finally, we present an application of our method to study ChIP-chip data sets in the context of an eukaryotic organism [33].

6.2.4.5. Learning context-free grammars (CFG)

Focusing rather on learning the structure than the language, we develop an approach of CFG learning based on recoding repeated words. To handle large sequences such as genomes, efficient data structures and algorithms have to be used. To detect and score repeats, we are using suffix arrays which need to be regularly updated after each rewriting of a repeated word. We proposed an incremental update algorithm of suffix arrays 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. Our implementation exhibits a significant speed-up compared to the construction from scratch at each step [12]. In collaboration with the Natural Language Processing Group from Universidad Nacional de Córdoba, we have applied the algorithm on the smallest coding grammar issue, studying new scores for choosing the words and their occurrences to be rewrited.

6.3. Systems biology: analysing data and modeling interactions

This axis strives to build dynamical systems that model interactions implied in biological processes such as metabolism, development and differentiation, signaling, etc.. It both adresses medium scale modeling with differential equations and large scale modeling using model reduction techniques and logical constraints.

6.3.1. Algorithms for the analysis of large-scale models

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

The analysis of static and heterogenous large-scale data is the main question of integrative biology. We pushed forward an automatic reasoning approach which allows the confrontation of observations with the network topology given by interactions graphs. In collaboration with Potsdam, we use Answer Set Programming for this purpose.

6.3.1.1. Curating a Large-scale Regulatory Network by Evaluating its Consistency with Expression Datasets

Regulatory networks are generally analysed by *in silico* simulation of network component fluctuations under perturbations. Many difficulties have to be considered such as the incomplete state-of-art of regulatory knowledge, the large-scale of regulatory models, heterogeneity in the available data and the sometimes violated assumption that mRNA expression is correlated to protein activity. We have proposed a method using a simple *consistency rule* that allows large-scale network analysis using small – but reliable – expression datasets.

We have developed a BioQuali plugin for the Cytoscape environment, designed to facilitate automatic reasoning on regulatory networks. The BioQuali plugin enhances user-friendly conversions of regulatory networks (including reference databases) into signed directed graphs [16]. Highlighting inconsistent regions in the network or predicting which products in the network need to be up or down regulated (active or inactive) to *globally* explain experimental data are basic functions of the package. We tested this approach with the transcriptional network of *E. coli* (1763 products and 4491 interactions) extracted from the RegulonDB database [36]. We proved that confronting our predictions with mRNA expression experiments enables determining missing post-transcriptional interactions in the model. After correction of the model, we calculated 502 gene-expression predictions (starting from 0.5% of observations) that correspond to nearly 30% of the network products predicted to change considerably under the analysed condition. These predictions were validated with microarray outputs, obtaining an agreement of 80%. This percentage is comparable to the one obtained by other methods working on *E. coli* data [72], [73], [77], and considerable, since we used only a transcriptional model without including metabolic regulations.

6.3.1.2. Relating inter-patient gene copy numbers variations with gene expression via gene influence networks

During tumorigenesis, genetic aberrations arise and may deeply affect the tumoral cell physiology. It has been partially demonstrated that an increase of gene copy numbers induces higher expression; but this effect is less clear for small genetic modifications. To study it, we used the Bioquali approach to perform the integration of CGH and expression data together with an influence graph derived from biological knowledge [31]. Interindividual variations in gene copy number and in expression allow to attack tumor varability and ultimately adresses the problem of individual-centered therapeutics. We tested this approach on Ewing tumor data. It allowed the definition of new biological hypotheses that were validated by comparison with random permutation of the initial data sets.

6.3.1.3. Knowledge based identification of essential signaling from genome-scale siRNA experiments

A systems biology interpretation of genome-scale RNA interference (RNAi) experiments is complicated by experimental variability and network signaling robustness. Over representation approaches (ORA), such as the hypergeometric or z-score, are an established statistical framework used to associate RNA interference effectors to biologically annotated gene sets or pathways. These methods, however, have known limitations: they can miss partial pathway activation, and cannot take advantage of interactome knowledge. In [4] we present a novel ORA, protein interaction permutation analysis (PIPA), that takes advantage of canonical pathways and established protein interactions to identify pathways enriched for protein interactions connecting RNAi hits. As a result we identify pathways and signaling hypotheses that are statistically enriched to effect cell growth in human cell lines. We used PIPA to analyze genome-scale siRNA cell growth screens performed in HeLa and TOV cell lines, showing that interacting gene pair siRNA hits are more reproducible than single gene hits. Using protein interactions, PIPA identifies enriched pathways not found using the standard Hypergeometric analysis including the FAK cytoskeletal remodeling pathway.

6.3.2. Construction and analysis of signalling and metabolic pathways

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

The previous section tackle the analysis of large-scale high-throughput static data. For the analysis of time-series data, we refined our strategy of automatic reasoning by developping abstract differential models. Our main goal is not to build full parametrized differential models (which would require a much large amount of data), but to reason over classes of models in order to understand which conclusion on active pathways can be deduced from available data. We assume that dynamics of biological networks is hierarchical, involving many separated time scales and have developed a dedicated mathematical methodology: it relies on model reduction and comparison techniques, within and between various levels of descriptions of biological networks [14]. We coupled this hierarchical approach together with sensitivity analysis and fitting under constraints to perform our conclusions.

6.3.2.1. Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation

In recent years, quantitative gene expression data have become available for the segment determination process in the Drosophila blastoderm, revealing a specific instance of canalization. We used a predictive dynamical model of gene regulation to study the effect of Bicoid variation on the downstream gap genes. The model correctly predicts the reduced variation of the gap gene expression patterns and allows the characterization of the canalizing mechanism. We show that the canalization is the result of specific regulatory interactions among the zygotic gap genes. We demonstrate the validity of this explanation by showing that variation is increased in embryos mutant for two gap genes, Krüppel and knirps, disproving competing proposals that canalization is due to an undiscovered morphogen, or that it does not take place at all [21]. In an accompanying article in PLoS Computational Biology [20], [28], we show that cross regulation between the gap genes causes their expression to approach dynamical attractors, reducing initial variation and providing a robust output. These results demonstrate that the Bicoid gradient is not sufficient to produce gap gene borders having the low variance observed, and instead this low variance is generated by gap gene cross regulation.

