INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team symbiose

SYstèmes et Modèles BIOlogiques, BIOinformatique et SEquences

Rennes

THEME BIO

Activity Report

2006

Table of contents

1. Team

The Symbiose project has been created in 2002. Its general purpose concerns bioinformatics, that is, modelling 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.

Head of project
Jacques Nicolas [Research Scientist (Inria)]

Administrative assistant
Marie-Noëlle Georgeault [Administrative Assistant (Inria)]

Inria staff members
François Coste [Research Scientist]
Ovidiu Radulescu [Associate Professor, Research Scientist, since Oct. 2005 – *délégation*]
Nicola Yanev [Visiting Scientist, apr 2005 / apr 2006 - poste d'accueil]

CNRS staff members
Dominique Lavenier [Research Director, HdR]
Anne Siegel [Research Scientist]

Faculty members
Rumen Andonov [Professor, univ. Rennes 1, HdR]
Catherine Belleannée [Associate Professor, univ. Rennes 1]
Michel Le Borgne [Associate Professor, univ. de Rennes 1]
Israël-César Lerman [Emeritus Professor, univ. Rennes 1, HdR]
Basavanneppa Tallur [Associate Professor, univ. Rennes 1, HdR]
Raoul Vorc'h [Associate Professor, univ. Rennes 1]
Giuseppe Lancia [Visiting scientist, univ. Udine, Italy, 1 month]

Research scientists (partners)
Laurence Duval [Assistant professor, ENSAI, Bruz]
Stéphane Rubini [Assistant professor, univ. Bretagne Ouest]

Post-doctoral fellow
Pierre PeterLongo [Post-doctoral fellow Inria since oct. 2006]

Ph. D. students
Mathieu Giraud [Ph. D. Student (AMC)]
Goulven Kerbellec [Ph. D. Student Inria/Région]
Sébastien Tempel [Ph. D. Student MENRT]
Philippe Veber [Ph. D. Student Inria]
Guillaume Collet [Ph. D. Student MENRT since oct. 2006]
Carito Guziolowski-Vargas [Ph. D. Student Bourse Conicyt/Ambassade de France/INRIA since oct. 2006]
Noël Malod-Dognin [Ph. D. Student Inria/Region since dec. 2006]
Van Hoa Nguyen [Ph. D. Student Inria/CORDI since dec. 2006]
Thibault Hénin [Ph. D. Student ENS]
Jérémy Gruel [Ph. D. Student INSERM]
Pierre Blavy [Ph. D. Student INRA]

Technical staff members (Ouest-Genopole bioinformatics computing center)
Hugues Leroy [Technical staff (Inria)]
Anthony Assi [Junior technical staff(Inria contract genopole)]
Gregory Ranchy [Junior technical staff (Inria contract genopole)]
Laetitia Guillot [Junior technical staff (Inria contract genopole)]
Sophie Roucan [Junior technical staff (Inria contract genopole)]

Olivier Filangi [Junior technical staff (Inria contract genopole)]
Gilles Georges [Junior technical staff]
Christine Rousseau [Junior technical staff (ANR contract Modulome)]
François Moreews [Junior technical staff (at 20%, national program Inra/Sigenae)]
Patrick Durand [Visiting Scientist, until oct. 2006 (Inria contract)]
Anne-Sophie Valin [Junior technical staff,until may 2006 (Inria contract genopole)]

Graduate student interns

Mathieu Morvan [Master 2 Info, Rennes / A. Siegel]
Nicolas Bitouzé [1st year ENS Cachan / F. Coste]
Yves-Pol Denielou [Master 2 Info, Rennes /F. Coste]
Thi Hong Hanh Hoang [4th year INSA / F. Coste]
Hélène Darolles [Master2 Info, Paris / D. Lavenier]
François-Julien Bourget-Marbaud [4th year INSA / D. Lavenier]
Mai Fei [INRIA Internship with Honk-Hong Univ., 6 months / R. Andonov]
Stefan Gerdgikov [INRIA Internship with Sofia Univ. / R. Andonov]
Carito Guziolowski-Vargas [Master 2 bioinformatics/ A.Siegel]
Nadia Bentayeb [Master 1 bioinformatics/ A. Siegel]
Clement Chatelain [1st year ENS Cachan/ A. Siegel]
Jeremy Gruel [Master 2 bioinformatics/ P. Veber]
Laurent Chamoin [Master 1 bioinformatics/ M Le Borgne]
Thibault Hénin [Master2 informatique/ O. Radulescu]
Anne-Cécile Morin [Master2 informatique École Polytechnique Nantes I.-C. Lerman]
Mickaël Taillé [Master2 informatique École Polytechnique Nantes I.-C. Lerman]

2. Overall Objectives

2.1. A Bioinformatics center

Bioinformatics has a quite large acceptation and we first delimit the definition we use in our framework: it specifies research at the interface between computer science and molecular biology (also called computational biology) and not all "standard" informatics that is necessary to manage biological data on a daily basis. However, it is hard to achieve in depth research in this domain without "biocomputing", that is, participating to services of the second kind with biologists. This is why we have decided to create a Bioinformatics Centre, with a research team, Symbiose, leaning back against a bioinformatics platform, Genouest (or the converse..) and an active participation in education (master program in bioinformatics). This report is mainly on the research project. The activity of the bioinformatics platform is described in the section 5.1.

Symbiose is a project of bioinformatics, devoted to modelling and analysing genomic data (DNA sequences, expression of RNA, proteins and metabolites). It is interested in trying to solve real biological questions in cooperation with biological labs and to combine multiple research fields in computer science towards this goal (with the hope to get also some return in computer science).

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 optimisation, analyses of systems of qualitative equations and formal language modelling. Our goal is to push forward the range of applicability of these methods, by designing **dedicated machines**.

2.2. Scientific axes

The *Scientific axes* on which the project focuses derive from our choice on modelling complex biological systems in a discrete framework, while managing efficiency issues. More precisely, the project links together three main directions.

2.2.1. Analysis of structures in sequences

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 try to answer three types of problems: the design of grammatical models on biological sequences; efficient filtering and model matching in data banks; maching learning of grammatical models from sequences.
We have an interest in both theoretical questions (language representations, search space) and practical questions (how to implement efficient parsers, how to infer language representations from a sample of sequences?). We follow a combinatorial approach. Corresponding disciplinary fields are algorithmic on words, machine learning, data analysis and combinatorial optimization.

2.2.2. Dedicated hardware architectures for bioinformatics

The fast access to millions of genomic objects has become a central scientific challenge. We investigate the usage of parallelism to speed up computations in genomics. Topics of interest range in intensive sequence comparisons to pattern or model matching, including structure prediction. We work on the design of hardware architectures tailored to the treatment of such applications. It is mainly based on the study of reconfigurable machines employing Field Programmable Logical Arrays (FPGA). Other activities concern GRID computing and parallelization of optimization algorithms.

2.2.3. Gene expression data: analysis and network modelling

The first purpose of analysis of biological sequences is to characterize each gene individually and to explore gene regulations by means of identifying regulatory cis-elements. But the ultimate goal, for the biologist, is to explain how the combination of genetic and metabolic interactions determines the phenotype which is observed at the molecular level, particularly in case of diseases. The scarcity of quantitative data on biological phenomena implies the use of qualitative models. Our approach is based on the definition of graph models of biological networks and the derivation of discrete or differential models for explaining and predicting (in a broad meaning) the behavior of the biological system. This research is rooted in various fields: data analysis, graph theory, discrete event systems, qualitative theory of differential systems.

3. Scientific Foundations

3.1. Bioinformatics

We study models on the macromolecular level of life (DNA, RNA, protein or metabolic molecules). The aim is to understand the structure, the activity, and more generally, the interactions and dynamics that may exist between components, for a general mechanism or a particular metabolic pathway. It is possible to distinguish four classes of studies (for more information, see for instance the introductory part of [94]) :

- *Data collecting.* The main unsolved research issue is the reconstruction of a sequence from its fragments after sequencing and/or mass fingerprinting. Finishing an assembly remains a hard task. There exists a renewal of interest in this area due to the multiple sources of data and to the raise of metagenomics (considering several genomes simultaneously).
- *Data and Knowledge management.* It is actually a major issue. Information is produced in a highly distributed way, in each laboratory. Normalization of data, 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.

- *Analysis of similarities/differences.* 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 sequences or structures. More macroscopic studies are also possible, involving more complex operations on genomes such as permutations. Once sequences have been compared, phylogenies, that is, trees tracing back the evolution of genes, may be built from a set of induced distances. A more recent track considers 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.
  Our own work in this area concerns the aspects involving intensive computing, classification and proteins comparison.
- *Functional and structural analysis of genomic data.* It is a wide domain, that aims at extracting biological knowledge from Xome studies, where X varies from genes to metabolites. It covers the search for genes and active functional sites, the determination of spatial structures, and, more recently, the study of interactions between macromolecules and with metabolites, particularly in regulation mechanisms. Biological sequences, as regards to DNA, RNA or proteins, must verify a number of important constraints with respect to the structure, the function or the activity that this sequence must exert. These constraints result in the conservation during evolution of "patterns" more or less precise and complex[1]. Complexity can range from the presence of given letters at given positions in the sequence, to long distance relations between words, due to spatial folding of the molecules, with phenomena of symmetry, copy, approximation, etc. The conservation of patterns not only makes it possible to characterize a family of sequences, but also to explain to a certain extent the structure/function relations. These patterns, made up manually or automatically, are then placed at the disposal of the community in banks like Prosite or eMOTIF for proteins [2] or TRRD for DNA, or through prediction programs for biologically important sites (intron/exon transition, open reading frames, etc.).

Their knowledge can be used in multiple applications in biology: characterization of families of proteins (many laboratories are studying a particular subfamily of proteins and can then amplify their discoveries by seeking in public banks all proteins matching the patterns found); characterization of genes expression (patterns located upstream genes provide important information on the probable localization of genes and the regulation of their expression level); protein annotation, i.e. to get clues on the functional family, the activity or the localization of a new protein.

3.2. Syntactical Analysis of sequences

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

3.2.1. Formal Languages and biological sequences

Sequences are considered as words on an alphabet of nucleic or amino acids. The set of superimposed structural and functional constraints leads to the formation of a true language whose knowledge would enable to predict the properties of the sequences. The theory of languages formalizes the basic concepts underlying the studied phenomena (degree of expressivity, complexity of the analysis, associated automata, algebra on languages). Still very few authors have explored this paradigm. It can be studied from two points of view:

- A fundamental point of view, where the goal is to define and study the most adapted classes of formal languages for the description of observed natural phenomena. The splicing systems of Head [80], or H-systems, reproducing the phenomenon of crossing over, represent one of the most

[1] we also use the term "signature" to specify that these patterns are not linked to consensus and can have an arbitrary complexity.

[2] http://www.expasy.org/prosite, http://motif.stanford.edu/emotif

fertile formalism in this respect. Language theorists like A. Salomaa and Gh. Paun [102] also explored standard questions (complexity, decidability, stable languages, etc) when faced with natural operations on biological sequences (inversion, transposition, copy, deletion, etc) and proposed in particular a model called Sticker-system based on the operation of complementarity as it occurs in Watson Crick pairings [86]. They aim at developing systems having the power of Turing Machines, in the line of works on DNA-computing, which is a bit different from the issue of deciding the class of languages necessary to describe biological structures. The current agreement is that the necessary expressivity is the class of "mildly context sensitive" languages, well-known in natural language analysis. For example Y. Kobayashi and T. Yokomori modeled and predicted the secondary structures of RNAs using Tree Adjoining Grammars (TAGs) [127]. The most complete work in this field seems due to D. Searls [113], [114] ;

- A more practical point of view, where 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 [63]. D. Searls proposed a more systematic approach based on logical grammars and a parser, Genlang [66]. Genlang remains still rarely used in the community of biologists, probably because it requires advanced competences in languages. We started our own work from this solution, keeping in mind the need for better accessibility of the model to biologists.

In practice, the biologist is often unable to provide sufficient models. To assist him in building relevant models necessitates the development of machine learning techniques.

3.2.2. Pattern Discovery

Because of its practical importance and the increasing quantity of available data, a number of pattern discovery methods have emerged since a few years. Particularly, due to the massive production of expression data from DNA chips, lots of papers have been proposed on pattern discovery in promoter sequences. Reviews of the field are available in [53] or [83]. The first criterion to classify methods is the type and expressivity of patterns they look for. 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, one finds Hidden Markov Models and stochastic automata, which have very good performances, but where classically the structure is fixed and learning is achieved on the parameters of the distribution. Thus, they are more related to the first type of representation. Distributional representations are expressed via various modalities : consensus matrices (probability of occurrence of each letter at each position), profiles (taking into account gaps), weight matrices (quantity of information at each position and contribution of each letter). At the algorithmic level, alignments play a fundamental role. One scans for short words in the sequences, then alignments are carried out by dynamic programming around these "anchoring" points. The production of "blocks" is typical of this approach [82]. A simplified search of patterns can be done after alignment, the variable intervals between subpatterns having been decided. Most powerful programs in this field are currently Gibbs Motif Sampler, a Bayesian procedure building a consensus matrix by Gibbs sampling with organism-specific higher order models (Markov chain) for prior frequencies estimate [93], Toucan, proposing a complete workbench for regulatory sequence analysis and a Gibbs sampler, Motif Sampler, and Meta-Meme, building a Markov network combining such matrices, produced by EM (Expectation-Maximization) algorithm.

The linguistic representation, which corresponds to our own work, generally rests on regular expressions. Algorithms use combinatorial enumeration in a partially ordered space. Among the most applied in this field, one finds the Pratt program [52], using principles very close to those found in the work of M.-F. Sagot and A. Viari [110]. Another track explores variations on the search for cliques in a graph [89], [56].

