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

2007

Table of contents

1. Team

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

Head of project

Jacques Nicolas [CR1 Inria]

Administrative assistant

Marie-Noëlle Georgeault [AI Inria]

Research scientist

Rumen Andonov [Professor, Univ. Rennes 1, HdR]
Catherine Belleannée [Associate Professor, Univ. Rennes 1]
François Coste [CR1 Inria]
Dominique Lavenier [DR2 Cnrs, HdR]
Michel Le Borgne [Associate Professor, Univ. Rennes 1]
Israël-César Lerman [Emeritus Professor, Univ. Rennes 1, HdR]
Ovidiu Radulescu [Associate Professor, *delegation* until Oct. 2007, external collaborator afterwards, HdR]
Anne Siegel [CR1 Cnrs]
Basavanneppa Tallur [Associate Professor, Univ. Rennes 1, HdR]
Raoul Vorc'h [Associate Professor, Univ. Rennes 1]

External collaborators

Laurence Duval [Assistant professor, ENSAI, Bruz]
Stéphane Rubini [Assistant professor, Univ. Bretagne Ouest]

Post-doctoral fellow

Pierre Peterlongo [ARC FLASH]

Ph. D. students

Pierre Blavy [INRA]
Guillaume Collet [MENRT]
Matthias Gallé [student intern between May 2007 and Nov., Inria/CORDI since Nov. 2007]
Jérémy Gruel [Inserm/Région]
Carito Guziolowski [Conicyt/Ambassade de France/Inria]
Thibaut Henin [ENS/MENRT]
Goulven Kerbellec [Inria/Région]
Noël Malod-Dognin [Inria/Region]
Van Hoa Nguyen [Inria/CORDI]
Guillaume Rizk [MENRT since Sept. 2007]
Sébastien Tempel [ATER until Oct. 2007]
Philippe Veber [Inria until Nov. 2007]

Technical staff(GenOuest, Bioinformatics Computing Center of Ouest Genopole)

Hugues Leroy [IR Inria]
Olivier Collin [IR Cnrs since Sept. 2007]
Anthony Assi [IE (Inria contract genopole)]
Annabel Bourdé [IE (Inria contract genopole) since June 2007]
Olivier Filangi [IE (Inria contract genopole)]
Laetitia Guillot [IE (Inria contract genopole)]
Gregory Ranchy [IE (Inria contract genopole) until July 2007]
Sophie Roucan [IE (Inria contract genopole)]

Technical staff (Symbiose)

Alexandre Cornu [IE (ANR contract Para) since Oct. 2007]
Gilles Georges [IE (ANR contract PARA) until June 2007]
Julien Jacques [European project ACGT]
Christine Rousseau [IE (ANR contract Modulome)]

Technical staff(Inra, special accomodation contract with Inria, one day/week for collaborative work)

François Moreews [INRA, national program Sigenae]
Fabrice Legeai [INRA since July 2007]

Visiting scientist

Nicola Yanev [2 months, project RILA and ANR PROTEUS]

Student interns

Ricardo Ascensio [SUPELEC/D. Lavenier]
Laurent Chamoin [M2 Bioinformatique /A. Siegel]
Sarah Dagher [ESIB-USJ, Liban/ A. Assi]
Tony Dang [National University of Singapore, Singapore/A. Assi, O. Collin, H. Leroy]
Corentin Duperreray [IFSIC/A. Assi]
Jessie Mahé [INSA Rennes/F.Coste]
Debasis Panda [6 months, Internship INRIA, IBAB Bangalore/O. Radulescu]
Karim Tazibt [M1 Bioinformatique /F. Moreews]
Xianchyn Ye [6 months, Internship INRIA, Institute of Computing Technology, Chine/D. Lavenier]

2. Overall Objectives

2.1. A Bioinformatics center

Bioinformatics has a quite large acceptation and we first delimit the definition we use in our framework: it specifies research at the interface between computer science and molecular biology and not all "standard" informatics that is necessary to manage biological data on a daily basis. However, it is hard to achieve in depth research in this domain without "biocomputing", that is, participating to services of the second kind with biologists. In order to favour cooperative studies with biological labs we have decided to create a Bioinformatics Center, with a research team, **Symbiose**, leaning back against a bioinformatics platform **GenOuest** (or the converse...). This report is mainly focused on the research project. The activity of the bioinformatics platform being essentially an activity of development and transfer, it is described in the section 5.1.

Our research specificities include our interest in **large scale studies** (genomes, proteomes or regulation networks) and **discrete methods** necessary to handle the associated complexity. Our methods relate on discrete optimization, analysis of systems of qualitative equations and formal language modelling. Our goal is to push forward the range of applicability of these methods, by exploring the impact of **specialized machines or algorithms**.

2.2. Scientific axes

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

2.2.1. Modelling sequence/structure relationships

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

2.2.2. Systems biology: network modelling and analysis

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

2.2.3. Optimized algorithms on parallel specialized architectures

We investigate the practical 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 codesign of algorithms and hardware architectures tailored to the treatment of such applications. It is based on the study of reconfigurable machines employing Field Programmable Logical Arrays (FPGA) or fast components such as flash memories or Graphical Processing Units.

2.3. Highlights of the year

- Three PhD thesis are ending this year.

  S. Tempel has worked on transposons, mobile genetic elements that are the main source of variation in many genomes. The thesis proposes a general method to split a family of related DNA sequences into repeated domains. It led to a new nomenclature of helitrons, the most widely spread transposable element in A. thaliana, which has been published in Gene. S. Tempel has currently a post-doctorate position at Mountain views, CA, in the laboratory of J. Jurka, responsible of one of the main international bank on repeated genomic elements, Repbase.

  P. Veber has worked on the modelling of large scale regulation networks. He particularly addressed the issue of diagnosing incompatibilities between expression data and a signed graph of cellular regulations. A solution has been developed based on Answer Set Programming that allowed considering up to several thousands genes and regulations on *E.Coli* and Yeast data. This study demonstrates the great usefulness of this framework for providing a consistency check, proposing corrective actions and deriving new predictions. It is accepted for publication in *BMC Bioinformatics*. P. Veber has currently a post-doctorate position at Potsdam, Germany, in the laboratory of T. Schaub, a leading team in Answer Set Programming.

  G. Kerbellec has elaborated a grammatical inference method for modelling functional or structural families of proteins. Given a sample of sequences belonging to the family (learning from positive instances only), the method relies on a characterization step, which can be formalized as a search under constraints of the best partial local multiple alignment of the sequences (PLMA), and a generalization step, in which the PLMA is used to infer non-deterministic finite state automata according to the Minimum Description Length principle. It is the first operational grammatical inference method available for biologists to date.

- The staff of the bioinformatics platform GenOuest has been greatly reenforced by two new research engineers' positions. GenOuest is now labelled as a national RIO platform and as a platform of

GIS IBiSA. It has offered this year many new tools that result directly from transfers from research prototypes, a very good achievement with respect to its mission to provide innovative software in bioinformatics. Particularly, each of the previously described thesis has been extended by a software package accessible on the web server and connected to the main databanks. **Domain organizer** proposes a synthetic view of a set of related DNA sequences by providing both a segmentation of them into domains and a classification on the basis of these domains. **Bioquali** is a python plug-in dedicated to the validation and the prediction of qualitative data coupled with a biological network represented as an interaction graph. **ProtomataLearner** is a software package for the inference of finite state automata on protein sequences. It provides an interesting tool for the discovery of new proteins from sequences that deciphers both the active positions in proteins of the studied family and its possible subclasses.

3. Scientific Foundations

3.1. A quick introduction to bioinformatics

Studying life at macromolecular level (DNA, RNA, protein or metabolic molecules) supposes a number of studies of different kinds (for more information, see for instance the introductory part of [101]) :

- *Data collecting.* The main unsolved research issue is the reconstruction of a sequence from its fragments after sequencing and/or mass fingerprinting. Finishing an assembly remains a hard task. There exists a renewal of interest in this area due to the multiple sources of data and to the raise of metagenomics (considering several genomes simultaneously).
- *Data and Knowledge management.* It is actually a major issue. Information is produced in a highly distributed way, in each laboratory. Normalization of data, structuring of data banks, detection of redundancies and inconsistencies, integration of several sources of data and knowledge, extraction of knowledge from texts, all these are very crucial tasks for bioinformatics.
- *Analysis of similarities/differences.* Referring to a set of already known sequences is the most important method for studying new sequences, in the search for homologies. The basic issue is the alignment of a set of sequences, where one is looking for a global correspondence between positions of each sequence. A more complex issue consists in aligning structures. More macroscopic studies are also possible, involving more complex operations on genomes such as permutations. Once sequences have been compared, phylogenies, that is, trees tracing back the evolution of genes, may be built from a set of induced distances. A more recent track considers Single Nucleotide Polymorphism data, which correspond to mutations observed at given positions in a sequence with respect to a population. Analysing this type of data and relating them to phenotypic data leads to new research issues.
- *From structural analysis to systems biology.* This large domain aims at extracting biological knowledge from Xome studies, where X varies from genes to metabolites. Biological sequences and networks of components in the cell must verify a number of important constraints with respect to dynamic stable and accessible conformations, and functional mechanisms. These constraints result in the conservation during evolution of "patterns" and types of interactions to be deciphered. The main point becomes the study of life as a system, abstracted in a network of components governed by interaction laws.

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 [86], 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 [106] 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 [92]. The current agreement is that the necessary expressivity is the class of "mildly context sensitive" languages, well-known in natural language analysis [126]. The most complete work in this field seems due to D. Searls [114], [115] ;
- 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 [71]. D. Searls proposed a more systematic approach based on logical grammars and a parser, Genlang [76]. 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 [62] or [89]. 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. 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 [88]. Most powerful programs in this field are currently bayesian procedure building a consensus matrix by Gibbs sampling with organism-specific higher order models (Markov chain) for prior frequencies estimate [100], 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 [61], using principles very close to those found in the work of M.-F. Sagot and A. Viari [111]. Another track explores variations on the search for cliques in a graph [95], [64].

Even if results obtained so far are interesting in a number of cases, we think that there is a fundamental limitation to current studies: 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. 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.

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, and from a more practical point of view (applied Machine Learning), focusing more on the algorithms and their performances. 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 [80]. 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. When available, domain knowledge may help to design empirically a structure but, in practice, it is often very simple (e.g. 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).

A grammatical inference problem is an optimization problem involving the choice of a) a relevant alphabet and a class of languages; b) a class of representations for the languages and a definition of the hypothesis space; c) a search algorithm using the hypothesis space properties and available bias (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 [112]. In the Symbiose project, we are studying more specifically how grammatical inference algorithms may be applied to bioinformatics, focusing on how to introduce biological bias and on how to obtain explicit representations.

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

3.3. Modelling and analysing genetic networks

3.3.1. Biological context

Once genes and proteins have been identified, it becomes possible to deal 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 emergence of new high throughput techniques (DNA chips for gene activity, Chip on Chip for DNA/protein interactions, mass spectroscopy for protein interactions). allowing massive observations on the whole cell of tissue level.

3.3.2. Systems biology: models and data

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

Cellular interactions' modelling is an old domain in biology, initiated by people interested in the dynamics of enzymes systems [90]. Models for genetic networks appeared as soon as gene interactions were discovered. The simplest static model consists in an oriented labeled graph, with labels + for activation or - for inhibition. Such graph representations are used to store known interactions in general databases. They are also the framework of Bayesian representations, used to infer gene networks from micro-array data. However, this technique appears to be incomplete without the support of literature information [120].

The *dynamical framework* includes simulations and prediction of behaviours; models can be either qualitative or quantitative, as reviewed in [74], [69], [96]. 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 [77], [122], [121]. The properties of continuous models can be studied with convex analysis, linear and non-linear control techniques [78], [87], [105], [55]. Stochastic models transform reaction rates into probabilities and concentrations into numbers of molecules, allowing to understand how noise influences a system [91]. Finally, in discrete models each component is assumed to have a small number of qualitative states, and the regulatory interactions are described by discrete functions. Relevant discrete frameworks can be boolean [94], [113], logical [93], [110], or Petri networks [68]. Piecewise linear differential models [75], [81] try to build a bridge between continuous and discrete models.

