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

2005

Table of contents

1. Team

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

Head of project

Jacques Nicolas [Research Scientist (Inria)]

Administrative assistant

Marie-Noëlle Georgault [Administrative Assistant (Inria)]

Inria staff members

François Coste [Research Scientist]
Ovidiu Radulescu [Research Scientist, since Oct. 2005 – *délégation*]

CNRS staff members

Dominique Lavenier [Research Director]
Anne Siegel [Research Scientist]
Nicola Yanev [Visiting Scientist, since Apr. 2005 – *poste d'accueil*]

Faculty members

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

Research scientists (partners)

Yves Bastide [Assistant professor, Ensar, Rennes (until Sept. 2005)]
Laurence Duval [Assistant professor, ENSAI, Bruz]
Daniel Fredouille [Post-doctoral Fellow, Aberdeen]
Stéphane Rubini [Assistant professor, univ. Bretagne Ouest]

Post-doctoral fellow

Xianyang Jiang [Post-doctoral fellow (Inria) until Sept. 2005]

Ph. D. students

Andre Floëter [Ph. D. Student (cotutored Potsdam univ.)]
Mathieu Giraud [Ph. D. Student (AMC)]
Stéphane Guyetant [Ph. D. Student BDI CNRS/Région]
Goulven Kerbellec [Ph. D. Student Inria/Région]
Ingrid Jacquemin [Ph. D. Student, Assistant professor]
Marie Lahaye [Ph. D. Student MENRT from 1st Oct. 05 until 21st Oct. 05]
Aurélien Leroux [Ph. D. Student Inria/Région until Mar. 05]
Sébastien Tempel [Ph. D. Student MENRT]
Philippe Veber [Ph. D. Student Inria]

Technical staff members (Ouest-Genopole bioinformatics computing center)

Hugues Leroy [Technical staff (Inria)]
Anthony Assi [Junior technical staff since Oct. 05]
Patrick Durand [Visiting Scientist (Inria contract)]
Esther Kaboré [Junior technical staff (Inria contract genopole) until June 05]
Gilles Georges [Junior technical staff]

Emmanuelle Morin [Junior technical staff (Inria contract genopole) until Nov. 05]
Gregory Ranchy [Junior technical staff (Inria contract genopole)]
Elodie Retout [Junior technical staff (at 20%, national program Inra/Sigenae), until Nov. 05]
Anne-Sophie Valin [Junior technical staff, (Inria contract genopole)]

Visiting scientist

Robin Gras [Visiting scientist, SIB, 3 months]

Graduate student interns

Rawad Assaf [3rd year Engineer, Libya / S. Tempel]
Rozenn Bouville [Master 2 bioinfo, Rennes / D. Lavenier]
Guillaume Collet [Master 2 Info, Rennes / R. Andonov]
Guillaume Couteau [1st year ENS Cachan / F. Coste]
Nicolas Deroche [4th year INSA / M. Giraud]
Nader Jalloul [4th year Esib, Beyrouth/ H. Leroy]
Ravi Kant [Master in Bioinformatics, Bengalore, India / P. Durand]
Marie Lahaye [Master 2 Info, Insa / F. Coste]
Nicolas Lê [Ingénieur, Nantes / I.C. Lerman]
Mathieu Morvan [Master 2 Info, Rennes / A. Siegel]
Van Hoa Nguyen [Master 2 Info, IFI Hanoi / D. Lavenier]
Tat Duong Nhat Quang [4th year INSA, Rennes / D. Lavenier]
Mickael Paty [Ingénieur ESAIP, Angers / D. Lavenier]
Alexandre Petriaux [Ingénieur, Nantes / I.C. Lerman]

Hommage

[*Marie Lahaye, qui commençait sa thèse dans l'équipe, est décédée dans un accident de la route en octobre. L'ensemble de l'équipe symbiose souhaite ici rendre hommage à sa vivacité d'esprit et à sa gentillesse. Marie avait su rapidement se faire apprécier pendant son de Master parmi nous et augurait d'un avenir plein de promesses. Nous nous associons ici à la peine de ses parents.*]

2. Overall Objectives

2.1. A project in Bioinformatics

Symbiose is a project of bio-informatics, devoted to modeling and analysing of genomic (DNA sequences) and post-genomic data (expression of RNA, proteins and metabolites). This project is deeply multidisciplinary, 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 or proteomes) and complex pattern filtering methods on sets of sequences. Two main tracks are studied: modeling with formal languages and development of dedicated machines. Other more transversal themes such as gene networks modeling and classification are also described in the document.

2.2. Scientific axes

Bioinformatics has a quite large meaning and we first delimit the restricted meaning 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. Note however that our experience – common to many bioinformaticians – is that it is hard to achieve in depth research in this domain without "biocomputing", that is, participating to services of the second kind with biologists.

The *Scientific axes* on which the project focuses derive from our choice on modeling complex biological systems in a linguistic and logical framework. 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

1. the design of grammatical models on biological sequences;
2. efficient model matching in data banks;
3. 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. Parallelism 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 modeling

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 limit ourselves to the study of the macromolecular level of life, that is all studies analyzing 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 such 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 [116]) :

- *Data collecting.* It seems that little research is needed at this level. The main unsolved 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, and this is an area for many research works. 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.
- *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.

Our work mainly addresses this last track. We are also interested in the analysis of similarities/differences between sequences, for the aspects of intensive computing, classification and protein threading.

3.1.1. Biological interest of pattern discovery

Due to its importance in the project, we give some details on the biological motivation of the pattern discovery issue in sequences. 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. For instance, patterns have been found in proteins determining an immune response (T-cells), or in promoter regions of DNA regulating the development of yeast. Of course, artifacts remain possible and a return to biological experimentation remains necessary to validate observed patterns. 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 [3], 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. One of the major interest lies in the characterization of families of proteins. Many laboratories are indeed studying a particular family of proteins, that is interesting because of its structure, function or its implication in a pathological mechanism. Working on some proteins, they can then amplify their discoveries by seeking in public banks all proteins matching the patterns found. Regarding DNA, located upstream genes, the discovery of patterns associated with areas might provide important information both on the probable localization of genes and their expression level. Another interest is to be able to carry out more reliable multiple alignments on the sequences (provided that the method of identification of patterns precisely does not rest on a multiple alignment method!). Finally, these patterns help in protein annotation, i.e. to get clues on the functional family, the activity or the localization of a new

[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

[3] http://dragon.bionet.nsc.ru/trrd

protein. This work is complex, because one has to take into account several sources of information and because proteins present most of the time several domains (frequently three or more) with a pattern combinatorics leading to the specific function. Note that manual annotation, that was until recently conducted by hand for high quality bases like SwissProt, is no more possible due to the size of the banks, and that obtaining an automatic annotation process of good quality is crucial for genomics.

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 [102], or H-systems, reproducing the phenomenon of crossing over, represent one of the most fertile formalism in this respect. Language theorists like A. Salomaa and Gh. Paun [125] 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 [108]. 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) [146]. The most complete work in this field seems due to D. Searls [134], [135] ;
- 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 [83]. D. Searls proposed a more systematic approach based on logical grammars and a parser, Genlang [88]. 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 [73] or [105]. 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 [104]. 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 [115], 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 [72], using principles very close to those found in the work of M.-F. Sagot and A. Viari [131]. Another track explores variations on the search for cliques in a graph [112], [76].