Even if results obtained so far are interesting in a number of cases, we think that there is a fundamental limitation to current studies: they all remain rather strongly dependent on the concept of position. It is primarily the presence at a given position of some class of letters which will lead to the prediction. However it is clear that relations exist between various sites – sometimes distant on the sequence – and play an important biological role. Some recent methods do consider distantly related patterns. There is no doubt that this issue will be fundamental in the next years. A purely statistical learning seems to have reached its limits here, because of the multiplication of parameters to be adjusted. The theoretical framework of formal languages, where one can seek to optimize this time the complexity of the representation (parsimony principle), seems to us more adapted. We are engaged in this research track, where pattern discovery becomes language learning. This does not preclude the use of statistical techniques that are essential for the treatment of real, noisy data, but our main contribution will be in the field of grammatical inference.

3.2.3. Machine Learning and Grammatical Inference

Machine Learning is a research field devoted to studying the design and analysis of algorithms for making predictions about the future based on past experiences. Taking roots in Artificial Intelligence and Statistics, it focuses on the study of learning algorithms inspired as well by a cognitive view of natural learning from experience as by statistical techniques for fitting model parameters to data. Research is achieved from a theoretical point of view (Computational Learning Theory), studying learnability criteria and learnable classes of function within these criteria, and from a more practical point of view (applied Machine Learning), focusing more on the algorithms and their performances measured on real or simulated tasks. Recent techniques mix both points of view, like for example, *boosting* techniques (allowing good performances from initial weak learner) or the development of *support vector machines* (applying structural risk minimization principle from statistical learning theory). Integrating statistical tools is a growing trend: one can cite reinforcement learning, classification or statistical physics and also research in neural networks or hidden Markov models (HMM). The problem of comparing and integrating these symbolic and numerical approaches has been extensively studied [70].

Hidden Markov models are ubiquitous in bioinformatics. They 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 contrast, determining the structure remains a difficult task. When available, domain knowledge may help to design empirically a structure but, in practice, the structure used is often very simple (e.g. left-right models like Profile HMM) and the discriminative power of HMM relies essentially on its parameter choice.

In the Symbiose project, we are studying this problem in the more general framework of Grammatical Inference. Grammatical Inference, variously referred to as automata induction, grammar induction, and automatic language acquisition, refers to the process of learning grammars and languages from sequences. Let us notice that the emphasis is not only on learning language (i.e. a set of sequences) but also on learning grammars (i.e. structural representations of the sequences of the language).

Traditionally, Grammatical Inference has been studied by researchers in several research communities including: Information Theory, Formal Languages, Automata Theory, Computational Linguistics, Pattern Recognition, etc. The grammatical inference community organize itself around its main conferences (e.g., the International Colloquium on Grammatical Inference, since 1993) and workshops. Japan, USA, Australia, Spain, Netherlands and France (with teams in St Etienne, Lille, Marseille, Rennes, Lannion) are among the most represented countries in this tight community.

A grammatical inference problem involves 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 (knowledge) about the domain 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 [111]. 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.

3.3. Modelling and analyzing genetic networks

3.3.1. Biological context

The genomes of multiple species being sequenced, a main question arises, dealing with integrative biology: how is genetic information used so that a given organism is able to develop and survive? Differences on a single gene may explain some simple (or Mendelian) characters as monogenetic diseases, color phenotypes, etc. However, a major part of phenotypic characters derive from the combined action of many genes. These interactions lead to complex genetic models for phenotypic characters, especially if one takes into account the influence of the environment on the character.

Networks are natural models for gene interactions: they appear to be abstract enough to be formalized while enabling to represent the complexity of a biological organism. In this framework, dynamics is essential: an organism cannot be understood without considering its development; similarly, the functions of a network cannot be separated from its dynamics.

Technically, this global point of view is motivated by the recent emergence of new high throughput techniques (DNA chips for gene activity, Chip on Chip for DNA/protein interactions, mass spectroscopy for protein interactions). A novel approach of molecular biological phenomena underlies these techniques: simultaneous observations on a mass of genes are available and the system itself has to be modeled. This contrasts sharply with the traditional approach in biology that focuses on isolated molecular interactions.

3.3.2. Systems biology: models and data

The field of *systems biology* appeared as a response to increasing need for analytical approaches in molecular biology. Its goals include modelling interactions, understanding the behaviour of a system from the interplay of its components, confronting the prediction of the model to data, and inferring models from data. Solutions to these challenges are often interdisciplinary.

modelling cellular interactions is an old domain of biology, initiated by biologists interested in the dynamics of enzymes systems [84]. Models for genetic networks appeared as soon as gene interactions were discovered. The simplest static model consists in modelling a genetic network as an oriented graph, with labels + (activation) or - (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. However, this technique appears to be incomplete without the support of literature information [119].

The dynamical framework includes simulations and prediction of behaviours; models can be either qualitative or quantitative, as reviewed in [64], [61], [90]. 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 [46], [67], [122], [97], [121]. The properties of continuous models can be studied with convex analysis, linear and non-linear control techniques [68], [81], [101], [45]. Stochastic models transform reaction rates into probabilities and concentrations into numbers of molecules, allowing to understand how noise influences a system [108], [85]. Finally, in the discrete models, each component is assumed to have a small number of qualitative states, and the regulatory interactions are described by discrete functions. Relevant discrete frameworks can be boolean [88], [112], logical [87], [109], or Petri networks [96], [60]. The bridge between continuous and discrete models is made by piecewise linear differential models [65], [71].

Each of these methods addresses in complementary ways 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 [116], [120], [117], [61]. They represent powerful tools to acquire a fine grained knowledge of the system at hand, but 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 to a limited number of variables.

Model identification addresses a different objective, that is, to form or modify a model consistently with a set of data. A first framework for identification consists in building models from scratch, using statistical techniques such as Bayesian networks [69], [98] or kernels [125]; these are particularly accurate when large amounts of data are available. Another efficient approach formalizes a priori knowledge as partially specified models. Fitting models to data is obtained by means of various techniques, depending on the class of models, that can be discrete [48], [128], [55], [109], continuous [47], [51], [90] or hybrid [57], [91]. Qualitative reasoning, hybrid system, constraint programming or model-checking allow either to identify a subset of active processes explaining experimental time-series data [48], [128], [55], [109] or to correct the models and infer some parameters from data [47], [59]. The identification methods are limited to a few dozen components. Model correction or parameter regression can cope with up to hundreds of products [59] provided that the biomolecular mechanisms and supplied kinetic data are accurate enough.

3.3.3. Qualitative data

Qualitative data such as DNA microarrays data cannot be easily used in most of the frameworks described above for two main reasons. First, the model-based identification approach has difficulties to take into account the errors and the variability that commonly affect measured expression levels in DNA microarrays. Secondly, time series data is not easily available and in many situations (for instance disease studies on clinical tissues) microarrays provide static data, meaning that they inform more on steady state shifts under perturbations than on the dynamics of the system.

The philosophy of our project is to develop techniques around network modelling, using models adapted to the kind of observations available with the biological techniques at hand. The methods we develop have two characteristics:

- Our models integrate simultaneously a biochemical (metabolic or signalling) component and a genetic component. Genetic actors are activated in the framework of complex metabolic or signaling pathways, that have their own dynamics. Contrary to simple organisms, in pluricellular organisms, biochemical phenomena have a real influence on genetic interactions, and need to be modeled precisely. Our goal is to understand better the relations between these two components.
- We follow a qualitative modelling approach, using either discrete event networks or qualitative equations derived from differential models.

3.4. Parallelism

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

Mixing parallelism and genomics is both motivated by the large volume of data to handle and by the complexity of certain algorithms. First, there are data coming from intensive genome sequencing. Today, (october 2005) about 300 genomes – including the human genome – are completely sequenced, and there exist more than 1000 other sequencing projects (see *Genomes online database*[3]). All these data are stored into huge data bases whose volume approximatively doubles every year. The growth is exponential and there is no reason to expect any decline in the next few years.

Thus, 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) can have some related functionality. Even if this axiom may not be true, it can give precious clues for further investigations.

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

The first algorithms for comparing genomic sequences have been developed in the seventies. They were essentially based on dynamic programming technics [99], [115]. Then, with the increasing growth of data, faster algorithms have been designed to drastically speed-up the search. The Blast software [118] 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. The LASSAP software developed by JJ Codani, Inria [75] has been designed in that direction: it parallelizes a standard suite of bioinformatics tools dedicated to intensive genomic computations.

Other ways of research have also been investigated to speed-up the search in large genomic banks, in particular dedicated hardware machines. Several research prototypes such as SAMBA [78], BISP [62], HSCAN [76] or BioScan [123], have been proposed, leading today to powerful commercial products: BioXL, DECYPHER and GeneMatcher coming respectively from Compugen ltd. TimeLogic and Paracel [4].

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.

But 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 in the protein folding research activity [50]. As a matter of fact, predicting the 3D structure of a protein from its amino acid sequence is an extremely difficult challenge, both in term of modelling and computation time. The problem is investigated following many ways ranging from *de novo* folding prediction to protein threading technics [94]. The first method tries to predict the spatial organization of a protein using only the sequence information. The second method tries to match an unknown protein sequence to a known 3D protein structure. The underlying algorithms are NP-complete and require both combinatorial optimization and parallelization approaches to calculate a solution in a reasonable amount of time.

4. Application Domains

4.1. Application Domains

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

The main stakes of bioinformatics are to assist in the processes of discovering prognostic, diagnostic and therapeutic targets and the understanding of biological mechanisms. This covers in practice a great variety of works.

The local context of OUEST-genopole provides us with a lot of collaborations with biology laboratories. We emphasize here three types of applications with major achievements in the project.

- **Targeted gene discovery** is studied with a syntactical approach. Models are built for proteins or promoters and then searched in whole genomes. We have for instance been able to discover new beta-defensins, a family of anti-microbial peptides, in the human genome with such a strategy.

[4] http://www.compugen.co.il/, http://www.timelogic.com, http://www.paracel.com

- **Whole genome analysis** is made practical through dedicated data structures and reconfigurable architectures. We have for instance proposed Blast comparisons on the human genome in 1 minute, built a software for bacterial genome fragmentation, GenoFrag, that helps to study genomes variations via Long Range PCR, and studied the occurrence of retro-transposons, a family of mobile genomic units, in the genome of *Arabidopsis thaliana*.
- **Genomic/metabolic interaction networks** are modeled in eukaryote organisms. We are studying genes and metabolites involved in the lipogenesis (chickens) and in TGF-beta-regulation in association with hepatocellular carcinomas (human). Recently we have included the known regulation model of E.Coli in our studies.

5. Software

5.1. Genouest, the Bioinformatics computing center of Ouest-Genopole

The bioinformatics platform [5] is linked to the Symbiose project, and propose a complete set of tools and databases for biologists and bioinformaticians. This platform received the national RIO label in December 2003 and has been a platform of University of Rennes 1 since this year. In 2006 the RIO label has been renew. O. Collin, from Station Biologique located in Roscoff, and H. Leroy are in charge of the boarding committee of the platform. We also participate to the National Network of bioinformatics platforms (ReNaBi, H. Leroy, J. Nicolas) and to the RIO national evaluation of these platform. The platform is supported by several contracts : CNRG 2005 and 2006, Région Bretagne 2005 and 2006.

We have organized the 4th annual meeting of the Bioinformatics platform of OUEST-genopole® : This meeting [6] held at Irisa, Rennes, on october 24 2006. Invited speakers included A. De Daruvar (Bordeaux, Bioinformatics centre), Catherine Letondal (Pasteur computing centre), Philippe Picouet and Xavier Bailly (Brest, Ecole Normale Supérieure Télécoms). We have also organized a calendar of training courses on bioinformatics tools (Ouest genopole and UMR 6026). [Sophie Roucan].

The research team Symbiose transfers his results to Genouest: all its developments are progressively made available within this platform of services in biocomputing. It allows our team to filter from routine service requests new subjects of research with a good relevance in biology and conversely to offer a better usability of research prototypes to biological labs. We propose original tools for complex filtering of sequences. This includes GenoFrag for PCR Scanning, Wapam, STAN and ModelDesigner for pattern matching, and a set of pattern discovery algorithms. A first version of a graphical analyser for regulatory and metabolic networks is also available. This year, the genouest.org web site has been rewritten using a content management software (Spip), with a complete reorganisation of pages under five tabs : tools, databases, training, help, general information. A RSS news feed is also available. A new hardware platform has been installed during the first semester.

5.2. Bioinformatics Toolbox

Participants: Sophie Roucan [correspondant], Anthony Assi, Laetitia Guillot, Grégory Ranchy, Anne-Sophie Valin, Dominique Lavenier, Hugues Leroy, Jacques Nicolas.

The toolbox [7] groups together accesses to standard tools (e.g. GCG package) and adapted softwares tailored to biologists needs collected in Ouest-genopole. One of the most recurrent demand is the possibility to make a Blast against a personal bank. This tool allows to perform a more relevant and faster search in this context. The main activity concerns the generation of primers.

[5] http://genouest.org/

[6] http://www.irisa.fr/videos/genopole/bioInfo2006/index.html

[7] http://genouest.org/tools.php

5.2.1. Specific primers http://genouest.org/primers.php

CAPS Tags. CAPS means Cleaved Amplified Polymorphic Sequence. The goal of this tool is to highlight differences between two related sequences. First, we virtually digest the two sequences with Emboss restrict program, secondly we align them with Multalign. We display single enzyme cuts, taking into account the gaps appeared in the alignment. Differences are validated with the alignment, in this case a difference is a potential SNP.