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 [118], [69]. 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 in practice to a small number of variables.

Model identification addresses a different objective, that is, to build or update a model consistently with a set of data. When large amounts of data are available, models can be built using statistical techniques such as Bayesian networks [79] or kernels [124]. 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 [58], [110], continuous [57], [96] or hybrid [65]. Qualitative reasoning, hybrid system, constraint programming or model-checking allow either to identify a subset of active processes explaining experimental time-series data [58], [110] or to correct the models and infer some parameters from data [57], [67]. Identification methods are limited to a few dozen components. Model correction or parameter regression can cope with up to hundreds of products [67] 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 are absent in many situations (for instance disease studies on clinical tissues), meaning that they inform more on steady state shifts under perturbations than on the dynamics of the system.

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

- Our models integrate simultaneously a biochemical (metabolic or signalling) component and a genetic component. Indeed, in pluricellular organisms, genetic actors are activated in the framework of complex metabolic or signaling pathways that have their own dynamics. It is our goal to better understand the relations between these two components.
- We follow a qualitative modelling approach, using either discrete event networks or qualitative equations derived from differential models.

3.4. Optimal algorithms or parallel specialized architectures

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

Mixing parallel computing and genomics is both motivated by the large volume of data to handle and by the complexity of certain algorithms. Today, (october 2007) about 650 genomes – including the human genome – are completely sequenced, and there exist more than 3000 other sequencing projects (see *Genomes online database*[1]). 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 since the cost of sequencing is greatly decreasing and new needs appear such as close species sequencing for a better isolation of important conserved structures or metagenomics for ecological studies.

The problem is to efficiently explore these banks, and extract relevant informations. A routine activity is to perform content-based searches related to unknown DNA or protein sequences: the goal is to detect similar objects in the banks. The basic assumption is that two sequences sharing any similarities (identical characters) 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 [103], [117]. Then, with the increasing growth of data, faster algorithms have been designed to drastically speed-up the search. The Blast software [119] 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 [82] 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 [85], BISP [70], HSCAN [84] or BioScan [123], have been proposed, leading today to powerful commercial products: BioXL, DECYPHER and GeneMatcher coming respectively from Compugen ltd. TimeLogic and Paracel [2].

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

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

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

Note that the large amount of genomic data is not the only motivation for parallelizing computations. The complexity of certain algorithms is also another strong motivation, especially for the analysis of structures in sequences [BMW03]. For instance, predicting the 3D structure of a protein from its amino acid sequence is an extremely difficult challenge, both in term of modelling and computation time. The problem is investigated following many ways ranging from *de novo* folding prediction to protein threading technics [101]. The underlying algorithms are NP-complete and require both combinatorial optimization and parallelization approaches to calculate a solution in a reasonable amount of time.

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

4. Application Domains

4.1. Application Domains

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

The main stakes of bioinformatics are to assist biologists in the processes of discovering prognostic, diagnostic and therapeutic targets and the understanding of biological mechanisms. 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 for instance proposed Blast comparisons on the human genome in 1 minute, built a software for bacterial genome fragmentation, GenoFrag, that helps to study genomes variations via Long Range PCR, and studied the occurrence of retro-transposons, a family of mobile genomic units, in the genome of *Arabidopsis thaliana*.
- **Genomic/metabolic interaction networks** are modeled in eukaryote organisms. We consider large scale models described by their interaction graphs. As an application, we are studying genes and metabolites involved in the lipogenesis (chickens) and in TGF-beta-regulation in association with hepatocellular carcinomas (human). We test our analysis algorithms on the transcriptional regulation networks of *E. Coli* and *S. Cerevisiae*.

5. Software

5.1. Introduction

The software section of Symbiose needs a short introduction for a better understanding of the specific organization we have set up in order to favour the scientific relations between biologists and computer scientists. Most of the prototypes built during our researches are transferred on the bioinformatics platform GenOuest for further development and integration in a suitable environment for biologists. However, GenOuest has its own activity in relation with the service it has to offer and shares also studies with INRA bioinformatics platforms. This section contains three parts in accordance with this organization scheme :

- the first subsection introduces the context of GenOuest and general elements of its activity;
- the second subsection presents the main results of the platform for this year on its own projects and in collaboration with Inra;
- the last section presents the yearly activity of transfer of the platform, in conjunction with Symbiose.

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

Participants: Olivier Collin, Hugues Leroy, Anthony Assi, Annabel Bourdé, Laetitia Guillot, Olivier Filangi, Grégory Ranchy, Sophie Roucan, Sarah Dagher, Tony Dang, Corentin Duperray.

The bioinformatics platform GenOuest acts as a facility and software tool provider for the analysis of genomic data generated by numerous laboratories (55) of OUEST-genopole®. The platform provides at first computing power but also a comprehensive list of software dedicated to sequence analysis. On a national level, the platform is developing an expertise in the field of pattern matching and pattern discovery tool. This activity results from a tight interaction with the Symbiose research team. This platform first received a national RIO label in 2003. This label has been renewed in 2006. The platform is supported by several contracts : CNRG 2005 and 2006, Région Bretagne 2006 and 2007.

One of the main characteristics of the platform is a strong interaction with biological labs. This leads to a fruitful partnership where new tools can originate from direct demands to the platform. Fostering this partnership requires a strong implication of the GenOuest engineers who participate to numerous meetings with end users. The platform plays an important role in the technological transfer process of new tools elaborated by the three bioinformatics research teams of OUEST-genopole® (Symbiose; LINA - Nantes; LERIA - Angers). These tools are transferred on the platform and online access via the web server is provided for the whole scientific community.

Main evolutions of the year have been:

- The GenOuest platform has initiated the first steps of the implementation of a quality management system (QMS) since 10 months. The initial goal is to pass the ISO 9001 certification for the first semester of 2008.
- The GenOuest platform has proposed training sessions attended by a broad audience ranging from beginners to more accustomed users of bioinformatics tools. The subjects of the training sessions were sequence analysis, motif discovery, phylogenetics. More than 120 persons have attended the training sessions.

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

The platfom is involved in different coordination activities at a national level. H. Leroy is a member of the ReNaBi (Reseau National des plates-formes Bio-informatiques) steering board since 2005. The GenOuest platform is in charge of the BioMAJ project, a joint project with INRA Jouy and INRA Toulouse.

5.2.1. *Bioinformatics services hosting*

The GenOuest bioinformatics platform is hosting bioinformatics services developed by research teams who are not able to provide the required computing power. Among these services are

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

5.2.2. *Databases maintenance*

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. Between two major releases, minor updates and corrections are also downloaded and installed in order to maintain up-to-date databases. Databases and tools are accessible on the web [3].

5.2.3. *Test framework for web platform*

A bioinformatics Web Platform gives access to specific services for a community of bioinformaticians and biologists. We need to reduce repetitive consuming test tasks needed to guaranty the quality of the services provided by bioinformatics portal.

Based on existing testing frameworks, we have developed a prototype of an automated system and functional test engine with various functionalities (service availability, scheduling, evaluation, load test...).

5.2.4. *Toolbox*

The toolbox [4] provides access to standard bioinformatics tools (e.g. EMBOSS package) and adapted softwares tailored to OUEST-genopole biologists needs. Besides these generic tools, some custom services are provided. For example, one of the most recurrent demand is the possibility to make a Blast comparison against a personal bank. This tool allows to perform a more relevant and faster search in this context. The tools are accessible by two ways: an anonymous way (with the web server) and an authenticated way (direct access to the computation cluster).

5.2.5. *Generation of primers*

The accurate determination of primers is of a key importance for the biologist carrying experiments. The platform has setup a set of tools dedicated to this kind of analysis[5].

GenoFrag. The goal of GenoFrag is to deal with Whole Genome PCR Scanning (WGPS), a means for analysing 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 [59], [54]. A graphical interface is present on the Ouest-genopole bioinformatics platform server [6]. 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.

CAPS Tags (Cleaved Amplified Polymorphic Sequence). virtually digests the two sequences with the Emboss restrict program, align them with Multalign and display single enzyme cuts, taking into account the gaps appeared in the alignment. Differences are validated with the alignment, in this case a difference is a potential SNP.

[3] http://genouest.org/

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

[5] http://genouest.org/primers.php

[6] http://genouest.org/genofrag.php

Degenerate primers. A way to look for new genes is to use degenerate primers. Data are a set of protein sequences, from different species, with the same biological function. They are aligned with Multalign and the consensus of longest fragments with few ambiguous amino acids are degenerated from the 3' end. We have developed a module, working with codon usage tables that reverse translate protein sequences in nucleic sequences with a bounded degeneration cost.

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

5.2.6. Pattern matching and pattern discovery

This activity is one strategic field where the platform is trying to develop its specificity on an international level. Our goal is to be recognized as an expertise resource center in this bioinformatics domain. Particularly, three algorithms have been developed in-house by the Symbiose research team. STAN is based on a suffix tree data structure. This tool scans complete genomes or user submitted sequences. The patterns are represented in the form of a grammar. WAPAM is a tool to parse for protein patterns expressed by weighted automata in banks (nucleic acids or proteins), complete genomes and personal sequences. The upgraded web interface of WAPAM allows to execute pattern searches on GenOuest servers.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. An additional tool, Protomata-Scan, is available on the platform since September 2007. It is based on an extension of WAPAM named WASCAN which is able to handle the full expressiveness of automata. This program scans data banks with respect to automata generated by the Protomata-Learner program (see Section 5.4).

The pattern discovery section of the web site proposes several pattern discovery algorithms [7]. To facilitate the interpretation and validation of results, we propose a toolbox with various modules: pattern matching in public databanks, visualization, statistical analysis, filtering. The implementation programming languages are Python, PHP and JavaScript.

5.3. New activities of the bioinformatics platform in 2007

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

5.3.1. National Project BioMAJ (BIOlogie Mise A Jour)

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

A data integration process is obviously an essential preliminary step. This represents a major challenge and bottleneck in bioinformatics. These biological data banks contain a mass of heterogeneous data (all in different formats, several Tera bytes). These banks, after their recovery, must undergo various post treatments more or less personalized upstream of their use via various bioinformatics software (blast, SRS, emboss, gcg, ...). The banks frequency update is variable, according to the source, from daily to several times per year. Moreover, the nature and the number of banks are in constant evolution; the data between sources are cross-linked. The maintenance task is complex and heavy. A first stake consists in automating the process of updating the data banks for the administrator. Another significant stake to resolve is for the "quality" of service, providing to the users a clear vision of the integrity of data (state, exact origin, ...) constitutive of their workspaces.

BioMAJ is a joint development between three bioinfomatics platforms : INRA Toulouse (David Allouche), INRA Jouy-en-Josas (Christophe Caron) and our platform. BioMAJ is written using state-of-the-art technologies (java, xml, ..) and is based on a parametrisable workflow engine. Post processes are written for the usual formats (gcg, blast, srs, ...) and are easily customisable at user's needs. BioMAJ has been relased under an opensource licence in September 2007. The software can be downloaded from the BioMAJ website [8]

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

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

5.3.2. Geo Exporter: automating publishing of microarray data to Gene Expression Omnibus

In collaboration with the Institute of Cancer Research in Sutton, UK, we have developed a software named GEO EXPORTER which allows validating, extraction and automatic publication of microarray data stored in BASE (BioArray Software Environment), a widly used open source microarray data management system, to Gene Expression Omnibus.

GEO Exporter features can be divided according to the submission process steps: data selection, local data validation, generation of an exported data file (SOFT), remote validation at GEO, remote submission and submission management.

The program is freely distributed under GPL Licence. The software source, documentation, install instructions and demo are available at [9].

This software has been used to publish human microarray experiment results [97].

5.4. Activity of transfer from Symbiose to GenOuest

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

5.4.1. LLAhclust and CHAVLh: unsupervised hierarchical classification

We propose a LLA hierarchical clustering principles-based approach for unsupervised hierarchical classification. A dedicated software CHAVLh (Classification Hiérarchique par Analyse de la Vraisemblance des Liens en cas de variables hétérogènes) has been developed and registered at the APP agency (2005, I-C. Lerman, P. Peter (LINA), H. Leredde (Paris Nord)).