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 [92].

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 [132]. 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. Modeling 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, 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 modeling 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.

Modeling cellular interactions is an old domain of biology, initiated by biologists interested in the dynamics of enzymes systems [106]. Models for genetic networks appeared as soon as gene interactions were discovered. The simplest static model consists in modeling 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 microarray data. However, this technique appears to be incomplete without the support of literature information [140].

The dynamical framework includes simulations and prediction of behaviours; models can be either qualitative or quantitative, as reviewed in [85], [81], [113]. 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 [67], [89], [143], [120], [142]. The properties of continuous models can be studied with convex analysis, linear and non-linear control techniques [90], [103], [124], [66]. Stochastic models transform reaction rates into probabilities and concentrations into numbers of molecules, allowing to understand how noise influences a system [129], [107]. 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 [110], [133], logical [109], [130], or Petri networks [118], [80]. The bridge between continuous and discrete models is made by piecewise linear differential models [86], [93].

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 [137], [141], [138], [81]. 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 [91], [122] or kernels [145]; 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 [69], [147], [75], [130], continuous [68], [71], [113] or hybrid [77], [114]. Qualitative reasoning, hybrid system, constraint programming or model-checking allow either to identify a subset of active processes explaining experimental time-series data [69], [147], [75], [130] or to correct the models and infer some parameters from data [68], [79]. The identification methods are limited to a few dozen components. Model correction or parameter regression can cope with up to hundreds of products [79] 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 modeling, 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 modeling approach, using either discrete event networks or qualitative 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 onlines database*[4]). 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.

The first algorithms for comparing genomic sequences have been developed in the seventies. They were essentially based on dynamic programming technics [123], [136]. Then, with the increasing growth of data, faster algorithms have been designed to drastically speed-up the search. The Blast software [139] 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 [97] 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 [100], BISP [82], HSCAN [98] or BioScan [144], have been proposed, leading today to powerful commercial products: BioXL, DECYPHER and GeneMatcher coming respectively from Compugen ltd.[5], TimeLogic[6] and Paracel[7].

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 [70]. 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 modeling and computation time. The problem is investigated following many ways ranging from *de novo* folding prediction to protein threading technics [116]. 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] http://www.genomesonline.org/

[5] http://www.compugen.co.il/

[6] http://www.timelogic.com

[7] http://www.paracel.com

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.
- **Whole genome analysis** is made practical through dedicated data structures and reconfigurable architectures. We have thus 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).

5. Software

5.1. Bioinformatics computing center of Ouest-Genopole

The team transfers his results to the bioinformatics computing center of Ouest-Genopole. All our developments are progressively made available within a service platform. This platform has a strategic role in the genopole, offering access to various softwares and databases. It allows our team to filter from routine service requests new subjects of research with a good relevance in biology. 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.

5.2. Bioinformatics Toolbox

Participants: Emmanuelle Morin [correspondant], Esther Kaboré, Grégory Ranchy, Anne-Sophie Valin, Dominique Lavenier, Hugues Leroy, Jacques Nicolas.

The toolbox [8] 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.

5.2.1. *Specific primers*

CAPS Tags. CAPS[9] 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

[8] http://idefix.univ-rennes1.fr:8080/Serveur-GPO/rubrique.php3?id_rubrique=1

[9] http://idefix.univ-rennes1.fr:8080/Serveur-GPO/rubrique.php3?id_rubrique=1

appeared in the alignment. Differences are validated with the alignment, in this case a difference is a potential SNP.

Degenerate primers[10]. 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. Access: [11].

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 [2] [65]. A graphical interface is present on the Ouest-genopole bioinformatics platform server [12]. 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. Tools for Databases

Participants: Esther Kaboré [correspondant], Hugues Leroy, Emmanuelle Morin, Yves Bastide, Elodie Retout.

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. These public databanks are available for GCG programs, a package for sequence analysis installed on the platform. A Rsync server has been also set up and maintains partial mirrors of our banks in other sites (Angers, Roscoff, InnovaProteomics, Ifremer Brest) for Blast and motif search tools. Databases and tools are accessible on the web server [13] under Banks item. We are setting up an environment for building specialized databases. The main goal of this work is to enable a custom view on public data tailored of a specific laboratory. Then, we make available dedicated tools for this database. An example of realization for this work is the oysters database. This database contains about 7000 sequences which represent about 20 oysters' subspecies. We can blast any subset of this specialized database against public databanks like GenBank, or a set of sequences against the specialized database.

We have also set up the BioArray Software Environment (BASE)[14] which is a comprehensive database server to manage the massive amounts of data generated by microarray analysis. .

5.4. Pattern matching

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

[10] http://idefix.univ-rennes1.fr:8080/Serveur-GPO/rubrique.php3?id_rubrique=1
[11] http://idefix.univ-rennes1.fr:8080/Serveur-GPO/rubrique.php3?id_rubrique=1
[12] http://idefix.univ-rennes1.fr:8080/Serveur-GPO/rubrique.php3?id_rubrique=1
[13] http://genouest.org/
[14] http://idefix.univ-rennes1.fr:8080/www-base

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[15] 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 proteic patterns expressed by weighted automata. Proteic databanks (like Swiss-Prot or TrEMBL) or nucleic databanks (like genbank), or complete genome can be parsed. The 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 [16]

5.5. Pattern discovery

Participants: Anne-Sophie Valin [correspondant], Emmanuelle Morin, Jacques Nicolas.

A Web platform grouping six pattern discovery algorithms is available for all french academic laboratories [17]. 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.

The implementation programming languages are Python, PHP and JavaScript.

6. New Results

6.1. Linguistic analysis of sequences

Two types of works are carried out within the framework of linguistic analysis of sequences. The first type of work aims at helping a biologist that designs a model for his family of interest. Our purpose is to make the model operational. This will help the biologist 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, Catherine Belleannée, Patrick Durand, Mathieu Giraud, Emmanuelle Morin, Gregory Ranchy, Élodie Retout, Sébastien Tempel, Anne-Sophie Valin, Raoul Vorc'h.

6.1.1.1. Logical grammars

We study the modeling of sequences with logical grammars in the line of Searls' work, in order to propose an expressive language to search for complex motif biological sequences. We have specified a language, Logol, that allows to write a particular form of Definite Clause Grammars, namely String Variable Grammars.

Our objective is to make this logical grammar formalism accessible to the biologist, so that with minimum training he can design and test his own models [56]. We propose the design of graphical models which are translated in terms of logical grammars. We began the conception of an additional graphical interface to show parsing results inside the initial model, that is, to show graphically how the model matches each sequence [84]. These works also require to adapt expressiveness to biological specificities – to deal with helix structures for instance.

[15] http://idefix.univ-rennes1.fr:8080/Serveur-GPO/rubrique.php3?id_rubrique=1

[16] http://idefix.univ-rennes1.fr:8080/PatternMatching/

[17] http://idefix.univ-rennes1.fr:8080/PatternDiscovery/

The main difficulty is then to propose a compromise between expressiveness and complexity for developing efficient analyzers on complete chromosomes. To achieve this, we rely on a lexical analysis based on suffix trees.