Degenerate primers. A way to look for new genes is to use degenerate primers. Data are a set of protein sequences, from different species, with the same biological function. We align this set of sequences with Multalign. We extract from the calculated consensus sequence longest fragments with few ambiguous amino acids. After manual validation of one or several fragments, we degenerate each fragment from the 3' end. We have developed a module, working with degenerate alphabet and codon usage tables, who reverse translate protein sequences in nucleic sequences, computing and bounding a degeneration cost.

Microsatellite primers. Microsatellites are shorty repeated sequences that are primers markers in genome mapping. Data are a set of nucleic sequences in Fasta format. We use Sputnik to find microsatellites of chosen length in these sequences. Then we try to design PCR primers in the sequences containing a microsatellite with primer3.

5.2.2. GenoFrag

The goal of GenoFrag is to deal with Whole Genome PCR Scanning (WGPS), a means for analyzing bacterial genome plasticity. This software is developed for the design of optimized primers for Long-Range PCR on whole genomes. GenoFrag initially seeks all the potential primers on a chromosome. Then it calculates the best distribution of the primer pairs, thanks to combinatorial optimization algorithms. It was tested on *Staphylococcus aureus* strains but can be used for other bacterial or viral species [49], [44]. A graphical interface is present on the Ouest-genopole bioinformatics platform server [9]. GenoFrag helps to design very good primers for PCR, thus avoiding checking primers and PCR conditions. This software is dedicated to biologists interested in bacterial genome variability analysis.

5.3. Databases and Data Analysis

Participants: Anthony Assi [correspondant], Olivier Filangi, Sophie Roucan, Hugues Leroy.

Genomic databases, including complete genomes such as the human genome, have been set up in an effort to help biologists in their research. Most of these databases are publicly available for consulting.

We automatically retrieve new releases when major updates for these databanks become available. Between two major releases, minor updates and corrections are also retrieved and installed in order to maintain up-to-date databases. Databases and tools are accessible on the web server [10] under Banks item.
National Project Biomaj (BIOlogie Mise A Jour) :

Biological knowledge, in proteomics and genomics context is mainly based on transitive bioinformatics analyzes 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.

A data integration process is obviously an essential preliminary step. This represents a major challenge and bottleneck in bioinformatics. These biological data banks contain a mass of heterogeneous data (all in different formats) and very bulky (Tera bytes). These banks, after their recovery, must undergo various post treatments more or less personalized upstream of their use via various bioinformatics software (blast, SRS, emboss, gcg, ...). The banks frequency update scale is variable, and may vary, according to the source, from daily to several times per year. With the growing number of complete genomes and others genomics data sources increase rapidly. Moreover, the nature and the number of the banks are in constant evolution; the data between sources

[9] http://genouest.org/genofrag.php
[10] http://genouest.org/

are cross-linked. The maintenance task is complex and heavy. A first stake consists in automating the process of updating the data banks for the administrator. Another significant stake to resolve is for the "quality" of service, providing to the users a clear vision of the integrity of data (state, exact origin, ...) constitutive of their workspaces.

Biomaj is a joint development between three bioinfomatics platforms : INRA Toulouse (David Allouche), INRA Jouy-en-Josas (Christophe Caron) and our platform genouest.org. Biomaj is written using state-of-the-art technologies (java, xml, ..) and is based on a parametrisable workflow engine. Post processes are written for the usual formats (gcg, blast, srs, ...) and are easily customisable at user's needs. Biomaj is currently under heavy tests and will be relased under an opensource licence in May 2007.

5.4. Pattern matching

Participants: Patrick Durand [correspondant], Anne-Sophie Valin, Mathieu Giraud, Jacques Nicolas, Gregory Ranchy, Catherine Belleannée.

Four pattern matching algorithms are available on the bioinformatics platform server. Two of them allow complex requests, STAN (Suffix Tree ANalyser) and WAPAM (Weighted Automata Pattern Matching). STAN is based on a suffix tree data structure. This tool scans complete genomes or sequence user. The patterns are represented in the form of a grammar. WAPAM is a tool to parse for protein patterns expressed by weighted automata. Proteic databanks (like Swiss-Prot or TrEMBL) or nucleic databanks (like genbank), or complete genome can be parsed. The upgraded web interface of WAPAM allows to execute pattern searches on Ouest Genopole servers or RDisk hardware. In both cases, the input patterns can be more complex than the usual regular patterns, such as PROSITE ones, since errors (substitutions and indels) and gaps of any size can be defined. In addition STAN provides string variables. The users are thus able to define precise, and possibly complex, signatures of biological functions. The implementation programming languages are OCaml, C, Prolog, Python, PHP and JavaScript. The platform is available for all french academic laboratories [11]

5.5. Pattern discovery

Participants: Laetitia Guillot [correspondant], Jacques Nicolas.

A Web platform grouping six pattern discovery algorithms is available for all french academic laboratories [12]. It allows a more reliable and faster pattern discovery process by comparing and by associating the results of all the available methods. To facilitate the interpretation and validation of results, we propose a a toolbox with various modules: pattern matching in public databanks, visualization, statistical analysis, filtering. This year, Smile was replaced by its upgraded version Risotto.

The implementation programming languages are Python, PHP and JavaScript.

5.6. Qualitative models on interaction graphs

Participants: Michel Le Borgne [correspondant], Philippe Veber.

Pyquali is a Python module dedicated to computations on qualitative models represented by interaction graph. Nodes of these graphs represents chemical species and arrows are labeled by $\{+,-\}$ representing influence of the variation of a specie on another specie. This variation occurs during an equilibrium shift. An efficient representation of qualitative equations have been developped. Various basic computations can be performed such that test of coherence of a model, test of coherence between a model and observations. Hard component which are qualitative predictions can be computed. A mesure of discordence between a qualitative model and observations is also prodided. The scalability of this software was demonstrated in a work on E.Coli regulation model. This model includes more than 1000 species and numerous edges.

[11] http://genouest.org/patternmatching.php

[12] http://genouest.org/patterndiscovery.php

6. New Results

6.1. Linguistic analysis of sequences

Two types of works are carried out within the framework of linguistic analysis of sequences. We first aim at helping a biologist that designs a model for his family of interest. Our purpose is to make the model operational. This will serve to both validate his/her model with respect to a set of sequences and to find new candidates in public sequence data banks.

The second type of work aims at helping a biologist wishing to build a model of his/her family of interest. Our purpose is then to infer a model from sequences.

6.1.1. Analysis by logical grammars

Participants: Jacques Nicolas [correspondant], Catherine Belleannée, Patrick Durand, Mathieu Giraud, Gregory Ranchy, Sébastien Tempel, Anne-Sophie Valin.

6.1.1.1. STAN and WAPAM

We have designed two parsers, STAN and WAPAM, able of treating Prosite expressions and elementary repetitions with substitution costs.

The parser STAN (Suffix Tree ANalyzer) is now freely available on the Genouest web of the bioinformatics platform. It allows matching patterns with string variables and errors on whole genomes [100]. It was used on the whole genome Arabidopsis thaliana (collaboration with UMR 6553) for a systematical analysis of a family of transposons [23], [36]. We propose a new definition of domain in DNA sequences, reflecting the presence of elementary modules repeated and composed to shape genomes. J. Nicolas coordinates a national ANR project, Modulome, including three biological labs on this subject.

WAPAM and pattern discovery softwares were used on the dog and rat genomes (collaboration with UMR 6061) [105], [74]. The olfactory receptors (OR) are genes devoted to the recognition of particular molecular substances. Biologists previously known 639 ORs located inside a 1.5x assembly [106]. In 2003, a 7x shotgun was conducted on the dog, but the first draft of the new assembly was only published in August 2004. In order to prevent the biologists for waiting the complete assembly, we developed a method which aims to directly analyze the sequenced runs. A pattern discovery step allowed to discover relevant patterns for OR and a very small subset of the runs was selected with the WAPAM tool, which keeps the sequences presenting the patterns expressed by weighted automata [72]. After assembly and cleaning, more than 400 new ORs were discovered and are further investigated by the biologists. This method allowed to spare the global assembly time while producing more sensitive results. Perspectives concern the conception of a tailored assembling algorithm.

6.1.1.2. Pattern matching with logical grammars

A more ambitious platform to search for motif within both DNA and protein sequences is under development. It is based on previous works made within the team in order to propose an expressive language to search for complex motif in biological sequences in the line of Searls' work. The language, called Logol allows to write a particular form of Definite Clause Grammars, namely String Variable Grammars. As for now, the system is capable of locating a Logol-based motif within a DNA (or protein) sequences database directly uses Prolog and can only be used by computer scientists.

The project's main goal is to provide the scientific community, both biologists and computer scientists involved in biological sequence analysis, with ModelDesigner, a graphical programming environment to search for Logol-based motifs. It is based on a client-server architecture which consists of two clients and one server modules. A first client module, ModelBuilder, allows a user to graphically create a motif without any particular knowledge of the underlying Logol grammar. Then, the user can run his/her motif against a database of sequences of his/her choice; the ModelDesigner platform also proposes a default set of sequences databases. The execution process, which may be computationally expensive, is delegated to the server module of ModelDesigner. Then, as soon as results are produced, the user can analyse them in the second client module, ModelAnalyser. Both ModelBuilder and ModelAnalyser runs on the user's computer, whereas the ModelDesigner server is installed on a separate, and more powerful, computer.

To achieve efficiency, we rely on a lexical analysis based on suffix trees. The entire platform is written using Java-based technologies. ModelDesigner server module uses a proprietary Sicstus Prolog server.

6.1.2. Genome Visualization

Participants: Patrick Durand [correspondant], Mathieu Giraud, Dominique Lavenier, Goulven Kerbellec, Hugues Leroy, Jacques Nicolas, Gregory Ranchy, Anne Siegel, Sebastien Tempel, Anne-Sophie Valin, Philippe Veber.

We have created a new genome sequence visualization method. Called pyramid diagram, or pygram, it aims at abstracting the organization of the repeated structures in genomic sequences. The pygram is created with the idea of visualizing all exact maximal repeats (MR) located either within single or multiple sequences without producing any link between pairs of MR. By choosing to highlight all eR in that way, a pygram not only display all the possible repeats of sub-sequences, it also reveals their hierarchical organization throughout the genome sequence.

We have implemented a prototype viewer forming an MR visualization tool associated to an MR querying tool.

First applications on Virus and Archaea genomes have prove that Pygram is a novel promising visualization technique. It is well suited to display the complex organization of repeated sequences within a single genome sequence or between sequences. The prototype we have developed achieve good linear performance, with respect to the sequence size as well as the number of MR to handle. In contrast with existing similar tools, pygram does not rely on the display of pairs of repeats. As a first immediate consequence, it produces a better view of repeated sequences at all level, from the entire genome sequence down to the nucleotide level.

6.1.3. Grammatical Inference

Participants: François Coste, Jacques Nicolas, Ingrid Jacquemin, Goulven Kerbellec, Nicolas Bitouzé, Yves-Pol Denielou, Thi Hong Hanh Hoang.

Our objective is to learn grammars for syntactically modeling a functional or structural family of bio-sequences. In practice, our experience showed that available sets of such sequences containing exceptions (sequences outside the family, or sequences in the family but for another account than the others) are very common. To tackle this problem when learning automata [27], we have introduced the definition of sub-sequence exception and proposed an algorithm for detecting them. This has been implemented in a new stand-alone generalization module with visualization features [40]. We have also proposed a new heuristic characterization algorithm, based on cliques of similar fragments, which is better suited for heterogeneous families. A first validation of the approach has been done successfully on the MIP (major intrinsic proteins) and the challenging TNF (tumor necrosis factors) protein families.

On the conceptual side, we have formalized the characterization problem as an exact problem of optimization under constraints with interesting links with classical alignments of sequences [39].

Ordered alphabets. We have also proposed the inference of automata using ordered alphabets based on a lattice that orders groups of amino acids according to their physico-chemical properties. An inference algorithm (SDTM) has been implemented in this framework [33]. The algorithm uses sequential machines in which the focus is on transitions (close to a Mealy machine) and computes best local alignments between pairs of proteins according to a score based on the order defined by the lattice and on the statistical properties of the given set of proteins. Experiments on artificial sequences and protein sequences (toxins) have shown the interest of the approach.

6.2. Gene expression data: analyzing data and modelling interactions

The purpose of this axis is to contribute to gene ad metabolite expression data analysis. The final goal is to build dynamical systems that model interactions implied in biological process.

Two kinds of analysis are investigated. First, analyzing gene expression data deals with a classification problem (how can one identify families of genes that are co-regulated?). Second, gene expression data provide information on the whole dynamics of gene networks which may be checked with respect to a model.

6.2.1. Classification

Participants: Israël-César Lerman, Jacques Nicolas, Basavanneppa Tallur, André Floeter, Yves Bastide.

This section includes various problems in unsupervised classification based on LLA (Likelihood Linkage Analysis, CHAVL program) as well as supervised classification relevant to the discrimination by decision trees.

Quality of association rules in Data Mining. One fundamental objective in Data Mining consists in defining rule-relevant measures. Relative to a rule (implication) A->B, such implication index (measure) evaluates in a certain way the propensity of B, knowing A. A non symmetrical nature is required for this index. The LLA approach provides fruitful probabilistic indices for measuring the rule interest. However, a local definition depending solely on the rule to be evaluated becomes non discriminant for large data bases. In these conditions we propose a discriminant extension of the probabilistic indices obtained with respect to a set of potential interest rules. This work has been performed in collaboration with J. Azé of the LRI laboratory (Univ. Paris Sud) [19].

