INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team symbiose

SYstèmes et Modèles BIOlogiques, BIOinformatique et SEquences

Rennes

THEME 3A

Table of contents

1. Team

The Symbiose project has been created on the 1st January 2002. Its general purpose concerns the field of bioinformatics, that is, modeling and analysis of 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. In the current Inria classification, it is linked to program 3A: Knowledge Bases, relatively to the final goal to improve the access, to explicit and discover some knowledge from data. The future classification will be more suited for our project, under the program "Computer Science applied to Life Sciences".

Head of project

Jacques Nicolas [CR Inria]

Administrative assistant

Marie-Noëlle Georgault [AA Inria]

Inria staff members

Rumen Andonov [Prof., univ. Valenciennes (on secondment to Inria until September 03)]
François Coste [CR Inria]
Michel Le Borgne [MC, univ. de Rennes 1 (on secondment to Inria until September 03)]

CNRS staff members

Dominique Lavenier [DR CNRS]
Frédéric Raimbault [MC, univ. Bretagne Sud (on secondment to CNRS)]
Anne Siegel [CR CNRS]

Faculty members

Catherine Belleannée [MC, univ. Rennes 1]
Israël-César Lerman [Prof., univ. Rennes 1]
Basavanneppa Tallur [MC, univ. Rennes 1]
Raoul Vorc'h [MC, univ. Rennes 1]

Research scientists (partners)

Stéphane Rubini [MC, univ. Bretagne Ouest]
Rumen Andonov [Prof., univ. Valenciennes (since September 03)]

Ph. D. students

Andre Floëter [Ph. D. student (cotutored Potsdam univ.)]
Daniel Fredouille [Teaching Assistant]
Mathieu Giraud [Ph. D. student (AMC)]
Stéphane Guyetant [Ph. D. student BDI CNRS/Région]
Ingrid Jacquemin [Ph. D. student MENRT]
Aurélien Leroux [Ph. D. student Inria/Région]
Yoann Mescam [Ph. D. student Inria (cofunded SIB Genève)]
Sébastien Tempel [Ph. D. student MENRT, since October 03]

Technical staff members

Yves Bastide [Project technical staff (Inria, European contract Stressgenes)]
Esther Kaboré [Project technical staff (Inria, Inria/Région contract genopole)]
Hugues Leroy [Engineer Inria, at 30%]
Michel Mac Wing [Project technical staff (Inria, ACI GénoGrid)]
Emanuelle Morin [Project technical staff (Inria, Inria contract genopole)]
Elodie Retout [Project technical staff (Inra, national program Inra/Agenae)]
Anne-Sophie Valin [Project technical staff, since July 03 (Inria, Inria contract genopole)]

Visiting scientist

Nicolas Yanev [Visiting scientist, university of Sofia (Bulgaria)]

Graduate student interns

Véronique Adoue [DESS CCI / I.-C. Lerman]
Olivier Brobecker [DEA Informatique Univ. Rennes 1 / F. Coste]
Liza Courbot [DESS-CCI / C. Belleannée]
Mehrez Douaihy [ESIB, université St Joseph, Beyrouth / H. Leroy, M. MacWing]
Estelle Gabarron [DEA d'Informatique, ENSTB, Brest / B. Tallur]
Maximilian Haussler [Diplom Informatik Berlin (Potsdam) / J. Nicolas]
Boris Idmont [DEA Informatique Univ. Rennes 1 / F. Coste]
Maud Jacquinot [DESS Bioinformatique, Lille / M. Le Borgne]
S. Karthikeyan [IBAB, Bangalore, Inde / D. Lavenier]
Anna Maria Khoury [ESIB, université St Joseph, Beyrouth / H. Leroy, R. Andonov]
Olivier Pareige [DEA Informatique Univ. Rennes 1 / J. Nicolas]
Sébastien Tempel [DEA Génomique et Informatique Rennes / J. Nicolas]

2. Overall Objectives

2.1. A project in Bioinformatics

We are interested in two types of data: sequences (DNA -genomes or SNP data-, RNA or proteins) coming from public databanks and experimental data generated from post-genomic studies. The first type may be represented with words on a finite alphabet (4 to 20 letters). For the second one, raw data are images corresponding to expression levels of genes or mass spectra of proteins. For expression levels, the current technology offers mostly qualitative data.

Our research specificities include our interest in large scale studies (genomes or proteomes) and pattern discovery methods on sets of sequences. Two main tracks are studied: modeling with formal languages and development of dedicated machines. Other emerging or more transversal themes such as gene networks modeling and classification are also described in the document.

2.2. Scientific axes

The *Scientific axes* on which the project focuses derive from our choice on modeling complex biological systems in a linguistic and logical framework. More precisely, the project links together three main directions.

2.2.1. Linguistic analysis of sequences

This track concerns the search for relevant (e. g. functional) spatial or logical structures in macromolecules, either with intent to model specific spatial structures (secondary structures, disulfide bounds ...) or general biological mechanisms (transposition, frameshift ...). We tackle these problems in the framework of language theory, with an interest in both theoretical questions (language representations, search space) and practical questions (how to implement efficient parsers, how to infer language representations from a sample of sequences?). We follow a global combinatorial approach, that is, we rely on the counting of similar structures to cluster or to characterize instead of trying to estimate or adapt parameters in fixed models. Corresponding disciplinary fields are machine learning, data analysis and algorithmic on words.

2.2.2. Gene expression data: analysis and network modeling

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

system and to infer hypotheses from observations and models. This research is rooted in various fields: data analysis, graph theory, discrete event systems, qualitative theory of differential systems.

2.2.3. Parallelism

The bioinformatics treatments which we have just described require a very high computational power, competiting with the daily high throughput of genomic data. The fast access to millions of genomic objects is thus becoming a central scientific challenge. The main purpose of this direction of research is to parallelize such treatments in order to provide a significant speed-up. Implementations range from parallel computers to hardware accelerators, including grid technology.

3. Scientific Foundations

3.1. Bioinformatics

Bioinformatics has a quite large meaning and we first delimit the restricted meaning we use in our framework: we use it to specify research at the interface between computer science and molecular biology (also called computational biology) and not all "standard" informatics that is necessary to manage biological data on a daily base. Note however that our experience – common to many bioinformaticians – is that it is hard to achieve indepth research in this domain without "biocomputing", that is, participating to services of the second kind with biologists.

From the biology point of view, the main stakes of bioinformatics are to assist in the processes of discovering prognostic, diagnostic and therapeutic targets and the understanding of biological mechanisms. This covers in practice a great variety of works and we limit ourselves to the study of the macromolecular level of life, that is all studies analyzing DNA, RNA, protein or metabolic molecules. The aim is to understand the structure, the activity, and more generally, the interactions and dynamics that may exist between such components, for a general mechanism or a particular metabolic pathway. It is possible to distinguish four classes of studies (for more information, see for instance the introductory part of [81]) :

- *Data collecting.* It seems that very little research is needed at this level. The main unsolved issues are the reconstruction of a sequence from its fragments after sequencing or mass fingerprinting. Some statistical problems also exist for normalization of expression data, but these do not seem to involve new theoretical research.
- *Data and Knowledge management.* It is actually a major issue. Informations are produced in a highly distributed way, in each laboratory. Normalization of data, structuration 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. Most of the efforts to progress in the analysis of biological mechanisms are still spent in the phase of collecting and assembling high quality data. Major progress is under way with the development of ontologies and the application of XML methodology.
- *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. However, more macroscopic studies are possible, involving more complex operations on genomes such as permutations. Once sequences have been compared, phylogenies, that is, trees tracing back the evolution of genes, may be built from a set of induced distances, and this is an area for many research works. A more recent track considers on the contrary data expressing differences between individuals. Indeed, it is now possible to produce Single Nucleotide Polymorphism data intensively, which correspond to mutations observed at given positions in a sequence with respect to a population. Analyzing this type of data and relating them to phenotypic data leads to new research issues.

- *Functional and structural analysis of genomic data.* It is a wide domain, that aims at extracting biological knowledge from Xome studies, where X varies from genes to metabolites. It covers the search for genes and active functional sites, the determination of spatial structures, and, more recently, the study of interactions between macromolecules and with metabolites, particularly in regulation mechanisms.

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

3.1.1. Biological interest of pattern discovery

Due to its importance in the project, we give some details on the biological motivation of the pattern discovery issue in sequences. Biological sequences, as regards to DNA, RNA or proteins, must verify a number of important constraints with respect to the structure, the function or the activity that this sequence must exert. These constraints result in the conservation during evolution of "patterns" more or less precise and complex[1]. Complexity can range from the presence of given letters at given positions in the sequence, to long distance relations between words, due to spatial folding of the molecules, with phenomena of symmetry, copy, approximation, etc.

The conservation of patterns not only makes it possible to characterize a family of sequences, but also to explain to a certain extent the structure/function relations. For instance, patterns have been found in proteins determining an immune response (T-cells), or in promoter regions of DNA regulating the development of yeast. Of course, artefacts of low complexity sequences (with a lot of repeats) or of sequences fortuitously preserved during evolution remain possible and a return to biological experimentation remains necessary to validate observed patterns. These patterns, made up manually or automatically, are then placed at the disposal of the community in banks like Prosite or eMOTIF for proteins[2] or TRRD for DNA[3], or through prediction programs for biologically important sites (intron/exon transition, open reading frames, etc.).

Their knowledge can be used in multiple applications in biology. One of the major interest lies in the characterization of families of proteins. Many laboratories are indeed specialized in the study of a particular family of proteins, that are interesting because of their structure, localization, function or their implication in a pathological mechanism. Working on some proteins, they can then amplify their discoveries by seeking in public banks all proteins answering the patterns found. Regarding DNA, it is rather important areas of regulation, located upstream genes, which benefit from some degree of conservation, and the discovery of patterns associated with these areas might provide important information both on the probable localization of genes and their expression level. Another interest is to be able to carry out more reliable multiple alignments on the sequences (provided that the method of identification of patterns precisely does not rest on a multiple alignment method!). Finally, these patterns help in protein annotation, i.e. to get clues on the functional family, the activity or the localization of a new protein. This work is complex, because one has to take into account several sources of information and because proteins present most of the time several domains (frequently three or more) with a pattern combinatorics leading to the specific function. Note that manual annotation, that was until recently conducted by hand for high quality bases like SwissProt, is no more possible due to the size of the banks, and that obtaining an automatic annotation process of good quality is crucial for genomic.