6.3.2.2. In silico investigation of ADAM12 effect on TGF-beta receptors trafficking

The transforming growth factor beta (TGF-beta) is known to have multiple effects, including differentiation, proliferation and apoptosis. However the underlying mechanisms remain poorly understood. The regulation and effect of TGF-beta signaling is complex and highly depends on specific protein context. In collaboration with INSERM and supported by a co-tutored PhD thesis (J. Gruel) [1], we have recently shown that the disintegrin and metalloproteinase ADAM12 interacts in liver with TGF-beta receptors and modulates their trafficking among membranes, a crucial point in TGF-beta signaling and development of fibrosis. In [15], we aimed to better understand how ADAM12 impacts on TGF-beta receptors trafficking and TGF-beta signaling. We extracted qualitative biological observations from experimental data and defined a family of models producing a behavior compatible with the presence of ADAM12. We computationally explored the properties of this family of models which allowed us to make novel predictions (increases TGF-beta receptors internalization rate between the cell surface and the endosomal membrane, modifies TGF-beta signaling shape favoring a permanent response. Alltogether, confronting differential models with qualitative biological observations, we obtained predictions giving new insights into the role of ADAM12 in TGF-beta signaling and hepatic fibrosis process.

6.3.2.3. Regulation of 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. In [7] our purpose was to identify the hierarchy of importance amongst pathways involved in fatty acid (FA) metabolism and their regulators in the control of hepatic FA composition. A step-by-step procedure was used in which a very simple model was completed by additional pathways until the model fitted correctly the measured quantities of FA in the liver during fasting in PPAR-knockout (KO) mice and wild-type mice. The resulting model included FA uptake by the liver, FA oxidation, elongation and desaturation of FA.

From the model analysis we concluded that PPAR had a strong effect on FA oxidation. In PPAR-knockout mice, FA uptake was identified as the main pathway responsible for FA variation in the liver. The models showed that FA were oxidized at a constant and small rate, whereas desaturation of FA also occurred during fasting. The latter observation was rather unexpected, but was confirmed experimentally by the measurement of delta-6-desaturase mRNA using real-time quantitative PCR (QPCR). These results confirm that mathematical models can be a useful tool in identifying new biological hypotheses and nutritional routes in metabolism.

6.3.2.4. Metabolic Flexibility of the Mammary Gland in Lactating Dairy Cows

In 2002, Van Milgen proposed a stoichiometric model to study the metabolism of the ruminant mammary gland. It includes reactions for lactose synthesis, milk protein, fatty acids and glycerol of triglycerides. A total of 10 metabolites involved in intermediary metabolism were used to describe 92 reactions, including those yielding or using ATP, cofactors, CO2, O2, and NH3. The model was applied to data from mammary gland balance studies carried out in dairy cows. We got a complete partition of nutrients measured in the balance studies including for each pathway the ATP production or cost, CO2 production and O2 consumption. The rules applied to find the partitioning were based on the hypotheses that there is no accumulation of intermediary metabolites in the mammary gland and that there is no deficiency of ATP or co-factors. Finally, we develop a new automatic tool using Flux Balance Analysis theory, suited to analyze data arising in nutrition studies. Preliminary results on this work have been discussed in [39].

6.3.3. Hierarchical models for complex biological systems

Participants: Ovidiu Radulescu [correspondant], Jérémie Bourdon, Michel Le Borgne, Nolwenn Le Meur, Jérémy Gruel.

In order to better understand the relations between logical/discrete models and continuous models, we investigated the effect of noise and time in models.

6.3.3.1. Hybrid stochastic simplifications for multiscale gene networks

Stochastic simulation of gene networks by Markov processes has important applications in molecular biology. The complexity of exact simulation algorithms scales with the number of discrete jumps to be performed and approximate schemes using a reduced number of simulated discrete events are necessary. Also, answering important questions about the relation between network topology and intrinsic noise generation and propagation should be based on general mathematical results. We proposed a unified framework for simplification of Markov models of multiscale networks dynamics. We discuss several possible hybrid simplifications, and provide algorithms to obtain them from pure jump processes. In hybrid simplifications, some components are discrete and evolve by jumps, while other components are continuous. Hybrid simplifications are obtained by partial Kramers-Moyal expansion which is equivalent to the application of the central limit theorem to a submodel. By averaging and variable aggregation we drastically reduce simulation time and eliminate non-critical reactions. The simplified models reproduce with good accuracy the stochastic properties of the gene networks, including waiting times in intermittence phenomena, fluctuation amplitudes and stationary distributions. Hybrid simplifications can be used for onion-like (multi-layered) approaches to multi-scale biochemical systems, in which various descriptions are used at various scales. Sets of discrete and continuous variables are treated with different methods and are coupled together in a physically justified approach [10].

6.3.3.2. The effect of time parameters in discrete modeling of gene networks

Several extensions of René Thomas' asynchronous logical approach have been proposed to better fit real biological dynamical systems: components may reach different discrete expression levels, depending on the status of other components acting as activators or inhibitors. In contrast, some fine-grained propositions are modelling the evolution of chemical concentrations through differential equation systems. Hybrid paradigms try to escape oversimplifications of logical models and the inaptitude of differential models to tackle to large real networks. Particularly, time delays are introduced in logical abstractions to pass from an expression level to next. Such delays are unknown new parameters added to the model. Then hybrid model-checking techniques are used to exhibit properties about the dynamical behaviour of the network. We have described a whole pipelined process which orchestrates the following stages: model conversion from a Piece-wise Affine Differential Equation (PADE) modelization scheme into a discretized model with attractors, focus on characterized subgraphs through a graph simplification step based on probabilistic criteria, conversion of the subgraphs into Parametric Hybrid Linear Automata, inference of dynamical properties through hybrid model-checking techniques. The publication [3] is the outcome of a methodological investigation launched to cope with the genetic regulation network involved during *Escherichia coli* carbon deprivation. We retrieved a remarkable cycle already exhibited by a previous analysis of the PADE.