Analysis of genotypes. We have proposed a method for the classification of SNP's data allowing some degree of interpretation of clusters [20]

Integrating CHAVL in the R environment. Around the objective of integrating the software CHAVL in the R environment collaboration has been established with the "École polytechnique de l'université de Nantes" (P. Peter). The implementation of the "Informational dissimilarity" of the LLA method has been validated . This general form of index enables pairwise comparison between complex objects described by a mixing of heterogeneous descriptive attributes. Therefore, the hierarchical classification functions provided in the R environment can be employed with this new family of indices.

6.2.2. Modelling genetic networks inside metabolic or signaling pathways

Participants: Michel Le Borgne, Ovidiu Radulescu, Anne Siegel, Philippe Veber, Carito Guziolowski, Nadia Betayeb, Clément Chatelain, Pierre Blavy.

Our biologist collaborators are concerned with biological systems that are regulated by genetic processes:

- the lipid metabolism in the liver of chicken is studied at the Animal Genetics Lab. (Inra, Rennes) in order to understand the genetic origin of fatting state;
- the signaling of TGF-beta in liver cancer (a molecule with a major influence on the expansion of the fibrosis) is studied in the U456 Lab. (Inserm Rennes);
- The Ewing inducible cellular model is a cell line characterized by a malignant genomic translocation and appearance of a chimeric gene EWS/FLI-1 whose activity leads to the uncontrolled cell growth.

In all systems, datasets provide information on the simultaneous states of hundreds of molecules and gene activity. Mastering such streams of data means finding a biological meaning to these data. Our approach implies building of models of biological networks, validation of models and observation explanation. We also develop new methods, suitable for qualitative models, to design new relevant experimentations. More fundamental studies were also made on modular aspects and compositionality of models. The goal is to master more and more complex models in terms of number of chemical species and interactions between species.

In building models, we have used existing database containing biological knowledge on interactions formalized in a form very close to our models.Our methods exploit the interaction graph associated with a differential model. This graph describes the qualitative constraints of the model such as *such as increases or decreases of the concentration of a product.*

The ability of qualitative models based on equilibrium shift to answer biological questions has been explored.The scalability of our methods has been demonstrated this real biological model. The resulting algorithms were implemented and tested on E.Coli model. They are available as a Python module.

Fundamental studies on modularity and compositionality of models are conducted on models based on differential equations.Qualitative models are derived from differential ones. Robustness of models is also an important property in founding a qualitative approach: we need, to some extend, properties which are independant of numerical values. Model simplification is also very important in the process of building and studying biological models. Model reduction is an interesting method to obtain more simple models.

Main contributions in this framework are the following:

- *Qualitative differential models and equilibrium shifts*. We introduce an approach to test the compatibility between differential data and knowledge on genetic and metabolic interactions. A behavioral model is represented by a labeled oriented interaction graph. The predictions of the behavioral model are compared with experimental data. We exploit a system of qualitative equations deduced from the interaction graph, which is linear in the sign algebra. We show how to partially solve the qualitative system. We also identified incompatibilities between the model and the data. Independently, we detect competitions in the biological process that is modeled. This approach can be used for the analysis of transcriptomic, metabolic or proteomic data [22].
- *Methods and algorithms for qualitative models* We found a new representation of qualitative equations based on finite field. This representation allows for representing complete sets of solutions of qualitative equations introduced in the previous work. With this new tool it is possible to address biological questions: is the model a coherent one, is it coherent with new observations? If negative answers are given by the algorithms, it is possible to isolate parts of the model which are incoherent with respect to observations or by themselves. It is also possible to compute the distance from observations and the admissible values of the model and to compute the correcting values. A measure of the efficiency of a set of observable variables, as a set test for validating a model, is proposed. Predictions of the model (with or without observations) can also be computed. Algorithms are now implemented and give basic tools for working with differential qualitative models from validation of models to the design of new experimentations.[24]
- *Application to the regulation of E. Coli* . We test the approach of [22] on the large scale network of transcriptional regulations of E. Coli (based on the database RegulonDB) and tested its coherence with our methods inspired by qualitative analysis. The graph contains about a thousand of nodes.
  - E. Coli transcriptional regulatory network was found incompatible. We were able to propose corrections: incompatibility problem was caused by a lack of interaction; we added interactions related to sigma factors.
  - This extended network was found to be incoherent with a dataset provided by regulonDB (stationanry phase experimental data). We proved that this was caused an error in RegulonDB (litterature results had not been correctly transfered in the database).
  - Given the coherent network of interaction and a compatible set of 40 experimental data, we were able to predit the variation of 401 new products of the system (26% of the system). We compared this prediction with micro-array data sets [31].
- *Response of interaction networks.* At many levels of organization, molecular biology interactions can be described as networks. These can be genetic, metabolic or mixed regulatory networks, or protein interaction networks. In absence of precise quantitative information on these networks or in the presence of overwhelming complexity we hope to find in topology hints for the understanding of functionality. Using concepts borrowed from electrical networks, we introduced a mathematical framework for such discussions. We investigated how the steady state of an interaction network responds to a change in the external conditions. The linear response solution has a graph theoretical interpretation as path series. The coefficients of the series are path that can be related to loop decomposition of the graph. This generalizes Mason-Coates graph approaches from linear electric

networks. We also show the usefulness of the concept of graph boundary. We apply our findings to specific biological examples, including lipid metabolism [21].

- *Modularity, hierarchical models, and robustness of complex biological systems* We have shown that these general, fundamental principles of biological systems, are in fact related and we have developed mathematical tools for their study. The main idea is that a study of robustness should use hierarchies of models. At least two levels of complexity should be present in this hierarchy: a level of low complexity, the abstract model and a level of high complexity, the extended model. The abstract model is the simplest model that still reproduces the behaviour of the biological system. It can be obtained by model reduction techniques. The extended model represents the level of variability, where different perturbations act. The mapping between the extended and the abstract model is a projection from high-dimension to low dimension that concentrates (in the Gromov/Talagrand sense) for robust properties. The model reduction method that we use is modular, being based on the selection of a set of variables to be eliminated. Depending on the type of property that we study, several choices are possible: rapid variables, Gale-Nikaido univoque response modules, monotone response modules. The robustness ideas have been applied to a study of NFkB response.[35], [30]

6.3. Parallelism and optimization

Participants: Rumen Andonov, Dominique Lavenier, Mathieu Giraud, Hugues Leroy, Stéphane Rubini, Pierre PeterLongo, Gilles Georges, Nicolas Yanev, Guillaume Collet, Mai Fei.

The parallelism axis mainly focuses on two activities:

- the design of specialized parallel machines for scanning genomic banks in relation with axis 6.1;
- the modelling and parallelization of optimization problems.

6.3.1. Specialized architectures for scanning and processing genomic banks

Participants: Mathieu Giraud, Dominique Lavenier, Stéphane Rubini, Philippe Veber, Gilles Georges.

BLAST [41], [42] has steadily become the reference software for exploring genomic banks. Large databases can be quickly and easily screened to detect similarity with a query sequence. This type of algorithm, and many other algorithms such as PATTERNHUNTER [95] or CHAOS [54], proceed in two steps: first they seek for anchors, then they extend them into alignments. The load balancing between this two tasks depends on the quality of the anchors. Since the alignment extension can be time consuming, the goal is to limit the number of hits by providing anchors of good quality.

More generally, the problem of mining genomic banks is either bounded by the data access (the time for scanning all the bank) or the computation time (the time to detect good anchors). We address this problem following two complementary ways: (1) speeding-up the anchor detection using reconfigurable hardware; (2) speeding-up the data access using parallel disk architectures and indexing techniques. We have developing two hardware prototypes: the RDISK system and the ReMiX systems. Both are parallel and reconfigurable systems. RDisk is developed since 2001, and ReMiX since September 2003.

With the increasing amount of available complete genomes, the need for inter or intra genome comparison is now a reality. However processing such a volume of data is extremely time consumming, and supercomputer manufacturers now propose to include accelerator boards in their machines. Through the ANR PARA project, in cooperation with the BULL R&D team (Les Clayes sous bois), we are currently designing a reconfigurable accelerator tightly interconnected to their system.

6.3.1.1. RDISK project: filtering genomic banks with reconfigurable disks

The central idea of the RDISK project is to directly filter genomic data at the disk output, in order to provide the host computer with only relevant data. The challenge is to process data at the output rate of the disk and to forward only a low percentage of the database together with anchoring informations. Previous attempts are motivated by a major trend: hard disk controllers are designed with an increasing amount of general purpose processing power and on-chip memory and filtering the data by pushing computation closer to the storage system is becoming an attractive solution for providing reduction in data movement through the I/O system.

Instead of an embedded processor we propose to connect a reconfigurable system based on a low cost FPGA component to the hard disk. The main advantage is that the anchoring-search algorithm can be highly parallelized on simple hardware structures [77], allowing on-the-fly filtering of the genomic data.

Another point to consider is the time for accessing the genomic data. The quantity of data transmitted to the processor is expected to be low and it is likely to have no data to process. The complete system is thus made of a front-end computer connected to a bunch of hard disks (a 48-node system) coupled to reconfigurable processing and interconnected through an Ethernet local network. Depending of the type of query, an adequate hardware filter is first downloaded to the FPGA component before scanning the banks. The filtering occurs locally and results are send back to the front-end computer for further post-processing. As an example, when performing complex motif extraction, the RDISK system has shown performances equivalent to a 192 PC cluster [73], [79].

Since 2005, the prototype is fully operational. An effort has been made to make it available to the scientific community. More precisely, we have implemented a complex motif search service (WAPAM) based on weighted automaton. This service is now available through the Ouest-Genopole bioinformatics platform [105], [74].

6.3.1.2. *ReMiX project: Reconfigurable memory for indexing huge volume of data*

ReMiX project: Reconfigurable memory for indexing huge volume of data

Compared to the previous project, the ReMiX project goes one step further by addressing the data access problem. The idea, here, is not to duplicate disk accesses, but to propose a hardware mechanism allowing fast random accesses to Gbytes of data. In that way, indexing techniques accessing only a fraction of the bank become highly efficient.

In the ReMiX architecture, hard drives are replaced by FLASH memories whose access time are 2 or 3 orders of magnitude shorter. In the same way, data bandwidth is increased by accessing simultaneously a large number of FLASH memories. As in the RDISK project, data are processed on-the-fly by reconfigurable hardware directly connected to the memory.

Note that the reconfigurable index memory does not fit in the addressing space of the processor but it is indirectly accessed by specific queries. The reconfigurable index memory does not hold any cache hierarchy, and therefore memory accesses do not have to worry about the data locality. In 2005, we have assembled and tested the ReMIX prototype. It is composed of a small cluster of five PCs. One acts as a front-end machine and the four others are the processing nodes , each one housing two PCI boards of 64 Gbytes of memory. The whole system hold 512 Gbytes of FLASH memory.

In 2006, two genomic applications have been implemented, illustrating the potential of the ReMIX concept. The first one deals with the search into large DNA banks. In that case, the FLASH memory contains the full bank index, allowing fast retreival compared to traditionnal approaches (speed up ranging from 20 to 50 has been measured). The second application is related to intensive comparison of a large set of proteins against the human genome. The two data set are fully indexed and compared using the FLASH memory as a temporary storage support. Again, high performances have been exhibited: plugging only one ReMIX board on a standard PC decreases the computation time by about 50.

6.3.1.3. *High Performance Reconfigurable Supercomputing*

For the last two or three years, processor performance growth have been limited due to the difficulty of steadily increasing the clock frequency. One response to continue to provide computer power has been the launching of dual core chips, and probably, in a very next future, quad and octo core processors. Another alternative is to enhance the traditionnal processing units with reconfigurable resources able to customize specific treatments. FPGA component offer today consequent processing power, but one of the challenge is to have these resources fed from the main memory at a very high speed. The PCI express interface, and especially the second generation, can achieved this goal by providing an agregated bandwidth of 10 Gbytes/sec.

In cooperation with the BULL R&D team (Les Clayes sous bois) we are currently designing a reconfigurable accelerator focussing on I/O performances. A genomic application, well suited for FPGA implementation, will serve as a benchmark for validating the concept.

6.3.2. Combinatorial optimization approach for solving protein threading problem

Participants: Rumen Andonov, Dominique Lavenier, Hugues Leroy, Nicola Yanev.

Protein folding is one of the most extensively studied problems in computational biology. The problem can be simply stated as follows: given a protein sequence, which is a string over the 20-letter amino acid alphabet, determine the positions of each amino acid atom when the protein assumes its 3D folded shape. In case of remote homologues, one of the most promising approaches is protein threading, i.e., one tries to align a query protein sequence with a set of 3D structures to check whether the sequence might be compatible with one of the structures.

We can summarize our contributions in this theme as follows. PTP has been shown to be equivalent to finding the augmented minimal path in a graph with a particular topology associated to any 3D protein structure. Several mathematical formulations for this problem have been proposed in terms of mixed integer programming models (MIP) [43]. These models were solved by the package CPLEX of ILOG and very interesting properties have been observed. The most amazing observation is that for almost all instances (more than 95% and even for polytopes with more than 10^{46} vertices), the LP relaxation of the MIP models is integer-valued, thus providing optimal threading. Moreover, when the LP relaxation is not integer, its value is a relatively good approximation of the integer solution. Our approach was proven to be significantly faster than the popular in the literature B&B approach for solving PTP [92]. Moreover, the integer programming model has been proven faster than the MIP model used in the package RAPTOR [124] which was well ranked among all non-meta servers in CAFASP3 and in CASP6 (Critical Assessment of Structure Prediction).

A first direction of improvement was oriented on parallelizing the software FROST (Fold Recognition-Oriented Search Tool) which was developed few years ago by our partners from MIG, Jouy en Josas. FROST uses a database of about 1200 known 3D structures. Computing the associated distributions of scores used to take about 40 days on a 2.4 GHz computer. On a cluster of 12 PCs, computing the score distributions takes now about three days which represents a parallelization efficiency of about 1 [103].