3.2. Syntactical Analysis of sequences

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

3.2.1. Formal Languages and biological sequences

From the point of view of sequences, 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

[1] we also use the term "signature" to specify that these patterns are not linked to consensus and can have an arbitrary complexity.
[2] http://www.expasy.org/prosite, http://motif.stanford.edu/emotif
[3] http://dragon.bionet.nsc.ru/trrd

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 [69], 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 [88] 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 [74]. They aim at developing systems having the power of Turing Machines, in the line of works on DNA-computing, which is a bit different from the issue of deciding the class of languages necessary to describe biological structures. The current agreement is that the necessary expressivity is the class of "mildly context sensitive" languages, well-known in natural language analysis. For example Y. Kobayashi and T. Yokomori modeled and predicted the secondary structures of RNAs using Tree Adjoining Grammars (TAGs) [105]. The most complete work in this field seems due to D. Searls [94][95] ;
- 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 [59]. D. Searls proposed a more systematic approach based on logical grammars and a parser, Genlang [64]. 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 [55] or [71]. The first criterion to classify methods is the type and expressivity of patterns they look for. One can primarily represent a language either within a probabilistic framework, by a distribution on the set of possible words, or within a formal languages framework, by a production system of the set of accepted words. At the frontier, one finds Hidden Markov Models and stochastic automata, which have very good performances, but where classically the structure is fixed and learning is achieved on the parameters of the distribution. Thus, they are more related to the first type of representation. Distributional representations are expressed via various modalities : consensus matrices (probability of occurrence of each letter at each position), profiles (taking into account gaps), weight matrices (quantity of information at each position and contribution of each letter). At the algorithmic level, alignments play a fundamental role in general. 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 [70]. A simplified search of patterns can be done after alignment, the variable intervals between subpatterns having been decided. Most powerful programs in this field are currently Gibbs Motif Sampler, a Bayesian procedure building a consensus matrix by Gibbs sampling with organism-specific higher order models (Markov chain) for prior frequencies estimate [80], Toucan, proposing a complete workbench for regulatory sequence analysis and a Gibbs sampler,

Motif Sampler, and Meta-Meme, building a Markov network combining such matrices, produced by EM (Expectation-Maximization) algorithm.

The linguistic representation, which corresponds to our own work, generally rests on regular expressions. Algorithms use combinatorial enumeration in a partially ordered space. Among the most applied in this field, one finds the Pratt program [54], using principles very close to those found in the work of M.-F. Sagot and A. Viari [92]. Another track explores variations on the search for cliques in a graph [77][57].

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

3.2.3. Machine Learning and Grammatical Inference

Machine Learning is a research field devoted to studying the design and analysis of algorithms for making predictions about the future based on past experiences. Taking roots in Artificial Intelligence and Statistics, it focuses on the study of learning algorithms inspired as well by a cognitive view of natural learning from experience as by statistical techniques for fitting model parameters to data. Research is achieved from a theoretical point of view (Computational Learning Theory), studying learnability criteria and learnable classes of function within these criteria, and from a more practical point of view (applied Machine Learning), focusing more on the algorithms and their performances measured on real or simulated tasks. Recent success in the field comes mainly from research applying theoretical ideas for the design of new algorithms, 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 in Machine Learning: 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 [6].

Hidden Markov models have become a major concern in bioinformatics. A hidden Markov model contains the mathematical structure of a (hidden) Markov chain with each state associated with a distinct independent and identically distributed (IID) or a stationary random process. Estimation of the parameters following maximum likelihood or related principles has been extensively studied and good algorithms relying on dynamic programming techniques are now available. In contrast, determining the structure remains a difficult task. When available, domain knowledge may help to design empirically a structure but, in practice, the structure used is often very simple (e.g. left-right models like Profile HMM) and the discriminative power of HMM relies essentially on its parameter choice.

Nevertheless, knowing the real underlying structure of such models would enable to get more accurate models and also more explicit ones, providing new insights for the application. In the Symbiose project, we are studying this problem in the more general framework of Grammatical Inference. Grammatical Inference, variously referred to as automata induction, grammar induction, and automatic language acquisition, refers to the process of learning grammars and languages from sequences. Let us notice that the emphasis is not only on learning language (i.e. a set of sequences) but also on learning grammars (i.e. structural representations of the sequences of the language).

Traditionally, Grammatical Inference has been studied by researchers in several research communities including: Information Theory, Formal Languages, Automata Theory, Computational Linguistics, Pattern Recognition, etc. The grammatical inference community has begun to organize itself around its main conferences (e.g., the International Colloquium on Grammatical Inference, since 1993) and workshops: a homepage providing a centralized resource information on Grammatical Inference and its applications is now available[4] and an official steering group representing the international community has been created during ICGI'02. Japan, USA, Australia, Spain, Netherlands and France (with teams in St Etienne, Lille, Marseille, Rennes, Lannion) are among the most represented countries in this tight community.

A grammatical inference problem involves the choice of a) a relevant alphabet and a class of languages; b) a class of representations for the languages and a definition of the hypothesis space; c) a search algorithm using the hypothesis space properties and available bias (knowledge) about the domain to find the "best" solution in the search space.

State of the art in grammatical inference is mostly about learning the class of regular languages (at the same level of complexity than HMM structures) for which positive theoretical results and practical algorithms have been obtained. Some results have also been obtained on (sub-)classes of context-free languages [93]. In the Symbiose project, we are studying more specifically how grammatical inference algorithms may be applied to bioinformatics, focusing on how to introduce biological bias and on how to obtain explicit representations.

3.3. Modeling and analyzing genetic networks

3.3.1. Biological context

The genomes of multiple species being sequenced, a main question arises, dealing with integrative biology: how is a 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 appears to be necessary: an organism cannot be understood without considering its development; similarly, the functions of a network cannot be separated from its dynamics.

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

3.3.2. A literature review

Modeling cellular interactions is an old domain of biology, initiated by biologists interested in the dynamics of enzymes systems [73]. Models for genetic networks appeared as soon as gene interactions were discovered. The simplest static model consists in modeling a genetic network as an oriented graph, with labels + (activation) or - (inhibition). Such graph representations are used to store known interactions in general databases. They are also the framework of Bayesian representations, used to infer gene networks from microarray data. However, this technique appears to be incomplete without the support of literature information [101].

Boolean electronic circuits inspired one of the oldest dynamical models [75]: each gene is represented by a boolean variable, that depends on the other variables. In this framework, multi-valued models based on piecewise linear differential equations, were developed and are improved even nowadays [61]. They have proved to be good at studying dynamical properties such as stationary states or limit cycles [97], and allow the analysis and simulation of genetic networks of about 30 genes [62].

[4] http://eurise.univ-st-etienne.fr/gi/

French research on genetic networks is mainly located in Grenoble (Inria, Helix project), Marseille (IBDM, IML) and Evry, mostly dealing with logical models. Few are concerned with applications that imply biochemical pathways. Such an approach is the purpose of some international projects, in Israel [65] or US. However, the tools and methods developed there do not fit with the means (human and financial) of the local biological teams we work with.

Our purpose is to develop bioinformatics methods in order to gain explanations on the behavior of metabolic and signaling pathways interacting with gene networks. Our final goal is to identify genetic actors that have significant effect on the pathway. Such a purpose is motivated by the biological context of OUEST genopole research, mostly concerned with pluricellular organisms (chicken, human). In such applications, genetic actors are activated in the framework of complex metabolic or signaling pathways, that have their own dynamics. In monocellular organisms, biochemical phenomena underlying the action of one gene on another gene can be ignored or roughly modeled. In pluricellular organisms, these biochemical phenomena have a real influence on genetic interactions, and need to be modeled precisely.

3.3.3. Building and analyzing the models

A model is an abstract representation of a biological phenomenon which mimics the behavior of the phenomenon and is suitable for explaining it. An important point is that quantitative data are rather poor in biology: this fuzzy information narrows the set of classes of models that can be used. Currently, several classes of models are distinguished: computer models, graph models, discrete models (static or dynamics), differential models, hybrid models.

We build computer models based on an object-relational approach. They are close to the usual description of biological phenomena, but they have to contain enough information to allow the derivation of various mathematical models. We develop them with a pragmatic approach, that is, with no aim to get a comprehensive data base on interactions. Our approach is based on a mixing of interaction models stored in a data base and partial network models stored in a library of pre-built models.

The computer model is the starting point of the derivation of various mathematical models, based on different interpretations. Let us point out what kind of results we expect. When dealing with models, almost everybody expects simulations (quantitative or qualitative), given some kind of predictions on the behavior of some variables. Our aim is rather to investigate qualitative modeling for explanation, prediction and inference. This is important, especially in signaling modeling, since biological signaling networks are very intricate. For instance, the same signaling molecule may have two opposite effects, depending on the context. Both graph and differential models will be developed and studied in this framework:

- A non negligible part of the biological knowledge is in the following form: “such product increases or decreases the concentration of such product”. The derivative of a graph model allows one to fully exploit this kind of data and extract more information on the network behavior.
- Simple qualitative reasoning reaches quite quickly its limits. These limits should be overcome by the use of qualitative reasoning based on differential models [89]. This qualitative theory of differential equations has a long history going back to Poincaré. It is a mature theory with rather deep results. It has to be applied to biological models. Such differential models should allow structural analysis (coupled variables, stable domains).

3.4. Parallelism

Key words: *parallel architectures*, *grids*, *dedicated architectures*, *reconfigurable architectures*.

Mixing parallelism and genomics is both motivated by the large volume of data to handle and by the complexity of certain algorithms. First, there are data coming from intensive genome sequencing. Today (by the end of 2003), about 160 genomes – including the human genome – are completely sequenced, and there exist more than 800 other sequencing projects (see *Genomes onlines database*[5]). All these data are stored into

[5] http://ergo.integratedgenomics.com/GOLD

huge data bases whose volume approximatively doubles every year. The growth is exponential and there is no reason to expect any decline in the next few years.

Thus, the problem is to efficiently explore these banks, and extract relevant informations. A routine activity is to perform content-based searches related to unknown DNA or protein sequences: the goal is to detect similar objects in the banks. The basic assumption is that two sequences sharing any similarities (identical characters) can have some related functionality. Even if this axiom may not be true, it can give precious clues for further investigations.

The first algorithms for comparing genomic sequences have been developed in the seventies. They were essentially based on dynamic programming technics [86][96]. Then, with the increasing growth of data, faster algorithms have been designed to drastically speed-up the search. The Blast software [98] 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 [66] has been designed in that direction: it parallelizes a standard suite of bioinformatics tools dedicated to intensive genomic computations.

Other way 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 [7], BISP [58], HSCAN [67] or BioScan [103], have been proposed, leading today to powerful commercial products: BioXL, DECYPHER and GeneMatcher coming respectively from Compugen ltd.[6], TimeLogic[7] and Paracel[8].