6.1.1.2. STAN and WAPAM

We have designed two parsers, STAN and WAPAM, able of treating Prosite expressions and elementary repetitions with substitution costs. Details are given in Section 5.5.

The software STAN (Suffix Tree ANalyzer) was used on the whole genome Arabidopsis thaliana (collaboration with UMR 6553) [4] for a systematical analysis of a family of transposons [27]. We propose a new definition of domain in DNA sequences, reflecting the presence of elementary modules repeated and composed to shape genomes.

WAPAM and pattern discovery softwares were used on the dog and rat genomes (collaboration with UMR 6061) [28], [35]. 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 [127]. 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 [95]. 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.

6.1.1.3. Pattern matching

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. The language, called Logol allows to write a particular form of Definite Clause Grammars, namely String Variable Grammars. As for now, the system 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.

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 (eMR) located either within single or multiple sequences without producing any link between pairs of eMR. By choosing to highlight all eMR 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 eMR visualization tool associated to an eMR 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 liner performance, with respect to the sequence size as well as the number of eMR 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. We are currently under way to provide a Web service access to the pygram visualization tool. It is planned to be available soon on the Ouest Genopole bioinformatics platform.

6.1.3. Grammatical Inference

Participants: François Coste, Jacques Nicolas, Ingrid Jacquemin, Goulven Kerbellec, Aurélien Leroux, Marie Lahaye.

6.1.3.1. Fundamental results

Search space for noisy data. Handling noise in the labeling of training sequences is an important issue when learning from biological data. We have characterized the search space for state merging algorithms with respect to an upper bound of the labeling noise rate [47].

Context-free language competition. In order to compare different grammatical inference algorithms and to gain insight into the current state-of-the-art of context-free grammatical inference algorithms, we organized the Omphalos competition. At the end of the competition, we analyzed the lessons learned and proposed an improved measure of the difficulty of the task [31].

6.1.3.2. Characterization of proteins sequences

Similar fragment merging approach to learn automata on proteins. We have improved and refined our work on learning automata for the characterization of proteins by the similar fragment merging approach [33], [46]. Speed of the implementation has been improved in a new prototype named *Protomata-L*. A likelihood ratio test has been designed for the detection of physico-chemical properties based on the frequencies of the amino acids at each position. A new procedure for the classification of proteins according their distance to an automaton has been introduced. The distance is given by the minimal sum of the substitution costs (as given by a substitution matrix like Blosum62) needed for sequence acceptation. In contrast to classical statistical settings, this procedure allows to handle unpredictable variations of the protein family outside the model.

We have investigated a new field of application for the similar fragment merging approach: the definition of new cores for the protein threading problem. The problem is to characterize proteins sequences sharing the same (known) 3D topology. New formulations of the problem have been proposed introducing promising perspectives [50].

Ordered alphabets. A thesis on the inference of automata using ordered alphabets has been defended by Aurélien Leroux [15]. He proposes a Taylor based lattice, which orders groups of amino acids according to their physico-chemical properties. An inference algorithm (SDTM) has been implemented in this framework. 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.

Learning Interactions. Motivated by the issue of predicting cysteins bonds within proteins, we have considered the inductive logic programming approach to characterize the neighborhood of cysteins involved in a disulfide bond. We used Progol [121] on a sequence of windows extracted around the cysteins. The system generalizes these examples and produces an explicit, general rule, which can be used to identify future examples. For the prediction of the oxidation state of cysteins, we achieved a recognition rate as good as 90%

with only 12 learned rules [37], [36], [58]. Other results are presented in the Ph-D thesis of I. Jacquemin, which includes an unsuccessful attempt to use the grammatical inference approach in this context and some recent results on prediction of matching pairs [14].

6.1.3.3. Characterization of genomic sequences: motif discovery on promoter sequences

This domain has been very active since the availability of transcriptomic data. An in depth bibliography review of it has been produced [48]. The report describes algorithmic details of main *ab initio* methods of prediction involved in the control of gene expression. We have also developed a tool, Jannotatix, simplifying the investigation of such methods.

6.2. Gene expression data: analyzing data and modeling 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.

6.2.1.1. Unsupervised and supervised classifications

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.

Comparing partitions. One of the main features of the LLA classification methodology consists of comparison between combinatorial structures on a finite set, generally provided by an object set. The association coefficient can be interpreted in terms of a very general notion of correlation. Its construction is essentially probabilistic. Comparing partitions corresponds to the case where the category sets of the different variables are neither weighted nor structured by their relationships. Several distance indices between close partitions in terms of transfers have been studied [87]. For some respects, this research has led to collaboration with the École nationale supérieure des télécommunications (Paris) and with the Institut de Mathématiques de Luminy (Marseille).

Hierarchical classification of very large data under contiguity constraints. With G. Douaire (Agrocampus, Rennes), we have studied a new strategy of class construction in the contiguity graph of a segmented image. The latter graph becomes not connected by adopting threshold value on the dissimilarities between single elements. The recursive formula is employed in order to update the dissimilarities between classes. Different versions of this formula are taken into account by association with different criteria.

Integrating CHAVL in a new 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 [52]. 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.

6.2.1.2. Metabolic pathways

The study of metabolic concentration data is the subject of a collaboration with the university of Potsdam and of a co-tutored thesis (A. Floeter). Metabolic concentration data has been provided by the Max Planck Institute of Golm (Berlin), based on Mass Spectrometry and Gas Chromatography. We have proposed a method for stable states extraction in metabolite concentration data, based upon a global analysis of Decision Forests learned on every possible threshold and evaluated with a function combining comprehensibility and robustness. Hidden states have been discovered for some variables and simple conditional dependencies networks have been drawn from the study of attributes associated to learned decision trees. The thesis will be defended in January 2006 [12].

6.2.2. Modeling genetic networks inside metabolic or signaling pathways

Participants: Michel Le Borgne, Ovidiu Radulescu, Anne Siegel, Philippe Veber, Yves Bastide, Mathieu Morvan.

Our biologist collaborators are concerned with biological systems that are regulated by genetic process. First, 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. Second, 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). In both systems, datasets provide information on the simultaneous states of hundreds of molecules. If the system is first modeled thanks to a formal organization of the available knowledge, then a dynamical model shall be built. The comparison between the predictions of the model and the experimental datasets allows to validate the model or explain the data and to propose new relevant experimentations.

Biological applications we are interested in lead to build models that integrate simultaneously a biochemical (metabolic or signaling) component and a genetic component. Our goal is to understand better the relations between these two components. Our methods exploit the interaction graph associated with a differential model. This graph describes the qualitative constraints of the model such as *such product increases or decreases the concentration of such product*. Main contributions in this framework are the following:

- *Qualitative differential models and equilibrium shifts*. Qualitative models appear to be appropriate to study genetically regulated metabolisms such as lipid metabolism. 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 [30].
- *Discrete event networks* . We study the combinatorics of incoming signals in discrete event networks. We explore the specificities of the Signal language and the associated model-checking tool-based Sigali, which manipulates ILTS: Implicit Labeled Transition Systems (which can be seen as an equational representation of an automaton). Both Signal and Sigali are developed since many years at Irisa for the study of real time systems (in the range from electronics to avionics systems), and are based on the "synchronized data flow" model. We have introduced a coding of models in Sigali, providing an efficient representation of qualitative systems. We show through several applications that this representation is a relevant tool for the understanding and testing of large and complex biological networks [43].
- *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 [29].