Heterogeneous description by several variable types can be handled by CHAVLh. Mathematically, the descriptive variables are interpreted in terms of relations on the object set. This software can be distinguised with respect to several items: 1. Dealing with descriptive variables of any types (numerical, categorical of any sort). 2. In the LLA method, probabilistic similarities, called likelihood of the link similarities, are established in order to compare mutually descriptive variables, described objects, variable classes or object classes. 3. In the classification tree built with the hierarchical clustering algorithm, the most "significant" levels - giving relevant partitions- and the most "significant" nodes - giving relevant aggregations - are detected.

This year, an ergonomic version called *LLAhclust* has been implemented in the R environment and is now available on the platform [10] hanks to a collaboration with I. Kojadinovic and P. Peter (École Polytechnique de l'Université de Nantes). This tool is also proposed in the framework of the european project ACGT.

5.4.2. Learning automata on sequences: Protomata Learner

Protomata Learner is a program for the inference of automata from samples of (unaligned) sequences belonging to a structural or functional family of proteins. Infered automata can be used to get new insights into the family, when classical multiple sequence alignments are insufficient, or to search for new family members in sequence data banks, with the advantage of a finer level of expressivity than classical sequence patterns (such as PSSM, Profile HMM, or Prosite Patterns). Its main interest is to allow the representation of heterogeneous sequence families. To learn such expressive models, Protomata Learner is based on the PhD work of G. Kerbellec and uses a sequence driven strategy based on the detection of strong local conservation and state merging generalization under preservation constraints[11]. It can be accessed on the GenOuest bioinformatics platform [12] where it offers an intuitive environment. A graphical viewer for the generated automata and the corresponding alignments gives a convenient way to analyse the relationship between sequences. When an interesting automata is found, it can be used to scan the sequences data banks, thanks to direct links to Protomata-Scan program.

[9] http://www.sigenae.org/geo_exporter

[10] http://cran.r-project.org/src/contrib/Descriptions/LLAhclust.html

[11] Protomata Learner is deposited with the *Agence pour la Protection des Programmes*(APP) under Cecill licence.

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

Protomata Learner is written in Python and C, the web interface has been implemented with Python, PHP and JavaScript.

5.4.3. *Domain Organizer*

Discovering the underlying organization of any genome or sequence is of high importance in order to improve our understanding of living organisms.*Domain organizer* is a new tool dedicated to the analysis of repeated elements along biological sequences, resulting from the PhD work of S. Tempel. It relies on the association of language analysis, optimization and classification tools that allow identification, characterization and graphical representation of the combinations of elementary domains that make up each sequence of a given family. This tool is available for the community on the web site [13] of the GenOuest platform.

5.4.4. *Network analysis: Pyquali and Bioquali*

Pyquali is a Python module dedicated to computations on qualitative models represented by interaction graph. Nodes of these graphs represents chemical species and arrows are labeled by $\{+,-\}$ representing influence of the variation of a specie on another specie. An efficient representation of qualitative equations as ternary decision diagrams have been developped in this module. The python interface allows a user-friendly manipulation of equations. The module is available on GenOuest website [14]. Most useful applications of pyquali are accessible through a web access named Bioquali [15]. It is dedicated to the analysis of network by following the rule which tells that a change on a molecule should be explained by the change of at least one of its direct predecessors in the graph that represents the influences on the network. Bioquali is dedicated to the following features: testing the self-consistency of a network, testing the consistency between a network and steady state data (such as expression data), diagnosing the location of errors when inconsistency occurs, predicting the variation of non-observed products when the network is coherent.

6. New Results

6.1. Modelling sequence/structure relationships

Two lines of research are carried out within the framework of linguistic analysis of sequences. The first one aims at the expression of complex models on sequences. This will serve biologists to both validate his/her model with respect to whole genomes or set of sequences and to find new candidates in sequence data banks. The second complementary line of research aims at the automatic inference of such models from a set of sequences sharing a functional or structural property. In addition to finding new candidates, the goal is then to assist the biologist to gain new knowledge on the sequence family by the discovery of explicit models showing up the important parts of the sequences and their organization.

6.1.1. *Finding modules in sequences*

Participants: Rumen Andonov, François Coste, Jacques Nicolas [correspondant], Dominique Lavenier, Israël-César Lerman, Anne Siegel, Basavanneppa Tallur, Sebastien Tempel, Phillipe Veber, Christine Rousseau, Grégory Ranchy.

J. Nicolas coordinates the national ANR project *Modulome* that aims at modelling the structure of genomes in terms of assembly of «modules» that may be copied and move inside or between genomes. This is supported by three applications on genomic mobile elements in cooperation with URGI/Inra Versailles, LME/Ifremer Brest and LEPG/CNRS Tours. We have worked this year on the characterization of two types of modules. Corresponding softwares have been transferred on the bioinformatics platform.

Dynamics of transposable element Helitron.

[13] http://genouest.org/patterndiscovery.php
[14] http://genoweb.univ-rennes1.fr/Serveur-GPO/outils/help/datafiles_test/pyquali.tgz
[15] http://genoweb.univ-rennes1.fr/Serveur-GPO/outils_acces.php3?id_syndic=226

The first study concerned the analysis of the dynamics of transposable element Helitron, a recently discovered type of transposon most widely spreaded in the genome of Arabidopsis thaliana. It has been concluded by the defense of S. Tempel's PhD [12]. The thesis studies three relations between helitrons and their host genome: their mode of invasion, the modularity of their internal sequence and their impact on close genes. We have produced a formal grammatical model of these elements, made of two extremities separated by a highly variable sequence of bounded size that led to the proposition of a new nomenclature of helitrons in Arabidopsis based on their extremities instead of their global sequence [29]. A method and a software, DomainOrganizer, allows to establish the domains' composition. It detects first domains' borders from a multiple alignment and provides a list of putative domains. From this list it searches an optimal coding of sequences in the family as concatenation of sequences extracted in the list [40]. We formulate the coding of the family as a (discrete) optimization problem, and express it as a 0/1 linear program. Sequences are then clustered and visualized with respect to domains.

Analysis of CRISPR.

The second study concerned the analysis of an intriguing genetic structure in bacteria and archaebacteria, CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats). These are formed by a repetitive skeleton including genetic material that seems imported from viruses and plasmid. CRISPR are likely to represent a fundamental immune system in prokaryotes. However, a systematic study of this family was in need of relevant models. We have formalized an important and general notion in this respect, locality, which delimits with a single parameter the distribution of occurrences of repeated units in a sequence. We have also improved our genome sequence visualization method, Pygram, which is now able to display the hierarchical organization of repeated structures on multiple genomic sequences and to synchronize the search on these sequences.

6.1.2. Matching sequences with automata and logical grammars

Participants: François Coste, Jacques Nicolas [correspondant], Catherine Belleannée, Pierre Peterlongo, Thibaut Henin, Goulven Kerbellec, Sébastien Tempel, Jessie Mahé.

Fractal structures inside DNA.

The Chaos Game representation has been used by several laboratories to get a global view of repeats in genomes. Each DNA sequence generated nice pictures with different non random fractal structures. A fundamental question is to know "where do these structures come from?", and "how can we caracterize them?". We have shown that observed structures come from the presence of regular languages in DNA sequences. We have also proposed a method to quantify this presence and to characterize these structures with probabilistic automata. We have shown that this structures are in fact sufficient to discriminate genomes [33]. Finally, the deeper understanding of the relation between the image level and the language level led us to introduce some variants on different problems.

Beyong regular patterns.

An ambitious platform to search for complex models 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 (Stan and Wapam) that goes beyond pattern matching in biological sequences and study modelling needs of biologists at the level of whole genomes. Project's concerns are expressiveness, efficiency and ergonomy.

Wapam has been extended to handle the full expressiveness of automata. This new version, named Wascan will be accessible on the GenOuest bioinformatics platform. It can already be used through the Protomata-Scan interface on the GenOuest platform to scan databases with automata produced by Protomata-Learner [50].

Our main research focus this year was on the specification of the modelling language. The language, called Logol allows writing a particular form of Definite Clause Grammars, namely String Variable Grammars (SVG) in the line of Searls' work. We did deeply revisit SVG modelling, notably including in our models the fact that observed sequences are variants that have evolved from a common ancestor. The difficulty of modelling and detecting such repeats is that the reference ancestor sequence may well not appear at all in the analysed sequence. (e.g. AATT may be an observed string made of two copies, AA and TT, of a same ancestor TA).

Formally, Logol is a dedicated constrained string language that considers two levels for chains: a level which designates abstract chains, and a level which designate concrete subsequences [46].

6.1.3. Protein structure alignment: algorithms and applications

Participants: Rumen Andonov [correspondant], Nicola Yanev, Guillaume Collet, Noël Malod-Dognin, Phillippe Veber.

Two typical problems related with protein structure are Fold Prediction/ Recognition and Fold Comparison. By their nature, 3D computational problems are inherently more complex than the similar 1D ones for which efficient solutions have been yet developed. A theoretical basis that can provide rigourous support in understanding models for structure prediction and analysis is almost non-existent, as the problems are blend of continuous, geometric and combinatorial, discrete-mathematics. We focus notably on creating efficient exact algorithms for solving integer programming problems of proved interest: sequence/structure alignment (protein threading problem-PTP), and structure/structure comparison.

Sequence/structure alignment.

We have been working on PTP since 2001 in a very narrow collaboration with two teams [14]. The first one is MIG from INRA, Jouy-en-Josas, and the second one is the Operation Research Departement from the University of Sofia, Bulgaria whose leader, Prof. Nicola Yanev is regularly visiting Symbiose. MIG created FROST the only french software participating in the prestigious competition CAFASP (Critical Assessment of Fully Automated Structure Prediction) which holds every two years. The dedicated optimization algorithm in this tool is currently based on the Lagrangian approach developed by our group [125].

The weak point of FROST and other existing methods is that it permits *only* global alignment of a sequence with a structure, that is, allowing omissions of blocks during the alignment process. Designing algorithms for flexible structures alignment is our current focus of interest. Some ideas to tackle this problem were presented in [72], based on MIP models for semi-global and local sequence/structure alignment.

Structure/structure comparison.

A multitude of measures have been proposed to quantify the similarity between protein 3-D structure. Among these measures, contact map overlap (CMO) maximization deserved sustained attention during past decade because it offers a fine estimation of the natural homology relation between proteins. Despite this large involvement of the bioinformatics and computer science community, the performance of known algorithms remains modest. Due to the complexity of the problem, they got stuck on relatively small instances and are not applicable for large scale comparison. Our results offer a clear improvement over past methods in this respect. We present a new integer programming model for CMO and propose an exact B&B algorithm with bounds computed by solving Lagrangian relaxation. The efficiency of the approach is demonstrated on a popular small benchmark (Skolnick set, 40 domains). On this set our algorithm significantly outperforms the best existing exact algorithms, and yet provides lower and upper bounds of better quality. Some hard CMO instances have been solved for the first time and within reasonable time limits. From the values of the running time and the relative gap (relative difference between upper and lower bounds), we obtained the right classification for this test. These encouraging result led us to design a harder benchmark to better assess the classication capability of our approach. We constructed a large scale set of 300 protein domains (a subset of ASTRAL database) that we have called Proteus_300. Using the relative gap of any of the 44850 couples as a similarity measure, we obtained a classication in very good agreement with SCOP. Our algorithm provides thus a powerful classication tool for large structure databases [52], [44].

6.1.4. Learning automata and grammars on protein sequences

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

This year, links between alignment and automata construction have been clarified. We formalized the partial alignment needed to build automata as an independant constrained matching optimization problem and we proposed a new heuristic algorithm to solve it [31]. This new algorithm is faster than the previous heuristic and the semi-exact algorithm proposed in [73]: it allows to handle efficiently classical real datasets. We proposed a Minimum Description Length measure to choose the best (non deterministic) automata with respect to the sequences sample. As a result, we are able to propose to the biologists Protomata-Learner, a program for the inference of automata on protein sequences that has been transferred in the GenOuest platform 5.4.

We began also this year a new line of research on learning more expressive grammars such as context-free grammars. Focusing rather on learning the structure than the language, we developed an approach based on recoding repeated elements. To handle large sequences such as genomes, efficient data structures and algorithms have to be used. To detect and score repeats, we used suffix arrays which needed to be regularly updated after each rewriting of repeat. We proposed and implemented efficient in-place algorithms to update suffix-arrays after each word recoding [48], [32].