6.3.3.3. Multiclock discrete models of biological systems

As each signal within a pathway follows its own clock, we have introduced multi-clock technique to model the dynamics of biological interactions. Discrete models do not contain any specification on the order of the transitions, which are usually defered to the simulator. We have proposed a new formalism to include timing specifications in the models. It is inspired by the formal models underlying real time programming languages such as Esterel, Lustre and Signal. In this approach, the notion of time refers the logical time used in computer science: it does not correspond to the duration of events but to their relative sequencing. This allows the description of several biological signals with different clocks, i.e., multiclock systems. One main improvement of our formalism is its capacity to support model-checking technique for properties involving biological entities and reaction time.

We validated our approach on published cell cycle models and worked on the influence of the EGF and TGF-β pathways in controlling cell proliferation and consequently tumor progression in the liver. We have evaluated the robustness of hepatocellular carcinoma cell line by using data from RNA interference experiments to constraint our model. This might help identify pathway checkpoints and buffering effects between different paths of the EGF and TGF-β pathways network, allowing to design news markers and new therapeutically targets for hepatocellular carcinomas [38], [44], [55].

6.4. Miscellaneous

We have a small theoretical research activity that is a by-product of our main work and is regrouped in this section.

6.4.1. Development of Optimization techniques : the knapsack problem

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). The result in [25] 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.

6.4.2. Quality of association rules in Data Mining

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. 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 [19].

7. Other Grants and Activities

7.1. Regional initiatives

7.1.1. Biogenouest

Biogenouest, the 8th national genopole, funded in 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 of academic labs, federated through a GIS structure (Scientific Interest Groupment, 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.

Biogenouest 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, participate in the monthly meetings of the Biogenouest committee. O. Collin is now in charge of the newly set-up bioinformatics axis that will coordinate the actions of 3 bioinformatics platforms of Biogenouest (Nantes, Rennes, Roscoff).

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:

- IRCCyN, Ecole Centrale de Nantes, ph-D thesis co-advisment (J. Bourdon)
- LINA. Modelling biological networks (J. Bourdon, A. Siegel)
- 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).
- MIcroenvironnement et CAncer, UPRES UA 3889 (Thierry Guillaudeux) : Tumor Necrosis Factor (F. Coste)
- INRA UMR Bio3P (Eric Grenier) Characterization of RBP1/SPRYSEC gene family (F. Coste, G. Collet)
- UMR BiO3P - INRA, Agrocampus Ouest, Rennes, France. key words: miRNA, GPU (D. Lavenier)
- UMR 6197 LM2E Ifremer, Centre de Brest, Brest, France . Key words: Primer generation (D. Lavenier)
- LINA (D. Eveillard, I. Rusu): Average-case analysis and Gene Networks analysis (J. Bourdon).
- IRCCyN (O. Roux, T. Merle): Temporal properties of Gene Networks (J. Bourdon).
- Institut du Thorax, U 915 (S. Carat, R. Houlgatte): Analysis of ChIP-chip data (J. Bourdon).

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)
- LiX, INRIA project team AMIB, statistical significance for motif finding algorithms (J. Bourdon)
- IML, Marseille and LIRMM, Montpellier: substitutive dynamical systems (A. Siegel).
- Helix inria team: biological sequences filtering (P. Peterlongo).
- URGI (Unité de Recherche Génomique Info), Versailles: LTR retrotransposons in plant genomes (C. Belleannée).
- Laboratoire d'Informatique de Nantes Atlantique, Equipe COD (COnnaissances et Décision), Site Polytech' Nantes, Hiérarchie Implicative Orientée (I.C. Lerman).
- Laboratoire d'Informatique, de Modélisation et d Optimisation des Systèmes (LIMOS), Université d'Auvergne, Qualité des Règles d'Association (I.C. Lerman).
- Colorado State University, Dept. Computer Science, Fort Collins, USA Key words: parallelism, RNA folfing (D. Lavenier)
- GICC UMR CNRS 6239, Tours (Génétique Immunothérapie Chimie et Cancer): logical models and characterization of transposons, RNA folding, GPU (C. Belleannée, D. Lavenier, J. Nicolas)
- EPI AMIB (M. Régnier): Average-case analysis and statistical tests (J. Bourdon).

7.2.1. National projects of the GenOuest platform

Participants: Olivier Collin, Hugues Leroy, François Moreews, Jacques Nicolas, Anthony Bretaudeau, Aurélien Roult, Olivier Sallou.

7.2.1.1. BioSide

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

7.2.1.2. MobyleNet

MobyleNet is a joined project between several bioinformatics platforms aiming to develop a bioinformatics web portal. The MobyleNet project involves 8 sites, 5 IBISA platforms, and 3 supporting sites external to this process, either providers of services and/or developers of the core MobyleNet framework. It thus involves seven physical bioinformatics platforms (nodes) distributed over France and having different skills that range from genomics, microarrays, sequence analysis, phylogeny, structural bioinformatics to chemoinformatics and different focuses from microorganism, plants to pharmacology and cancer.

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

GRISBI is an IBISA joined initiative started in oct. 2008 between 6 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 resources of the 6 platforms with grid software components and coordination tools: computing and storage hardware resources, but also database and software resources.

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 studying with UMR Inra 598 the modeling of fatty acids metabolism (see Sec. 6.3). We have also developed sigReannot, an oligo-set re-annotation pipeline based on similarities with the Ensembl transcripts and Unigene clusters [8].

7.2.2. ANR Projects

7.2.2.1. 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 [9].

7.2.2.2. *BioWIC: Bioinformatics Workflows for Intensive Computation*

Participants: Dominique Lavenier, Rumen Andonov, Olivier Collin, Guillaume Collet, Guillaume Launay, Odile Rousselet, Fabrice Legeai, Alexandre Cornu, Guillaume Rizk, Van Hoa Nguyen.

The increasing flow of genomic data provided by the steadily improvement of new biotechnologies cannot be now efficiently exploited without a systematic in silico analysis. Data need to be filtered, curated, classified, annotated, validated, etc., to be actively used in a discovery process.

Such treatments can be very critical, especially in terms of time. The design of complex pipelines is a tedious and error-prone activity which requires consequent human resources. The execution time of several bioinformatics program can also be a major bottleneck when huge amount of data need to be processed. The BioWIC environment aims to save time in both directions.