- Interaction graphs are deduced from interaction information that are spread in the literature. The content of specialized biological literature related to regulation of metabolic and signaling process have been carefully studied. Though there exist many experiments to gather such knowledge, we have not found in existing databases the information that we need to build interaction graphs. Hence, we have created a database GARDON from specialized literature informations related to this classification of interactions (behavioral versus mechanistic). Its purpose is not to be a comprehensive database of interactions but a repository of knowledge. 250 papers were read about the regulation of lipid metabolism in liver and 1900 genetic interactions were extracted from these papers; about 100 papers were read on the signaling of TGF-β, providing 350 interactions about this system [53].

6.3. Parallelism and optimization

Participants: Rumen Andonov, Dominique Lavenier, Mathieu Giraud, Xianyang Jiang, Hugues Leroy, Stéphane Rubini, Nicolas Yanev.

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 modeling and parallelization of optimization problems.

6.3.1. Specialized architectures for scanning genomic banks

Participants: Mathieu Giraud, Xianyang Jiang, Dominique Lavenier, Stéphane Rubini, Philippe Veber.

BLAST [63], [64] 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 [117] or CHAOS [74], 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 are currently 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. We now detail these 2 projects.

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. The idea of attaching computation capabilities near the disk for providing on-the-fly data filtering is not new. SmartDisk [119], Active Disk [62] or IDISK [111] are examples of such investigation. All of them are motivated by a major trend: hard disk controllers are designed with an increasing amount of general purpose processing power and on-chip memory. Thus, 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 [99], 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. To reach a good balancing between the post-processing and the filtering process, several disks are attached to the processor. The complete system is thus made of a front-end computer connected to a bunch of hard disks coupled to reconfigurable processing and interconnected through a local network – in our case an Ethernet 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.

In 2004, a 48-node system has been assembled and successfully tested. As an example, when performing complex motif extraction, the RDISK system has shown performances equivalent to a 192 PC cluster [94], [96], [101].

In 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 West Genopole bioinformatics platform. It allows biologists to query protein or DNA databases with a PROSITE-like motif. Compared with other approaches, the algorithm can deal with several errors, leading to a much sensitive investigation [21], [28], [35], [56].

6.3.1.2. 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 to access 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 four PCs (3Ghz, 2Gbytes of RAM), each one housing two PCI boards of 64 Giga bytes of FLASH memory. The first application is currently being implemented and tested on this machine. Preliminary results confirm the expected performances.

6.3.1.3. Epsilon-transitions removal for automata in FPGA

Weighted Finite Automata (WFA) are used for pattern matching in genomic databanks with substitution costs. They can be efficiently parsed using reconfigurable hardware as in the application WAPAM. Some biological applications require that one doubles every regular transition with an epsilon-transition to model deletions.

Critical paths in the FPGA prevent chains of epsilon-transitions from being arbitrarily large, so the epsilon-transitions must be removed, but generic removal algorithms produce too many new transitions. We proposed an analysis of the removal under a condition of path-equivalence. We obtained a constructive way to remove the epsilon-transitions on the linear-shaped parts of a WFA and an optimal bound of the number of new transitions produced. Other new results include bounds on algorithms producing automata with a limited number of consecutive epsilon-transitions [22], [13].

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 to the above problem 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 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) [1], [16]. 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 (more than 95%) of the instances, the LP relaxation of the MIP models is integer-valued, thus providing optimal threading. This is true even for polytopes with more than 10^{46} vertices. Moreover, when the LP relaxation is not integer, its value is a relatively good approximation of the integer solution. Our approach was prove significantly faster than the popular in the literature branch and bound approach for solving PTP.

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, each one associated with empirically determined score distributions. Computing these distributions used to take about 40 days on a 2.4 GHz computer. We succeed to redesign and structure FROST in modules and independent tasks [39]. On a cluster of 12 PCs, computing the score distributions takes now about three days which represents a parallelization efficiency of about 1.

A second direction of improvement focused on accelerating the resolution of the PTP underlying optimization problems. 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 branch-and-bound algorithms using LP relaxation, one can design more efficient special-purpose algorithms. Our results in this direction are extremely encouraging and suggest the Lagrangian relaxation is much faster than the general purpose methods[44].

All our results clearly illustrate that it is possible to solve real-life (biological) instances in a reasonable amount of time. These results also show that one of the most promising approaches in solving this problem is using advanced mathematical programming (MIP) models for PTP.

6.3.3. Comparative genomics of bacteria using LR-PCR

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

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

A practical way to carry out genome plasticity analysis of bacteria – without a systematic sequencing of all the available strains – 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 – or non amplified – segments will be assigned to each bacterium strain. The final step performs a global analysis of all profiles.

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: they must not include short palindromes (to avoid hairpin loops), they must contain a good balance between AT and CG nucleotides (for stability purpose), *etc.* Practically, the bacterium genome is split into a few number of linear segments, called domains [2]. Thus, 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. Given a domain, i.e. a DNA sequence ranging from a few 100 Kpb to a few Mbp, together with all potential primer positions, we need to cover it with a sequence of overlapping segments of nearly identical size. Two cases have been considered. In the first one we search for a sequence of overlapping segments, each one of size in the interval $[\underline{L}, \overline{L}]$ and as close as possible to a *given* ideal size L. In the second case L is considered as *unknown* and we look for L^*, $\underline{L} \leq L^* \leq \overline{L}$, such that the best segmentation with respect to it is of minimal error. In both cases, we solved the problem by dedicated graph algorithms (see [65] for details), allowing a short computation time (1-2 minutes).

This research is an active collaboration with Y. Leloir and N. Ben Zacour from the Inra Ensar UMR 1055 microbiology laboratory, Rennes (see [2], [54]). Implementation of the two algorithms have been performed and packaged into the GenoFrag software (see Section 5.2).

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 [126]. 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 [78].

In [18], we survey different constructions and properties of some multiple tilings of the space that can be associated with beta-numeration and substitutions. It is indeed possible, generalizing Rauzy's and Thurston's constructions, to associate in a natural way either with a Pisot number β (of degree d) or with a Pisot substitution σ (on d letters) some compact basic tiles that are the closure of their interior, that have non-zero measure and a fractal boundary; they are attractors of some graph-directed Iterated Function System. We know that some translates of these prototiles under a Delone set Γ (provided by β or σ) cover $\mathbb{R}^{d-1}$; it is conjectured that this multiple tiling is indeed a tiling (which might be either periodic or self-replicating according to the translation set Γ). This conjecture is known as the Pisot conjecture and can also be reformulated in spectral terms: the associated dynamical systems have pure discrete spectrum. We detail the known constructions for these tilings, their main properties, some applications, and focus on some equivalent formulations of the Pisot conjecture, in the theory of quasicrystals for instance. We state in particular for Pisot substitutions a finiteness property analogous to the well-known (F) property in beta-numeration, which is a sufficient condition to get a tiling [32]. The non-unit case is studied in [40].

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.

7.1.2. Bioinformatics Computing Center

Participants: Esther Kaboré, Hugues Leroy, Emmanuelle Morin, Grégory Ranchy, Anne-Sophie Valin, Jacques Nicolas.