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

But the large amount of genomic data is not the only motivation for parallelizing computations. The complexity of certain algorithms is also another strong motivation, especially in the protein folding research activity [53]. As a matter of fact, predicting the 3D structure of a protein from its amino acid sequence is an extremely difficult challenge, both in term of modeling and computation time. The problem is investigated following many ways ranging from *de novo* folding prediction to protein threading technics [81]. The first method tries to predict the spatial organization of a protein using only the sequence information. The second method tries to match an unknown protein sequence to a known 3D protein structure. The underlying algorithms are NP-complete and require both combinatorial optimization and parallelization approaches to calculate a solution in a reasonable amount of time.

4. Application Domains

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

Since the Symbiose project is focused on the field of bioinformatics, its natural application domain concerns all the standard applications of genomics: discovery of diagnostic and prognostic markers and of therapeutic targets. The understanding of the mechanisms of life is the more general underlying goal of all these studies.

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.

[6] http://www.compugen.co.il/
[7] http://www.timelogic.com
[8] http://www.paracel.com

- **Targeted gene discovery** is studied with a syntactical approach. Models are built for proteins or promoters and then searched in whole genomes. We have for instance been able to discover new beta-defensins, a family of anti-microbial peptides, in the human genome with such a strategy.
- **Whole genome analysis** is made practical through dedicated data structures and reconfigurable architectures. We have thus proposed Blast comparisons on the human genome in 1 minute, built a software for bacterial genome fragmentation, GenoFrag, that helps to study genomes variations via Long Range PCR, and studied the occurrence of retro-transposons, a family of mobile genomic units, in the genome of *Arabidopsis thaliana*.
- **Genomic/metabolic interaction networks** are modeled in eukaryotic organisms. We are studying genes and metabolites involved in the lipogenesis (chickens) and in TGF-beta-regulation in association with hepatocellular carcinomas (human).

5. Software

5.1. GenoFrag

Participants: Dominique Lavenier [correspondant], Emmanuelle Morin, Rumen Andonov, Nicolas Yanev.

The goal of GenoFrag is to deal with Whole Genome PCR Scanning (WGPS), a means for analyzing bacterial genome plasticity. This software is developed for the design of optimized primers for Long-Range PCR on whole genomes. GenoFrag initially seeks all the potential primers on a chromosome. Then it calculates the best distribution of the primer pairs, thanks to combinatorial optimization algorithms. It was tested on *Staphylococcus aureus* strains but can be used for other bacterial or viral species [23]. A graphical interface is present on the Ouest-genopole® bioinformatics platform server[9]. GenoFrag helps to design very good primers for PCR, thus avoiding checking primers and PCR conditions. This software is dedicated to biologists interested in bacterial genome variability analysis.

5.2. Pattern discovery platform

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

An inventory of already existing pattern discovery algorithms was carried out, and algorithms were developed within the project. We thus could select a panel of algorithms finding patterns whose expressivity covers a wide spectra. With the purpose to make all these algorithms accessible to computer scientists and to biologists, a Web platform grouping pattern discovery algorithms was realized, and six algorithms are available at the moment[10]. For biologists, it allows a more reliable and faster pattern search by comparing and by associating the results of all the available methods. For computer scientists it is useful to compare objectively the performances of algorithms. To facilitate the interpretation of the patterns found and to refine the search, we integrated a tool base with various modules: pattern matching in public databases, vizualisation, statistical analysis, filtering.

A module is currently under development, called STAN for Suffix Tree ANalyser. It is a syntactic analyzer applicable to a whole genome. It allows pattern matching on a chromosome using a suffix tree data structure. The selected programming languages are Python, PHP et JavaScript. The platform is available for all laboratories in OUEST-genopole®. It should be opened soon to a larger public.

5.3. Tools for Databases

Participants: Esther Kaboré [correspondant], Jacques Nicolas, Emmanuelle Morin, Yves Bastide.

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.

[9] http://genouest.no-ip.org/Services/GenoFrag
[10] http://idefix.univ-rennes1.fr:8080/PatternDiscovery/

We automatically retrieve new releases when major updates for these databanks become available. Between two major releases of certain databanks, minor updates and corrections are out at regular intervals which are also retrieved and installed in order to maintain up-to-date databases. These public databanks are available for GCG programs, a package for sequence analysis installed on the platform. An Rsync server has been also set up and maintains partial mirrors of our banks in other sites (Angers, Roscoff) for Blast and motif search tools. Databases and tools are accessible on the web server[11] under Banks item. We are setting up an environment for building specialized databases. The main goal of this work is to enable a specialized view on public data for users. A specialized database can be built around a specific species, or topic. Then, we make available dedicated tools for this database. An example of realization for this work is the oysters database. This database contains about 7000 sequences which represent about 20 oysters' subspecies. We can blast any subset of this specialized database against public databanks like GenBank, or blast a set of any sequences against the specialized database.

5.4. Genetic networks: Garmen and Gardon

Participants: Michel Le Borgne [correspondant], Anne Siegel, Maud Jacquinot.

For supporting our research in modeling and analysis of dynamical systems of biological metabolic and genetic interactions, two softwares have been developed and are still under development. The first one, named GARMeN (Graphical Analyzer for Regulatory and MEtabolic Networks) implements a first version of a computer model of interaction and the derivation of a graph model. The second, named GARDON, implements a data base of interaction.

In GARMeN, the models are specified using a declarative language. The user has to enumerate the descriptions of the various interactions of interest occurring in a cellular localization (cytoplasm, mitochondria, nucleus, etc). Localizations are implemented as scopes in traditional programming languages. A compiler builds the object based model and a graph generator builds a graph representation of the biological networks. Various tools, based on different graph traversals, can be used to display subgraphs representing interesting subnetworks of interactions such as, for example, all interactions regulated by a transcription factor. More involved algorithms generate explanations for some observations. Experiment results, with the associated hypothesis on the molecular environment, can be taken into account. Algorithms for the generation of explanations are not yet satisfactory and more research are planned to improve them.

The purpose of the data base of interaction (GARDON) is not to be a comprehensive data base of interactions, even on a precise domain. It is rather considered as a repository of knowledge on some interactions. Another part of the knowledge will consist of libraries of computer models. The development of a modeling engine mixing the two sources of informations to build new models, is under study.

6. New Results

6.1. Linguistic analysis of sequences

Two types of works are carried out within the framework of linguistic analysis of sequences. The first situation concerns a biologist designing a model for his family of interest. Our purpose is to make the model operational. This will help the biologist to both validate his model on his sequences and to find new candidates in public sequence data banks.

The second situation concerns a biologist knowing no model for his biological family of interest. Our purpose is then to infer a model from sequences. More specifically, the goal of our research is to prove tractability of recognition and discovery of complex signatures relevant for some biological problem.

6.1.1. Analysis by logical grammars

Participants: Catherine Belleannée, Jacques Nicolas, Raoul Vorc'h, Liza Courbot, Sébastien Tempel.

[11] http://genouest.no-ip.org/

We study the modeling of sequences with logical grammars, following Searls' work. Actually, he extended DCGs (Definite Clause Grammars) with new concepts such as string variables and morphisms.

One of the first objective is to make the logical grammar formalism accessible to the biologist, so that with minimum training he can design and test his own models [22]. Because the biologist is usually not familiar with grammars, this supposes the design of visual programming interfaces. The graphical model is then translated in terms of logical grammar. We began the conception of an additional graphical interface to show parsing results "into the initial model", that is, to show graphically how the model matches each sequence [60]. These works also require to adapt expressiveness to biological specificities – to deal with helix structure for instance. Consequently, the design of models for DNA, RNA or proteins needs specific expressions because of the different types of handled structures, even if the underlying analyzer remains the same. The main difficulty is then to propose a compromise between expressiveness and complexity for developing efficient analyzers. Particularly, we want this tool to be able of treating genomes or complete chromosomes. To achieve this, we rely on a lexical analysis based on a suffix tree data structure, offering particularly flexible possibilities of calculation [87]. This leads to a first tool for analysis, able of treating Prosite expressions and elementary repetitions with substitution costs. An experimentation was carried out on the genome of Arabidopsis Thaliana[5] analysing systematically for a retrotransposons family [102].

6.1.2. Grammatical Inference

Participants: François Coste, Jacques Nicolas, Daniel Fredouille.

We have continued our work on the search space for the inference of non deterministic representations by means of state-merging methods formalized in [41]. All the work descibed in this section is contained in the Ph-D work of D. Fredouille [11]. From these theoretical results, we have developed an inference algorithm visiting the space of unambiguous automata. This algorithm has been compared experimentally on artificial data with the best known state-merging algorithms (for deterministic representations, for different versions of the EDSM heuristic [79] and of the hill-climbing heuristic) and with the algorithm DeLeTe II [63]. Our conclusion positions each algorithm with respect to the kind of target languages it is best suitable for [25].

We have considered the integration of background knowledge into automata inference algorithms [24]. The goal of this integration is to improve the convergence of algorithms thanks to this knowledge and to allow the inferred automata to be interpretable by an expert of the application domain. Two kinds of knowledge have been considered:

- The first one is a formalization of syntactic constraints on strings that belong to the target language. The proposed method enables to exclude a (possibly infinite) set of strings from the inferred language. This knowledge can also be interpreted in terms of domain, i.e. a specification of a language including the target language.
- The second one reflects more semantic contraints. It is represented by the association of types on symbols of the considered strings. The typing semantics is integrated as constraints on the structure of the inferred automata, leading possibly to the inference of interpretable representations. This typing notion is an extension of the work [78]. Our work as been to enable to consider non trivial typing functions, extending both the existing formalism, algorithms to integrate this knowledge into inference, and the understanding of the notion of typing.

6.1.3. Grammatical inference for the characterization of genomic sequences

Participants: François Coste, Jacques Nicolas, Daniel Fredouille, Aurélien Leroux, Ingrid Jacquemin, Yoann Mescam.

We investigate the application of machine learning methods for the characterization of sets of genomic sequences. We focus on the discovery of explicit characterizations. In contrast with classical algorithms trying to identify significant common subsequences (or motifs), our work aims at learning more expressive models in order to better understand the organization of such subsequences.

6.1.3.1. Structured Motif discovery

For biologists, a motif is often a single word which can be found in biological sequences. When such a word is present in a set of related sequences (with mutations on it), one can define a consensus sequence which can be considered to have a biological meaning. We have already described the need for more complex motifs such as String Variable Grammars for the analysis of biological sequences. Our goal is to infer such motifs from a given sequence set, and because it is unrealistic to achieve that by exact methods, we will use a metaheuristical approach. As a first experiment, we have built a generator which produces sequences containing linked dyads. Dyads are simple motifs made of two sub-words; in linked dyads these subwords are linked by a constraint (for example a morphism between sub-words) which could be detected by statistical means (covariation). An adaptation of D.Hernandez's algorithm MoDEL, using an evolutionary exploratory strategy, shows good results on our artificial sequences and the next step will be to characterize structures in real biological sequences [15].