Partnaires of the BioWIC project are the CAIRN INRIA group (Rennes), the MIG INRA group (Jouy-en-Josas) and the ELIAUS group (University of Perpignan).

7.2.2.3. *Modulome: Identifying and displaying modules in genomic sequences*

Participants: Catherine Belleannée, François Coste, Dominique Lavenier, Jacques Nicolas, Pierre Peter-Longo, Christine Rousseau.

This ANR project, Modulome [10], 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: URGI/Inra Versailles, LME/Ifremer Brest and GICC/CNRS Tours. See details in section 6.2.1.

7.2.2.4. *LepidOLF*

This ANR project is funded by the “blanc” program of the ANR. It started on october 2009. It is leaded by E. Jacquin-Joly from UMR PISC 1272 and implies INRIA Rennes Bretagne Atlantique (F. Coste) and UMR GABI INRA (P. Martin). LepidOLF ANR project aims at better understanding olfactory mechanisms in insects. The goal is to establish the antennal transcriptome of the cotton leafworm Spodoptera littoralis, a noctuid representative of crop pest insects. This antennal gene repertoire will then be used to identify for the first time the complete repertoire of olfactory genes (among them olfactory receptors) in a Lepidoptera crop pest. This repertoire will be used to design the first available ’antennal’ microarray. This array will be used to establish the molecular signature of different functional types of sensilla and will also offer us the possibility to investigate the mechanisms of olfactory plasticity at the peripheral level.

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 6.3.1

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

[10] http://www.irisa.fr/symbiose/projets/Modulome/

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, Basavanneppa Tallur, Julien Jacques, Jacques Nicolas.

The project [11] 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).

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

Participants: Rumen Andonov, Nicola Yanev.

This program is managed by the French Ministry of Foreign Affairs. 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.

We also collaborate with South-West University, Blagoevgrad in Bulgaria. This collaboration is supported by a bulgarian project DVU/01/197.

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

Participants: François Coste [correspondant], Matthias Gallé.

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

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 6.2.4). 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.

7.3.1.4. 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.

7.3.1.5. Freie Universität Berlin

Co-tutored Ph-D thesis with J. Bourdon.

7.3.2. Visiting scientists

7.3.2.1. Visitors

The following scientists visited the Symbiose project.

- Gabriel Infante-Lopez (from Cordoba, Argentina), 2 weeks visiting.
- Joerg Thuswaldner (Leoben, Austria), 1 week visiting.

7.3.2.2. Scientific visit exchanges

- C. Guziolowski, Chile university, Santiago de Chile, three monthes visiting, Jaunary 2010.
- Guillaume Risk, Colorado State University, four monthes visiting, October 2009 - January 2010.
- O. Radulescu, Bangalore, India, April 2009
- O. Radulescu, U.Chicago, USA, April 2009
- A. Siegel, Zun-Yat Tsen university, Guanzhou, China, July 2009.
- F. Coste, Cordoba, Argentina, July 2009.
- A. Siegel, University of Leoben, February 2009.

8. Dissemination

8.1. Leadership within scientific community

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

- Scientific Advisory Board of Institut Génétique et Développement, INSERM [J. Nicolas], then Scientific Advisory Board of ITMO Genetics Genomics and Bioinformatics. J. Nicolas has been in charge of the think group on the bioinformatics strategy at Inserm and coordinated a white paper on the definition of a national strategy in this area. It is involved in the new institute GGB of the National Alliance for Health and Life Sciences.
- Scientific Advisory Board of University of Rennes 1 [O. Radulescu].
- Member of the Evaluation Comitee of Inria [A. Siegel]
- Scientific Advisory Board of Biogenouest [O. Collin, J. Nicolas].
- Editorial Board of *Mathématiques et Sciences Humaines* [I.-C. Lerman].
- Scientific Advisory Board of Bioinfapa, INRA (Bioinformatics for Animal genomics) [J. Nicolas].

- Reviewer for Algorithms (F. Coste, M. Gallé), BMC Bioinformatics (A. Siegel, J. Nicolas), RAIRO (A. Siegel), Annales Institut Fourier (A. Siegel).

8.1.2. Jury of PhD Theses

- Jury, PhD Thesis of Hélène Daviet, *ClassAdd, une procédure de sélection de variables basée sur une troncature k-additive de l'information mutuelle et sur une Classification Ascendante Hiérarchique en prétraitement* , Université de Nantes [I.C. Lerman].
- President, PhD Thesis of Maroun Ojail, *Plateforme reconfigurable pour terminaux mobiles multi-standards et multimodes* [D. Lavenier]
- Jury, Habilitation Thesis of Laurent-Stéphane Didier [D. Lavenier]
- Referee, Ph-D thesis of Marc Ferré, *Analyse bioinformatique du protéome mitochondrial et du spectre des mutations de la protéine OPA1* [D. Lavenier]
- Jury, Ph-D thesis of Gwenaël Kervizic [A. Siegel]

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. Master 1 & 2 Modeling biological systems. (R. Andonov, P. Peterlongo, O. Radulescu, A. Siegel)
2. Bioinformatics, Biomedical technology, Ecole SupÃ©rieure d'Electronique de l'Ouest, Angers (D. Lavenier)
3. Formation continue - Université de Rennes - Initiation à R et Bioconductor (N. Le Meur)
4. Workshop - Centre de Recherche en Cancérologie de Marseille - FlowCore: A Bioconductor package for high throughput flow cytometry (N. Le Meur)
5. Cours DSS M2RI (Données Séquentielles et Symboliques, Master 2 Recherche en Informatique), Université de Rennes 1 (F. Coste)
6. School on Combinatorial Automata and Number Theory - Liège (A. Siegel)

Popular scientific work : Journal Doc Sciences for french secondary schools [29].