Five technical platforms funded by a state and regional contract have been defined within the framework of *OUEST-genopole*.

The bioinformatics platform is linked to the Symbiose project, and propose a complete set of tools and databases for biologists and bioinformaticians. The web site is http://genouest.org/. This platform received the national RIO label in December 2003 and has been a platform of University of Rennes 1 since this year.

Olivier Collin, from Roscoff, and Hugues Leroy are in charge of the boarding committee of the platform. Training courses have been carried out (Wisconsin package, etc.) and three engineers, recruited on a fixed duration work contract, ensure the management of databases, softwares and the communication with the biologists of OUEST-genopole, enabling thus inter-disciplinary cooperations.

The platform is supported by a contract CNRG 2004 and a contract from Region Bretagne 2005.

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, MathResoGen, VICANNE. These contracts are detailed heraafter.
- ARC Inria (Action recherche concertée) Integrated Biological Networks. This contracts is detailed hereafter.

7.2.1. Sigenae and Genanimal

Participants: Elodie Retout, 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 (Apis-Gene, Cipa). 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, E. Retout, 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. Our contribution will start in 2006, focusing on pattern discovery in promoter regions [48].

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[18] 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).

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 – at a hardware level – 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 [128].

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). The Symbiose project is both involved in the design of the hardware platform and the indexation of genomic data.

7.2.4. Project MathResoGen: Mathematical models for networks dynamics

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

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).

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

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.

7.2.5. *Project VicAnne: animation of community of biological networks*

Participants: Michel Le Borgne, 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.

7.2.6. *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. *European Project DEISA*

In the Genomics Joint Research Activity (JRA) of the European DEISA grid project we propose to implement two genomic time-consuming calculations that show extremely good properties for an efficient gridification. The gridification is mainly thought in term of a distribution of independent tasks over the grid.

- Application 1: High-throughput identification of new human mitochondrial proteins by in silico comparative genomics. the objective is to identify unknown nuclear mitochondrial genes in order to better understand mitochondrial diseases.
- Application 2: Large scale microbial genome re-annotation. The objective is to re-annotate all known prokaryotic genomes (194) using the AGMIAL platform developed at the MIG INRA lab, and provide to the scientific community unified data mining bioinformatics tools to explore this huge amount of data.

This is a joined work with the LAMIH laboratory (Valenciennes), the INRA MIG team, (Jouy-en-Josas) and the INSERM E0018 team (CHU, Angers).

7.3.2. Integrated Project ACGT

We have worked this year on the preparation of an european IP, ACGT (Advanced Clinico-Genomics Trials on Cancer), which is in the final negociation phase. 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 (tumor growth simulation) and visualization (of genomic data).

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 (Y. Bastide, M. Le Borgne, J. Nicolas, A. Siegel, P. Veber).
- Agrocampus-Rennes (G. Douaire): Ascendant hierarchical classification applied to image segmentation (I.-C. Lerman).
- École Polytechnique de l'Université de Nantes : integration of CHAVL in a R environment (I.-C. Lerman).
- GURIH: Micro-environnent cellulaire moléculaire, Medecine faculty, Rennes: Characterization and modelization of the TNF (Tumor Necrosis Factor) ligands and receptors families (F. Coste, G. Kerbellec).
- Inra Rennes - Technologie Laitière - Microbiologie : Study of Staphylococcus aureus genome plasticity; GenoFrag (R. Andonov, D. Lavenier).
- Inserm U456 (Détoxication et réparation tissulaire). Study of gene regulations in TGF-beta signalling in liver cancer (M. Le Borgne, A. Siegel, P. Veber).
- Inserm U625 GERHM Rennes : Human defensins (J. Nicolas, G. Ranchy)
- Irmar, Rennes : Mathematical modeling of lipogenesis (A. Siegel, M. Le Borgne, P. Veber).
- VALORIA, UBS, Vannes: ReMIX project (R. Andonov).
- UMR-CNRS 6026 (Equipe Structure et Dynamique des Macromolécules) : Study of the structure of MIP proteins (F. Coste, G. Kerbellec), aquaporins (G. Ranchy).
- UMR-CNRS 6061 - Génétique et Développement : Olfactive receptors of dog and rat (M. Giraud, E. Morin, J. Nicolas, E. Retout, A.-S. Valin).
- 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).

7.5. National collaborations

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

- ADAGE, Loria, Nancy (L. Noé, G. Kucherov): Sequence indexation (M. Giraud)
- CEA, Saclay (N. Ventroux): Reconfigurable computing (D. Lavenier).
- ENST, Paris (L. Denoeud): behavior of association coefficients between partitions (I.-C. Lerman).
- ESSCA, Angers (K. Bachar) : Ascendant hierarchical classification applied to image segmentation (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).
- IML, Marseille (P. Arnoux, A. Guenoche): behavior of association coefficients between partitions (I.-C. Lerman); substitutive dynamical systems (A. Siegel).

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 Modeling): mathematical modeling of bio-molecular networks (A. Siegel, O. Radulescu). This cooperation is reenforced by an intership Conycit/Inria program.
- Germany, Postdam university. Learning in metabolic pathway. A co-tutored Ph-D thesis started in 2002 and will be defended in Jan. 2006.
- Malta, Department of Computer Science & AI, University of Malta. Searching for smallest consistent deterministic automata (F. Coste).
- Switzerland, University of Geneva (SIB). Motif discovery with metaheuristics (Y. Mescam, J. Nicolas, R. Andonov and F. Coste).

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.

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.

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

This program is managed by the French Ministry of Foreign Affairs [19]. 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. P. Zhang and Dr. X. Liu (Beijing, ICT, Chinese Academy of Science). Visiting from 29/08 to 3/09. Contract PRA SI04-04 : Algorithms and Architectures for bioinformatics 2005 - 2006 .
- Prof. N. Yanev, Sofia Univ., Bulgaria. One year CNRS position.

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

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 Modeling, 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. Open Days event

Participants: Mathieu Giraud, Grégory Ranchy, Anne-Sophie Valin, Nicolas Deroche, Dominique Lavenier.

As the IRISA was celebrating its 30th anniversary, a three-days event was organized in october 2005. Symbiose already participated in the yearly "Fête de la science" event, but it was the first time the team massively involves in public communication.

The main work was the design of three algorithmic puzzles on sequence assembly, pattern discovery and sequence classifications. Wooden puzzles were hand crafted, and an internship created a Flash game.

Additionally, we designed a few posters explaining genomics basics and we proposed a demonstration of the Rdisk machine with the key message *find your surname in the genome*.

A covering on this event with links to the puzzles, the posters, an the applet is available on the web [20].

8.1.2. Third 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 18th octobre 2005. Invited speakers includ D. Shemen (Bordeaux, Bioinformatics plateform), F. Plewniak (Strasbourg bioinformatics platform), L. Duquene (national coordination of bioinformatics platforms) and K. Wostencroft (orygrid project, univ. Manchester).

8.1.3. BioInfoOuest thematic-day conferences

The Symbiose project regularly organizes thematic-day conferences on bioinformatics subjects[21]. 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. Two thematic-day conferences were organized during the year 2004-2005, about 3D modeling (Marie Chabert, Alexandre G. de Brevern, Benoit Masquida) and polymorphism (Fabrice Fouchet, Mickael Guedj, Alain Vignal).