6.1.3.2. Inference of automata on protein sequences

We study how to adapt the classical grammatical inference state merging approach for its use on protein sequences. A new kind of heuristics based on significant similarity of subsequence have been proposed to guide the learning algorithm [72]. Similarity measures are based on physico-chemical properties of amino acids. We are also developing a more fundamental approach for taking into account these properties. We propose a new merging scheme, working on transitions instead of states and introducing a partial order on the alphabet corresponding to a lattice structure built on the set of amino acids.

6.1.3.3. Learning Language Control

Motivated by the issue of predicting cysteins bonds within proteins, we have studied the application of the control language framework proposed by Takada [100]. The idea is to build first a universal model using a formal grammar, able to recognize bonds between any cystein-pairs, and then, to control the model application through a second simpler grammar that may be automatically learned using grammatical inference.

Preliminary experiments have been made, studying various classical regular inference algorithms[29][44]. Our main conclusion is that heuristics used in classical algorithms are not appropriate for our problem. Future work is needed to refine existing algorithms and to design new heuristics. Disulfide bonds prediction appears to be a new challenging problem for testing machine learning methods.

6.2. Gene expression data: analyzing data and modeling interactions

The purpose of this axis is to contribute to an emerging research field, that is, gene expression data analysis. The final goal is to build dynamical systems that model interactions implied in biological process.

Two kinds of gene expression data analysis are investigated. First, analyzing gene expression data deals with a classification problem (how can one identify families of genes that are co-regulated?) [36]. Second, gene expression data provide information on the whole dynamics of gene networks, leading to modeling for interactions.

6.2.1. Classification

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

Our work may be placed in the general context of the interaction between the non-metric, combinatorial and statistical classification on the one hand, and a set of fundamental algorithmic problems encountered in complex data analysis, on the other. Classification comprises the un-supervised classification based on LLA (likelihood Link Analysis, CHAVL program) as well as supervised classification relevant to the discrimination by decision trees.

6.2.1.1. Hierarchical classification of very large data under contiguity constraints

One basic principle of hierarchical classification consists in agglomerating step by step the most reciprocal neighbor couples of classes. The neighboring concept is provided by a dissimilarity measure between disjoint subsets of the set to be clustered. In many large data sets applications like image segmentation, an additional

contiguity constraint must be satisfied between merged clusters. In collaboration with K. Bachar (ESSCA, Angers), the hierarchical classification algorithm CAHCVR (Classification Ascendante Hiérarchique sous Contraintes utilisant les Voisins Réciproques), supported by an efficient software, has been set up. It is theoretically and experimentally established that our algorithm is linear with respect to the size of the object set. The linear increase is lowered by adopting an original strategy of multiple aggregation, instead of a binary one. We study the behavior of two types of criteria for class association. The former is given by the classical inertia variation (Ward criterion) and the latter is provided by a parametrized family of the criteria of the LLA (Analyse de la Vraisemblance des Liens (AVL)[16]). New results have been obtained relative to an adequate parametrization in image segmentation [21].

6.2.1.2. Quality of association rules in Data Mining

One fundamental objective in Data Mining consists in defining rule-relevant measures. Relative to a rule (implication) A->B, where A and B are conditions on a given data base, such implication index (measure) evaluates in a certain way the propensity of B, knowing A. A non symmetrical nature is required for this index. The LLA approach provides fruitful probabilistic indices for measuring the rule interest. However, a local definition depending solely on the rule to be evaluated becomes non discriminant for large data bases. In these conditions we propose a discriminant extension of the probabilistic indices. The latter are obtained with respect to a set of potential interest rules. We show that the new probabilistic index remains discriminant on large data sets. The limit value for infinite sets, enables to situate the relative interest of a given rule in a set of potential rules. This work has been performed in collaboration with J. Azé of the LRI laboratory (Univ. Paris Sud) [32].

6.2.1.3. Seeking for genetic discriminant factors in the iron homeostasis

This research is a collaboration with J. Mosser (UMR-CNRS 6061), Y. Deugnier and V. Dehais (CHU Pontchaillou). The general problem consists of determining genetic patterns which can influence the hemochromatose pathology (iron overload). The data are provided by a sample (training set) of 1000 healthy breton subjects. On each individual 5 numerical biological variables, defining the iron balance, are measured and 20 SNPs (Single Nucleotid Polymorphism) are observed on each individual. Correspondence between different classifications, using LLA clustering, enables to build homogenous group. A work on class "explanation" by the descriptive variables, numerical or qualitative, has been produced [48].

6.2.1.4. Analysis of microarray data

Advances in microarray technologies have triggered the production of a huge quantity of data and clustering has become a standard tool for analyzing such data. We have been working for adapting the CHAVL program (based on LLA method) to cluster the gene expression data and to produce convenient visualization tools. This has been applied to data from the Stressgenes project.

6.2.1.5. Towards the identification of metabolic pathways

The study of metabolic concentration data is the subject of a collaboration with the university of Potsdam and of a co-tutored thesis (A. Floeter's thesis, co-tutored by J. Nicolas and T. Schaub). Metabolic concentration data are provided by the Max Planck Institute of Berlin, based on Mass Spectrometry and Gas Chromatography. The long term goal is to understand metabolite dynamics. We have achieved this year a first step towards this goal, proposing a method for stable states extraction in metabolite concentration data [33][26][42]. This method is looking for a characterization of significant thresholds for metabolite variables, based upon a global analysis of Decision Forests learned on every possible threshold and evaluated with a function combining comprehensibility and robustness. Hidden states have been discovered for some variables that are currently under investigation.

6.2.2. Modeling genetic networks inside metabolic or signaling pathways

Participants: Michel Le Borgne, Anne Siegel, Elodie Retout.

Our objective of modeling molecular interactions involved in a biological phenomenon is developed along three directions: using models supposes to know 1) how to build the models 2) how to analyze the models 3) how to confront the models with experimental evidences.

6.2.2.1. Building the models

Models we build are based on the object approach, which is closest to the description of biological phenomena in the literature.

We develop computer model building with a pragmatic approach, that is, a tool devoted to biological studies based on modeling, far from a comprehensive data base on interactions. Hence, our approach is based on a mixing of interaction models stored in a data base,and partial network models stored in a library of pre-built models. This implies the development of a data model and the develoment of a language to describe models. First versions of such data model and language were defined this year. The mixing of different sources of information on models gives rise to various problems which will be solved by a building engine, under development. As an example, the use of many different names for the same product is very common in biology despite the work done on normalization and ontologies. This is an obstacle when one wants to chain interaction models automatically.

6.2.2.2. Mathematical models

The object model is the base that is used to build various mathematical models.

The derivative of a graph model has been done and some simple analysis, based on graph theory, already implemented. This work has been done around the development of a software tool GARMeN (Graphical Analyzer for Regulation and Metabolic Networks) (see Section 5.4).

We are also aware of the limits of simple qualitative reasoning. For this reason we initiated a collaboration with mathematicians (E. Pecou in Dijon, O. Radulescu at Irmar, Rennes) on qualitative reasoning based on differential models.

6.2.2.3. Biological applications

Our models are applied to two biological problems.

First, the lipid metabolism in the liver of chicken, studied at the Animal Genetics Lab. (Inra, Rennes): fatting state (leading to obesity when pushed to extreme) is one of these complex characters, steered by hundreds of genes that may appear in different organs (liver, brain, etc). The genetic origin of this character and its variations is studied in the chicken, with a special interest to the liver. A model of the lipid metabolism is under construction in collaboration with the biologists of Inra. Such a model aims first at connecting the elements in the models and predicting new properties that will give rise to new experiments. Secondly, the model aims at interpreting the DNA chip data (predicting the gene network involved in the response to precise factors). Algorithms on graph models are under development to answer these questions.

Our second application deals with modeling the signaling of TGF-beta in the liver cancer. Indeed, many liver diseases are associated with huge changes in the micro-environment; these changes imply the expansion of a fibrosis. If the aggressions are too numerous, fibrosis gives rise to a cirrhosis. However, in France, 90 % of liver cancers follow a cirrhosis. Liver cancer is associated with the deregulation of a molecule named TGF-beta, that also has a major influence on the expansion of the fibrosis. In the U456 Lab. (Inserm Rennes), this molecule is studied intensively. The Symbiose projects works in collaboration with this team to build a model of the signaling of TGF-beta. Such a model will allow to analyze gene networks that depend on TGF-beta, thanks to DNA-chips.

6.3. Parallelism

The parallelism axis mainly focuses on two activities: (1) the design of specialized parallel machines for scanning genomic banks (RDISK project); (2) the parallelization of protein threading algorithms, and their deployment on a grid.

6.3.1. RDISK project: filtering genomic banks with reconfigurable disks

Participants: Dominique Lavenier, Stéphane Guyetant, Mathieu Giraud, Frédéric Raimbault, Stéphane Rubini.

Genomic databases are growing exponentially: the number of genomic sequences (mostly DNA and RNA sequences) is doubling every year. The GenBank release of August 2003 contains more than 27 millions

of sequences, representing 39 billions of nucleotides [52]. This bank is routinely and daily scrutinized by thousands of researchers. Actually, a common task of the molecular biology is to try to assign a function to an unknown gene or an unknown protein. More precisely, proteins are synthesized within the cells of plants and animals. To be active, a protein must adopt a specific 3D shape related to its sequence of amino acids. The shape is important because it determines the function of the protein, and how it interacts with other molecules. It is assumed that two proteins with identical functions may have similar 3D structures, leading to a similar sequence of amino acids. Even if this hypothesis is not always verified, a great number of algorithms have been proposed to quickly extract significant alignments from the banks, i.e. sequences (or portions of sequences) having a high similarity with a query sequence.

BLAST [50][51] has steadily become the reference software for detecting similarities in genomic banks. The algorithm is fast: the main idea is that a statistically significant alignment is likely to contain a high-scoring pair of aligned words. Thus, the algorithm first detects small anchors of full similarity (identical words of W characters between the query and the sequences of the bank), then try to extend in either direction in an attempt to generate an alignment with a score greater than a threshold value. The larger anchor, the faster algorithm, but the smaller sensitivity: sequences without (at least) a common word of W characters (anchor) are not reported. This type of algorithm, and many other algorithms such as PATTERNHUNTER [82] or CHAOS [56], proceed in two steps: first they seek for anchors, then they extend them into alignments. The load balancing between these two tasks depends on the quality of the anchors. Since the alignment extension can be time consuming, the goal is to limit the number of hits by providing anchors of good quality, i.e. reflecting a good probability of generating a significant alignment. Unfortunately, the more complex the anchor detection, the longer will be the computation time.