8.3. Conference and workshop committees, invited conferences

8.3.1. International invited conferences

- N. Le Meur, M Le Borgne, J Gruel and N Théret, Hong Kong.
- O. Radulescu, 2nd International workshop on model reduction in reactive flows Notre Dame USA.
- O Radulescu, NIH-INRIA meeting, Roquencourt, June 2009

8.3.2. National invited conferences

- F. Coste, Séminaire Combinatoire et Algorithmes, LITIS, Université de Rouen
- D. Lavenier, Gen2Bio
- D. Lavenier, TeraTec 2009, HPC: new challenges for bio

- D. Lavenier, GenOuest bioinformatics platform workshop
- D. Lavenier, CNRS Winter School, HPC on didcated Hardware
- N. Le Meur. Centre de Recherche en Cancérologie de Marseille
- N. Le Meur. Workshop Cell cycle and signal transduction: a biological and mathematical perspective Université Rennes 1, November 2-3, 2009
- N. Le Meur, F. Hahne, D. Sarkar and R. Gentleman. UseR 2009, Rennes.
- A. Siegel, Groupe de travail de systèmes dynamiques, Université paris XI.
- A. Siegel, Séminaire de théorie des nombres, Université de Grenoble.

8.3.3. Conference committees

- EGC 2009 (Extraction et Gestion de Connaissances, Strasbourg, Janvier 2009) (I.-C. Lerman)
- JOBIM, Nantes, (J. Bourdon, O. Collin, Organization Comitee), (F. Coste)
- JOBIM sattelite "Modélisation dynamique et simulation des réseaux biologiques", Nantes, (J. Bourdon, A. Siegel)
- JOBIM satellite "Jobim côté calcul" le 8 juin à Nantes (O. Collin)
- Parallel Bio-Computing 2009 (P. Peterlongo)
- International Conference on Field Programmable Logic and Applications (FPL) (D. Lavenier)
- International Conference on Engineering of Reconfigurable Systems and Algorithms (ERSA) (D. Lavenier)
- Southern Programmable Logic Conference (SPL) (D. Lavenier)
- EuroPar: Workshop on Highly Parallel Processing on a Chip (HPPC) (D. Lavenier)
- Workshop on Using Emerging Parallel Architectures for Computational Science (ICCS) (D. Lavenier)
- ACM International Conference on Computing Frontiers (UCHPC Workshop) (CF) (D. Lavenier)
- International Conference on ReConFigurable Computing and FPGAs (ReConFig) (D. Lavenier)
- ParCo: Parallel Computing with FPGAs (ParaFPGA) (D. Lavenier)
- SYMPosium en Architectures nouvelles de machines (SympA) (D. Lavenier)
- ACML'09 (The 1st Asian Conference on Machine Learning, November 2-4, 2009, Nanjing, China) (F. Coste)
- Cap 2009 (F. Coste)
- Steering Committee ICGI (F. Coste)
- Scientific committee of Zulu competition on active learning [12]: () (F. Coste)
- Organisation committee Biograle 2008 (O.Collin)
- Organisation committee Biograle 2009 (O.Collin, P. Peterlongo)
- Organisation committee ISyIP (O.Collin, F. Legeai)
- Organization of the conference Numeration: Mathematics and Computer Science, Marseille, 2009 (A. Siegel)

8.3.4. Seventh meeting of the Bioinformatics platform of Biogenouest

The meeting held at Irisa, Rennes on Oct. 26, 2009. It was organized by Olivier Collin. ([13]). Invited speakers included Véronique Receveur-Bréchot (CNRS Marseille) and Pierre Tufféry (Inserm Paris).

[12] http://cian.univ-st-etienne.fr/zulu/

[13] http://genouest.org/

8.3.5. BioInfoOuest thematic-day conferences

The Symbiose project regularly organizes thematic-day conferences on bioinformatics subjects[14]. 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.

- *Variability*. This conference day was organized by S. Blachon. Invited speakers were Anne Lopes (CEA), Jean - Jacques Kupiec (ENS, Paris) and Nicolas Voyer (Mitsubishi R&D).
- *New Generation Sequencing*. This conference day was organized by P. Peterlongo. Invited speakers were Thomas Le Calvez (Univ. rennes 1), Eric Rivals (LIRMM, Montpellier), Patrick Wincker (Genoscope CEA), Pierre Taberlet (Laboratoire d'Ecologie Alpine, Grenoble), Peter von Dassow (Station Bio Roscoff).

As a complement to these thematic days, the **team seminar** is held on a bi-weekly basis. 21 talks were given in this framework during the year 2009.

8.4. Theses defenses

8.4.1. PhD thesis, Van Hoa Nguyen

Parallel Intensive Genomic Sequence Comparison [2] The thesis was supervised by D. Lavenier.

The sequence comparison process is one of the main bioinformatics task. The new sequencing technologies lead to a fast increasing of genomic data and strengthen the need of fast and efficient tools to perform this task.

In this thesis, a new algorithm for intensive sequence comparison is proposed. It has been specifically designed to exploit all forms of parallelism of today microprocessors (SIMD instructions, multi-core architecture). This algorithm is also well suited for hardware accelerators such as FPGA or GPU boards.

The algorithm has been implemented into the PLAST software (Parallel Local Alignment Search Tool). Different versions are available according to the data to process (protein and/or DNA). A MPI version has also been developed. According to the nature of the data and the type of technologies, speedup from 3 to 20 has been measured compared with the reference software, BLAST, with the same level of quality.

8.4.2. PhD thesis, Jérémie Gruel

From biological data to molecular modelisation; application to TGF-beta signaling and hepatic fibrosis [1] The thesis was co-advised by Michel Le Borgne and Nathalie Théret.

Hepatic fibrosis is a complex pathology mainly due to chronic viral or toxic aggressions. In this context, hepatic stellate cells are the main producers of the excess of extra-cellular matrix modifying liver normal activity. The main pro-fibrotic cytokine is the TGF-beta and a perturbation of the TGF-beta signaling pathways is observed in hepatic stellate cells in the pathological context. The thesis aimed at a better understanding of these cellular regulation perturbations.

In the first part of this work, we studied the impact of the protein ADAM12 on TGF-beta receptors trafficking. Using a combination of differential models and qualitative experimental data, we have obtained predictions about the modifications induced by ADAM12 in this system.

In the second part of this work, we studied the transcriptional signatures present in the promoters of cytokine regulated genes, including TGF-beta. We have proposed an algorithm allowing to gather genes potentially sharing the same expression regulation mechanisms than a given gene of interest.