8.1.4. Symbiose Seminar

The Symbiose seminar is held on a weekly basis. 15 talks were given in this framework during the year 2004-2005. 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:

- ASI: Third International Meeting. Analyse Statistique Implicative, Palerme, october 2005 (I.-C. Lerman, program committee).

[20] http://www.irisa.fr/symbiose/symbiologik

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

- CAP: Conférence francophone sur l'apprentissage automatique, Nice, june 2005 (F. Coste, programm committe).
- EGC 2006: 6èmes Journées sur l'Extraction et Gestion des Connaissances, Lille, january 17-20 2006 (I.-C. Lerman, program committee).
- ERSA 2005 : International Conference on Engineering of Reconfigurable Systems and Algorithms, june 27-30 2005, Las Vegas, nevada, USA (D. Lavenier, program committee).
- FPL 2005: International Conference on Field Programmable Logic and Applications, Tampere, Finland, on August 24-26, 2005 (D. Lavenier, program committee).
- JOBIM'2005, Lyon. (J. Nicolas, Steering Committee).
- OGSB: Journée Ontologie, Grille et intégration Sémantique pour la Biologie, Lyon, july 2005 (D. Lavenier, program committee).
- SYMPA 2005 : Sympositium en Architecture de Machines, Nantes, april 2005 (D. Lavenier, President).
- Workshop on Computational Methods in Bioinformatics, 12th Portuguese Conference in Artificial Intelligence, december 5-82005 Covilhã, Portugal (I.-C. Lerman, program committee).
- Workshop VicAnne on Aspects stochastiques de la modelisation des reseaux de regulation, Nice, january 2005 (A. Siegel).
- Workshop Substitutions et automorphismes de groupes libres, Marseille, april 2005 (A. Siegel).

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).
- Edition of a CD under the patronage of the Classification Society of North America, including the book of I.C.Lerman "Classification et Analyse Ordinale des Données" (editor: F. Murtao, Univ. of London).

8.1.7. Miscellaneous administrative functions

- Jury of the Habilitation-thesis of J.P. Diguet, UBS, Lorient 09-2005. Jury of the Ph-D Thesis of M. Giraud (Rennes, 12-2005), E. Guerin (Rennes, 12-2005), L. Noé (Nancy, 09-2005), Y. Thoma (EPFL, Lausanne, 03-2005) [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. Rossignol (Rennes, 10-2005) [I.-C. Lerman].
- Referee of the Ph-D thesis of Sanjay Pande, Myosore university, India (B. Tallur).
- Jury of the SPECIF price for the best ph-D thesis (D. Lavenier).
- Organizing commitee of *Porte ouvertes de l'Irisa* (D. Lavenier).
- Scientific commitee of the french ministry program ANR *Calcul Intensif et Grille* (D. Lavenier).
- Scientific commitee of the french ministry program ACI *IMPBIO* (D. Lavenier).
- Scientific Advisory Board: MIA Department, INRA and Ouest Genopole (J. Nicolas).
- Steering Committee of ICGI (F. Coste).

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 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 Mathematics. (B. Tallur)
4. Master 2 libanese university, Beyrouth (H. Leroy)
5. Specialized trainings: IFI summer school, Dalat, Vietnam (D. Lavenier).
6. IRISA bioinformatics Intern school (R. Andovov, F. Coste, M. Giraud, M. le Borgne, J. Nicolas, A. Siegel, N. Yanev).

8.3. Conference and workshop committees, invited conferences

8.3.1. Meetings

We attended the following meetings:

- CAp: Conférence francophone sur l'apprentissage automatique, Nice, 06-2005 (F. Coste, G. Kerbellec).
- ECCS: European Conference on Complex Systems, Paris, 11-2005 (M. Le Borgne, O. Radulescu, P. Veber).
- ECML: European conference on machine learning, Porto, 10-2005 (F. Coste, G. Kerbellec).
- EJC 2005: École Jeunes Chercheurs algorithmique et calcul formel (M. Giraud).
- JOBIM 2005, Lyon (F. Coste, M. Giraud, D. Lavenier).
- MajecStic 2005 : Manifestation des Jeunes Chercheurs francophones dans les domaines des STIC, Rennes, 11-2005. (M. Giraud).
- Modélisation dynamique et analyse des réseaux de régulations biologiques, Marseille, 05-2005 (M. Le Borgne, A. Siegel, P. Veber).
- Numeration, Tilings, Substitutions Days, Grenoble, 03-2005, (A. Siegel).
- P&NT 05: International Conference on Probability and Number Theory 2005, Kanazawa, Japan, 06-2005) (A. Siegel).
- PReMI'05: International conference on Pattern Recognition and Machine Intelligence, Inde, 12-2005 (B. Tallur).
- SFC:12-èmes Rencontres de la Société Francophone de Classification, Montréal, 05-2005 (I.-C. Lerman, B. Tallur).
- SYMPA 2005 : Symposium en Architectures de Machines, Nantes, 04-2005 (D. Lavenier).
- Workshop Spectre des systèmes adiques, Marseille, 01-2005 (A. Siegel).
- Workshop Substitutions et automorphismes de groupes libres, Marseille, 04-2005 (A. Siegel).
- XIII Colloque Element transposables, Orsay, 07-2005 (S. Tempel).

8.3.2. International invited conferences

- Beijing, ICT, Chinese Academy of Science, *Dedicated Architectures for Bioinformatics Applications* [D. Lavenier].
- Vienna, Austria, Technical university of Vienna, *Arithmetic conditions for discrete space filling* [A. Siegel].

8.3.3. National invited conferences

- Angers, LERIA, *Grilles et Applications Génomiques* [D. Lavenier].
- Bordeaux, *Organisation Modulaire des séquences d'ADN répétées : Application à l'étude d'un hélitron non autonomes dans le génome d'Arabidopsis thaliana*|S. tempel].
- Le Croisic, Workshop NP-PAR' 05 : Résolution Parallèle des Problèmes NP-complets, *FROST (Fold Recognition-Oriented Search Tool) : Re-conçu et distribué* [R. Andonov].
- Lille, AS Indexation de texte et découverte de motifs, *STAN (Suffix Tree ANalyser): un outil de recherche de motifs dans les génomes* [A.-S. Valin, G. Ranchy].
- Lille, Journée RNG, *La plate-forme GenOuest : services et activités de recherche* [H. Leroy, A.-S. Valin].
- Lille, LIFL seminar, *Automates pondérés sur FPGA et epsilon-transitions* [M. Giraud].
- Lyon, Journée GrilBIO, ACI GénoGRID, *Une grille expérimentale pour la génomique* [D. Lavenier].
- Montpellier, École Jeunes Chercheurs algorithmique et calcul formel (EJC 2005), *Suppression d'epsilon-transitions dans les automates pondérés* [M. Giraud].
- Montpellier, LIRMM seminar, *Recherche de motifs par automates sur FPGA et assemblage ciblé*[M. Giraud].
- Marseille, Workshop Spectre des systèmes adiques, *Systèmes substitutifs à spectre discret: conditions explicites* [A. Siegel].
- Marseille, Workshop Substitutions et automorphismes de groupes libres, *Fractals de Rauzy* [A. Siegel].
- Nantes, carrefour Ouest-Génopole, *Identification bio-informatique d'une famille de gènes sur un génome non assemblé : application au répertoire des gènes olfactifs canins* [M. Giraud].
- Nice, workshop Apprentissage automatique et bioinformatique, Cap'05, *A Similar Fragments Merging Approach to Learn Automata on Proteins* [G. Kerbellec].
- Orsay, *XIII Colloque Element transposables*, *Organisation Modulaire des séquences d'ADN répétées : Application à l'étude d'un hélitron non autonomes dans le génome d'Arabidopsis thaliana* [S. Tempel].
- Rennes, IRMAR, Workshop Sur les exemples de Lattes, *Fractals de Rauzy* [A. Siegel].
- Rouen, LIFAR seminar, *Recherches de motifs par automates pondérés et découverte de gènes* [M. Giraud].
- OuestChips, Rennes, *Modélisation qualitative et données expérimentales* [M. Le Borgne].