The central idea of the RDISK project is to hardwire the first step into a reconfigurable system. More precisely, we directly filter the genomic data at the disk output, in order to provide the host computer with only relevant data. The challenge is to process data at the output rate of the disk and to forward only a low percentage of the database together with anchoring informations. The filter implemented on the reconfigurable system depends both on the data to process (DNA, protein) and the application (similarity search, motif search, etc.).

The idea of attaching computation capabilities near the disk for providing on-the-fly data filtering is not new. The SmartDisk project [85], the Active Disk project [47] or the IDISK project [76] are examples of such investigation. All of them are motivated by a major trend: hard disk controllers are designed with an increasing amount of general purpose processing power and on-chip memory. Today, most of the processing power is devoted to disk scheduling and other duties, but the next generation of controllers will contain powerful embedded processors, able to perform extra tasks [99]. Compared to these projects, we differ by the type of processing power we attach to the hard disk. Instead of an embedded processor we propose to connect a reconfigurable system based on a low cost FPGA component. The main advantage is that the anchoring-search algorithm can be highly parallelized on simple hardware structures [68], allowing on-the-fly filtering of the genomic data.

Another point to consider is the time for accessing the genomic data. The goal is to design efficient filters ; therefore, the quantity of data transmitted to the processor is expected to be low. In that case, the processor is likely to be in a starving situation, with no data to process. The computation time is thus bounded by the time to access and filter the data coming from the disk. To reach a good balancing between the post-processing and the filtering process, several disks are attached to the processor, and a reconfigurable processing system is joined to each disk. The complete system is thus made of a front-end computer connected to a bunch of hard disks coupled to reconfigurable processing and interconnected through a local network – in our case an Ethernet network. Depending on the type of the query, an adequate hardware filter is first downloaded to the FPGA component before scanning the banks. The filtering occurs locally and results are sent back to the front-end computer for further post-processing.

A prototype board has been designed and successfully tested. By the end of 2003, a reconfigurable parallel disk system of 48 boards has been assembled. Genomic applications range from similarity search to complex motif extractions based on regular expression, or context free grammar [30][27][28][43].

This project is a joint collaboration with S. Derrien and L. Lhours (R2D2 project) and L. Amsaleg (TexMex project) from Irisa.

6.3.2. Protein 3D Structure Prediction via Threading

Participants: Rumen Andonov, Dominique Lavenier, Hugues Leroy, Michel Mac Wing, Nicolas Yanev.

The objective of *in silico* functional analysis techniques is to determine the function of the product of identified genes. They are based on the concept of homology. Two genes (proteins) are homologous if they are related by descent from a common ancestral gene. Homologous proteins share close 3D structures and, in most of the time, they have similar functions. The most straightforward way to infer a homology relationship between two proteins consists of comparing their amino acid sequences by well known techniques such as BLAST or FASTA. Sequence comparison methods are very fast but they suffer from a drawback: at large evolutionary distance, it can be difficult to detect any sequence similarity.

One of the alternative strategies for improving the detection of remote homologs, consists of development of fold recognition (also called threading) methods. These methods are essentially based on three observations: i) it is well known that 3D structures are better conserved than their amino acid sequences; ii) it is now widely accepted that the number of different 3D structures (folds) that exist in nature is limited; iii) the populations of different known folds are heterogeneous and it is accepted that we know a sizable fraction of all existing folds.

The *protein threading problem* consists of testing whether a target sequence *query* is likely to fold into each member in turn of a library of representative folds *cores* by searching for an *alignment* which minimizes a suitable *score function*. This problem is proved to be NP-complete optimization problem and is widely recognized as one of the most important challenges in computational biology.

Currently we follow two complementary axis in this research domain. The first is proposing novel approaches to do protein 3D structure prediction via threading. Our recent study in this direction proves that Mixed Integer Programming (MIP) models are very successful for solving the PTP problem [104][90][49]. The MIP model significantly outperforms the dedicated *branch&bound* algorithm currently used by the community in the domain. In addition, these results show that in practice the problem can be easier than in theory and that it is possible to solve real-life (biological) instances in reasonable time. However, the huge size of the MIP models for PTP seriously limits the size of solvable instances. To overcome this drawback, we propose a divide-and-conquer method based on various splitting strategies. We show that splitting the problem in subproblems and solving them separately yields an important reduction of the solution time. Our second research axis consists of deriving an efficient parallel algorithm with rare communications based on the splitting strategies [37][19][39][38][12].

Next step will be a complete integration of our MIP model in the fold recognition package (FROST) developed by the researchers from Inra/MIG [84][83], and its deployment on a grid (GenoGRID project, see Section 8.2).

6.4. Other contributions

6.4.1. Comparative genomic of bacteria using LR-PCR

Participants: Rumen Andonov, Dominique Lavenier, Nicolas Yanev.

The comparative genomic aims to study the genome variations between different species or different *versions* of the same organism. Here, we consider bacterium strains, and more precisely, the pathogenic bacteria *Staphylococcus aureus*. Strains of S. aureus, a Gram positive pathogenic bacterium, are genomically and phenotypically highly heterogeneous.

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

corresponding to the amplified – or non amplified – segments will be assigned to each bacterium strain. The final step is to perform a global analysis of all the profiles.

The segments to be amplified are determined using a reference bacterium strain. The goal is to cover the genome with overlapping segments of nearly identical size, knowing that the segments locations are constrained by starting and ending-primers. The primers are short synthetic oligonucleotides used in the PCR. They are designed to have a sequence which is the reverse complement of a target DNA region from which the amplification starts. They have to respect certain constraints: they must not include short palindrome sequences (to avoid hairpin loops), they must contain a good balance between AT and CG nucleotides (for stability purpose), they must not have a *similar* sequence upstream and downstream the primer site, *etc*. Getting all these criteria together leads to selecting specific sites in the genome to initiate the LR-PCR.

Practically, the bacterium genome is split into a few number of linear segments, called domains [13]. For example the N315 *S. aureus* reference strain, whose size is equal to 2.8 Mbp, has been split into 5 domains; the size of the larger one is equal to 1.3 Mbp. Thus, the problem of segmenting a complete bacterial genome is reduced to split each domain into segments of nearly identical size. Along a domain, there are specific positions (i.e. small 25 DNA character string) corresponding to all possible primer sites. The overlapping segments can only start and end at these positions. If we assume, for the sake of simplicity, that a solution is made of a list of N segments, and that each segment can take only P different positions, then the number of possibilities is equal to P^N. Finding the best one when N is large is clearly a combinatorial problem (N>100).

We have explored various approaches for solving this problem [20]. Given a domain, i.e. a DNA sequence ranging from a few 100 Kpb to a few Mbp, together with all potential primer positions, we need to cover it with a sequence of overlapping segments of nearly identical size. Two cases have been considered. In the first one we search for a sequence of overlapping segments, each one of size in the interval $[\underline{L}, \overline{L}]$ and as close as possible to a *given* ideal size L. In the second case L is considered as *unknown* and we look for L^*, $\underline{L} \leq L^* \leq \overline{L}$, such that the best segmentation with respect to it is of minimal error. In both cases, we solved the problem by dedicated graph algorithms (see [40] for details), allowing a short computation time (1-2 minutes).

This research is an active collaboration with Y. Leloir and N. Ben Zacour from the Inra Ensar UMR 1055 microbiology laboratory, Rennes. Implementation of the two algorithms have been performed and packaged into the GenoFrag software (see Section 5.1).

6.4.2. Iterated morphisms

Participant: Anne Siegel.

The present work is concerned with the continuity of the research of A. Siegel, started before she arrived in the Symbiose project. This work is not concerned with bioinformatics.

Iterated morphisms of the free monoid are very simple combinatorial objects which produce infinite sequences by replacing iteratively letters with words [91]. In [14], a formalism for a notion of two-dimensional iterated morphisms is introduced. It is shown that they can be iterated by using local rules, and that they generate two-dimensional patterns related to discrete approximations of irrational planes with algebraic parameters. Such a two-dimensional iterated morphism can be associated with any usual Pisot unimodular one-dimensional iterated morphism over a three-letter alphabet.

One-dimensional iterated morphisms also generate symbolic dynamical system with specific ergodic properties[35]. In some specific case (unimodular morphism of Pisot type), these dynamical properties can be interpreted into geometrical properties, in the framework of Rauzy fractal, that is, a self-similar compact subset of the Euclidean space. In [18], the dynamical properties of iterated morphism of Pisot type are generalized to non-unimodular morphisms. More precisely, the combinatorial condition of *strong coincidence* is proved to be a sufficient condition for the dynamical system associated with a non-unimodular iterated morphism of Pisot type to be measure-theoretically isomorphic with an exchange of domain on a set called the *Rauzy fractal of the iterated morphism*. The Rauzy fractal is a self-similar compact subset of the product of an Euclidean space with finite extensions of p-adic fields. As a consequence, every substitutive dynamical system of Pisot type is

a finite extension of its maximal equicontinuous factor. This maximal factor contains a p-adic translation if and only if the incidence matrix of the substitution is nilpotent modulo p.

7. Contracts and Grants with Industry

We have no contracts in relation with industry.

8. Other Grants and Activities

8.1. Regional initiatives

8.1.1. OUEST-genopole

OUEST-genopole, is the seventh national genopole (the first one is Evry genopole and the eigth and last one is Institut Pasteur). It has been created in January 2002, for an initial duration of 2 years and offers particularly unique competences in the field of marine genomics. More generally, it acts as a strategic project for higher formation and research in the field of life sciences, bioinformatics, and for the economic development in the fields of *marine sciences*, *agriculture and food processing* and *human health*. It is a network, federated through a GIS structure (Scientific Interest Groupment), of the various academic organisms involved in these fields (Inra, Inserm, Ifremer, Inria, CNRS, Universities of Rennes, Nantes, Brest and Angers) in western France (régions Bretagne and Pays de la Loire). A network of technological platforms is proposed to all members. The director is M. Renard (Inra Le Rheu). The co-director and representative for Inria is C. Labit. J. Nicolas is responsible for the Bioinformatics field and takes part in the monthly meetings of the scientific committee. He also takes part in the national bioinformatics committee of the network of genopoles (RNG). The bioinformatics platform is developed in the Symbiose project with a complete set of tools and data bases for biologists and bioinformaticians. The web site is http://genouest.no-ip.org/. A general presentation of the platform has been done in October in Lyon[46].

8.1.2. Bioinformatics Platform

Participants: Esther Kaboré, Hugues Leroy, Michel Mac Wing, Emmanuelle Morin, Anne-Sophie Valin, Jacques Nicolas.