[14] http://www.irisa.fr/sci-events/seminars/bioinfo

9. Bibliography

Year Publications

Doctoral Dissertations and Habilitation Theses

[1] J. GRUEL. *From biological data to molecular modelisation; application to TGF-beta signaling and hepatic fibrosis*, Université de Rennes1, 2009, Ph. D. Thesis.

[2] V. H. NGUYEN. *Traitement parallèle des comparaisons intensives de séquences génomiques*, Université de Rennes 1, 2009, http://tel.archives-ouvertes.fr/tel-00435792/fr/, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

[3] J. AHMAD, J. BOURDON, D. EVEILLARD, J. FROMENTIN, O. ROUX, C. SINOQUET. *Temporal constraints of a gene regulatory network: refining a qualitative simulation*, in "BioSystems, Special Issue on Gene Regulatory Networks", 2009, 10.1016/j.biosystems.2009.05.002, http://hal.archives-ouvertes.fr/hal-00423353/en/.

[4] A. BANKHEAD, I. SACH, C. NI, N. LE MEUR, M. KRUGER, M. FERRER, R. GENTLEMAN, C. ROHL. *Knowledge based identification of essential signaling from genome-scale siRNA experiments.*, in "BMC Syst Biol", vol. 3, 2009, 80, http://hal.inria.fr/inria-00426749/en/.

[5] 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, 2009, p. 435-447, http://hal.inria.fr/inria-00359115/en/.

[6] Y. BIGOT, S. RENAULT, J. NICOLAS, C. MOUNDRAS, M.-V. DEMATTEI, S. SAMAIN, D. K. BIDESHI, B. A. FEDERICI. *Symbiotic Virus at the Evolutionary Intersection of Three Types of Large DNA Viruses; Iridoviruses, Ascoviruses, and Ichnoviruses*, in "PLoS ONE", vol. 4, n^{o} 7, 07 2009, e6397, http://dx.doi.org/10.1371/journal.pone.0006397.

[7] P. BLAVY, F. GONDRET, H. GUILLOU, S. LAGARRIGUE, P. MARTIN, J. VAN MILGEN, O. RADULESCU, A. SIEGEL. *A minimal model for hepatic fatty acid balance during fasting: Application to PPAR alpha-deficient mice*, in "Journal of Theoretical Biology", vol. 261, 2009, p. 266-278, http://hal.inria.fr/inria-00429806/en/.

[8] P. CASEL, F. MOREEWS, S. LAGARRIGUE, C. KLOPP. *sigReannot: an oligo-set re-annotation pipeline based on similarities with the Ensembl transcripts and Unigene clusters*, in "BMC Proc", vol. 3 Suppl 4, 2009, S3.

[9] R. CHIKHI, D. LAVENIER. *Paired-end read length lower bounds for genome re-sequencing*, in "Bmc Bioinformatics", 10 2009, http://hal.inria.fr/inria-00426856/en/.

[10] A. CRUDU, A. DEBUSSCHE, O. RADULESCU. *Hybrid stochastic simplifications for multiscale gene networks*, in "Bmc Systems Biology", vol. 3, 2009, 89, http://hal.inria.fr/inria-00431227/en/.

[11] E. FLEURY, A. HUVET, C. LELONG, J. DE LORGERIL, V. BOULO, Y. GUEGUEN, E. BACHERE, A. TANGUY, D. MORAGA, C. FABIOUX, P. LINDEQUE, J. SHAW, R. REINHARDT, P. PRUNET, G. DAVEY, S. LAPEGUE, C. SAUVAGE, C. CORPOREAU, J. MOAL, F. GAVORY, P. WINCKER, F. MOREEWS, C. KLOPP, M. MATHIEU, P. BOUDRY, P. FAVREL. *Generation and analysis of a 29,745 unique Expressed Sequence Tags*

from the Pacific oyster (Crassostrea gigas) assembled into a publicly accessible database: the GigasDatabase, in "BMC Genomics", vol. 10, Jul 2009, 341.

[12] M. GALLE, P. PETERLONGO, F. COSTE. *In-place update of suffix array while recoding words*, in "International Journal of Foundations of Computer Science", 2009, http://hal.inria.fr/inria-00430406/en/.

[13] A. GATTIKER, L. HERMIDA, R. LIECHTI, I. XENARIOS, O. COLLIN, J. ROUGEMONT, M. PRIMIG. *MIMAS 3.0 is a Multiomics Information Management and Annotation System*, in "BMC Bioinformatics", vol. 10, n^{o} 1, 2009, 151, http://www.biomedcentral.com/1471-2105/10/151.

[14] A. GORBAN, O. RADULESCU, A. ZINOVYEV. *Asymptotology of Chemical Reaction Networks*, in "Chemical Engineering Science", 2009, http://hal.inria.fr/inria-00431225/en/.

[15] J. GRUEL, M. LE BORGNE, N. LE MEUR, N. THÉRET. *In silico investigation of ADAM12 effect on TGF-beta receptors trafficking.*, in "BMC Res Notes", vol. 2, 2009, 193, http://dx.doi.org/10.1186/1756-0500-2-193.

[16] C. GUZIOLOWSKI, A. BOURDÉ, F. MOREEWS, A. SIEGEL. *BioQuali Cytoscape plugin: analysing the global consistency of regulatory networks*, in "Bmc Genomics", vol. 26, 2009, 244, http://hal.inria.fr/inria-00429804/en/.

[17] F. HAHNE, N. LE MEUR, R. R. BRINKMAN, B. ELLIS, P. HAALAND, D. SARKAR, J. SPIDLEN, E. STRAIN, R. GENTLEMAN. *flowCore: a Bioconductor package for high throughput flow cytometry.*, in "Bmc Bioinformatics", vol. 10, 2009, 106, http://hal.inria.fr/inria-00426746/en/.

[18] F. LEGEAI, S. SHIGENOBU, J.-P. GAUTHIER, J. COLBOURNE, C. RISPE, O. COLLIN, S. RICHARDS, A. WILSON, D. TAGU. *AphidBase: A centralized bioinformatic resource for annotation of the pea aphid genome*, in "Insect Molecular Biology", 2010, sous presse.