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", Special Issue on Computational Molecular Biology/Bioinformatics, Eds. H. Greenberg, D. Gusfield, Y. Xu, W. Hart, M. Vingro, 2004.

[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] 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.

[5] R. GRAS, D. HERNANDEZ, P. HERNANDEZ, N. ZANGGE, Y. MESCAM, J. FREY, O. MARTIN, J. NICOLAS, R. D. APPEL. *Cooperative Metaheuristics for Exploring Proteomic Data*, in "Artif. Intell. Rev.", vol. 20, n^o 1-2, 2003, p. 95–120.

[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", to appear.

[9] 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.

[10] 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", to appear.

Books and Monographs

[11] D. LAVENIER, A. AUGUIN. *Architectures des Ordinateurs*, vol. 24, n^o 6, Technique et Science Informatiques, 2005.

Doctoral dissertations and Habilitation theses

[12] A. FLOTER. *Analysing biological expression data based on decision tree induction*, Ph. D. Thesis, Postdam University and Université de Rennes 1, 2005.

[13] M. GIRAUD. *Architectures reconfigurables pour la recherche par automates de motifs dans les séquences génomiques*, Ph. D. Thesis, Université de Rennes 1, 2005.

[14] I. JACQUEMIN. *Découverte de motifs relationels en bioinformatique: application à la prédiction de ponts disulfures*, Ph. D. Thesis, Université de Rennes 1, 2005.

[15] A. LEROUX. *Inférence grammaticale sur des alphabets ordonnés : application à la découverte de motifs dans des familles de protéines*, Ph. D. Thesis, Université de Rennes 1, 2005.

Articles in refereed journals and book chapters

[16] R. ANDONOV, S. BALEV, N. YANEV. *Parallel Computing for Bioinformatics and Computational Biology*, A. Zomaya (edt.), chap. Ch. 18: High Performance alignment methods for protein threading, Wiley & Sons, 2005, p. 429-460.

[17] R. ANDONOV, D. LAVENIER, P. VEBER, N. YANEV. *Dynamic programming for LR-PCR segmentation of bacterium genomes*, in "Concurrency and Computations: Practice and Experience", vol. 17, n^o 14, 2005, p. 1657-1668.

[18] V. BERTHÉ, A. SIEGEL. *Tilings associated with beta-numeration and substitutions*, in "INTEGERS (Electronic Journal of Combinatorial Number Theory)", vol. 5, n^o 3, 2005, A2.

[19] R. DAVID, D. LAVENIER, S. PILLEMENT. *Du micro-processeur au circuit FPGA*, in "TSI", vol. 24, n^o 4, 2005.

[20] 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", to appear.

[21] M. GIRAUD, S. GUYÉTANT, D. LAVENIER. *Encodage Linéaire d'automates pondérés. Filtrage de motifs génomiques et application sur l'architecture prototype R-disk*, in "TSI", vol. 24, n^o 6, 2005.

[22] M. GIRAUD, D. LAVENIER. *Dealing with hardware space limits when removing epsilon-transitions in a genomic weighted finite automaton*, in "Journal of Automata, Languages and Combinatorics (JALC)", 2005, in press.

[23] 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.

[24] D. LAVENIER, M. GIRAUD. *Reconfigurable Computing – Accelerating Computation with Field-Programmable Gate Arrays*, chap. Bioinformatics Applications, M. B. Gokhale and P-S. Graham (edts),

Springer, 2005.

[25] 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", to appear.

[26] A. MEIL, P. DURAND, J. WOJCIK. *PIMWalker: visualizing protein interaction networks using the HUPO PSI Molecular Interaction Format*, in "Applied Bioinformatics", vol. 4, n^o 2, 2005, p. 137–139.

[27] 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", to appear.

[28] 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.

[29] O. RADULESCU, S. LAGUARRIGUE, A. SIEGEL, M. LE BORGNE, P. VEBER. *Topology and linear response of interaction networks in molecular biology*, in "Royal Society Interface", to appear.

[30] 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", to appear.

[31] B. STARKIE, F. COSTE, M. VAN ZAANEN. *Progressing the State-of-the art in Grammatical Inference by Competition*, in "AI Communications", accepted for publication, to appear.

Publications in Conferences and Workshops

[32] V. BERTHÉ, A. SIEGEL. *Finiteness properties for substitution dynamical systems*, in "Numeration, Tilings, Substitutions Days, Grenoble, France", 2005.

[33] 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.

[34] M. GIRAUD, L. NOE, G. KUCHEROV, D. LAVENIER. *Recherches de motifs et de similarités en bioinformatique : modélisations, solutions logicielles et matérielles (tutoriel)*, in "Majestic 2005, Rennes", 2005.

[35] 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.

[36] I. JACQUEMIN, J. NICOLAS. *Disulfide bonds prediction using inductive logic programming*, in "Workshop on Constraint Based Methods for Bioinformatics, WCB, Sitges, Spain", 2005, p. 56-65.

[37] I. JACQUEMIN, J. NICOLAS. *Modélisation de cystéines oxydées à l'aide de la programmation logique inductive*, in "JOBIM 2005, Lyon, France", 2005, p. 331-340.

[38] I.-C. LERMAN. *Une forme unifiée pour les indices de discrimination de classes. Application en cas de données génotypiques*, in "Comptes rendus des 12-èmes Rencontres de la Société Francophone de Classification, Montréal, Canada", V. MAKARENKOV, G. CUCUMEL, F.-J. LAPOINTE (editors). , 2005, p. 186-190.

[39] 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.

[40] A. SIEGEL. *Beta-numeration and Rauzy fractals for non unit Pisot numbers*, in "International Conference on Probability and Number Theory 2005 (P&NT 05), Kanazawa, Japan", 2005.

[41] B. TALLUR. *Analyse des données incomplètes avec l'application aux expériences biopuces*, in "Comptes rendus des 12-èmes Rencontres de la Société Francophone de Classification, Montréal, Canada", V. MAKARENKOV, G. CUCUMEL, F.-J. LAPOINTE (editors). , 2005.

[42] B. TALLUR. *The linear factorial smoothing for the analysis of incomplete data*, in "PReMI'05", S. KPAL, ET AL. (editors). , LNCS, Springer-Verlag, 2005.