Five technical platforms funded by a state and regional contract have been defined within the framework of the *OUEST-genopole*, at five different locations in the western part of France. These are:

1. A DNA chips platform in Nantes (IFR26, Inserm unit 533).
2. A proteomics platform in Rennes (Inserm U435).
3. A gene sequencing platform in Roscoff (CNRS, marine biology research station) and in Le Rheu (Inra, UMR 118).
4. An Innovative Protein Array discovery platform for vectorology in Nantes (IFR 26) and functional exploration in Rennes (IFR 91, 97, 98).
5. A Bioinformatics platform, based on a parallel computer facility (SunFire 12K and 6800), with secondary centers in Brest, Nantes, Roscoff and Angers.

O. Collin, from Roscoff, and H. Leroy are in charge of the boarding committee of the platform, until the arrival of an engineer, specially assigned to this task. Training courses have been carried out (Wisconsin package, etc.) and two engineers, recruited on a fixed duration work contract, have recently strengthened the staff in order to ensure the management of data and softwares, enabling thus inter-disciplinary potential cooperations.

8.1.3. Agenae

Participants: Elodie Retout, Jacques Nicolas.

The AGENAE 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). This research program is driven by a scientific interest group formed for 5 years. It associates public research organizations (Inra, Cirad) and professional structures (Apis-Gene, Cipa). At the international level, a privileged partner is the American ARS (Agricultural Research Service) which develops a comparable project.

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

In the midst of the project, the role of E. Retout, working in the Symbiose project, is to take part in the construction of the SIGENAE informations system, sequence and expressivity database, to develop tools of clones management and transcriptome analysis, and to make comparative mapping between similar species.

This year, as part of the AGENAE project, E. Retout was especially fastened to the following missions of development:

- tweaking and maintenance of a module to publish private sequences to public databases;
- realization of a SIGENAE data-gathering module;
- statistical analysis of sequence quality in return of sequencing;
- working on revision of SIGENAE system to assemble and to annotate EST;
- realization of plug-ins for a software of microarrays analysis, BASE.

E. Retout also provides the coherence of developments of Stressgenes project in relation with AGENAE ones. In addition to these missions of development, E. Retout has taken part in a training mission (support working-out and speech) near the biologists from Rennes implied in the AGENAE program.

Moreover, as part of the Symbiose project, E. Retout was in charge of drawing human and mouse defensins on comparative maps. This work progressed in two steps:

- searching, choice and implementation of a web-based tool to view comparisons of genetic and physical maps, CMAP;
- development of scripts to retrieve and to format different data needed for building comparative maps.

8.2. National initiatives

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

- CNRS Specific Action "Machine Learning and Bioinformatics', Working group: Machine Learning and sequences (F. Coste, J. Nicolas).
- Action interEPST "motifs finding in biological sequences" (J. Nicolas)
- Working group GiGn: "grid for genomics" of IMPG (D. Lavenier).
- Working group "Genetic networks" of IMPG (M. Le Borgne, A. Siegel).
- National contracts GENOTO3D, ReMiX, GenoGRID, RDISK. These contracts are detailed in the following.

8.2.1. Project GENOTO3D

Participants: Jacques Nicolas, François Coste, Rumen Andonov.

The goal of the GENOTO3D project is to develop and integrate machine learning approaches for the protein tertiary structure prediction task. The prediction of the three-dimensional structure of protein is a great challenge both for the difficulty of the task and for the importance of the problem with applications in many fields (biology, genetics, drug design, etc.). Although the experimental approach is very expensive, an increasing number of structures are available in the Protein Data Bank PDB[12] which may be used by programs to predict the structure of a query protein sequence. The GENOTO3D project proposes to use numerical and symbolic machine learning approaches to predict long-term dependencies - which are still badly exploited by the classical prediction methods - and a divide-and-conquer strategy to integrate the different prediction levels in a single model.

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

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

Participants: Jacques Nicolas, Dominique Lavenier, Frédéric Raimbault, Stéphane Rubini.

Indexing is a well-known technique that accelerates searches within large volumes of data such as the ones needed by applications related to genomics, to content-based image or text retrieval. Very large indexes (larger than the main memory capacities) need to be stored on the hard disk drives. In that case, the design of indexes is fully disk-oriented, since minimizing disk I/Os is the key point to reduce response times. Therefore, such indexes are concerned with low level notions such as pages, fill-factors, tracks, cylinders, etc. In addition, such disk-oriented design indirectly impacts the search algorithms that navigate within the index since they have to favor sequential patterns both for processing data in individual disk pages and for fetching disk pages, avoiding as much as possible any random access to data.

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

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

It is important to point out that this new memory architecture is far from being a simple memory extension to substantially increase the memory capacity of a standard computer. The reasons are the following:

- The reconfigurable index memory is not a simple storage device. It is enhanced with additional reconfigurable hardware resources for tailoring its use according to the index characteristics and to the data it manipulates.
- The reconfigurable index memory does not fit in the addressing space of the processor. It is indirectly accessed by specific queries submitted by the processor in order to execute crucial and costly indexing subroutines.

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

- The reconfigurable index memory does not hold any cache hierarchy, and therefore memory accesses do not have to worry about the data locality. Memory read operations have a unique cost, whatever the memory address, and whatever the previous memory accesses.

Experimentation on this platform will be carried out with three application domains where huge volume of data are manipulated: genomic bank search, content-based image retrieval, and text information retrieval in heterogeneous XML knowledge databases [17][34].

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

8.2.3. Project GénoGRID: An experimental grid for genomic application

Participants: Dominique Lavenier, Hugues Leroy, Rumen Andonov, Frédéric Raimbault, Michel Mac Wing.

The GénoGRID[13] project aims to experiment with a grid of parallel computers for time-consuming genomic computations [31][45]. The computing and data resources belong to genomics or bioinformatics centers spread over the western part of France, and are interconnected through the Renater and the Megalis high speed French networks. The access to the grid is secured and restricted to authentified users.

The project mainly includes three different aspects:

1. *A secure and interactive access to the grid* The idea is that a biologist can access the grid as simply as he can access a standard web site. The only difference is that he must be recognized by the system. A connection is thus established through a secured portal by means of a Certificate Authority protocol. Once connected, a list of applications is proposed according to the user identification. Running an application is done by filling up one or several forms to tune the application parameters and to provide access path to the data. A job control panel allows the biologist to follow interactively the progress of the jobs.
2. *A transparent use of the resources* The grid is composed of several parallel computers (nodes) geographically dispatched in the western part of France. Since they are located in genomics or bioinformatics centers, the main genomic banks and software are available on different nodes. On the GénoGRID system, running an application across the grid consists of: (1) splitting the application into independent batches, (2) selecting the nodes having the right resources (data and/or software), (3) broadcasting the request to those nodes, (4) allocating the batches to the nodes according to their loads. Actually, the last operation is performed dynamically: every node runs a distributed algorithm based on a consensus protocol mechanism. From the user side, the allocation of the grid resources is entirely transparent and fully fault tolerant.
3. *The "gridification" of a few genomic applications* The purpose is to validate our approach with real genomic time-consuming problems. Three applications have been selected as a first experiment: The first one deals with intensive sequence comparison such as data bank to data bank comparison implying sensitive search. The second one concerns the implementation of a protein threading algorithm. The third one is related to the detection of repeat sequences inside full genomes. These applications share the extremely interesting property that they can easily be cut into independent tasks, leading to a very efficient degree of parallelism across the grid.

D. Lavenier is the coordinator of this three year project (January 2002 - December 2004) funded by the French ministry of research (Ministry Grant (ACI) GRID program). The Symbiose project is involved in the grid deployment, the design of the portal and the gridification of a very time consuming application: protein threading.

[13] GenoGRID web site: http://genogrid.no-ip.org

8.2.4. RDISK: Reconfigurable DISK

Participants: Dominique Lavenier, Mathieu Giraud, Stéphane Guyetant, Frédéric Raimbault, Stéphane Rubini.

This 2 year contract (January 2003 - December 2003) has been funded by the French ministry of research under a specific bioinformatics program. It aims to develop a specialized architecture for contend-based genomic data extraction. Data are stored on a parallel disk system and on-the-fly filtered at the disk output. Thus, only relevant data are propagated to the main processor memory for further processing. Filters can be on-line configured according to the database scanning (similarity search, motif search, etc,). More details can be found in Section 6.3.

8.3. European initiatives: Stressgenes

Participants: Yves Bastide, Jacques Nicolas.

In November of 2001 started the European contract STRESSGENES (Q5RS-2001-02211, *Quality of Life and Management of Living Resources Area 5.1.2*), a functional genomic approach to measuring stress in fish aquaculture. This is a partnership between Inra (Scribe) in Rennes, Inria, the university of Aberdeen, the National University of Ireland (Galway), the university of Liverpool, the Natural Environment Research Council (Ambleside), and Uppsala university. The overall aim of this study is to identify in fish candidate genes associated with resistance to stress conditions and thus provide the physiological and genetic basis for new marker-assisted selection strategies. Using new genomic tools, particularly microarray technology, and a well-known model species (rainbow trout), this study should lead to characterization of stress-responsive genes as potential candidate gene markers.

Our task is to provide other partners with the technology and tools for:

1. experimental data management,
2. searching and extraction of useful data in public banks,
3. exploitation of data retrieved from both sources.

Last year, we installed a mailing list and a web-accessible shared files system, and provided an electronic laboratory notebook. We implemented this year the centralized database for storing DNA sequencing informations. Each partner can upload his sequences, which are then compared between each others and with public databanks (genetic as well as proteic) and annotated.

Another tool we developed allows the comparison of genes expression levels. Genes from stressed samples, in different tissues and at different time points, are compared with genes from control specimens. This tool allows the selection of interesting genes, that is, genes varying significantly among different sample pools, and also invariant genes to be used as control data. This selection will serve to manufacture DNA microarrays for the actual research. Here, false positives are less inconvenient than false negatives. Hence, we used simple thresholds on the log-ratios of expression levels; the users have to input previously normalized expression data. No statistical tool for lowering the false discovery rate was implemented. We also manually increased these sets of interesting genes thanks to standard bioinformatics softwares.

Yearly progress reports are available on the Stressgenes web site.

8.4. 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:

- Project Equipage of Valoria (UBS). Collaboration with the Symbiose, R2D2 and TexMex projects in the framework of the ReMIX contract.
- Inra Rennes - Technologie Laitière - Microbiologie : study of Staphylococcus aureus genome plasticity (R. Andonov, D. Lavenier).