[19] 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/.

[20] M. MANU, S. SURKOVA, A. V. SPIROV, V. GURSKY, H. JANSSENS, A.-R. KIM, O. RADULESCU, C. E. VANARIO-ALONSO, D. SHARP, M. SAMSONOVA, J. REINITZ. *Canalization of gene expression and domain shifts in the Drosophila blastoderm by dynamical attractors*, in "PLoS Computational Biology", vol. 5, 2009, 3, http://hal.inria.fr/inria-00431228/en/.

[21] M. MANU, S. SURKOVA, A. V. SPIROV, V. GURSKY, H. JANSSENS, A.-R. KIM, C. E. VANARIO-ALONSO, O. RADULESCU, D. H. SHARP, M. SAMSONOVA, J. REINITZ. *Canalization of gene expression in the Drosophila blastoderm by gap genes cross regulation*, in "Plos Biology", vol. 7, 2009, 3, http://hal.inria.fr/inria-00431229/en/.

[22] V. H. NGUYEN, D. LAVENIER. *PLAST: parallel local alignment search tool for database comparison*, in "Bmc Bioinformatics", vol. 10, 10 2009, 329, http://hal.inria.fr/inria-00425301/en/.

[23] P. PETERLONGO, F. COSTE, M. GALLE. *In-place update of suffix array while recoding words*, in "International Journal of Foundations of Computer Science (IJFCS)", 2010.

[24] P. PETERLONGO, G. A. T. SACOMOTO, A. P. DO LAGO, N. PISANTI, M.-F. SAGOT. *Lossless filter for multiple repeats with bounded edit distance.*, in "Algorithms for Molecular Biology :: Amb", vol. 4, 2009, 3, http://hal.inria.fr/inria-00425377/en/.

[25] 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/.

[26] C. ROUSSEAU, M. GONNET, M. LEROMANCER, J. NICOLAS. *CRISPI: a CRISPR Interactive database*, in "Bioinformatics", vol. 24, n^{o} 25, 2009, p. 3317-3318, http://dx.doi.org/10.1093/bioinformatics/btp586.

[27] A. SIEGEL, J. THUSWALDNER. *Topological properties of Rauzy fractals*, in "Mémoires de la SMF", 2010.

[28] S. VAKULENKO, M. MANU, J. REINITZ, O. RADULESCU. *Size Regulation in the Segmentation of Drosophila: Interacting Interfaces between Localized Domains of Gene Expression Ensure Robust Spatial Patterning*, in "Phys.Rev.Lett.", vol. 103, 2009, 168102, http://hal.inria.fr/inria-00431224/en/.

Articles in Non Peer-Reviewed Journal

[29] J. NICOLAS. *Décoder le vivant*, in "Doc Sciences", 2009, p. 26-33, http://hal.inria.fr/inria-00438517/en/.

[30] A. SIEGEL, B. ADAMCZEWSKI, W. STEINER. *Présentation de "Numeration: Mathematics and Computer Science" (CIRM, 2009)*, in "Actes des rencontres du CIRM", vol. 1, 2009, p. 1-2, http://hal.inria.fr/inria-00429811/en/.

International Peer-Reviewed Conference/Proceedings

[31] S. BLACHON, G. STOLL, C. GUZIOLOWSKI, A. ZINOVYEV, E. BARILLOT, A. SIEGEL, O. RADULESCU. *Method for Relating Inter-patient Gene Copy Numbers Variations with Gene Expression via Gene Influence Networks*, in "AIAI 2009 Workshop – Workshop on Biomedical Informatics and Intelligent Approaches in the Support of Genomic Medicine (BMIINT), Grèce", vol. 475, 04 2009, p. 72-87, http://hal.inria.fr/inria-00429801/en/.

[32] J. BOURDON, I. RUSU. *Statistical Properties of Factor Oracles*, in "COMBINATORIAL PATTERN MATCHING, France", vol. LNBI 5577, Springer Verlag, 2009, p. 326-338, http://hal.archives-ouvertes.fr/hal-00415952/en/.

[33] S. CARAT, R. HOULGATTE, J. BOURDON. *A statistical method for PWM clustering*, in "Moscow Conference on Computational Molecular Biology, Russie", 2009, http://hal.archives-ouvertes.fr/hal-00415980/en/.

[34] R. CHIKHI, D. LAVENIER. *Paired-end reaf legth lower bounds for genome resequensing*, in "ISMB / ECCB, Stockholm", 2009.

[35] G. COLLET, R. ANDONOV, N. YANEV, J.-F. GIBRAT. *Protein Threading*, in "8th Cologne-Twente Workshop on Graphs and Combinatorial Optimization, France", 06 2009, http://hal.inria.fr/inria-00412642/en/.

[36] 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 CIBB 2008: Computational Intelligence Methods for Bioinformatics and

Biostatistics - Selected revised papers, Italie", vol. 5488, 2009, p. 144-155, http://hal.inria.fr/inria-00429809/en/.

[37] P. KUNTZ, I.-C. LERMAN. *Graph and Hierarchical-based approaches for structuring association rule sets*, in "11th IFCS International Conference 2009", H. BOCK (editor), Dresden University, 2009.

[38] N. LE MEUR, J. GRUEL, M. LE BORGNE, N. THÉRET. *Modeling the influence of EGF and TGF-b pathways in tumor progression of hepatocellular carcinoma.*, in "Asian Pacific Association for Study of the Liver, Hong-Kong", vol. 3, 2009, 64, FP107, http://hal.archives-ouvertes.fr/hal-00435766/en/.

[39] S. LEMOSQUET, J. BOURDON, J. GUINARD-FLAMENT, A. SIEGEL, J. VAN MILGEN. *A Generic Stochiometric Model to Analyse the Metabolic Flexibility of the Mammary Gland in Lactating Dairy Cows*, in "7th Workshop Modelling Nutrient Digestion and Utilization in Farm Animals, France", 2009, http://hal.archives-ouvertes.fr/hal-00416396/en/.

[40] F. MOREEWS. *A Framework for Building Reliable Distributed Bioinformatics Service Repositories*, in "ICWS 2009. IEEE International Conference on Web Services", July 2009, p. 1018–1019, http://dx.doi.org/10.1109/ICWS.2009.143.