6.2. Systems biology: analysing data and modelling interactions

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

6.2.1. Construction of models

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

Effect of ADAM12 in TFGβ signalling pathway.

This thematic is the subject of a collaboration with Inserm and a cotutored Ph-D thesis (J. Gruel). Hepatic stellate cells (HSC) are a major actor of liver fibrosis, which appears as a deregulation of an inflamatory reaction which frequently evolves into cirrhosis and liver cancer. Under various agressions from toxics (alcool) or virus (hepatitis virus C), HSC undergo a transformation from a quiescent state to an activated one. Their response to TGFβ is quite different from the response of quiescient cells or epithelial cells. Their proteasome is also different and in particular, a metallo-protease, ADAM12 is expressed specifically in these cells. Our biologist collegues from Inserm have shown that this protein binds to TGF$-\beta$ receptor II and this binding is correlated with highest levels of TGFβ signaling ([56]).

We have begun to model the traffic of TGFβ receptors in order to understand the effect of ADAM12. The model was specified in Garmen language, a modelling tool we developped in previous years. We had to write new code generators for Scilab for the simulations and Maple and Sage for algebraic computations. We were able to formulate different hypothesis on the action of ADAM12 on TGFβ receptor traffic. Using algebraic forms of the equations, it was possible to compute kinetic constants from measures in [56], giving three structural models. Comparison with other experimental measures ruled out two of these models. The remaining one predicts: (1) a sustained TGF$-\beta$ signaling after TGF$-\beta$ is washed out from the HSC; (2) a different ratio of endosomes/cell membrane TGF$-\beta$ receptors when ADAM12 is expressed. Experimentations are now underway, in the wet lab, and first results seem to confirm the first prediction.

In parallel, another study is underway to explore the specificity of the responce of HSC to five cytokines: IFN (inducing interferon), IFG (insulin like growth factor), TGF$-\beta$ (transforming growth factor), Leptine and PDGF (platelet derived growth factor). A gene expression profile as early (6h) response to each cytokine was established using micro-array techniques. Moreover, RT-PCR was used to measure the expression of 40 transcription factors in the same conditions. The goal of this study is to explore the peculiar signal pathways which are active in HSC. For this purpose, an influence graph was build from informations extracted from data bases. Statistical teatment of data is now finished and the data are under investigation. We use qualitative techniques similar to those used in the study of gene networks.

Fatty acid metabolism.

In collaboration with laboratories of INRA and supported by a cotutored Ph-D Thesis (ASC Inra program, P. Blavy), we investigated the regulation of fatty acids metabolism in hepatic cells. We initiated the construction of models at two levels of complexity: a low level *abstract model* and a high level *extended model*.

The abstract model was literature-minded. We have gathered the set of behaviors that should be satisfied by the model. In this list of behaviors, we have included metabolic and genetic data during a diet protocol on mice (provided by our collaborators). We first identified the molecules implied in the target mechanisms, then we reduced the list of molecules and interactions in order to obtain a *minimal model* coping with the initial set of behaviors. First fitting analyses with data showed that fatty acids have to be distinguished in the model depending on both their elongation and their saturation. Our aim is to obtain an explicit differential model and then to compare the prediction of this model to observation in other tissues, mostly adipocyte cells.

Complementary to this approach, we have initiated the construction of a large scale transcriptional interaction model related to the fatty acid metabolism. We aim at comparing this model to DNA-chip data on diet chicken produced by our collaborators, in order to identify missing parts in the model. We have created a curated biological interaction network data source for lipid metabolism. This network contains signed transcriptional information, that is, known (and signed)) influences between molecules; concretely it is modeled as a graph constituted of HUGO gene IDs and signed directed edges. Information in this network are deduced from 1. A manual curated annotation of the literature. 2. A curation of information contained in dedicated databasis (Ingenuity and transpath) 3. Public databases (Kegg). The model contains now interactions among 400 genes. We expect to gather interactions on up to 600 genes.

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

Participants: Ovidiu Radulescu [correspondant], Anne Siegel, Laurent Chamoin, Debasis Panda.

On the theoretical side, we have investigated the relation between robustness and hierarchies of model scontaining at least an abstract and an extended model. The abstract model is the simplest model that still reproduces the behaviour of the biological system. It can be obtained by model reduction techniques. The extended model represents the level of variability, where different perturbations act. The mapping between the extended and the abstract model is a projection from high-dimension to low dimension that concentrates (in the Gromov/Talagrand sense) for robust properties. The model reduction method that can be used is modular, being based on the selection of a set of variables to be eliminated. Depending on the type of property that is studied, several choices are possible: rapid variables, Gale-Nikaido univoque response modules, monotone response modules. This approach was studied and detailed in the *Habilitation à diriger des recherches* of O. Radulescu, defended by the end of 2006 [109].

Dynamical robustness of biological networks with hierarchical distribution of time scales.

We analysed the concepts of distributed robustness and r-robustness proposed by biologists to explain a variety of stability phenomena in molecular biology. We discussed the robustness of the relaxation time using a chemical reaction description of genetic and signalling networks. The following result holds for linear networks: for large multiscale systems with hierarchical distribution of time scales the variance of the inverse relaxation time (as well as the variance of the stationary rate) is much lower than the variance of the separate constants. Moreover, it can tend to 0 faster than $1/n$, where n is the number of reactions. We argue that similar phenomena are valid in the nonlinear case as well. As a numerical illustration we used a model of signalling network for the important transcription factor NFkB. [108], [83] [17], [18], [36]

Module reduction on MAPK network.

We have applied a model reduction method based on Gale-Nikaido univoque response module of a model of MAPK signal transduction published by Bhalla [60]. This model contains 80 products and 58 interactions and is too large complexity to be analysed. Our approach was to identify modules in this network that have a unique equilibrium state whatever the environment is, in order to reduce these modules to a unique node. A criterion to obtain such modules is the Gale-Nikaido univoque response theorem, leading to the search for negative subdeterminants in the matrix of the network. Meanwhile, computing the signs of all sub-determinants led to an untractable combinatorial issue. Then we reduced the search into biological submodules, and we used a stability rule in order to reduce the time of verification of the Gale-Nikaido property of 60%. The biological interpretation of this reduction remains to be done [47].

6.2.3. Algorithms for the analysis of models

Participants: Michel Le Borgne, Jacques Nicolas, Ovidiu Radulescu, Anne Siegel [correspondant], Carito Guziolowski, Thibaut Henin, Philippe Veber, Annabel Bourdé, Gregory Ranchy.

Diagnosis of large-scale models.

For a couple of years, one of the questions we have been investigating is whether graphs of transcriptional interactions are able to help biologists to analyse DNA data. We have modeled this question as a formal problem of compatibility between an interaction graph (representing the knowledge) and qualitative variation of a partial subset of nodes of the interaction graph (representing DNA chip data). We are know able to propose a method to validate and correct a model by using qualitative data, based on the following principle: "*during a steady state experimentation, every variation on a node should be explained by the variation of at least one predecessor*". This rule is transcribed as a set of constraints between the variations of the nodes and the signs of interactions. Then, analysing the network is nothing but studying the solutions to the constraints: (i) The *internal consistency* of the network corresponds to checking that the whole set of constraints have at least a solution. (ii) *Consistency between a network and datasets* corresponds to checking that a partial set of variations on node can be extended to a whole solution to the set of constraints. (iii) *Diagnosing* an inconsistent network means that if a system does not check the basic rule, we shall identify a subset of interactions and data that bear inconsistencies. (iv) *Predicting* new variations corresponds to identifying the variables that have the same sign in all solutions of the set of constraints.

We have developed two different approaches to study the system of constraints generated by a graph and datasets:

1. Reducing the system of constraints to an equivalent subsystem then studying the solutions to the new system of constraints with a ternary decision diagram. This approach was initiated in 2004 [116] and then implemented in a python plugin in order to manipulate efficiently the system of constraints [28] (see Sec. 5.4). We have proved that this approach is efficient on large scale regulation systems such as E. Coli transcriptional network [19], [37]: transcriptional regulations have been proved insufficient to explain steady state datasets on *E. Coli*, and we identified the part of the system that had to be corrected (by including post translation regulation) to explain the behavior of the whole system.

2. Using constraints solvers (collaboration with T.Schaub's lab, Potsdam university), we have studied the applicability of Answer Set Programming (ASP) to solve the qualitative constraints used for checking consistency between a regulatory model and experimental data. The conclusion drawn from this preliminary work is twofold : first ASP solvers can solve particularly large qualitative constraints, which occur when comparing an organism-wide regulatory network with expression data; this was demonstrated in [51] by considering genetic networks of several thousands of genes and regulations. Second, it appears that ASP, taken as a language, provides a simple and reasonably efficient way of developing and testing some extensions of our framework (in particular, stronger notions of consistency between models and data).

   This approach is complementary to the previous one since decision diagrams allow to study the full set of solutions to the system of constraints (percentage of solutions satisfying a given condition) but its size of application is limited (at most 100 nodes).

   We have illustrated the efficiency of this approach with a contribution to *sign inference* in transcriptional regulation network. Based on analyses with artificial datasets on *E. Coli* network, we have proved that we can expect to infer about 30% of the signs of the network from about 20 DNA-chip. We have then proposed an algorithm allowing to deal with errors, based on the verification of prediction by several different datasets. With this approach, we are able to infer correctly about 10% of the signs of *E. Coli* network [51]. From a technical point of view, we had to combine decisions diagrams and constraint solvers to acheive these tasks, due to the size of the reduced system of constraints and the type of questions we had about the network. The complementary between theses approaches is the subject the Ph-D thesis of P. Veber, defended by the end of 2007 [13].

Reasoning on networks with constraint solvers.

Automated reasoning on metabolic pathways is the subject of a collaboration with the university of Potsdam and of a co-tutored PhD thesis (T. Henin). We are also collaborating with UMR 6026 (Molecular and Cellular Interaction) on the modelling of the regulations of bacteria genes implied in oxydadive stress.

We use public metabolic pathway databases such as Biomodels and Reactome. We are looking for models that allow explanation of observed behaviours of biological cells, abduction of new reactions (in order to explain some non yet explainable observations) and easy querying by biologists. Answer Set Programming seems a good framework for addressing these various issues.

6.3. Optimized Algorithms on Parallel Specialized Architectures

6.3.1. ReMiX project: Reconfigurable memory for indexing genomic banks

Participants: Dominique Lavenier [correspondant], Pierre Peterlongo, Gilles Georges, Julien Jacques, Stéphane Rubini.

BLAST has steadily become the reference software for exploring genomic banks. This type of algorithm and many other algorithms such as PATTERNHUNTER [102] or CHAOS [63] 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.

More generally, the problem of mining genomic banks is either bounded by the data access (the time for scanning all the banks) or the computation time (the time to detect good anchors) [98]. We address this problem following two complementary ways: (1) speeding-up the anchor detection using reconfigurable hardware; (2) speeding-up the data access using indexing techniques ambedded with FLASH technology.

Compared to the previous RDISK project [21], 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 Gigabytes of data. In that way, indexing techniques accessing only a fraction of the bank become highly efficient.

In the ReMIX architecture, hard drives (RDISK project) 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 whole system hold 512 Gbytes of FLASH memory.

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

In 2007, we have addressed the problem of optimizing data indexing of huge banks into the ReMIX memory. This has been performed through an Arc INRIA (called FLASH). Our research has been oriented in direction of subset seeds, based on the work of L. Noé and G. Kucherov (LIFL). A subset seed is composed by non-consecutive flexible characters. This kind of seed leads to better results than standard seeds but their design is much more difficult. Thus, in order to quickly compute subset seeds or a set of subset seeds characteristics, we have built large bank of exhaustive alignments by intensive computing of Smith-Waterman alignments. This corpus, with the use of the YASS program developed in the Sequoia team (LIFL), guided us for constructing several efficient sets of subset seeds, having a sensibility slightly better that the one used by Blastp. Those seeds were implemented on the ReMIX machine. The tests performed allowed us to notice a 16X speed-up factor. This speed-up is obtained thanks to a 13 speed-up hardware factor, while the pure algorithmic approach leads to a 24% additional speed-up [35].