- Inra Rennes - Laboratoire de Génétique Animale : analysis of gene regulation involved in the lipid metabolism (M. Le Borgne, J. Nicolas, A. Siegel).
- Inra Rennes - Scribe : gene analysis implied in the trout response to stress (J. Nicolas, Y. Bastide).
- Ifremer, CNRS Roscoff, in the framework of Genogrid (H. Leroy, D. Lavenier).
- UMR-CNRS 6061 - Génétique et Développement : Statistical analysis of SNPs (I.-C. Lerman).
- UMR-CNRS 6026 (Equipe récepteurs et canaux membranaires) : study of the structure of MIP proteins (C. Belleannée, J. Nicolas, F. Coste, D. Lavenier).
- Hôpital de Pontchaillou, Yves Deugnier (I.-C. Lerman).
- Inserm U456 (Détoxication et réparation tissulaire). Study of gene regulations in TGF-beta signalling in liver cancer (M. Le Borgne, A. Siegel).
- Inserm U522. Liver transcriptome (J. Nicolas).
- Inserm U435 (groupe Germ). Family of defensines, intensive sequence matching (D. Lavenier, J. Nicolas).

8.5. National collaborations

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

- ABISS, Univ. Rouen: Laurent Mouchard (Ministry Grant (ACI) GénoGRID).
- EURISE, Univ. J. Monnet, St Etienne: C. de la Higuera, C. Kermorvant (D. Fredouille, F. Coste).
- Equipe Bioinformatique, Inra Toulouse: D. Kahn, E. Courcel (D. Lavenier, J. Nicolas).
- IMB, Dijon: E. Pécou (A. Siegel, M. Le Borgne).
- IML, Marseille: P. Arnoux (A. Siegel).
- LIH, Le Havre: F. Guinand, S. Balev (Ministry Grant (ACI) GénoGRID).
- LIRMM, Montpellier: V. Berthé (A. Siegel).
- MIG, Inra, Jouy en Josas: J.-F. Gibrat, A. Marin (Ministry Grant (ACI) GénoGRID).

8.6. International Collaborations

- University of Geneva (SIB). Y. Mescam, J. Nicolas, R. Andonov and F. Coste are involved in collaborations on motif discovery.
- University of Sofia (Bulgaria). We have a cooperation with the University of Sofia, Bulgaria in the framework of the exchange research program RILA'2003 "Programmes d'actions intégrées (PAI)" managed by the French Ministry of Foreign Affairs[14]. The project focusses on the application of combinatorial optimisation techniques in two different domains. The first domain is the so called "Protein Threading Problem" which aims at finding the three dimensional (3D) structure of a target sequence based on known 3D template structures. The second domain of application concerns the optimisation of the inference automata for discovering signatures of a sequence. Both domains are very rich in NP-hard problems and the goal of the project is to propose and to analyze new mathematical models allowing to accelerate the solution of these problems. Researchers participating in the project are R. Andonov, J. Nicolas, F. Coste and D. Lavenier.
- Postdam university (learning in metabolic pathway). A co-tutored Ph-D thesis started in 2002.
- Liverpool university (expression data analysis). The european contract Stressgenes provides the opportunity of contacts.

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

8.7. Visiting scientists

- Prof. Nicolas Yanev, Université de Sofia, Bulgarie. Visit during 2 months (invited visiting scientist at IFSIC) and 20 days (bilateral project RILA 2003).
- R. Gras and D. Hernandez (Swiss Institute of Bioinformatics) visited the Symbiose project during one week in April.

9. Dissemination

9.1. Leadership within scientific community

9.1.1. First meeting dealing with the Bioinformatics platform of OUEST-genopole

The first meeting dealing with the Bioinformatics platform of OUEST-genopole held at Irisa, Rennes, on 18th and 19th September 2003. About a hundred biologists and computer scientists, mostly from western France attended this meeting.

9.1.2. European "Stressgenes" meeting

A meeting of the European project Stressgenes was organized at Irisa, from the 21 to the 23 May 2003. A vocational training session was organized (Y. Bastide and E. Kaboré were involved in the teaching sessions).

9.1.3. BioInfoOuest thematic-day conferences

The Symbiose project regularly organizes thematic-day conferences on bioinformatics subjects[15]. 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. Four thematic-day conferences were organized during the year 2002-2003. 13 talks were given in this framework. The themes were Classification (F. Brucker, P. Bertrand, B. Tallur, I.-C. Lerman); Motif finding (G. Thijs, L. Marsan, J. Nicolas); Protein threading (C. Geourgeon, J.-F. Gibrat, R. Andonov) and genetic networks (H. De Jong, H. Geiselman, A. Cornish-Bowden, C. Dillman). The next thematic-day will be held in December 2003. The theme will be Grid Computing (D. Lavenier, V. Breton, C. Blanchet, W. Saurin).

9.1.4. Symbiose Seminar

The Symbiose seminar is held on a weekly basis. 18 talks were given in this framework during the year 2002-2003. 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.

9.1.5. Fête de la science 2003

The Symbiose project was in charge of a stand in Rennes in the framework of the national manifestation "La fête de la science 2003". The stand was organized around the theme of motif finding, using suffix trees.

9.1.6. Conferences, meetings and tutorial organization

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

- Third annual international workshop on Bioinformatics and Computational Biology in conjunction with the international conference on High Performance Computing; to be held in Hyderabad, December 2003 (R. Andonov, Program Committee).
- CAp03: Conférence d'Apprentissage Francophone (F. Coste, Program Committee)
- ERSA'03: Engineering of Reconfigurable Systems and Algorithms, Las Vegas, Nevad, USA (D. Lavenier, Steering Committee)
- EGC'2003: Extraction et Gestion des Connaissances, Lyon (I.-C. Lerman).
- FPL'03: Field Programmable Logic conference, Lisbon, Portugal (D. Lavenier, Program Committee)

[15] http://www.irisa.fr/symbiose/seminaire.htm

- ICGI: International Colloquium Grammatical Inference (F. Coste, Steering Committee)
- PPAM'2003: Fifth international conference on Parallel Procession and Applied Mathematics, Czestochowa, Poland (D. Lavenier, Program Committee).
- SFC 2003: 10-èmes Rencontres de la Société Francophone de Classification, Neuchâtel (I.-C. Lerman).
- Substitutions généralisées, pavages et numération (Research in Teams), Marseille (A. Siegel).
- SympAAA'2003: Symposium en Architectures Nouvelles de machines et Adéquation Algorithmes Architectures, La Colle sur Loup (D. Lavenier, Conference Chair).

9.1.7. Journal board

The members of Symbiose take part in the animation of the following journals:

- *La Revue de Modulad* (I.-C. Lerman et B. Tallur, lecture board).
- *Mathématiques et Sciences Humaines, Mathematics and Social Sciences* (I.-C. Lerman, editorial board).
- *RO-Operations Research* (I.-C. Lerman, editorial board).
- *Traitement du Signal* (D. Lavenier, editorial board).

9.1.8. Miscellaneous administrative functions

- Board of the Société francophone de classification (I. C. Lerman).
- Referee of the Ph-D thesis of G. Parmentier, juin 2003 (D. Lavenier), chair of the jury of the Ph-D thesis of R. David (D. Lavenier), jury of two Ph-D thesis, B. Cherfaoui and R. Bekkouche, december 2003 (D. Lavenier), jury of two Ph-D thesis (J. Nicolas). Chair of the jury of the Ph-D thesis of R. Priam, october 2003,(I.-C. Lerman), referee of the Ph-D thesis of J. Azé, december 2003 (I.-C. Lerman).
- Scienfic comittee of the BIA (Biométrie et Intelligence Artificielle) department of Inra (J. Nicolas).
- Steering of comitee of ICGI, responsible of Gowachin, a benchmark website for the evaluation of a grammatical inference programs (http://www.irisa.fr/Gowachin/) (F. Coste).

9.2. Faculty teaching

The Symbiose project is actively involved in the University bioinformatics teaching program. Especially, M. Le Borgne is in charge (for the computer science department) of studying the gaps in the bioinformatics teachings, in relation with the department of biology at the University of Rennes 1. Similarly, D. Lavenier shares the responsibility of the 5th year degree "Génomique et Informatique", in the teaching program Vie-Santé of the University of Rennes 1. The original aspect of this degree lies in the recruiting of both biologists and computer scientists: hence, national students holding a license degree either in biology or in computer science can apply for the master degree. Many members of the Symbiose project are involved in the teaching.

A master degree program of bioinformatics, oriented towards research, was proposed to the ministry of education. It should start in September 2004.

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

1. DEA Génomique et Informatique. (D. Lavenier, J. Nicolas, I.-C. Lerman, F. Coste, B. Tallur)
2. DEA IFSIC. (I.-C. Lerman, H. Leroy)
3. DESS Maths Appliquées. (B. Tallur)
4. DIIC. (I.-C. Lerman, B. Tallur)
5. Maîtrise de biochimie. (D. Lavenier)
6. INSA Rennes. (D. Lavenier, J. Nicolas)
7. Formation continue Inra. (D. Lavenier, J. Nicolas)
8. Specialized trainings Ecole chercheurs sur le "Data Mining" (B. Tallur), journées européennes "Stressgene" (B. Tallur), DEA Modelisation et Calcul Intensif, libanese universy (H. Leroy).

9.3. Conference and workshop committees, invited conferences

9.3.1. Meetings

We attended the following meetings:

- CAP'03 (Conférence d'Apprentissage Francophone), Laval (D. Fredouille, F. Coste, I. Jacquemin).
- Colloque CNRS "Interactions protéine-protéine: l'approche bioinformatique", Ecole Polytechnique, Palaiseau (J. Nicolas, invited talk).
- ECCB'2003 European Conference on Computatopnal Biology (F. Coste, S. Guyetant, I. Jacquemin, A. Siegel, M. Haeussler).
- ECML'03 (European Conference on Machine Learning), Dubrovnik, Croatie (D. Fredouille).
- EGC'2003: Extraction et Gestion des Connaissances, 23-24 janvier 2003, Lyon (I.-C. Lerman).
- ERSA'03, Engineering of Reconfigurable Systems and Algorithms, Las Vegas, Nevada, USA (D. Lavenier).
- EWM (European Women in Mathematics) (A. Siegel).
- HiCOMB'03 (High Performance Computational Biology), held in conjunction with 17th IPDPS, Nice (R. Andonov).
- Healthgrid, Lyon, 16-17/01/2003 (H. Leroy, M. MacWing, D. Lavenier).
- Journees de bilan de la bioinformatique des genopoles, Lyon-Gerland (J. Nicolas, invited talk).
- RenPar'15, La Colle-sur-Loup, 2003 (F. Raimbault, D. Lavenier).
- Rencontre Substitutions généralisées, pavages et numération, Marseille, 2003 (A. Siegel).
- Roadef 2003, Avignon, France, 2003 (D. Lavenier, R. Andonov).
- SympAAA'2003 9ème Symposium en Architectures de Machines et Adéquation Algorithme Architecture, La Colle sur Loup, 2003 (S. Guyetant, M. Giraud, D. Lavenier).
- SFC 2003: 10-èmes Rencontres de la Société Francophone de Classification, Neuchâtel (I.-C. Lerman, B. Tallur).