[41] V. H. NGUYEN, A. CORNU, D. LAVENIER. *Implementing Protein Seed-Based Comparison Algorithm on the SGI RASC-100 Platform*, in "International Parallel and Distributed Processing Symposium - Reconfigurable Architectures Workshop - IPDPS, Italie", 2009, p. 1-7, http://doi.ieeecomputersociety.org/10.1109/IPDPS.2009.5161206.

[42] P. PETERLONGO, J. NICOLAS, D. LAVENIER, R. VORC'H, J. QUERELLOU. *c-GAMMA: Comparative Genome Analysis of Molecular Markers*, in "Pattern Recognition in Bioinformatics, Royaume-Uni", 08 2009, http://hal.inria.fr/inria-00425373/en/.

[43] G. RIZK, D. LAVENIER. *GPU accelerated Rna folding algorithm*, in "9th International Conference on Computational Science, États-Unis d'Amérique", J. ALLEN, AL (editors), vol. 5544, 05 2009, 1031, http://hal.archives-ouvertes.fr/hal-00425543/en/.

Workshops without Proceedings

[44] N. LE MEUR, J. GRUEL, M. LE BORGNE, N. THÉRET. *Multiclock discrete models of the eukaryotic cell cycle.*, in "Journées Ouvertes en Biologie, Informatique et Mathématiques. Journée Satellite, France", 2009, http://hal.archives-ouvertes.fr/hal-00435771/en/.

[45] N. LE MEUR, F. HAHNE, D. SARKAR, B. ELLIS, J. SPIDLEN, R. R. BRINKMAN, R. GENTLEMAN. *High throughput flow cytometry analysis with Bioconductor*, in "Advanced Flow Cytometry Analysis 2009, France", 2009, http://hal.inria.fr/inria-00435735/en/.

[46] N. LE MEUR, F. HAHNE, D. SARKAR, B. ELLIS, J. SPIDLEN, R. R. BRINKMAN, R. GENTLEMAN. *High throughput flow cytometry analysis with Bioconductor.*, in "useR! 2009, France", 2009, http://hal.archives-ouvertes.fr/hal-00435777/en/.

[47] A. SIEGEL. *From pure discrete spectrum conditions to topological properties of self-affine tiles*, in "Aperiodic Order, Royaume-Uni", 2009, http://hal.inria.fr/inria-00429814/en/.

[48] A. SIEGEL. *Rational numbers with purely periodic beta-expansions*, in "Fractals and Tilings, Autriche", 2009, http://hal.inria.fr/inria-00429815/en/.

Scientific Books (or Scientific Book chapters)

[49] I.-C. LERMAN. *Analyse de la vraisemblance des liens relationnels: une méthodologie d'analyse classificatoire des données*, in "Apprentissage artificiel et fouille de données, RNTI A3", Y. BENNANI, E. VIENNET (editors), Cépaduès, 2009, p. 93-126.

[50] N. PISANTI, M. GIRAUD, P. PETERLONGO. *Filters and seeds approaches for fast homology searches in large datasets*, in "Algorithms in computational molecular biology", M. ELLOUMI, A. Y. ZOMAYA (editors), John Wiley & sons, 2009, http://hal.inria.fr/inria-00425370/en/.

Research Reports

[51] T. BITARD FEILDEL. *Création d'une bibliothèque de coeurs structuraux pour le protein threading. Utilisation des familles structurales.*, Université de Rennes 1, 2009, http://hal.inria.fr/inria-00435113/en/, Stage.

[52] G. COLLET, R. ANDONOV, N. YANEV, J.-F. GIBRAT. *Flexible Alignments for Protein Threading*, INRIA, 2009, http://hal.inria.fr/inria-00355546/en/, Research Report.

[53] F. COLOMBEL. *Identification in silico d'un nouveau gène de la famille des TNF (tumor necrosis factor)*, INSA, 2009, http://hal.inria.fr/inria-00431114/en/, Stage.

[54] I.-C. LERMAN, P. KUNTZ. *Directed binary hierarchies and directed ultrametrics*, n^{o} 6815, IRISA-INRIA, February 2009, Rapport de recherche.

[55] M. LE BORGNE. *Solving loosely coupled constraints*, IRISA / INRIA, 2009, http://hal.inria.fr/inria-00397098/en/, Research Report.

[56] I.-C. LERMAN, P. KUNTZ. *Directed binary hierarchies and directed ultrametrics*, IRISA / INRIA, 2009, http://hal.inria.fr/inria-00357532/en/, Research Report.

[57] I.-C. LERMAN, P. KUNTZ. *Directed binary hierarchies and directed ultrametrics*, IRISA / INRIA, 2009, http://hal.inria.fr/inria-00361727/en/, Research Report.

[58] N. MALOD-DOGNIN, R. ANDONOV, N. YANEV. *Maximum Cliques in Protein Structure Comparison*, IRISA / INRIA, 2009, http://hal.inria.fr/inria-00422198/en/, Research Report.

[59] N. MALOD-DOGNIN, R. ANDONOV, N. YANEV. *Solving Maximum Clique Problem for Protein Structure Similarity*, IRISA / INRIA, 2009, http://hal.inria.fr/inria-00356816/en/, Research Report.

[60] V. PICARD. *Pondération d'automates modélisant des familles de protéines et significativité des scores*, ENS Cachan antenne Bretagne, 2009, http://hal.inria.fr/inria-00431111/en/, Stage.

References in notes

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

[67] 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.

[68] 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.

[69] 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.

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

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

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

[82] 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.

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

[84] 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.

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

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

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

[88] 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.

[89] 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.

[90] 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.

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

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

[93] 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.

[94] 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/.

[95] 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.

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

[97] 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.

[98] 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.

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

[100] 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.

[101] 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.

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

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

[104] 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.

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

[106] T. SMITH, M. WATERMAN. *Identification of common molecular subsequences*, in "J. Mol. Biol.", n^{o} 147, 1981, p. 195-197.

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

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

[109] 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.

[110] 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.

[111] 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.

[112] 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.

[113] 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.

[114] 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.