6.3.2. Parallelization of bioinformatics computation

Participants: Dominique Lavenier [correspondant], Van Hoa Nguyen, Guillaume Rizk, Anthony Assi, Alexandre Cornu, Gilles Georges, Julien Jacques, Ricardo Ascensio, Xianchun Ye.

This section details the different hardware we are currently investigating for supporting parallelization of various bioinformatics computations. We successively present FPGA, GPU, multicore and cluster approaches.

FPGA.

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

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

In 2007, we have developped the concept of "reconfigurable interface" for automatically generating high performance PCIexpress channels between reconfigurable operators and host processor. From a high description language, based on smart transfer actors, a generator will deliver a synthetizable VHDL code able to achieve high speed communication trough DMA channels. In the context of the PARA project, a parallel operator made of 512 pico-processors dedicated to protein sequence comparison [99] will be connected to this "reconfigurable interface".

GPUs.

We have adapted the seed heuristic algorithm on one of the latest available graphics board (NVIDIA GeForce 8800 GTX). This adaptation is not straightforward. It needs to exhibit important low-level parallelism suitable to multithread execution. Starting from a matrix implementation on GPU (using the CUDA programming model), the seed-heuristics code has been modified to fit with the execution model of the GPU chip (a mutiprocessor SIMD architecture). Performances show a speed-up ranging from 5 to 10. These first experiments are very promising and a PhD thesis has been started on this topic (G. Rizk).

Multicore chip.

For the last 2-3 years, processor performance growth have been limited due to the difficulty of steadily increasing the clock frequency. Manufacturers and worldwild R&D team already envision multicore processors of hundred (or more) processors. These new architectures will really be efficient for high performance computation if code is targeted for these hardware structures.

Having these objectives in mind, we have revisited several bioinformatics codes in order to make them suitable for an efficient implementation on these architectures. In october 2007, we have started a collaboration with ICT (Institute for Computing Technology), Beijing, China, for implementing a parallel code [104] for intensive genomic sequence comparison on the Godson-T multicore chip developed in this Institute.

Cluster.

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

We have in charge parallelizing the OncoSimulator code on the ACGT grid, and more specifically on nodes having huge computational resources (cluster) either for providing a better interactivity (shorter response time) or for providing a better accuracy of the model in a reasonable amount of time.

6.3.3. Efficient algorithms based on classical optimization technics

Participants: Rumen Andonov, Nicola Yanev.

This research axis aims to design efficient algorithms based on classical optimization technics (like Dynamic Programing (DP), Branch&Bounds (B&B), Lagrangian Relaxation (LR), diverse Heuristics etc) for solving well known and largely studied optimization problem : the knapsack problem. Our last paper [45] presents a preprocessing procedure for the 0-1 multidimensional knapsack problem. First, a non-increasing sequence of upper bounds is generated by solving LP-relaxations. Then, a non-decreasing sequence of lower bounds is built using dynamic programming. The comparison of the two sequences allows either to prove that the best feasible solution obtained is optimal, or to fix a subset of variables to their optimal values. In addition, a heuristic solution is obtained. Computational experiments with a set of large-scale instances show the efficiency of our reduction scheme. Particularly, it is shown that our approach allows to reduce the CPU time of the leading commercial software CPLEX of ILOG.

6.4. Other contributions

6.4.1. 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 [107]. In some specific case (unimodular morphism of Pisot type), iterated morphisms can be understood in a geometrical framework, thanks to the construction of a fractal [66]. This year was dedicated to the study of the topology of fractals and the application of topological properties to number systems. By combining the self-similar structure of a Rauzy fractal and its tiling properties, we obtain a suitable partition and description of the boundary of the fractal. We derive characterization of topological properties of the fractal: box dimension and Haussdorf dimension of the boundary, connectivity, homeomorphism to a disk, uncontable fundamental group [38]. When the substitution is non unimodular (the dominant eigenvalue of its matrix is an algebraic number but it is not an integer number), the fractal is adjoined adic components. By describing the boundary of the resulting structure, we obtain a characterization of real number having a purely periodic greedy expansion in a non-integer basis [15] an we given an explicit computation of the largest rational with a purely periodic greedy-expansion in a given basis (two examples are given) [42]. The examples show that the behavior of purely periodic expansions in the non unit is slightly different from the unit case.

6.4.2. Quality of association rules in Data Mining.

Participant: Israël-César Lerman.

One main goal in image segmentation consists of splitting the set of the image pixels into homogeneous spatially connex parts. Hierarchical clustering under spatial contiguity constraint can contribute to a solution of this fundamental problem. Two types of criteria have been tested in the framework of the latter approach, the Ward Criterion and the LLA criterion. Accelerated algorithm using the nearest neighbors aggregation principle, has been established [30]. This work is a result of collaboration with K. Bachar (CREDO Laboratory, ESSC Angers).

Classical data classification is concerned by symmetrical association measures. However, in the field of Data Mining association rules, one fundamental objective consists in building asymmetrical association rule measures. Relative to a rule (implication) A->B, such implication index (measure) evaluates in a certain way the propensity of B, knowing A. The LLA approach provides fruitful probabilistic indices for measuring the rule interest. However, a local definition depending solely on the rule to be evaluated becomes non discriminant for large data bases. In these conditions we have proposed a discriminant extension of the probabilistic indices obtained with respect to a set of potential interest rules. This work has been performed in collaboration with J. Azé (LRI laboratory, University of Paris Sud) [22]. These ideas and developments have to be compared with other probabilistic approaches.

7. Other Grants and Activities

7.1. Regional initiatives

7.1.1. OUEST-genopole

OUEST-genopole, the eighth national genopole, funded in January 2002 acts as a strategic project for higher education and research in life sciences, bioinformatics, and for the economic development in the fields of *marine sciences*, *agriculture and food processing* and *human health*. It is a network, federated through a GIS structure (Scientific Interest Groupment), of the various academic organisms involved in these fields (More than 50 laboratories from Inra, Inserm, Ifremer, Inria, CNRS, Universities of Rennes, Nantes, Brest and Angers) in western France (Region Bretagne and Pays de la Loire). A network of technological platforms is proposed to all members.

OUEST-genopole is headed by M. Renard (Inra Le Rheu). Jacques Nicolas in charge of the bioinformatics research field, participates in the monthly meetings of the OUEST-genopole committee.

7.2. National initiatives

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

- National Inra project Genanimal, Sigenae, BioWorkFlow and BioMAJ detailed hereafter.
- National contracts Proteus, Modulome, PARA, SITCON, VICANNE, DyCoNum. These contracts are detailed heraafter.
- ARC Inria (Action recherche concertée) FLASH and MOCA. These contracts are detailed hereafter.

7.2.1. National projects of the GenOuest platform

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

BioSide.

BioSide is a collaboration with ENSTB-Brest (P. Picouet, S. Bigaret, P. Tanguy), Station Biologique Roscoff (FR2424 - X. Bailly, E. Corre, G. Le Corguille). BioSide is an environment where bioinformatic programs are thoroughly described through a set of metadata (including the semantics of their parameters and the execution policy) thus providing access to original (and not rewritten) programs. The BioSide graphical interface, a Java based interface, intuitive and didactical, allows the design, the execution and the storage of workflows (called scenarios in the BioSide terminology). Metadata is a major element of this platform since it is used both to provide high level help to the final user (help to parameterize programs) and to guarantee dynamic extensibility of BioSide. A standalone version is already available for phylogenetic programs.

BioWorkFlow.

BioWorkFlow is a collaboration with INRA Jouy (C. Caron, V. Martin, J.F. Gibrat), INRA Toulouse (J. Gouzy, S. Carrère), Institut Pasteur (C. Letondal), RPBS (P. Tuffery), ENST-Bretagne (P.Picouet), Station Biologique Roscoff (E.Corre), URGI (F. Samson), Laboratoire Statistique et Génome Evry (M. Hoebeke).

This project, funded by ReNaBi, initiated a reflexion concerning the setup of workflows at a national level. At first, the different solutions available for bioinformatics will be analysed and a prototype will be implemented with the best solution.

BioMAJ.

BioMAJ is a joint collaboration between INRA Toulouse (David Allouche), INRA Jouy (Christophe Caron) and IRISA. It aims at the developement of a tool dedicated to the management of public biological databanks. Managing several databanks on any site is currently a tedious task involving a lot of time and man power. BioMAJ cope with this problem and proposes a complete environment dedicated to these tasks. It allow to quickly set up new databases on any site and to manage them efficiently (reports, errors management, etc.). The resulting software has been made available for the community in September 2007.

Sigenae and Genanimal.

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

Symbiose collaborates to this program via an Inra engineer, F. Moreews, 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 collaborate on the modelling of fatty acids metabolism (see Sec. 6.2) and contract eQTL. QTL (Quantitative Trait Loci) are biomarkers of genomic regions responsible of a substantial part of variations deserved on a given character. The aim of the project eQTL is to relate QTL regions obtained by linkage analysis and regions obtained by transcriptomic studies, responsible of the regulation of a set of genes.

7.2.2. ANR Projects

7.2.2.1. *Proteus*

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

PROTEUS *Fold recognition and inverse folding problem : towards large scale protein structures prediction* is a three years ANR grant *Calcul intensif* (ANR-06-CIS6-008) managed by the CEA. It started in January 2007 and involves also: i) BIOS: Biocomputing and Structure at the Ecole Polytechnique ; ii) MIG (Mathematics, Informatics and Genome), INRA Jouy-en-Josas; iii) the Physics Laboratory at the Ecole Normale Supérieure of Lyon; is) the ABI team, University Pierre et Marie Curie, Paris 6. The main co-ordinator is T. Simonson from Ecole Polytechnique. The project focuses on fold recognition and inverse folding problem. Massive genome sequencing has revealed the sequences of millions of proteins, yet only 30.000 3D protein structures are known. The classic but difficult « fold recognition » problem requires identifying the 3D structure among a library of possible structures. A complementary approach turns the problem around, and poses the «inverse folding problem » : to enumerate all the amino acid sequences compatible with a given 3D structure. In PROTEUS we will solve both problems. On the one hand, we will predict the fold of all bacterial proteins of unknown structure (300.000 proteins), using new techniques developed by the participants. On the other hand, we will solve the inverse folding problem for 1300 folds, out of 2300 known today (SCOP database), using the emerging technique of directed evolution, which mimics the natural evolutionary process. The first year report of the project is available on its site (http://migale.jouy.inra.fr/proteus).

7.2.2.2. *PARA: Parallelism and Improvement of Application Performances*

Participants: Dominique Lavenier, Gilles Georges, Alexandre Cornu, Julien Jacques.

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

7.2.2.3. *Modulome: Parallelism and Improvement of Application Performances*

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

This ANR project, Modulome [16], aims at providing methods for the identification, visualization and formal modelling of the structure of genomes in terms of an assembly of nucleotides "modules" that are repeated along a genome or between several genomes. Three other teams of Biologists and bioinformaticians are involved in this project: LDGE (Dynamique du Génome et Evolution), Institut Jacques Monod, Paris; LEPG (Etude des Parasites Génétiques), Tours; LM2E (Microbiologie des Environnements Extrêmes), Brest

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

7.2.2.4. *Sitcon: modelling signal transduction induced by a chimeric oncogene*

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

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

7.2.2.5. *DyCoNum: diophantine, dynamical and combinatorial studies of several numeration systems*

Participant: Anne Siegel.

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

7.2.3. *Other projects*

7.2.3.1. *ACI VicAnne: animation of community of biological networks*

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

The French ministry of research (Ministry Grant (ACI) IMPBio program) funded a project named Vicanne aiming to support French workshops related to dynamics of biological networks from 2005 to 2007. Jean-Pierre Mazat (Université de Bordeaux II) is the coordinator of this project. Supported workshops are the epigenomic program (genopole Evry), three two-days working sessions on a specific theme, and a satellite workshop of the French conference of bioinformatics JOBIM.