A second direction of improvement focused on accelerating the resolution of the PTP underlying optimization solver. The advantage of MIP models is that their LP relaxations give the optimal solution for most of the real-life instances. Their drawback is their huge size (both number of variables and number of constraints) which makes even solving the LP relaxation slow. Instead of solving them by general-purpose B&B algorithms using LP relaxation, one can design more efficient special-purpose algorithms. based on the specific properties of the PTP problem: we have proposed two Lagrangian approaches, Lagrangian relaxation and cost splitting. These approaches are more powerful than the general integer programming and allow to solve huge instances[13], with solution space of size up to 10^{77}, within a few minutes. ork [25].

The above presented results confirm that integer programming approach is well suited to solve the protein threading problem. These results concern the global alignment of protein sequence and structure template. But the methods that have been developed can be adapted to other classes of matching problems arising in computational biology. Examples of such classes are semi-global alignment, where the structure is aligned to a part of the sequence (the case of multi-domain proteins), or local alignment, where a part of the structure is aligned to a part of the sequence. Problems of structure-structure comparison, for example contact map overlap, are also matching problems that can be treated with similar techniques. Solving these problems by Lagrangian approaches is work in progress.

6.3.3. Comparative genomics of bacteria using LR-PCR

Participants: Rumen Andonov, Dominique Lavenier, Philippe Veber, Nicola Yanev.

[13] Solution space size of 10^{40} corresponds to a MIP model with 4×10^4 constraints and 2×10^6 variables [126].

Comparative genomics aims to study genome variations between different species or different *versions* of the same organism. Here, we consider various strains of the pathogenic Gram positive bacteria *Staphylococcus aureus*.

A practical way to carry out genome plasticity analysis of bacteria – without a systematic sequencing – is to exploit the LR-PCR (Long Range Polymerase Chain Reaction) technique. The idea is to split the genomes of different strains into a large number of short segments, then to perform a LR-PCR on each segment. Depending on the reorganization, the deletion or the insertion of certain genomic zones, it is expected that a few segments will not be amplified by the LR-PCR. Thus a *profile* corresponding to the amplified segments will be assigned to each bacterium strain.

The goal is to cover the genome of a reference strain with overlapping segments of nearly identical size, constrained by starting and ending-primers. Primers are short synthetic oligonucleotides that have to respect certain constraints (no short palindromes, good balance between AT and CG nucleotides , *etc.* Practically, the bacterium genome is split into a few number of linear segments, called domains [49]. The problem of segmenting a complete bacterial genome is reduced to split each domain into segments of nearly identical size. Along a domain, there are specific positions (i.e. small 25 DNA character string) corresponding to all possible primer sites. The overlapping segments can only start and end at these positions. If we assume that a solution is made of a list of N segments, and that each segment can take only P different positions, then the number of possibilities equals P^N (N>100 in practice). We have explored various approaches for solving this problem by dedicated graph algorithms (see [44] for details), allowing a short computation time (1-2 minutes). Implementation of two algorithms have been performed and packaged into the GenoFrag software (see Section 5.2).

In 2006, the GenoFrag package has been enhanced with a new software (IThOS) for better selecting the primers. IThOS combines sequence indexing and thermodynamic evaluation to optimise oligonucleotide specificity. The whole genome sequence is first indexed using 6-base word. Then secondary hybridisation sites are searched for each primer candidate. For each position where a primer might hybridise, the thermodynamic stability of the duplex is calculated using the Nearest Neighbour Model and cut-off G^O values are then applied to select specific oligonucleotides. IThOS was found quicker and more sensitive than a BLASTn filter when tested for primers design dedicated to WGPS on a 0.5 Mb chromosomal portion of Staphylococcus aureus, a low G+C bacterium.

6.4. Other contributions: Iterated morphisms

Participant: Anne Siegel.

The present work is the continuation of part of A. Siegel research, started before she arrived in the Symbiose project and does not concern bioinformatics.

Iterated morphisms of the free monoid are very simple combinatorial objects which produce infinite sequences by replacing iteratively letters with words [104]. It naturally generates a minimal symbolic dynamical system that have many arithmetical, geometrical and dynamical properties. In some specific case (unimodular morphism of Pisot type), iterated morphisms can be understood in a geometrical framework, thanks to the construction of a Rauzy fractal, that is, a self-similar compact subset of the Euclidean space [58].

This year was devoted to the extension of representation of iterated morphisms in two different directions

- In [15], we introduce cancellation in iterated morphisms, leading to morphisms of the free group. We extend to automorphisms of free groups some results and constructions that classically hold for morphisms of the free monoid, i.e., so-called substitutions. A geometric representation of the attractive lamination of a class of automorphisms of the free group (irreducible with irreducible powers (*iwip*) automorphisms) is given in the case where the dilation coefficient of the automorphism is a unit Pisot number. We prove that in this case, the shift map associated with the attractive symbolic lamination is measure-theoretically isomorphic to a domain exchange on a self-similar Euclidean compact set. This set admits some specific symmetries, and is conjectured under the Pisot hypothesis to be a fundamental domain for a toral translation.

- In [26], we consider the infinite composition of a given number of morphisms, instead of the iteration of a unique given morphism. This corresponds to multidimensional continued fractions. In this framework, we consider the problem of generation of discrete planes using three-dimensional iterated morphisms. We give sufficient conditions to be sure to generate all of a discrete plane by a sequence of morphisms; these conditions, however, are not easy to check, even on simple examples.

7. Other Grants and Activities

7.1. Regional initiatives

7.1.1. OUEST-genopole

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

OUEST-genopole has a governing board. Michel Renard (Inra Le Rheu) is director and Claude Labit is president. Jacques Nicolas in charge of the bioinformatics field, participates in the monthly meetings of the OUEST-genopole committee.

7.2. National initiatives

The Symbiose project is involved in the following national collaboration programs:

- National Inra project Sigenae and Genanimal, detailed hereafter.
- National *contract Interface de la numération*, funded by the French ministry of research (Ministry Grant (ACI) Mathematical Interfaces program.
- National contracts GENOTO3D, ReMiX, GenoGRID, RDISK, Modulome, PARA, MathResoGen, VICANNE. These contracts are detailed heraafter.
- ARC Inria (Action recherche concertée) IBN and FLASH. These contracts are detailed hereafter.

7.2.1. Sigenae and Genanimal

Participants: François Moreews, Jacques Nicolas.

The SIGENAE program (Analysis of Breeding Animals' Genome) is an Inra national program with the ambition to develop generic steps and finalized research actions in the domain of animal genomics. It aims at identifying the expressed part of genome, developing the map-making of entire genomes and studying genetic diversity in animal populations in the midst of several species of breeding animals (pig, chicken, trout, cow). It associates public research organizations (Inra, Cirad) and professional structures. At the international level, a privileged partner is the American ARS (Agricultural Research Service) which develops a comparable project.

The transcriptome of three species (trout, chicken and pig), are studied in Rennes.

Symbiose collaborates to this program via an Inra engineer, F. Morrews, contributing to the Sigenae information system. The program is coordinated by Inra Toulouse. We are involved in this framework in a collaborative work with UMR Agrocampus-Inra 598: we participate to project MathResoGen (see next Section) and we have just started a contract in the genomic national program eQTL. QTL (Quantitative Trait Loci) are biomarkers of genomic regions responsible of a substantial part of variations deserved on a given character. The aim of the project eQTL is to relate QTL regions obtained by linkage analysis and regions obtained by transcriptomic studies, responsible of the regulation of a set of genes.

7.2.2. Project GenoTo3D

Participants: François Coste, Jacques Nicolas, Rumen Andonov, Nicola Yanev, Ingrid Jacquemin, Goulven Kerbellec, Marie Lahaye, Aurélien Leroux, Yoann Mescam.

The goal of GENOTO3D is to develop and integrate machine learning approaches for the protein tertiary structure prediction task. It is a great challenge both for the difficulty of the task and for its applications in many fields (biology, genetics, drug design, etc.). An increasing number of structures are available in the Protein Data Bank PDB [14] which may be used by programs to predict the structure of a query protein sequence. The GENOTO3D project proposes to use numerical and symbolic machine learning approaches to predict long-term dependencies - which are still badly exploited by the classical prediction methods - and a divide-and-conquer strategy to integrate the different prediction levels in a single model.

Yann Guermeur (Loria) is the coordinator of this 3 year project (October 2003 - October 2006) funded by the French ministry of research (Ministry Grant (ACI) Data Mass program). Involved teams are MODBIO (Loria, Nancy), Symbiose, Bioinformatique et RMN structurales (IBCP, UMR 5086, Lyon), BDA (LIF, Marseille), MAP (LIRMM, Montpellier), Mathématiques Informatique et Génome (Inra, Jouy-en-Josas).

A new ANR project (program "Calcul intensif et Simulation" will start at the end of GENOTO3D, called Proteus and working on protein folding prediction.

7.2.3. Project ReMiX: Reconfigurable Memory for Indexing Huge Amount of Data

Participants: Dominique Lavenier, Jacques Nicolas, Stéphane Rubini, Xianyang Jiang.

Indexing is a well-known technique that accelerates searches within large volumes of data such as the ones needed by applications related to genomics. Very large indexes (larger than the main memory capacities) need to be stored on the hard disk drives. In that case, the design of indexes is concerned with low level notions such as pages, fill-factors, tracks, cylinders, etc and indirectly impacts the search algorithms that navigate within the index.

The ReMiX project proposes the design of a dedicated and very large RAM index memory (several hundreds of Giga bytes, distributed among a cluster of PCs), big enough to entirely store huge indexes in main memory, avoiding the use of any disk. The use of an almost unlimited main memory raises completely new issues when designing indexes and allows to entirely revisit the principles that are at the root of almost all existing indexing strategies. Here, within this scheme, direct access to data, massive parallel processing, huge data redundancy, pre-computed structures, etc, can be advantageously promoted to speed-up the search.

In addition, the index memory uses reconfigurable hardware resources to tailor the memory management to best support the specific properties of each indexing scheme. It also offers the opportunity to implement – again, at the hardware level – algorithms having interesting potential parallelism for processing data directly from the output of the index memory. As an example, image indexing requires massive distance calculation between image descriptors: this kind of calculation can be directly performed by the reconfigurable index memory.

Experimentation on this platform will be carried out with three application domains where huge volume of data are manipulated: genomic bank search, content-based image retrieval, and text information retrieval in heterogeneous XML knowledge databases [107].

D. Lavenier is the coordinator of this 3 year project (October 2003 - October 2006) funded by the French ministry of research (Ministry Grant (ACI) Data Mass program). Symbiose is both involved in the design of the hardware platform and the indexation of genomic data.

7.2.4. Project ANR PARA: Parallelism and Improvement of Application Performances

Participants: Dominique Lavenier, Gilles Georges.

[14] http://www.rcsb.org/pdb/

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

7.2.5. Project ANR Modulome: Parallelism and Improvement of Application Performances

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

This ANR project, Modulome [15], aims at providing methods for the identification, visualization and formal modelling of the structure of genomes in terms of an assembly of nucleotides "modules" that are repeated along a genome or between several genomes. Combined together, these methods will provide an appropriate methodology for a fruitful production of hypotheses concerning genome organizations. The challenge is to allow the biologist to represent and reason on large genomic sequences in an abstract way, by segmenting them into modules and revealing their organization. It would thus be possible to get a more unified view of genomes and to discover new interesting structures (e.g. promoters or transposons) in genomes. Three other teams of Biologists and bioinformaticians are involved in this project: LDGE (Dynamique du Génome et Evolution), Institut Jacques Monod, Paris; LEPG (Etude des Parasites Génétiques), Tours; LM2E (Microbiologie des Environnements Extrêmes), Brest

7.2.6. Project MathResoGen: Mathematical models for networks dynamics

Participants: Michel Le Borgne, Ovidiu Radulescu, Anne Siegel, Philippe Veber.

The MathResoGen projects aims at developing mathematical methods to identify main actors in biological process regulated by a genetic network. Biologists, mathematicians and computer scientists are involved in this project: IRMAR (mathematics, Rennes), Symbiose project (computer science, Rennes), Comore project (computer science, Sophia-Antipolis), UMR ENSAR-INRA 598 (biology, Rennes), UMR CNRS 7000 (biology, CHU Pitié-Salpêtrière, Paris), Inserm U456 (biology, Rennes).

MathResoGen project study biological networks with mathematical qualitative dynamics tools, in order to understand the behavior and the properties of genetic regulations. Three biological applications will be studied in details: lipid metabolism in liver, signaling of TGF-β in liver cancer, induction of NFkB, a regulator of intro-cellular signaling and cell-cycle.

The project aims to answer to three specific questions related to biological networks regulated by genetic network:

- Existence of time scales, that will be study with singular perturbations.
- System complexity, with a hierarchical and modular approach.
- Stochasticity of biological process.

7.2.7. Project VicAnne: animation of community of biological networks

Participants: Michel Le Borgne, Ovidiu Radulescu, Anne Siegel, Philippe Veber.

The French ministry of research (Ministry Grant (ACI) IMPBio program) funded a project named Vicanne aiming to support French workshops related to dynamics of biological networks in 2005 and 2006. Jean-Pierre Mazat (Université de Bordeaux II) is the coordinator of this project. Symbiose team is in charge of the financial support. Supported workshops will be the epigenomic program (genopole Evry), three two-days working sessions on a specific theme in 2005, and a satellite workshop of the French conference of bioinformatics JOBIM.