[43] P. VEBER, M. LE BORGNE, A. SIEGEL, O. RADULESCU. *Complex Qualitative Models in Biology: a new approach*, in "European Conference on Complex Systems - ECCS'05, Paris, France", 2005.

[44] P. VEBER, N. YANEV, R. ANDONOV, V. POIRRIEZ. *Optimal protein threading by cost-splitting*, in "5th Workshop on Algorithms in Bioinformatics - WABI, Mallorca, Spain", R. CASADIO, G. MYERS (editors). , Lecture Notes in Bioinformatics, 3692, 2005, p. 365–375.

[45] N. VENTROUX, D. LAVENIER. *A Low Complex Scheduling Algorithm for Multi-Processor System-on-Chip*, in "PDCN'2005 Parallel and Distributed Computing and Networks, Innsbruck, Austria", 2005.

Internal Reports

[46] F. COSTE, G. KERBELLEC. *A Similar Fragments Merging Approach to Learn Automata on Proteins*, Technical report, n^o 5672, INRIA, 2005, http://www.inria.fr/rrrt/rr-5672.html.

[47] G. COUTEAU. *Apprentissage d'automates en présence de bruit de classification*, Internship report, IRISA, 2005.

[48] M. HAEUSSLER, J. NICOLAS. *Motif discovery on promoter sequences*, Research Report, n^o 5714, INRIA, 2005, http://www.inria.fr/rrrt/rr-5714.html.

[49] N. JALLOUL. *Une étude de cas pour la mise en place de web services*, Internship report, IRISA, 2005.

[50] M. LAHAYE. *Apprentissage de signatures de repliements de protéines*, Internship report, IRISA, 2005.

[51] I.-C. LERMAN. *Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. Application à des données génotypiques*, Publication Interne Irisa, n^o 1652, Irisa, 2005, http://www.irisa.fr/bibli/publi/pi/2004/1652/1652.html.

[52] N. Lê, A. Pertriaux. *Intégration de la méthode AVL dans le logiciel R*, Internship report, École polytechnique de l'université de Nantes, 2005.

Miscellaneous

[53] Y. Bastide, S. Laguarrigue, M. Le Borgne, A. Siegel, P. Veber, O. Radulescu, A. Le Bechec. *Une méthodologie pour l'analyse qualitative des réseaux biologiques : De la base de données à la vérification formelle*, 2005, JOBIM 2005, poster session.

[54] N. Ben Zakour, D. Lavenier, M. Gautier, Y. Leloir. *Utilisation de l'indexation de séquences et du calcul thermodynamique pour optimiser la spécificité des oligonucléotide*, 2005, JOBIM 2005, Poster session.

[55] P. Durand, L. Labarre, A. Meil, J. Wojcik. *GenoLink: discovering drug target proteins by exploring networks of heterogeneous biological data*, 2005, ERCIM News (60).

[56] P. Durand, D. Lavenier, M. Le Borgne, A. Siegel, P. Veber, J. Nicolas. *Applying Complex Models on Genomic Data*, 2005, ERCIM News (60).

[57] O. Glorieux, M. Ferré, A. Fouilloux, I. Dupays, D. Raux, D. Lavenier, Y. Malthièry, P. Raynier, Y. Tourmen. *Optimisation of the NCBI-BLAST code for high throughput in silico comparative genomics in the DEISA project*, 2005, JOBIM 2005, Poster session.

[58] I. Jacquemin, J. Nicolas. *Disulfide bonds prediction using inductive logic programming*, 2005, ECCB'05, poster session.

[59] L. Labarre, D. Vallenet, F. Boyer, A. Morgat, A. Viari, P. Durand, C. Médigue. *Syntonizer : Un outil pour l'exploration des synténies bactériennes*, 2005, JOBIM 2005, Poster session.

[60] D. Lavenier. *Accélérer l'exploration et l'analyse des textes des génomes*, 2005, http://interstices.info/display.jsp?id=c_9754&qs=
Interstices.

[61] S. Tempel, M. Giraud, I.-C. Lerman, I. Couée, A. El Amrani, J. Nicolas. *Organisation Modulaire des séquences d'ADN répétées : Application à l'étude d'un hélitron non autonome dans le génome d'Arabidopsis thaliana*, 2005, JOBIM 2005, Poster session.

Bibliography in notes

[62] A. Acharya, M. Uysal, J. Saltz. *Active Disks: Programming Model, Algorithms and Evaluation*, in "ASPLOS-VIII, San Jose, California", 1998.

[63] S. Altschul, W. Gish, W. Miller, E. Myers, D.J. Lipman. *Basic local alignment search tool*, in "J. Mol. Biol.", vol. 215, 1990.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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° Suppl 1, 2005, p. i19-i28.

[69] 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° 289-297, 2003.

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

[71] F. BOYER, A. VIARI. *Ab initio reconstruction of metabolic pathways*, in "Bioinformatics", vol. 19, n° suppl. 2, 2003.

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

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

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

[75] 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.

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

[77] 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.

[78] V. CANTERINI, A. SIEGEL. *Geometric representation of substitutions of Pisot type*, in "Trans. Amer. Math. Soc.", vol. 353, n° 12, 2001, p. 5121-5144.

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

[80] 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.

[81] 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.

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

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

[84] L. COURBOT. *Filtrage de données protéiques à l'aide d'un modèle syntaxique. Réalisation d'une application fonctionnelle*, Stage de DESS CCI, Université de Rennes1, Irisa, 2003.

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

[86] 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.

[87] L. DENOEUD, H. GARRETA, A. GUÉNOCHE. *Comparison of distance indices between partitions*, in "Conference on Applied Stochastic Models and Data Analysis, Brest, France", 2005.

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

[89] 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.

[90] D. FELL. *Understanding the Control of Metabolism*, Portland Press, London, 1997.

[91] 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.

[92] 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° 5, 1998, p. 517-572.

[93] 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.

[94] M. GIRAUD, D. LAVENIER. *Dealing with Size Limits in a Hardware Encoding of Weighted Finite Automata*, in "Workshop WATA 2004: Weighted Automata: Theory and Applications, Dresden, Germany", 2004.

[95] 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.

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

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

[98] 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.

[99] 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.

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

[101] 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.

[102] 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.

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

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

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

[106] 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.

[107] 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.

[108] 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.

[109] 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.

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

[111] K. KEETON, D. A. PATTERSON, J. M. HELLERSTEIN. *A Case for Intelligent Disks (IDISKs)*, in "SIGMOD Record", vol. 27, n^o 3, 1998.

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

[113] 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.

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

[115] 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.

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

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

[118] 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.

[119] G. MEMIK, M. KANDEMIR, A. CHOUDHARY. *Design and Evaluation of Smart Disk Architecture for DSS Commercial Workloads*, in "Proceedings of International Conference on Parallel Processing (ICPP), Toronto, Canada", 2000.

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

[121] S. MUGGLETON. *Inverse Entailment and Progol*, in "New Generation Computing, Special issue on Inductive Logic Programming", vol. 13, n^o 3-4, 1995, p. 245-286.

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

[123] 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.

[124] 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.

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

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

[127] 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.

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

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

[130] 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.

[131] 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.

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

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

[134] 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.

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

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

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

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

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

[140] 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.

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

[142] 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.

[143] 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.

[144] 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.

[145] 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.

[146] 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.

[147] 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.