7.2.3.2. *ARC MOCA: MOdularité, Compositionalité et Abstraction dans les réseaux géniques et protéiques*

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

The development of formal languages aimed at the description of biological systems leads to new tools for reasoning about these systems. modelling, model inference, parameter estimation, model validation and checking of properties can be performed in a semi-automatic process (computer assisted). However, the growing complexity of models needs more conceptual tools. Modular decomposition is a promising one. Module definition, module composition and module abstraction/refinement are central in this approach. The goal of this ARC (Action de Recherche Cooperative) is to explore new clues in this new domain.

7.2.3.3. *ARC FLASH: Seed Optimization and Indexing of Genomic Databases*

Participants: Dominique Lavenier, Pierre PeterLongo, Gilles Georges, Julien Jacques.

This project investigates the optimization of seeds in the context of large genomic database search using BLAST-like algorithms. Typically, seeds are optimized to reach a better search sensitivity. Here, for a given sensitivity, we want to investigate how seeds can be optimized to reduce the size of the database index. It includes (i) Theoretical study based on previous works of G. Kucherov and L. Noe (LIFL) to design new seeds; (ii) Test of these new seeds on a real challenging problem: comparing 700 000 proteins agains the Human genome (in cooperation with INSERM U694, Angers); (iii) Improve the ReMIX programming for speeding up the implementation of new reconfigurable operators based on these new seeds (in cooperation with the LESTER lab., Lorient).

7.3. European initiatives

7.3.1. *Integrated Project ACGT*

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

The project [17] aims at delivering the cancer research community an integrated CIT environment enabled by a powerful GRID infrastructure. Our contribution concerns parallelism (Grid development, tumor growth simulation) and data mining (integration of CHAVL in a R environment). In 2007, as a contribution to this program, a user-friendly version of the data analyses software CHAVL was developed (see Sec. 5.4).

7.3.2. *Egide project: Modularity, Abstraction and Robustness of Network Models in Molecular Biology*

Participant: Ovidiu Radulescu.

Principal Investigators of this 2007-2008 project are A.N. Gorban (Leicester, UK), O. Radulescu (Univ. Rennes, France), INRA Génétique animale (ASC project Rennes, Toulouse) and Institute Curie, Paris. This project involves several domains: mathematics, molecular biology and bio-informatics. Its purpose is to develop mathematical techniques for the analysis of large-scale non-linear differential equations systems. These systems should describe molecular biology networks. We propose to give an answer to the following questions: 1). How can the complexity of such a system be reduced and what is the minimal complexity of a model that overcomes the most important behaviors? 2). How can the robustness of a biological system be explained We propose to study a simplified model for fatty acids metabolism including metabolic, genetic and hormonal regulations.

7.3.3. *Bulgaria: exchange research program RILA (PAI)*

Participants: Rumen Andonov, Nicola Yanev.

This program is managed by the French Ministry of Foreign Affairs [18]. The project focusses on the application of combinatorial optimization 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 analyse new mathematical models allowing to accelerate the solution of these problems. This program involves R. Andonov, J. Nicolas, F. Coste and D. Lavenier.

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 (P. Blavy, O. Radulescu, A. Siegel).
- Inserm U456 (Détoxication et réparation tissulaire). Study of gene regulations in TGFβ signalling in liver cancer (J. Gruel, M. Le Borgne, O. Radulescu, P. Veber).
- École Polytechnique de l'Université de Nantes : integration of CHAVL in a R environment (I.-C. Lerman). Related to european project ACGT.
- École Supérieure des Sciences Commerciales d'Angers (CREDO laboratory) integration of CHAVL in a R environment (I.-C. Lerman)
- Laboratoire de Microbiologie UMR1253 STLO, INRA-Agrocampus, 65, rue de Saint Brieuc CS84215, 35042 Rennes cedex, France. Design of primers (D. Lavenier)
- UMR-CNRS 6026 Structure et Dynamique des Macromolécules, Rennes(C. Delamarche): Major Intrincsic Proteins (MIP) (F. Coste, G. Kerbellec, G. Ranchy)
- VALORIA, UBS, Vannes: ReMIX project (D. Lavenier, G. Georges, S.Rubini).
- LESTER, Lorient, ARC FLASH (D. Lavenier, G. Georges).
- UMR-CNRS 6553 - EcoBio : Arabidopsis thaliana transposons (J. Nicolas, S. Tempel), database on abyssal fungi (A. Assi).
- MIcroenvironnement et CAncer, UPRES UA 3889 (Thierry Guillaudeux) : Tumor Necrosis Factor (F. Coste, G. Kerbellec)

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

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

7.5. National collaborations

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

- MIG, Inra, Jouy en Josas (J.-F. Gibrat, A. Marin): Protein threading, Proteus (R. Andonov, F. Coste, D. Lavenier)
- Institut de Biologie et de Chimie des Protéines (R. Andonov, F. Coste)
- CEA, Saclay: Laboratoire d'Intégration des Systèmes et des Technologies, Service Architectures et Conceptions. Reconfigurable Architectures Reconfigurable computing (D. Lavenier).
- LEPG Tours (Y. Bigot) : ANR Modulome (J. Nicolas, C. Rousseau, F. Coste).
- URGI Versailles (H. Quesneville): ANR Modulome (J. Nicolas, C. Rousseau).
- IML, Marseille and LIRMM, Montpellier: substitutive dynamical systems (A. Siegel).
- LIFL, Lille (D. Lavenier, P. Peterlongo).
- Curie Institute, biology, medecine and bioinformatics, Paris (O. Radulescu, A. Siegel, M. Le Borgne)
- CNRS, biology, Paris : Alain Lilienbaum. (O. Radulescu)
- U.Paris 7, rheology : Sandra Lerouge. (O. Radulescu)
- ABI, Paris VI University (Joël Pothier): structural cores (F. Coste).

7.6. International cooperations

7.6.1. Bilateral cooperations

- Institute of Computing Technology, Beijing, Chine. Parallelization of bioinformatics algorithms onto multicore processors (D. Lavenier)
- Sofia University, Protein structure prediction (R. Andonov)
- US, Stony Brook University, Drosophila developmental biology: John Reinitz (O. Radulescu).
- Russia, St. Petersburg, mathematics: Sergei Vakulenko, modelling : Vitaly Gursky, bioinformatics: Maria Samsonova. (O. Radulescu).
- India, NCBS Bangalore, systems biology: Upi Bhalla, biophysics : Madan Rao. (O. Radulescu).
- UK, U.Leicester, applied mathematics: Alexander Gorban. (O. Radulescu).
- Germany, Postdam university, Prof. T. Schaub's lab. Logic programming and boolean constraint solving. (P. Veber, J. Nicolas, T. Henin).
- Greece, Institute of Communication and Computer Systems, National Technical University of Athens, cooperation: parallelization of cancer tumor growth (oncosimulator) (D. Lavenier, A. Assi, J. Jacques)
- Poland, Poznan Supercomputing and Networking Center: Gridification of the oncosimulator on the ACGT GRID (D. Lavenier, A. Assi, J. Jacques)

7.7. Visiting scientists

The following scientists visited the Symbiose project.

- Prof. Sergio Camiz (Sapienza University, Rome) visited I.-C. Lerman during the week 26-30 of March 2007. They worked on the subject of empirical behaviour of CHAVLh in case of heterogenous data.
- Prof. Shigeki Akiyama (Niigata university, Japan) visited A. Siegel during one week in Sept. 2007. They investigated non integer number systems properties.

- Associate Prof. Joerg Thuswaldner (Leoben university, Austria) visited A. Siegel during one week in March 2007. They investigated topological poperties of fractal sets.
- Professor Sergei Vakulenko, Institut typographique, Saint Petersburg, visited O. Radulescu during 30 days. They worked on network reduction problem.
- Professor Alexander Gorban, Department of Mathematics, Université de Leicester, visited O. Radulescu during 7 days. They worked on network reduction problem.
- Professor Radu Burlacu, Department of Mathematics, Institut d'Agronomie, Bucarest, visited O. Radulescu during 7 days. They investigated modelling question related to agriculture problems.

The Symbiose project supported the following scientific visits:

- Bulgaria: 14 days in Sofia, funded by the Hubert Curien French-Bulgarian partnership "RILA 2006" No 15071XF *Prédiction et comparaison des structures tridimensionnelles des protéines* [R. Andonov]
- Leceister, UK. Two one-week visits, collaboration with A. Gorban, funded by an Egide Project [O. Radulescu]
- United-States, university of Stony-Brook, Collaboration with John Reinitz, Apr. 2007 [O. Radulescu]
- Postdam university, Germany. 3 monthes visit to the group led by Prof. Torsten Schaub. The main research interest of the team is logic programming and boolean constraint solving. More precisely, the group works on the implementation of Answer Set Programming (ASP) solvers, and on the definition of so-called Action Languages (AL). ASP and AL are a family of languages for logic programming which are particularly suited to the specification of combinatorial and planning problems. Notably, T. Schaub's group has developed a solver which is one of the most efficient among currently available implementations. [P. Veber]
- Postdam university, Germany. Collaboration during 10 days with Torsten Schaub on constraint solver ASP [T. Henin].

8. Dissemination

8.1. Leadership within scientific community

8.1.1. Fête de la Science 2007

In the context of the *Fête de la science* O. Radulescu gave a lecture *Mathématiques des cellules* in Rennes, Oct. 2007.

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

The fifth meeting dealing with the Bioinformatics platform of OUEST-Genopole held at Irisa, Rennes on Oct. 23, 2007. It was organized by Sophie Roucan, Olivier Collin and Hugues Leroy. (http://genoweb.univ-rennes1.fr/Serveur-GPO/article.php3?id_article=510). Invited speakers included Benno Schwikowski (Institut Pasteur).

8.1.3. Tribute day to L. Trilling

A special meeting was organized to the tribute of L. Trilling at the occasion of its retiring from teaching duties. It held on June, 22nd 2007. J. Nicolas was the main organizer of this meeting. Invited speakers were J.P Verjus, J.P. Banâtre, J. Cohen, C. Desmoulins, P. Besnard, R. Allen and J. Nicolas.[19]).

[19] http://www.irisa.fr/videos/seminaire_Irisa/trilling/index.htm

8.1.4. VicAnne Meeting

In the framework of ACI VicAnne, P. Veber and A. Siegel organized a two-days meeting on biological networks. The subject was *Boolean networks and celular automata in molecular biology*. Invited speakers were Philippe Narbel (Labri, Bordeaux), Nazim Fates (LORIA, Nancy), Carsten Peterson (Lund University), Claudine Chaouiya (TAGC, Marseille), Vincent Fleury (GMCM, Université de Rennes). The first day was an free conference day. The second day was dedicated to discussions and working groups.

8.1.5. BioInfoOuest thematic-day conferences

The Symbiose project regularly organizes thematic-day conferences on bioinformatics subjects[20]. 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 2007, about

- *Transcriptional regulation sequences*. This day was organized by F. Coste. Invited speakers were Maximilian Hauessler (MSNC, Gif/Yvette), Jean Philippe Vert (Ecole des Mines de Paris, Centre for Computational Biology) and Jean Imbert (Centre de Recherche en Cancérologie de Marseille).
- *Boolean networks and celular automata in molecular biology*. See the previous paragraph.

8.1.6. Symbiose Seminar

The Symbiose seminar is held on a bi-weekly basis. 15 talks were given in this framework during the year 2007. 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.

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

- Scientific Advisory Board: BIOINFAPA, INRA (Bioinformatics for Animal genomics) and Ouest Genopole (J. Nicolas).
- Editorial Board of *Mathématiques et Sciences Humaines, Mathematics and Social Sciences* [I.-C. Lerman].
- Editorial Board of *La Revue de Modulad* [B. Tallur]
- Member of the Scientific commitee of the french ministry program ANR 2007 Architectures du futur [D. Lavenier].
- Scientific committee of The *Prix de thèse Gilles Kahn 2007* [D. Lavenier]
- Member of the Scientific & Technical Council of the European Morpheus project [D. Lavenier]

8.1.8. Jury of PhD Theses

- Jury of the PhD Thesis of Julien Ah-Pine, *Sur des aspects algébriques et combinatoires de l'analyse relationnelle. Applications en Classification Automatique, en Théorie du choix social et en théorie des tresses*, University of Paris 6, [I-C. Lerman, referee].
- Jury of the PhD Thesis of Adrien Goëffon *Nouvelles heuristiques de voisinage et mémétiques pour le problème Maximum de Parcimonie* [R. Andonov]
- Jury of the PhD Thesis of Laurent Gaubert, ENIB Brest *Auto-organisation et émergence dans les systèmes couplés, individuation de données issues de systèmes biologiques couplés*, Applied Mathematics [O. Radulescu]
- Jury of the PhD Thesis of PhD of Sabri Bayyouh, *Apprentissage par proportion analogique*, Enssat Lannion [J. Nicolas].