7.2.8. ARC Inria: MOCA: MOdularité, Compositionalité et Abstraction dans les réseaux géniques et protéiques

Participants: Michel Le Borgne, Ovidiu Radulescu, Anne Siegel.

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

The development of formal languages aimed at the description of biological systems leads to new tools for reasoning about these systems. modelling, model inference, parameter estimation, model validation and checking of properties can be performed in a semi-automatic process (computer assisted).

However, the growing complexity of models needs more conceptual tools. Modular decomposition is a promising one. Module definition, module composition and module abstraction/refinement are central in this approach. The goal of this ARC (Action de Recherche Cooperative) is to explore new clues in this new domain.

7.2.9. Project Sitcon: modelling signal transduction induced by a chimeric oncogene (beginning December 2006)

Participants: Ovidiu Radulescu, Michel Le Borgne, Anne Siegel.

French National Agency funded project (ANR grant "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:

- Reconstruct the interaction network through which the chimeric protein EWS/FLI-1 affects cell cycle and apoptosis;
- Determine the signal transduction pathways involved in passing the signal from EWS/FLI-1 to the cell cycle and apoptotic genes;
- Create a detailed model of the functioning of the identified pathways in the context of the Ewing system and understand the details of the difference in their behaviour in comparison with the normal context;
- Use modelling to propose new experiments in order to validate and improve the constructed models and perform these experiments.

7.2.10. Project DyCoNum: diophantine, dynamical and combinatorial studies of several numeration systems

Participant: Anne Siegel.

The French National Agency (ANR grant "Jeunes chercheurs" program) funded a project named DyCoNum aiming to consider by a transversal approach digital expansions in several number systems. This project will focus on integer base expansions, non-standard systems with integer base (signed digit expansions), beta-expansions and substitutive numeration systems, (generalized) continued fractions. Even if the variety of the tackled questions needs to master many techniques and methods, a unified dynamical approach allows exhibiting some general aspects and common investigations. This program involves W. Steiner and C. Frougny (LIAFA, Paris 7) and B. Adamczewski (Institut Camille Jordan, Lyon 1).

7.2.11. ARC Inria: Seed Optimization and Indexing of Genomic Databases

Participants: Dominique Lavenier, Pierre PeterLongo, Gilles Georges, Hélène Darolles.

This project investigates the optimization of seeds in the context of large genomic database search using BLAST-like algorithms.

Typically, seeds are optimized to reach a better search sensitivity. Here, for a given sensitivity, we want to investigate how seeds can be optimized to reduce the size of the database index.

On the ReMIX prototype, the database index is stored on a large FLASH memory. With the rapidely increasing of genomic databases, it is extremely important to manage indexes as small as possible.

Three complementary actions are investigated:

1. Theoretical study based on previous works of G. Kucherov and L. Noe (LIFL) to design new seeds;

2. Test of these new seeds on a real challenging problem: comparing 700 000 proteins agains the Human genome (in cooperation with INSERM U694, Angers);

3. Improve the ReMIX programming for speeding up the implementation of new reconfigurable operators based on these new seeds (in cooperation with the LESTER lab., Lorient).

7.2.12. ARC Inria: Integrated Biological Networks

Participants: Jacques Nicolas, François Coste, Dominique Lavenier, Michel Le Borgne, Anne Siegel.

Recently, concern with getting a deeper understanding on how elementary biological objects interact in the general context of a genome, cell or organism has led to the development of whole new areas of investigation by computational biologists called integrative of system biology.

Lack of enough or sufficiently clean data and lack of good models has slowed down the development of revolutionary new ways of considering such relations. The project is divided into three main deeply inter-related topics of investigation: exploration and analysis of the complex regulation motifs that represent important elements in any study of biochemical and evolutionary networks, genome dynamics, and genetic and biochemical networks.

7.3. European initiatives

7.3.1. Integrated Project ACGT

Start of a new european IP, ACGT (Advanced Clinico-Genomics Trials on Cancer), which is in the final negociation phase [16]. The project aims at delivering the cancer research community an integrated CIT environment enabled by a powerful GRID infrastructure. It will start in 2006 and our contribution will concern parallelism (Grid development, tumor growth simulation), visualization (of genomic data) and data mining (integration of CHAVL in a R environment).

7.4. Regional cooperations

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

- Agrocampus-Inra Rennes - Laboratoire de Génétique Animale: Analysis of gene regulation involved in the lipid metabolism (M. Le Borgne, J. Nicolas, O. Radulescu, A. Siegel, P. Veber).
- Inserm U456 (Détoxication et réparation tissulaire). Study of gene regulations in TGF-beta signalling in liver cancer (M. Le Borgne, O. Radulescu, A. Siegel, P. Veber).
- Irmar, Rennes : Mathematical modelling of lipogenesis (A. Siegel, M. Le Borgne, P. Veber).
- Agrocampus-Rennes (G. Douaire, K. Bachar): Ascendant hierarchical classification applied to satellite image segmentation (I.-C. Lerman).
- École Polytechnique de l'Université de Nantes : integration of CHAVL in a R environment (I.-C. Lerman). Related to european project ACGT.
- UMR-CNRS 6026 Structure et Dynamique des Macromolécules, Rennes(C. Delamarche): Major Intrincsic Proteins (MIP) (F. Coste, G. Kerbellec, G. Ranchy)
- UPRES EA 3889 Microenvironnement et Cancer (MICA), Rennes: tumor necrosis factor (TNF) proteins (F. Coste, G. Kerbellec)
- Inra Rennes - Technologie Laitière - Microbiologie : Study of Staphylococcus aureus genome plasticity; GenoFrag (R. Andonov, D. Lavenier).
- Inserm U625 GERHM Rennes : Human defensins (J. Nicolas, G. Ranchy)
- VALORIA, UBS, Vannes: ReMIX project (D. Lavenier, G. Georges, S.Rubini).
- CHU Angers, Inserm U694, Angers, ARC FLASH (D. Lavenier, G. Georges, P. Peterlongo)
- LESTER, Lorient, ARC FLASH (D. Lavenier, G. Georges, H. Darolles)

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

- UMR-CNRS 6026 (Equipe Structure et Dynamique des Macromolécules) : Study of the structure of MIP proteins (F. Coste, G. Kerbellec), aquaporins (G. Ranchy).
- UMR 6197 Laboratoire de microbiologie des environnements extrêmes Brest: Study for genomic diversity of virus and hyperthermophil plasmids (J. Nicolas, P. Durand)
- UMR-CNRS 6553 - EcoBio : Arabidopsis thaliana transposons (J. Nicolas, S. Tempel), database on abyssal fungi (A. Assi).

7.5. National collaborations

The Symbiose project has worked with and/or welcomed in Rennes the following french collaborators:

- CEA, Saclay (N. Ventroux): Reconfigurable computing (D. Lavenier).
- ENST, Paris (L. Denoeud): behavior of association coefficients between partitions (I.-C. Lerman).
- LIRMM, Montpellier (V. Berthé): substitutive dynamical systems (A. Siegel).
- MIG, Inra, Jouy en Josas (J.-F. Gibrat, A. Marin): Protein threading, GenoGRID (R. Andonov, F. Coste, D. Lavenier).
- LIRMM, Montpellier (V. Berthé): substitutive dynamical systems (A. Siegel).
- LIFL, Lille, ARC FLASH (D. Lavenier, P. Peterlongo, M. Giraud)
- Institut de Biologie et de Chimie des Protéines (R. Andonov, F. Coste)
- LIH, Lab. Informatique du Havre (R. Andonov)
- LAMIH, Valenciennes (R. Andonov)

7.6. International cooperations

7.6.1. Bilateral cooperations

- Australia, Brad Starkie (University of Newcastle, Melbourne Victoria) and Menno van Zaanen (Macquarie University, Sydney, Australia) : Context-free language learning difficulty and evaluation (F. Coste).
- Chili, A. Maass and E. Pecou (University of Chili, Center of Mathematical modelling): mathematical modelling of bio-molecular networks (A. Siegel, O. Radulescu). This cooperation is reenforced by an intership Conycit/Inria program.
- Germany, Postdam university. After a co-tutored Ph-D thesis on Learning in metabolic pathway, we are starting a new cooperation on the use of ASP logical framework on biological networks modelling.
- Malta, Department of Computer Science & AI, University of Malta. Searching for smallest consistent deterministic automata (F. Coste).
- LBIT, Université de Montréal, Canada. N. El Mabrouk and J.-E. Duchesnes: Fast and sensitive RNA search (M. Giraud has spent 4 months in LBIT).

7.6.2. Advanced Research Program China/France SI04-04

This two-years program is entitled *Algorithms and Architectures for bioinformatics* and started in 2005. it is funded with 15000 Euros. We have proposed a renewal of the program in 2007. Based on the need in bioinformatics and the experience of the Symbiose team and the NCIC team of ICT (Institute of Computing Technology, Beijing) the cooperation aims to combine the research advantages of both labs to explore dedicated reconfigurable architecture in bioinformatics. The goal is not only to explore the knowledge of the data and characteristics of algorithm, but also rely on new architectures and algorithms. Hence, the collaboration aims:

- To invent new indexing algorithm and extend such indexing algorithm to other possible applications, for example information security.
- To develop parallel architecture for indexing algorithm. The system should be a high performance and low cost dedicated system.
- To study reconfigurable computing dedicated to bioinformatics and its new applications. This is a long term research activity involving competences in hardware design together with genomic applications.

In this framework, D. Lavenier and G. Georges have been visiting scientists, ICT, Chinese Academy of Science (one week , december 2006)

In the framework of an INRIA partnership, R. Andonov has spend one week in the City University (may 2006, Hong Kong, INRIA workshop).

7.6.3. Bulgaria: exchange research program RILA'2003 (PAI)

This program is managed by the French Ministry of Foreign Affairs [17]. The project focusses on the application of combinatorial optimisation techniques in two different domains, Protein Threading and automata inference for discovering signatures of a sequence. Both domains are rich in NP-hard problems and the goal of the project is to propose and to analyze new mathematical models allowing to accelerate the solution of these problems. This program involves R. Andonov, J. Nicolas, F. Coste and D. Lavenier.

7.7. Visiting scientists

The following scientists visited the Symbiose project.

- Prof X. Zhao, Dr, X. Jiang [ICT, Beijing, Chinese Academy of Science, 20-24/11]
- Prof Giuseppe Lancia [univ. Udine, Italy, 1 month] .
- Prof. N. Yanev, Sofia Univ., Bulgaria.

The Symbiose project supported the following scientific visits:

- Austria: Invited visiting scientist, University of Leoben (one week, A. Siegel).
- Chile: Invited visiting scientist, Center of Mathematical modelling, University of Chile, Santiago de Chile (one month, A. Siegel).
- China: Visiting scientist, Chinese Academy of Science, Beijing (one week, D. Lavenier).
- Vietnam: IFI summer school, Dalat, *bioinformatics course* (one week, D. Lavenier).

8. Dissemination

8.1. Leadership within scientific community

8.1.1. "Fête de la Science" event

Participants: Mathieu Giraud, Sophie Roucan, Antony Assi, François Moreews, Laetitia Guillot, Hélène Darolles, Dominique Lavenier.

Following the success of the 2005 "open days event", the Symbiose project was one the teams selected to represent nationwide INRIA at the yearly event "Fete de la Science" in october 2006. In the "Jardin du Luxembourg" of Paris, we presented three bioinformatics wooden puzzles - sequence assembly, motif discovery and protein classification. Moreover, the Flash game of those puzzles was published on the dissemination website Interstices (http://www.interstices.info/symbiologik)

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

8.1.2. Fourth meeting dealing with the Bioinformatics platform of OUEST-genopole

The third meeting dealing with the Bioinformatics platform of OUEST-genopole held at Irisa, Rennes, on october 24 2006 (http://www.irisa.fr/videos/genopole/bioInfo2006/index.html). Invited speakers included A. De Daruvar (Bordeaux, Bioinformatics centre), Catherine Letondal (Pasteur computing centre), Philippe Picouet and Xavier Bailly (Brest, Ecole Normale Supérieure Télécoms)

8.1.3. BioInfoOuest thematic-day conferences

The Symbiose project regularly organizes thematic-day conferences on bioinformatics subjects[18]. The public of this thematic-day is made of computer scientists as well as biologists. Usually, this public gathers 50 persons (with 50 % of biologists) coming from all western France. Three thematic-day conferences were organized during the year 2005-2006, about

- Inductive Logic Programming (Chris Bryant, Luc De Raedt, Stephen Muggleton, Sébastien Ferrié,
- Genome Plasticity (Boris Vitzaner, Meriem El Karoui, Nouri Ben Zakour)
- Transposables elements (Christian Biemont, Pierre Capy, Marie Angèle Grandbastien)

8.1.4. Symbiose Seminar

The Symbiose seminar is held on a weekly basis. 14 talks were given in this framework during the year 2006. Invited speakers can be local speakers as well as national speakers. The public is usually made of the members of the Symbiose project. However, biologists, computer scientist (Irisa) or mathematicians (Irmar) often attend the seminar, depending on the subject of the conference.

8.1.5. Conferences, meetings and tutorial organization

The members of Symbiose were involved in the organization of the following meetings:

- Journées montoises d'informatique théorique, Rennes, August 2006 (A. Siegel)
- CAp 2006, Conférence Apprentissage, Trégastel 2006 (F. Coste)
- EGC 2006: 6èmes Journées sur l'Extraction et Gestion des Connaissances, Lille, january 17-20 2006 (I.-C. Lerman, program committee).