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

- D. Lavenier was a member of the Jury of the following PhD Theses
  Jeremy Detrey, *Arithmétique Réelle sur FPGA, virgule fixe, virgule flottante et système logarithmique*, ENS Lyon [Président].
  Arthur Segard *Conception d'une architecture pour le décodage LDPC : application au cas des codes DVB-S2* Université de Cergy Pontoise [Président].
  Marc Aubry, *Annotation fonctionnelle de groupes de gènes issus de l'analyse transcriptomique*, Université de Rennes 1.
  Alix Poungou, *Nanotechnologies et Architectures reconfigurables*, Université de Bretagne Occidentale [Referee]
  Antony Lebechec, *Gestion, Analyse et Intégration des données transciptomiques*, Université de Rennes 1
  Julien Pley, *Protocoles d'accord pour la gestion d'une grille de calcul dynamique*, Université de Rennes 1
  Emeric Sevin, *Annotation des petits éléments dans les génomes procaryotes: nouveaux outils informatiques et application à Sinorhizobium meliloti*
  Sandrine Pawlicki, *Approche bioinformatique des mécanismes d'agrégation et de polymérisation des protéines amyloïdes* Université de Rennes 1
  Elise Prieur, *Méthodes et structures de données pour l'indexation et la détection de répétitions dans les séquences biologiques*, Université de Rouen

8.2. Faculty teaching

Members of the Symbiose project are actively involved in the bioinformatics teaching program proposed by the University of Rennes 1. Furthermore, R. Andonov and D. Lavenier respectively share the responsibility of the 4th and 5th year bioinformatics master degrees, with biologist colleagues from the life science department *Vie-Agro-Santé*. The originality of this 2 year training program lies in recruiting both biologists and computer scientists.

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

1. Master 1 & 2 BioInformatics. (R. Andonov, D. Lavenier)
2. Master 2 Computer Science, IFSIC. (A. Assi, F. Coste, H. Leroy)
3. Master 2 Computer Science, ENST. (H. Leroy)
4. Master 2 Mathematics. (B. Tallur)
5. Master 2 modelisation and intensive computing Lebanese University of Beirut (H. Leroy)
6. Master Biotechnologie et Bio-santé, Université de Brest (O. Collin)
7. Ecole Supérieure Agronomique de Rennes (F. Coste)
8. Engineering school S3, ENST-Bretagne Brest (O. Collin)
9. Bioinformatics, ESEO Engineering school, Angers (D. Lavenier)
10. INRIA/EDF/CEA School, Oct. 2007 *Réduction de Modèles* (O. Radulescu)

8.3. Conference and workshop committees, invited conferences

8.3.1. Conference committees

- CAp 2007, Conférence Apprentissage (F. Coste, program committee)
- EGC'2007: Knowledge extraction and management (Extraction et Gestion des Connaissances 2007), January 2007, 23-26, Namur, Belgium (I.-C. Lerman, program committee)
- ERSA 2007: International Conference on Engineering of Reconfigurable Systems and Algorithms, june 25-28, 2007, Las Vegas, Nevada, USA (D. Lavenier, program committee).
- FPL 2007: International Conference on Field Programmable Logic and Applications, Amsterdam, Netherlands, 27-29 August 2007 (D. Lavenier, program committee).
- JOBIM 2007, Marseille, July 2007 (A. Siegel, program committee)
- MajeSTIC 2007: MAnifestation des JEunes Chercheurs en Sciences et Technologies de l'Information et de la Communication, Caen 29-31 octobre 2007, France. (R. Andonov)
- Research-In-Team Generalized substitution and tilings and numeration, Porquerolles, June 2007 (A. Siegel)
- ParaFPGA : Parallel Computing with FPGA's, symposium in conjunction with ParCo2007, 4-7 September 2007. (D. Lavenier, program committee)
- PBC 2007 : Parallel Bio-Computing, Workshop on Parallel Computational Biology, Gdansk, Poland, September 9-12, 2007 (D. Lavenier, program committee)
- RC education : The 2nd International Workshop on Reconfigurable Computing Education May 12, 2007, Porto Allegre, Brasil. (D. Lavenier, program committee)
- RIAMS 2007, Réseaux d'interaction : analyse, modélisation et simulation, Lyon, Novembre 2007 (A. Siegel, program committee)
- SPL 2007 : The III Southern Conference on Programmable Logic February 26-28, 2007 - Mar del Plata, Argentina (D. Lavenier, program committee)
- First Southern African Bioinformatics Workshop (F. Coste, Reviewer)

8.3.2. International invited conferences

- Euromech meeting on "Shear Banding", Londres, UK, Sept. 2007, (O. Radulescu)
- NIH-INRIA Biomedical and Life Sciences Computing Workshop, US, Apr. 2007, Bethesda, US. (O. Radulescu)
- Center for Developmental Genetics, University of Stony Brook, US, pr. 2007, Stony Brook, US (O. Radulescu)
- Max Planck Institute Potsdam-Golm, Potsdam, Germany, Jul. 2007, *Improving model consistency using qualitative equations* (P. Veber)
- DFG-Research Center Matheon, Freie Universität Berlin, Germany, Nov. 2007 *Improving model consistency using qualitative equations* (P.Veber)

8.3.3. National invited conferences

- Séminaire de Xerox XRCE, Grenoble (F. Coste)
- Séminaires Codes & Travaux, IRISA, Apr. 2007, *Introduction aux SGBDR et conception d'un système d'information avec MERISE* [A. Assi].
- Cinquièmes rencontres autour de la plateforme Bio-informatique, Oct. 2007 *P2P and Grid combination* [A. Assi].

- Cinquièmes rencontres autour de la plateforme Bio-informatique, Oct. 2007 *Mise en place d'un système de manage ment qualité pour la plate-forme GenOuest* [A. Bourdé]
- Séminaire des groupes de travail de INRIA, Oct. 2007 *Exemple concret de certification ISO 9001 sur les plates-formes technologiques de OUEST-genopole : intérêt, difficultés rencontrées, bénéfice retiré et retour d'expérience* [A. Bourdé].
- Cinquièmes rencontres autour de la plateforme Bio-informatique, Oct. 2007 *La plate-forme GenOuest* [O. Collin].
- Cinquièmes rencontres autour de la plateforme Bio-informatique, Oct. 2007, *BioMAJ : pour la mise à disposition des données biologiques* [O. Filangi]
- ENS Cachan, Antenne de Bretagne, *Architectures Matérielles Spécialisées pour la Bioinformatique*, Oct. 2007 [D. Lavenier].
- Ecole thématique Architectures des systèmes matériels enfouis et méthode de conception associées, *Architectures Spécialisées pour la bioinformatique* March 2007 [D. Lavenier],
- Journée en l'honneur de L. Trilling, Rennes, June 2007 *Expérimentation sur le vivant: de la souris à la puce* [J. Nicolas]
- Journée Nationale du Groupe de travail SDA2 du GDR Informatique-Mathématique, Oct. 2007, *Topologie de fractals et automates* [A. Siegel].
- Séminaire du Laboratoire Systèmes Navals Complexes, Toulon, Dec. 2007, *Analyses de réseaux grande-échelle* [A. Siegel].
- Séminaire de l'unité INSERM U456, Rennes, Feb. 2007, *Utiliser les données issues de la littérature : quand, comment et pour quoi faire ?* [P. Veber]
- Séminaire de bioinformatique du LINA, Nantes, June 2007, *Modélisation qualitative et inférence dans les réseaux de régulations* [P. Veber]

9. Bibliography

Major publications by the team in recent years

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

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

[3] F. COSTE, G. KERBELLEC. *A Similar Fragments Merging Approach to Learn Automata on Proteins*, in "European Conference on Machine Learning (ECML-2005), Porto, Portugal", J. GAMA, R. CAMACHO, P. BRAZDIL, A. JORGE, L. TORGO (editors), LNAI, vol. 3720, Springer, 2005, p. 522–529.

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

[5] 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.

[6] C. Guziolowski, P. Veber, M. Le Borgne, O. Radulescu, A. Siegel. *Checking Consistency Between Expression Data and Large Scale Regulatory Networks: A Case Study*, in "The Journal of Biological Physics and Chemistry", vol. 7, 2007, p. 37-43, http://hal.inria.fr/inria-00178914/en/.

[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", oct. 2005, http://dx.doi.org/10.1093/bioinformatics/bti710.

[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] S. Tempel, J. Nicolas, A. El Amrani, I. Couée. *Model-based Identification of Helitrons Results in a New Classification of Their Families in Arabidopsis thaliana*, in "Gene", vol. 403, 2007, 10.1016/j.gene.2007.06.030http://hal.inria.fr/inria-00180376/en/.

[11] P. Veber, M. Le Borgne, A. Siegel, S. Lagarrigue, O. Radulescu. *Complex Qualitative Models in Biology: A new approach*, in "Complexus", Doi: 10.1159/000093686, vol. 2, n^{o} 3-4, 2006, p. 140 – 151, http://content.karger.com/ProdukteDB/produkte.asp?Aktion=JournalHome&ProduktNr=227088.

Year Publications

Doctoral dissertations and Habilitation theses

[12] S. Tempel. *Dynamique des hélitrons dans le genome d'Arabidopsis thaliana : développement de nouvelles stratégies d'analyse des éléments transposables*, Ph. D. Thesis, Université de Rennes 1, 2007, http://tel.archives-ouvertes.fr/tel-00185256/en/.

[13] P. Veber. *Modélisation grande échelle de réseaux biologiques :vérification par contraintes booléennes de la cohérence des données*, Ph. D. Thesis, Université de Rennes 1, 2007, http://tel.archives-ouvertes.fr/tel-00185895/en/.

Articles in refereed journals and book chapters

[14] R. Andonov, G. Collet, J.-F. Gibrat, A. Marin, V. Poirriez, N. Yanev. *Recent advances in solving the protein threading problem*, in "Grid Computing for Bioinformatics and Computational Biology", E.-G. Talbi, A. Y. Zomaya (editors), Wiley, 2007, http://hal.inria.fr/inria-00180695/en/.

[15] V. Berthé, A. Siegel. *Purely Periodic beta-Expansions in the Pisot Non-unit Case*, in "Journal of Number Theory", vol. 127, 2007, p. 153-172, http://hal.inria.fr/inria-00181997/en/.

[16] G. Didier, C. Guziolowski. *Mapping sequences by parts*, in "Algorithms for Molecular Biology", vol. 2, 2007, 11, http://hal.inria.fr/inria-00180004/en/.

[17] A. Gorban, O. Radulescu. *Dynamic and static limitation in multiscale reaction networks, revisited*, in "Chemical Engineering Science", 2007, http://hal.inria.fr/inria-00184976/en/.

[18] A. Gorban, O. Radulescu. *Dynamical robustness of biological networks with hierarchical distribution of time scales*, in "IET Systems Biology", 2007, http://hal.inria.fr/inria-00181450/en/.

[19] C. Guziolowski, P. Veber, M. Le Borgne, O. Radulescu, A. Siegel. *Checking Consistency Between Expression Data and Large Scale Regulatory Networks: A Case Study*, in "The Journal of Biological Physics and Chemistry", vol. 7, 2007, p. 37-43, http://hal.inria.fr/inria-00178914/en/.

[20] D. Lavenier, G. Georges, X. Liu. *A Reconfigurable Index FLASH Memory tailored to Seed-Based Genomic Sequence Comparison Algorithms*, in "The Journal of VLSI Signal Processing", vol. 48, 2007, p. 255-269, http://hal.inria.fr/inria-00178314/en/.