8.1.6. Journal board

- Edition of a special number of the journal *TSI: Théorie des Systèmes Informatiques* named *Architecture des Machines* (D. Lavenier).
- Editorial Board of *La Revue de Modulad* (I.-C. Lerman, B. Tallur).
- Editorial Board of *Mathématiques et Sciences Humaines, Mathematics and Social Sciences* (I.-C. Lerman).
- Editor Special issue "IA et Bioinformatique", bulletin de l'AFIA (F. Coste and C. Froidevaux). To be published (December 2006).

8.1.7. Miscellaneous administrative functions

- Jury of the Habilitation-thesis of Vincent Poirriez, 6 decembre 2006, LAMIH, Institut des Sciences et Techniques de Valenciennes. "Contribution à la compréhension de problèmes d'optimisation combinatoire et études d'extensions de la méthode B". Jury of the Ph-D Thesis of Adrien Goëffon, 21 novembre 2006, "Nouvelles heuristiques de voisinage et mémétiques pour le problème maximum de parcimonie". [R. Andonov]

[18] http://www.irisa.fr/events/seminars/bioinfo/

- Jury of the PhD Thesis of Nicolas Ventroux, 19 Septembre 2006, "Contrôle en ligne des systèmes multiprocesseurs hétérogènes embarqués : élaboration et validation d une architecture"; Antoine Vernois, 11 septembre 2006, "Ordonnancement et réplication de données bioinformatiques dans un contexte de grille de calcul", Fleur Mougin, 1 décembre 2006, "Conception d'un modèle Web sémantique appliqué à la génomique fonctionnelle". [D. Lavenier].
- Jury of the Habilitation-thesis of R. Gras (Rennes). Jury of the Ph-D Thesis of J.-P. Forest (Orsay) and I. Jacquemin (Rennes, 12-2005) [J. Nicolas].
- Jury of the ph-D thesis of A. Leroux (Rennes, 06-2005) [F. Coste].
- Jury of the ph-D thesis of M. Manceny (Evry, 11-2006) [A. Siegel].
- Scientific commitee of the french ministry program ANR 2006 *Calcul Intensif et Simulation* and Architectures du futur (D. Lavenier).
- Scientific Advisory Board: BIOINFAPA, INRA (Bioinformatics for Animal genomics) and Ouest Genopole (J. Nicolas).

8.2. Faculty teaching

Members of the Symbiose project are actively involved in the bioinformatics teaching program proposed by the University of Rennes 1. Furthermore, R. Andonov and D. Lavenier respectively share the responsibility of 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 the usual teachings of the faculty members, the Symbiose project is involved in the following programs:

1. Master 1 & 2 BioInformatics. (R. Andonov, F. Coste, D. Lavenier, J. Nicolas, B. Tallur)
2. Master 2 Computer Science, IFSIC. (F. Coste, H. Leroy)
3. Master 2 Computer Science, ENST. (H. Leroy)
4. Master 2 Mathematics. (B. Tallur)
5. Master2 Molecular biology and Biochemistry (M. Le Borgne, D. Lavenier, R. Andonov)
6. Master 2 modelisation and intensive computing Lebanese University of Beirout (H. Leroy)
7. Bioinformatics, ESEO Engineering school, Angers (D. Lavenier)

8.3. Conference and workshop committees, invited conferences

8.3.1. Committees

- Journées montoises d'informatique théorique, Rennes, August 2006 (A. Siegel, program comittee, organization comittee)
- JOBIM 2006, Bordeaux, July 2006 (A. Siegel, steering comittee)
- Research-In-Team Generalized substitution and tilings and numeration, Marseille, March 2006 (A. Siegel)
- ICSIP: International Conference on Signal and Image Processing 2006(IEEE "Signal and Image Processing", www.icsip.org, B. Tallur, Technical committee)
- RIAMS 2006, Réseaux d'interaction : analyse, modélisation et simulation, Lyon, Novembre 2006 (A. Siegel, program committee)
- CAp 2006, Conférence Apprentissage, Trégastel 2006 (F. Coste, program committee)
- EGC 2006: 6èmes Journées sur l'Extraction et Gestion des Connaissances, Lille, january 17-20 2006 (I.-C. Lerman, program committee).

- ICGI, International Colloquim on Grammatical Inference, Tokyo 2006 (F. Coste, steering committee)
- ERSA 2006 : International Conference on Engineering of Reconfigurable Systems and Algorithms, june 16-20 2006, Las Vegas, Nevada, USA (D. Lavenier, program committee).
- FPL 2006: International Conference on Field Programmable Logic and Applications, Madrid, Spain, aug. 28-30, 2006 (D. Lavenier, program committee).
- OGSB 2006: Journée Ontologie, Grille et intégration Sémantique pour la Biologie, Bordeaux, jul. 2006 (D. Lavenier, program committee).
- SYMPA 2006 : Symposium en Architecture de Machines, Perpignan, oct. 3-6 2006 (D. Lavenier, steering committee).
- Workshop VicAnne on Aspects stochastiques de la modelisation des reseaux de regulation, Nice, january 2005 (A. Siegel).
- WRCE 2006 The 1st International Workshop on Reconfigurable Computing Education, mar. 1, 2006, Karlsruhe, Germany (D. Lavenier, prog. committee)
- HiPC High Performance Computing, dec. 2006, Bangalore (R. Andonov, prog. committee)
- SFC'2006: 13-èmes Rencontres de la Société Francophone de Classification, sep. 2006, Metz, France (I.-C. Lerman, program committee)

8.3.2. Meetings

We attended the following meetings:

- ICGI, International Colloquim on Grammatical Inference, Tokyo September 20-22 2006 (F. Coste).
- Grammatical inference: workshop on open problems and new directions, St Etienne, November 20-22 2006 (F.Coste, G. Kerbellec, G. Collet).
- CAp: Conférence francophone sur l'apprentissage automatique, Trégastel may 22-24 2006 (F. Coste, G. Kerbellec).
- JOBIM 2006, Bordeaux July 5-7 2006 (F. Coste, G. Kerbellec, P. Durand).
- International IEEE Conf. on Field Programmable Technology (FPT), Bangkok, Thailand, 2006 (D. Lavenier, G. Georges)
- Sympa 2006, Symposium en Architecture de Machines, Perpignan, France, 2006 (G. Georges, S. Rubini, D. Lavenier)
- 11th International Conf. on implementation and application of Automata, Taipei, Taiwan, 2006 (M. Giraud, P. Veber, D. Lavenier)
- 4th International Conf. on research, innovation & vision for the future, Ho Chi Minh Ville, Vietnam, 2006 (Van Hoa Nguyen, D. Lavenier)
- Roadef 2006, Lille, 2006 (G. Collet, N. Yanev, R. Andonov)
- JOBIM 2006, Bordeaux, 2006 (G. Collet, N. Yanev, R. Andonov)
- EGC'2006: Extraction et Gestion des Connaissances, 17-20 janvier 2006, Lille, France.
- Réunion nationale ALPHY/GTGC - Lyon (France) - jan. 23-24 2006 (P. Durand).
- ACI VicAnne et ARC MOCA, Lille, jun. 2006, Subject: Modularity in networks.
- Réunion Société Française de Biologie Théorique, Paris, feb. 2006,
- Summer school "Robustness and Noise in Gene Regulatory Networks", sept. 2005, Coquelles, France.

8.3.3. International invited conferences

- Fine-Grained Parallelism for Genomic Computation (invited speaker), SIAM Conf. on Parallel Processing for Scientific Computing, San Francisco, CA, 2006 [D. Lavenier]
- Beijing, ICT, Chinese Academy of Science, FPGA / FLASH Accelerator [D. Lavenier]
- Hong Kong, City University / INRIA workshop [R. Andonov]
- New qualitative approaches in molecular biology, Third Indo-French Bioinformatics Meeting (IFBM 2006),Bangalore, India [O. Radulescu].
- University of Neuchâtel, colloquium of mathematics jan. 2006 / Fractals, autosimilarités et combinatoire [A. Siegel]
- VicAnne Meeting, Paris, IHP, Abstraction, modularity, and compositionality of genetic and protein interaction networks [A. Siegel]
- Tutorial on Status and Trends in High Performance Computing and Networking, followed by hands on MPI programming, International meeting on Grid & Parallel Computing (http://www-lb.cams.aub.edu.lb/events/confs/parallel_04-01-06/index.html), January 4, 2006, American university of Beirut, Lebanon [H. Leroy].
- Status and Trends in High Performance Computing and Networking, atelier de calcul numérique intensif, ENIT-LAMSIN, Tunis, November 27, 2006 [H. Leroy].

8.3.4. National invited conferences

- Angers, LERIA, *Grilles et Applications Génomiques* [D. Lavenier].
- Troisième carrefour OUEST-genopole, Brest (01/2006) Modélisation in silico de la régulation génétique du métabolisme des lipides: Problèmes et méthodes [A. Siegel]
- Université Paris XI, groupe de travail Systèmes Dynamiques (05/2006) Fractals de Rauzy: combinatoires autosimilaires [A. Siegel]
- ENS Lyon, seminar of mathematics (04/2006) Exemple de représentation de la lamination attractive d'un automorphisme de groupe libre [A. Siegel]
- University of Montpellier, LIRMM (02/2006) Autosimilarité, Beta numeration, et approximation de plans discret [A. Siegel]

9. Bibliography

Major publications by the team in recent years

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

[2] N. Ben Zakour, M. Gautier, R. Andonov, D. Lavenier, P. Veber, A. Sorokin, Y. Leloir. *GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification*, in "Nucleic Acid Research", vol. 32, n^{o} 1, 2004.

[3] F. Coste, G. Kerbellec. *A Similar Fragments Merging Approach to Learn Automata on Proteins*, in "European Conference on Machine Learning (ECML-2005), Porto, Portugal", J. Gama, R. Camacho, P. Brazdil, A. Jorge, L. Torgo (editors). , LNAI, vol. 3720, Springer, 2005, p. 522–529.

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

[5] A. ELAMRANI, L. MARIE, A. AÏNOUCHE, J. NICOLAS, I. COUÉE. *Genome wide distribution and potential regulatory functions of AtATE, a novel miniature inverted-repeat transposable element that is present in the promoter region of one of the Arginine Decarboxylase genes in Arabidopsis thaliana*, in "Molecular Genetics and Genomics", vol. 267, 2001, p. 459-471.

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

[7] I.-C. LERMAN, F. ROUXEL. *Comparing classification tree structures: A special case of comparing q-ary relations I & II*, in "RAIRO Operations Research", vol. 33 & 34, 1999, p. 339-365 & 251-281.

[8] J. NICOLAS, P. DURAND, G. RANCHY, S. TEMPEL, A.-S. VALIN. *Suffix-Tree ANalyser (STAN): looking for nucleotidic and peptidic patterns in chromosomes.*, in "Bioinformatics", vol. 21, n^{o} 24, oct. 2005, p. 4408-10, http://dx.doi.org/10.1093/bioinformatics/bti710.

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

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

[11] A. SIEGEL, O. RADULESCU, M. L. BORGNE, P. VEBER, J. OUY, S. LAGUARRIGUE. *Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation network*, in "BioSystems", vol. 84, 2006, p. 153-174, http://dx.doi.org/10.1016/j.biosystems.2005.10.006.

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

Year Publications

Books and Monographs

[13] D. LAVENIER, M. DAUMAS. *Architectures des Ordinateurs*, vol. 25, n^{o} 6, Technique et Science Informatiques, 2006.

Articles in refereed journals and book chapters

[14] R. ANDONOV, S. BALEV, N. YANEV. *High Performance alignment methods for protein threading*, in Parallel Computing for Bioinformatics and Computational Biology, edited by Prof. Albert Zomaya, John Wiley & Sons Wiley-Interscience, 2006, p. 427-457.

[15] P. ARNOUX, V. BERTHÉ, A. HILLION, A. SIEGEL. *Fractal representation of the attractive lamination of an automorphism of the free group*, to appearin "Annales de l'Institut Fourier".

[16] P. DURAND, L. LABARRE, A. MEIL, J.-L. DIVOL, Y. VANDENBROUCK, A. VIARI, J. WOJCIK. *GenoLink: a graph-based querying and browsing system for investigating the function of genes and proteins*, in "BMC Bioinformatics", vol. 7, n^{o} 21, 2006, http://www.biomedcentral.com/1471-2105/7/21.

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

[18] Z. KOTE-JARAI, L. MATTHEWS, A. OSORIO, S. SHANLEY, I. GIDDINGS, F. MOREEWS, I. LOCKE, D. G. EVANS, D. ECCLES, R. D. WILLIAMS, M. GIROLAMI, C. CAMPBELL, R. EELES. *Accurate prediction of BRCA1 and BRCA2 heterozygous genotype using expression profiling after induced DNA damage.*, in "Clin Cancer Res", vol. 12, n^{o} 13, 2006, p. 3896-901.

[19] I.-C. LERMAN, J. AZÉ. *Quality Measures in Data Mining*, H. J. GUILLET (editor). , Studies in Computational Intelligence, to appear, vol. 43, chap. A New Probabilistic Measure of Interestingness for Association Rules, Based on the Likelihood of the Link, Springer, p. 207-236.

[20] I.-C. LERMAN. *Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. Application à des données génotypiques*, in "Revue de Statistique Appliquée", vol. 2, 2006, p. 33-63.

[21] O. RADULESCU, S. LAGUARRIGUE, A. SIEGEL, M. LE BORGNE, P. VEBER. *Topology and linear response of interaction networks in molecular biology*, in "Journal of The Royal Society Interface", vol. 3, n^{o} 6, 2006, p. 185 - 196.

[22] A. SIEGEL, O. RADULESCU, M. LE BORGNE, P. VEBER, J. OUY, S. LAGUARRIGUE. *Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation network*, in "BioSystems", vol. 84, 2006, p. 153-174, http://dx.doi.org/10.1016/j.biosystems.2005.10.006.

[23] S. TEMPEL, M. GIRAUD, D. LAVENIER, I.-C. LERMAN, A.-S. VALIN, I. COUEE, A. E. AMRANI, J. NICOLAS. *Domain organization within repeated DNA sequences: application to the study of a family of transposable elements.*, in "Bioinformatics", vol. 22, n^{o} 16, 2006, p. 1948 – 1954.

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

[25] N. YANEV, R. ANDONOV, P. VEBER, S. BALEV. *Lagrangian Approaches for a class of Matching Problems in Computational Biology*, in "Computers and Mathematics with Applications", to appear, also available as RR INRIA No 5973, Aug. 2006, 2006.

Publications in Conferences and Workshops

[26] P. ARNOUX, V. BERTHÉ, A. SIEGEL. *Finiteness properties for Pisot S-adic tilings*, in "Journées Montoises d'Informatique théorique, Rennes, France", JM'06, 2006.

[27] F. COSTE, G. KERBELLEC. *Learning Automata on Protein Sequences*, in "JOBIM'06, Bordeaux", 2006.

[28] G. GEORGES, S. DERRIEN, S. RUBINI, F. RAIMBAULT, L. AMSALEG. *ReMIX: une architecture pour la recherche dans les masses*, Sympa 2006, Symposium en Architecture de Machines, Perpignan, France, 2006, http://www.irisa.fr/symbiose/lavenier/Publications/Lav06cd.pdf.

[29] M. GIRAUD, P. VEBER. *Path-Equivalent Removals of Epsilon-Transitions in a Genomic Weighted Finite Automaton*, International Conference on implementation and application of Automata, Taipei, Taiwan, 2006.

[30] A. GORBAN, O. RADULESCU. *Concentration and spectral robustness of biological networks with hierarchical distribution of time scales*, in "European Conference on Complex Systems - ECCS'05, Paris, France", nov. 2005.

[31] C. GUZIOLOWSKI, P. VEBER, M. LE BORGNE, O. RADULESCU, A. SIEGEL. *Checking Consistency Between Expression Data and Large Scale Regulatory Networks: A Case Study*, in "Réseaux d'interaction : analyse, modélisation et simulation, Lyon, France", RIAMS'06, 2006.

[32] D. LAVENIER, X. XINCHUN, G. GEORGES. *Seed-based Genomic Sequence Comparison using a FPGA/FLASH Accelerator*, in "International IEEE Conference on Field Programmable Technology (FPT), Bangkok, Thailand", IEE, 2006, http://www.irisa.fr/symbiose/lavenier/Publications/Lav06ce.pdf.

[33] A. LEROUX, J. NICOLAS. *SDTM, une méthode d'inférence grammaticale pour la découverte de motifs dans des ensembles de protéines*, CAP 2006 Conf. francophone sur l'apprentissage automatique, 2006.

[34] V. H. NGUYEN, D. LAVENIER. *Recherche dans les banques d'ADN par indexation parallèle*, 4th International Conference on research, innovation & vision for the future, Ho Chi Minh Ville, Vietnam, 2006, http://www.irisa.fr/symbiose/lavenier/Publications/Lav06cb.pdf.

[35] O. RADULESCU, A. GORBAN, A. ZINOVYEV. *Hierarchies and modules in complex biological systems*, in "European Conference on Complex Systems - ECCS'06, Sabïd Business School, University of Oxford", 2006.

[36] S. TEMPEL, J. NICOLAS, I. COUEE, A. E. AMRANI. *The combinatorics of helitron termini in A. thaliana genome revealed strongly structured superfamilies*, in "1st international Conference Genomic Impact of Aukariotic Transposable Elements, Pacific Grove, CA, USA", The Asilomar Conference Center, 2006, http://www.girinst.org/conference/schedule.html.

[37] A.-S. VALIN, G. RANCHY, Y. MESCAM, P. DURAND, S. TEMPEL, J. NICOLAS. *Suffix Tree ANalyser (STAN): Un outil de recherche de motifs nucléiques et peptidiques dans les chromosomes*, in "Roadef2006", 2006, http://www.lifl.fr/ROADEF2006/programme.html.

Internal Reports

[38] N. BITOUZÉ. *A limited discrepancy approach of grammatical inference problems*, Internship report, IRISA, sep. 2006.

[39] Y.-P. DENIELOU. *Alignements multiples de séquences protéiques*, Internship report, IRISA, jun. 2006.

[40] T. H. H. HOANG. *Généralisation de familles de protéines*, Internship report, IRISA, aug. 2006.

References in notes

[41] S. ALTSCHUL, W. GISH, W. MILLER, E. MYERS, D.J. LIPMAN. *Basic local alignment search tool*, in "J. Mol. Biol.", vol. 215, 1990.

[42] S. ALTSCHUL, T. MADDEN, A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, D. LIPMAN. *Gapped Blast and PSI-Blast: a new generation of protein database search programs*, in "Nucleic Acids Research", vol. 27, n^{o} 17, 1997.

[43] R. ANDONOV, S. BALEV, N. YANEV. *Protein Threading Problem: From Mathematical Models to Parallel Implementations*, in "INFORMS Journal on Computing", Special Issue on Computational Molecular Biology/Bioinformatics, Eds. H. Greenberg, D. Gusfield, Y. Xu, W. Hart, M. Vingro, 2004.

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

[45] 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.

[46] B. BAKKER, P. MICHELS, F. OPPERDOES, H. WESTERHOOF. *Glycolysis in bloodstream from Trypanasoma brucei can be understood in terms of the kinetics of the glycotic enzymes*, in "J. Biol. Chem.", vol. 272, 1997, p. 3207-3215.

[47] 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.

[48] S. BAY, J. SHRAGER, A. POHORILLE, P. LANGLEY. *Revising regulatory networks: from expression data to linear causal models*, in "Journal of Biomedical Informatics", vol. 35, n^{o} 289-297, 2003.

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

[50] P. BOURNE, H. WEISSIG. *Structural Bioinformatics*, Wiley-Liss Inc., New Jersey, 2003.

[51] F. BOYER, A. VIARI. *Ab initio reconstruction of metabolic pathways*, in "Bioinformatics", vol. 19, n^{o} suppl. 2, 2003.

[52] 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.

[53] 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.

[54] M. BRUDNO, B. MORGENSTERN. *Fast and sensitive alignment of large genomic sequences*, in "Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB)", 2002.

[55] C. Bryant, S. Muggleton, S. Olivier, D. Kell, P. Reiser, R. King. *Combining inductive logic programming, active learning and robotics to discover the function of genes*, in "Electronic Transaction in Artificial Intellingence", vol. 5, 2001, p. 1-36.

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

[57] 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.

[58] V. Canterini, A. Siegel. *Geometric representation of substitutions of Pisot type*, in "Trans. Amer. Math. Soc.", vol. 353, n^{o} 12, 2001, p. 5121-5144.

[59] 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.

[60] C. Chaouiya, E. Remy, P. Ruet, D. Thieffry. *Qualitative Modelling of Genetic Networks: From Logical Regulatory Graphs to Standard Petri Nets*, in "Lecture Notes in Computer Science", vol. 3099, 2004, p. 137-156.

[61] 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.

[62] E. Chow, T. Hunkapiller, J. Peterson. *Biological Information Signal Processor*, in "ASAP", 1991, p. 144-160.

[63] 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.

[64] 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.

[65] 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.

[66] S. Dong, D. Searls. *Gene structure prediction by linguistic methods*, in "Genomics", vol. 23, 1994, p. 540-551.

[67] R. Eisenthal, A. Cornish-Bowden. *Propsects 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.

[68] D. Fell. *Understanding the Control of Metabolism*, Portland Press, London, 1997.

[69] 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.

[70] O. GASCUEL, B. BOUCHON-MEUNIER, G. CARAUX, P. GALLINARI, A. GUÉNOCHE, Y. GUERMEUR, Y. LECHEVALLIER, C. MARSALA, L. MICLET, J. NICOLAS, R. NOCK, M. RAMDANI, M. SEBAG, B. TALLUR, G. VENTURINI, P. VITTE. *Twelve numerical, symbolic and hybrid supervised classification methods*, in "Int. J. of Pattern Recognition and Artificial Intelligence", vol. 12, n^{o} 5, 1998, p. 517-572.

[71] 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.

[72] M. GIRAUD, D. LAVENIER. *Linear Encoding Scheme for Weighted Finite Automata*, in "CIAA 2004: Ninth International Conference on Implementation and Application of Automata, Queen's University, Kingston, Ontario, Canada", to be published in LNCS, 2004.

[73] M. GIRAUD, D. LAVENIER. *Workshop Weighted Finite Automata in Hardware for Approximate Pattern Marching*, EDAA PhD Forum at DATE (Poster), Paris, France, 2004.

[74] M. GIRAUD, P. QUIGNON, E. RETOUT, E. MORIN, A.-S. VALIN, D. LAVENIER, M. RIMBAULT, F. GALIBERT, J. NICOLAS. *Assemblage ciblé : recherche d'une famille de gènes sur un génome non assemblé*, in "JOBIM 2005, Lyon", 2005.

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

[76] 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.

[77] 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.

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

[79] S. GUYETANT. *Architecture parallèle reconfigurable pour le filtrage de banques de données non structurées ; application à la génomique.*, Ph. D. Thesis, IRISA, 2004, http://www.irisa.fr/bibli/publi/theses/2004/guyetant/guyetant.html.

[80] 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.

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

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

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

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

[85] 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.

[86] 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.

[87] P. KARP, M. RILEY, S. PALEY, A. PELLEGRI, M. KRUMMMENACKER. *Eco-Cyc: Encyclopedia of Escerichia Coli genes and metabolism*, in "Nucleic Acids Res.", vol. 24, 1996, p. 32-39.

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

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

[90] 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.

[91] P. LANGLEY, O. SHIRAN, J. SHRAGER, L. TODOROVSKI, A. POHORILLE. *Constructing explanatory process models from biological data and knowledge*, in "AI in Medicine", 2005.

[92] R. LATHROP. *The protein threading problem with sequence amino acid interaction preferences is NP-complete*, in "Protein Eng", vol. 7, 1994, p. 1059-1068.

[93] C. E. LAWRENCE, S. F. 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.

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

[95] B. MA, J. TROMP, M. LI. *PatternHunter: Faster And More Sensitive Homology Search*, in "Bioinformatics", vol. 18, n^{o} 3, 2002.

[96] H. MATSUNO, A. DOI, M. NAGASAKI, S. MIYANO. *Hybrid Petri net representation of gene regulatory network*, in "Pac Symp Biocomput.", vol. 5, 2000, p. 341-352.

[97] P. MENDES. *Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3*, in "Trends Biochem. Sci.", vol. 22, 1997, p. 36-363.

[98] I. NACHMAN, A. REGEV, N. FRIEDMAN. *Inferring quantitative models of regulatory networks from expression data*, in "Bioinformatics", vol. 20, 2004, p. i248 - i256.

[99] 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.

[100] J. NICOLAS, P. DURAND, G. RANCHY, S. TEMPEL, A.-S. VALIN. *Suffix-Tree ANalyser (STAN): looking for nucleotidic and peptidic patterns in chromosomes.*, in "Bioinformatics", vol. 21, n^{o} 24, oct. 2005, p. 4408-10, http://dx.doi.org/10.1093/bioinformatics/bti710.

[101] 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.

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

[103] V. POIRRIEZ, A. MARIN, R. ANDONOV, J.-F. GIBRAT. *FROST: Revisited and Distributed*, in "HiCOMB 2005, Fourth IEEE International Workshop on High Performance Computational Biology, Denver, USA", 2005.

[104] M. QUEFFÉLEC. *Substitution dynamical systems-spectral analysis*, Lecture Notes in Mathematics, 1294. Springer-Verlag, Berlin, 1987.

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

[106] P. QUIGNON, E. KIRKNESS, E. CADIEU, N. TOULEIMAT, R. GUYON, C. RENIER, C. HITTE, C. ANDRE, C. FRASER, F. GALIBERT. *Comparison of the canine and human olfactory receptor gene repertoires*, in "Genome Biology", vol. 4, 2003, R80.

[107] F. RAIMBAULT, D. LAVENIER. *Des machines reconfigurables orientées objet pour les applications spécifiques*, in "TSI", vol. 22, 2003, p. 759-782.

[108] C. RAO, D. WOLF, A. ARKIN. *Control exploitation and tolerance of intracellular noise*, in "Nature", vol. 420, 2002, p. 231-237.

[109] 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.

[110] 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.

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

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

[113] 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.

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

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

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

[117] C. SOULÉ. *Graphic Requirements for Multistationarity*, in "Complexus", vol. 1, n^{o} 123-133, 2003.

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

[119] 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.

[120] R. THOMAS. *Boolean formalization of genetic control circuits*, in "J. Theor. Biol.", vol. 42, 1973, p. 563-585.

[121] 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.

[122] 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.

[123] 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.

[124] J. XU, M. LI, G. LIN, D. KIM, Y. XU. *RAPTOR: optimal protein threading by linear programming.*, in "Journal of Bioinformatics and Computational Biology", vol. 1, n^{o} 1, 2003, p. 95–118.

[125] 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.

[126] N. YANEV, R. ANDONOV. *Parallel Divide and Conquer Approach for the Protein Threading Problem*, in "Concurrency and Computation: Practice and Experience", vol. 16, 2004, p. 961-974.

[127] 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.

[128] B. ZUPAN, I. BRATKO, J. DEMSAR, J. BECK, A. KUSPA, G. SHAUNLSKY. *Abductive inference of genetic networks*, in "Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine", Lecture Notes In Computer Science; Vol. 2101, 2001, p. 304 - 313.