[21] D. Lavenier, G. Mathieu, V. Philippe. *Path-equivalent developments in acycliv weighted automata*, in "International Journal of Foundations of Computer Science (IJFCS)", vol. 18, 2007, p. 799-812, http://hal.inria.fr/inria-00178293/en/.

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

[23] I.-C. Lerman, P. Peter. *Representation of Concept Description by Multivalued Taxonomic Preordonance Variables*, in "Selected Contributions in Data Analysis and Classification Studies in Classification, Data Analysis, and Knowledge organization", P. Brito, P. Bertrand, G. Cucumel, F. de Carvalho (editors), Studies in Classification, Data Analysis, and Knowledge organization, Springer, 2007, p. 271-284, http://hal.inria.fr/inria-00180101/en/.

[24] P. Peterlongo, J. Allali, M.-F. Sagot. *Indexing gapped-factors using a tree*, in "International Journal of Foundation of Computer Science", 2007, http://hal.inria.fr/inria-00179719/en/.

[25] P. Peterlongo, N. Pisanti, F. Boyer, A. Pereira Do Lago, M.-F. Sagot. *Lossless filter for multiple repetitions with Hamming distance*, in "Journal of Discrete Algorithms", 2007, http://hal.inria.fr/inria-00179731/en/.

[26] O. Radulescu, A. Muller, A. Crudu. *Théorèmes limites pour des processus de Markov à sauts. Synthèse des résultats et applications en biologie moleculaire*, in "TSI (Technique et Science Informatiques)", vol. 26, 2007, p. 443-469, http://hal.inria.fr/inria-00181451/en/.

[27] C. Rispe, M. Kutsukabe, V. Doublet, S. Hudaverdian, F. Legeai, J.-C. Simon, D. Tagu, T. Fukatsu. *Large Gene Family Expansion and Variable Selective Pressures for Cathepsin B in Aphids*, in "Molecular Biology and Evolution", 2007, http://hal.inria.fr/inria-00180087/en/.

[28] A. Siegel, C. Guziolowski, P. Veber, O. Radulescu, M. Le Borgne. *Optimiser un plan d'expérience à partir de modèles qualitatifs?*, in "Biofutur", vol. 275, 2007, p. 27-31, http://hal.inria.fr/inria-00178791/en/.

[29] S. Tempel, J. Nicolas, A. El Amrani, I. Couée. *Model-based Identification of Helitrons Results in a New Classification of Their Families in Arabidopsis thaliana*, in "Gene", vol. 403, 2007, http://hal.inria.fr/inria-00180376/en/.

Publications in Conferences and Workshops

[30] K. Bachar, I.-C. Lerman. *Comparaison de deux critères de la CAH sous contrainte de contiguïté spatiale*, in "39-èmes Journées de Statistique de la SFdS, Angers France", J.-M. Marion (editor), Université d'Angers, 2007, http://hal.inria.fr/inria-00180155/en/.

[31] F. Coste, G. Kerbellec. *Problème d'optimisation de recherche de cliques pour caractériser des familles de protéines*, in "ROADEF 2007, Grenoble France", J. C. Billaut (editor), Universités de Grenoble, 2007, p. 157–158, http://hal.inria.fr/inria-00180494/en/.

[32] F. Coste, P. Peterlongo, M. Gallé. *Mise à jour incrémentale de tableau des suffxes en cours de recodage*, in "Algorithmique, combinatoire du texte et applications en bio-informatique, Chessy France", 2007, http://hal.inria.fr/inria-00186352/en/.

[33] T. Henin. *Représentation par jeux du chaos de séquences d'ADN*, in "Algorithmique, combinatoire du texte et applications en bio-informatique, Chessy France", 09 2007, http://hal.archives-ouvertes.fr/hal-00181675/en/.

[34] I.-C. Lerman. *Sur les différentes expressions formelles d'une hiérarchie binaire symétrique ou implicative*, in "XIV Rencontres de la Société Francophone de Classification, Paris France", O. Hudry, I. Charon, G. Hébrail (editors), École Nationale des Télécommunications de Paris, École Nationale des Télécommunications de Paris, 2007, p. 139-142, http://hal.inria.fr/inria-00180133/en/.

[35] P. Peterlongo, L. Noé, D. Lavenier, G. Georges, J. Jacques, G. Kucherov, M. Giraud. *Protein similarity search with subset seeds on a dedicated reconfigurable hardware*, in "Parallel Bio-Computing, Gdansk, Pologne", 2007, http://hal.inria.fr/inria-00178325/en/.

[36] O. Radulescu, A. Zynoviev, A. Lilienbaum. *Model reduction and model comparison for NFkB signaling*, in "FOSBE'07: Foundations of Systems Biology in Engineering, Stuttgart", 2007.

[37] A. Siegel, C. Guziolowski, P. Veber, O. Radulescu, M. Le Borgne. *Qualitative response of interaction networks: application to the validation of biological models*, in "6th International Congress on Industrial and Applied Mathematics. ICIAM 07, Zurich Suisse", 2007, http://hal.inria.fr/inria-00178854/en/.

[38] A. Siegel, J. Thuswaldner. *Topological properties of central tiles for substitutions*, in "Journées de Numération, Graz Autriche", 2007, http://hal.inria.fr/inria-00180246/en/.

[39] P. Veber, S. Tempel, R. Andonov, D. Lavenier, J. Nicolas. *Détection de domaines dans des séquences génomiques : un problème de couverture optimale*, in "FRANCORO V/ROADEF 2007, Grenoble France", 02 2007, http://hal.archives-ouvertes.fr/hal-00186471/en/.

[40] P. Veber, S. Tempel, R. Andonov, D. Lavenier, J. Nicolas. *Détection de domaines dans des séquences génomiques : un problème de couverture optimale.*, in "Francoro / Roadef, Grenoble France", Presses Universitaires de Grenoble, G-SCOP (FRE 3028), 2007, http://hal.inria.fr/inria-00180392/en/.

[41] D. Wegener, T. Sengstag, S. Sfakianakis, S. Rüping, A. Assi. *GridR: An R-based grid-enabled tool for data analysis in ACGT clinico-genomics trials*, in "3rd IEEE International Conference on e-Science and Grid Computing, Bangalore Inde", IEEE (editor), 2007, http://hal.inria.fr/inria-00179078/en/.

Internal Reports

[42] S. Akiyama, G. Barat, V. Berthé, A. Siegel. *Boundary of central tiles associated with Pisot beta-numeration and purely periodic expansions*, Research Report, arXiv:0710.3584v1, 2007, http://arxiv.org/pdf/0710.3584.

[43] R. Andonov, G. Collet, J.-F. Gibrat, A. Marin, V. Poirriez, N. Yanev. *Recent Advances in Solving the Protein Threading Problem*, Research Report, 2007, http://hal.inria.fr/inria-00165274/en/.

[44] R. Andonov, N. Yanev, N. Malod-Dognin. *Towards Structural Classification of Proteins based on Contact Map Overlap*, Research Report, 2007, http://hal.inria.fr/inria-00192206/en/.

[45] S. Balev, N. Yanev, A. Fréville, R. Andonov. *A dynamic programming based reduction procedure fr the multidimensional 0-1 knapsack problem*, accepted for publication in European Journal of Operational Research, ELSEVIER, doi 10.1016/j.ejor.2006.02.058, 2007, 10.1016/j.ejor.2006.02.058http://hal.inria.fr/inria-00184771/en/.

[46] C. Belleannée, J. Nicolas. *Logol : Modelling evolving sequence families through a dedicated constrained string language*, Research Report, 2007, http://hal.inria.fr/inria-00186568/en/.

[47] L. Chamoin. *Recherche de modules dans le réseau MAPK*, Stage, 2007, http://hal.inria.fr/inria-00185032/en/.

[48] M. Gallé. *In-place updating suffix arrays for grammatical inference*, Stage, 2007, http://hal.inria.fr/inria-00186356/en/.

[49] M. Le Borgne, P. Veber. *Decision Diagrams for Qualitative Biological Models*, Research Report, 2007, http://hal.inria.fr/inria-00144597/en/.

[50] J. Mahé. *Parsing avec erreurs par un protomate*, Stage, 2007, http://hal.inria.fr/inria-00185428/en/.

[51] P. Veber, C. Guziolowski, M. Le Borgne, O. Radulescu, A. Siegel. *Inferring the role of transcription factors in regulatory networks*, accepted for publication in BMC Bioinformatics, Research Report, 2007, http://hal.inria.fr/inria-00185038/en/.

[52] N. Yanev, R. Andonov. *A Novel Algorithm for Finding Maximum Common Ordered Subgraph*, Research Report, 2007, http://hal.inria.fr/inria-00171780/en/.

Miscellaneous

[53] D. Allouche, A. Assi, Y. Beausse, C. Caron, O. Filangi, J.-M. Larre, H. Leroy, V. Martin. *BioMAJ : A workflow engine dedicated to biological bank management*, in JOBIM (Journées Ouvertes Biologie, Informatique et Mathématiques), poster session, 2007, http://hal.inria.fr/inria-00176506/en/.

References in notes

[54] 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.

[55] 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.

[56] A. ATFI, E. DUMONT, F. COLLAND, D. BONNIER, A. LHELGOUALCH, C. PRUNIER, N. FERRAND, B. CLÉMENT, U. WEWER, N. THÉRET. *The disintegrin and metalloproteinase ADAM12 contributes to TGF-beta signaling through interaction with the type II receptor*, in "J Cell Biol.", vol. 178, 2007, p. 201-208.

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

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

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

[60] BHALLA, RAM, IYENGAR. *MAP KinasePhosphatase As a Locus of Flexibility in a Mitogen-Activated Protein Kinase Signaling Network*, in "Science", vol. 297, August 2002.

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

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

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

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

[65] 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.

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

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

[68] 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.

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

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

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

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

[73] F. COSTE, G. KERBELLEC. *Learning Automata on Protein Sequences*, in "JOBIM, Bordeaux, France", A. DENISE, P. DURRENS, S. ROBIN, E. ROCHA, A. DE DARUVAR, A. GROPPI (editors), 2006, p. 199–210, http://hal.inria.fr/inria-00180429/en/.

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

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

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

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

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

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

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

[81] 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.

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

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

[84] 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.

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

[86] 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.

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

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

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

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

[91] 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.

[92] 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.

[93] 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.

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

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

[96] 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.

[97] Z. KOTE-JARAI, L. MATTHEWS, A. OSORIO, S. SHANLEY, I. GIDDINGS, F. MOREEWS, I. LOCKE, G. EVANS, D. ECCLES. *Accurate prediction of BRCA1 and BRCA2 heterozygous genotype using expression profiling after induced DNA damage.*, in "Clinical Cancer Research", 2006, http://hal.inria.fr/inria-00180301/en/.

[98] D. LAVENIER. *Fine-Grained Parallelism for Genomic Computation*, in "Conference on Parallel Processing for Scientific Computing, États-Unis d'Amérique", 02 2006, http://hal.archives-ouvertes.fr/hal-00179996/en/.

[99] D. LAVENIER, L. XINCHUN, G. GEORGES. *Seed-based Genomic Sequence Comparison using a FPGA/FLASH Accelerator*, in "International IEEE Conference on Field Programmable Technology, Thaïlande", 2006, http://fpt.selfip.org/fpt06/program.php, http://hal.archives-ouvertes.fr/hal-00179994/en/.

[100] 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.

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

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

[103] 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.

[104] V. H. NGUYEN, D. LAVENIER. *Recherche dans les banques d'ADN par indexation parallèle*, in "RVIF, HCM Viet Nam", 2006, http://hal.inria.fr/inria-00180372/en/.

[105] 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.

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

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

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

[109] O. RADULESCU. *Modèles mathématiques de la complexité en biologie moléculaire et en mécanique des fluides*, Ph. D. Thesis, Habilitation à diriger des Recherches. Université de rennes 1, 2006.

[110] 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.

[111] 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.

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

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

[114] 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.

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

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

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

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

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

[120] 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.

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

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

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

[124] 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.

[125] N. YANEV, R. ANDONOV, P. VEBER, S. BALEV. *Lagrangian Approaches for a class of Matching Problems in Computational Biology*, Research Report, n^{o} RR-5973, INRIA, 2006, http://hal.inria.fr/inria-00090635/en/.

[126] 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.