9.3.2. Invitations

The Symbiose project supported the following scientific visits:

- Invited visiting scientist at the Institut für Informatik of Potsdam and at Max Planck Golm (Berlin), 21/06 to 05/07/03 (J. Nicolas).
- Expert mission in bioinformatics for the French Ministry of Foreign Affairs, California, 20 to 27/09/03 (J. Nicolas).
- St Etienne, 3 days (F. Coste, invited visit).
- Montpellier (LIRMM), 2 days (A. Siegel, invited visit).

10. Bibliography

Major publications by the team in recent years

[1] R. ANDONOV, S. BALEV, S. RAJOPADHYE, N. YANEV. *Optimal semi-oblique tiling.* in « SPAA'01: Proceedings of the Thirteenth annual ACM Symposium on Parallel Algorithms and Architectures », ACM Press, pages 153–162, Crete Island, Greece, 2001.

[2] C. BELLEANNÉE, J. NICOLAS, R. VORC'H. *Vers un démonstrateur adaptatif.* J. SALLANTIN, J.-J. SZCZECINIARZ, editors, in « Le concept de preuve à la lumière de l'intelligence artificielle », series Nouvelle Encyclopédie Diderot, Presses Universitaires de France, 1999.

[3] F. COSTE, D. FREDOUILLE. *Efficient ambiguity detection in C-NFA, a step toward inference of non deterministic automata.* in « ICGI 2000, Grammatical inference: algorithms and applications », A. L. OLIVEIRA, editor, pages 25-38, Lisbonne, 2000.

[4] C. DELAMARCHE, P. GUERDOUX-JAMET, R. GRAS, J. NICOLAS. *A symbolic-numeric approach to find patterns in genomes : Application to the translation initiation sites of E. coli.* in « Biochimie », volume 81, 1999.

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

[6] 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 », number 5, volume 12, 1998, pages 517-572.

[7] P. GUERDOUX-JAMET, D. LAVENIER. *SAMBA: Hardware Accelerator for Biological Sequence Comparison.* in « CABIOS », number 6, volume 13, 1997, pages 609-615.

[8] D. LAVENIER, J. PACHERIE. *Parallel Processing for Scanning Genomic Data-Bases.* in « ParCo'97 (International Conference on Parallel Computing) », Bonn, Germany, 1997.

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

[10] B. TALLUR, J. NICOLAS, A. FROGER, D. THOMAS, C. DELAMARCHE. *Sequence classification of water channels and related proteins in view of functional predictions.* in « Theoretical Chemistry Accounts », volume 101, 1999, pages 77-81.

Doctoral dissertations and "Habilitation" theses

[11] D. FREDOUILLE. *Inférence d'automates finis non déterministes, par gestion de l'ambiguité, en vue d'application en bioinformatique.* Ph. D. Thesis, University of Rennes I, France, 2003.

Articles in referred journals and book chapters

[12] R. ANDONOV, S. BALEV, S. RAJOPADHYE, N. YANEV. *Optimal semi-oblique tiling.* in « IEEE Transactions on Parallel and Distributed Systems », number 9, volume 14, 2003, pages 944-960.

[13] 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 », 2003, to appear.

[14] V. BERTHÉ, P. ARNOUX, A. SIEGEL. *Two-dimensional iterated morphisms and discrete planes.* in « Theoretical Computer Science », 2003, to appear.

[15] R. GRAS, D. HERNANDEZ, P. HERNANDEZ, N. ZANGGER, Y. MESCAM, J. FREY, O. MARTIN, J. NICOLAS, R. APPEL. *Cooperative metaheuristics for exploring proteomic data.* in « Artificial Intelligence Review », number 1, volume 20, 2003, pages 95-120, Special issue on Life Science and AI.

[16] I. LERMAN, P. PETER. *Indice probabiliste de vraisemblance du lien entre objets quelconques analyse comparative entre deux approches.* in « Revue de Statistique Appliquée », volume LI(1), 2003, pages 5-35.

[17] F. RAIMBAULT, D. LAVENIER. *Des machines reconfigurables orientées objet pour les applications spécifiques.* in « TSI », volume 22, 2003, pages 759-782.

[18] A. SIEGEL. *Représentation des systèmes dynamiques substitutifs non unimodulaires.* in « Ergod. Th. and Dynam. Sys. », volume 23, 2003, pages 1247-1273.

[19] N. YANEV, R. ANDONOV. *Parallel Divide and Conquer Approach for Solving the Protein Threading Problem.* in « Concurrency and Computation: Practice and Experience », volume HiCOMB'03 special issue, S. Aluru and D. Bader (editors), 2003, to appear.

Publications in Conferences and Workshops

[20] R. ANDONOV, D. LAVENIER, N. YANEV, P. VEBER. *Fragmentation de génomes bactériens : deux approches d'optimisation combinatoire.* in « Roadef 2003 », Avignon, 2003.

[21] K. BACHAR, I. LERMAN. *Étude d'un comportement paramétré de CAHCVR sur des données réelles en imagerie numérique.* in « Méthodes et Perspectives en Classification, Comptes rendus des 10-èmes Rencontres de la Société Francophone de Classification », Y. D. PRESSES ACADÉMIQUES NEUCHÂTEL, G. M. (EDITORS), editors, pages 63-66, 2003.

[22] C. BELLEANNÉE, O. RIDOUX. *Programmation logique et calcul. Présentation d'un enseignement de maîtrise d'informatique.* in « Programmation en logique avec contraintes- JFPLC 2003 », LAVOISIER, editor, pages 49–62, 2003.

[23] N. BEN ZACOUR, M. GAUTIER, R. ANDONOV, D. LAVENIER, P. VEBER, A. SOROKIN, Y. LE LOIR. *GENOFRAG: a software to design primers optimized for whole genome scaning by long-range PCR amplification. Application to the study of Staphylococcus aureus genome plasticity.* in « ECCB'2003 European Conference on Computational Biology, Paris », Paris, 2003.

[24] F. COSTE, D. FREDOUILLE. *Introduction de connaissances structurelles et langagières pour l'apprentissage d'automates.* in « CAp'03 Conférence d'Apprentissage AFIA », Laval, 2003.

[25] F. COSTE, D. FREDOUILLE. *Unambiguous automata inference by means of state-merging methods.* in « ECML'03 », Dubrovnik, 2003.

[26] A. FLOETER, J. NICOLAS, T. SCHAUB, J. SELBIG. *Threshold Extraction in Metabolite Concentration Data.* in « GCB'03 German Conference on Bioinformatics », Munchen, 2003.

[27] M. GIRAUD, D. LAVENIER. *Réalisation matérielle d'automates pondérés pour la recherche de motifs génomiques.* in « SympAAA'2003 9ème Symposium en Architectures de Machines et Adéquation Algorithme Architecture », La Colle sur Loup, 2003.

[28] S. GUYÉTANT, D. LAVENIER. *Filtrage de bases de données sur le prototype RDIS.* in « SympAAA'2003 9ème Symposium en Architectures de Machines et Adéquation Algorithme Architecture », La Colle sur Loup, 2003.

[29] I. JACQUEMIN, J. NICOLAS. *Prediction de ponts disulfures par langages de controle.* in « CAp'03 Conférence d'Apprentissage AFIA », Laval, 2003.

[30] D. LAVENIER, S. GUYÉTANT, S. DERRIEN, S. RUBINI. *A reconfigurable parallel disk system for filtering genomic banks.* in « ERSA'03, Engineering of Reconfigurable Systems and Algorithms », Las Vegas, Nevada, USA, 2003.

[31] D. LAVENIER, H. LEROY, M. MAC WING, R. ANDONOV, M. HURFIN, P. RAIPIN-PARVEDY, L. MOUCHARD, F. GUINAND. *GénoGRID: an experimental grid for genomic applications.* in « HealthGrid 2003 », Lyon, 2003.

[32] I. LERMAN, J. AZÉ. *Une mesure probabiliste contextuelle discriminante de qualité des règles d'association.* in « EGC 2003, Extraction des Connaissances et Apprentissage, RSTI série RIA-ECA-Vol 17- n°1-2-3/2003 », D. B. M-S. HACID, editor, pages 247-262, 2003.

[33] J. NICOLAS. *Modélisation syntaxique d'interactions et apprentissage automatique.* in « Colloque CNRS "Interactions protéine-protéine : l'approche bioinformatique" », Ecole Polytechnique, Palaiseau, 2003, communication invitée.

[34] F. RAIMBAULT, D. LAVENIER. *Compilation d'un langage orienté objet pour une exécution répartie dans un composant reconfigurable : le parallélisme de classe.* in « RenPar'15, 15ème Rencontres Francophones du Parallélisme », La Colle sur Loup, 2003.

[35] A. SIEGEL. *Spectral theory for dynamical systems arisen by substitutions.* in « European Women in Mathematics », Marseille, 2003.

[36] B. TALLUR. *Étude d'un comportement paramétré de CAHCVR sur des données réelles en imagerie numérique.* in « Méthodes et Perspectives en Classification, Comptes rendus des 10-èmes Rencontres de la Société Francophone de Classification », Y. D. PRESSES ACADÉMIQUES NEUCHÂTEL, G. M. (EDITORS), editors, pages 185-188, 2003.

[37] N. YANEV, R. ANDONOV. *Solving the Protein Threading Problem in Parallel.* in « Workshop on HiCOMB'03, held in conjunction with 17th IPDPS Nice, France », 2003.

[38] N. YANEV, R. ANDONOV. *Une approche programmation linéaire pour la reconnaissance de repliements de protéines.* in « Roadef 2003 », Avignon, 2003.

Internal Reports

[39] R. ANDONOV, S. BALEV, N. YANEV. *Protein Threading: From Mathematical Models to Parallel Implementations.* Technical report, number PI 1552, IRISA, 2003.

[40] R. ANDONOV, N. YANEV, D. LAVENIER, P. VEBER. *Combinatorial approaches for segmenting bacterium genomes.* Technical report, number RR-4853, INRIA, 2003, http://www.inria.fr/rrrt/rr-4853.html.

[41] F. COSTE, D. FREDOUILLE. *What is the search space for the inference of non deterministic, unambiguous and deterministic automata ?.* Technical report, number rapport technique 4907, INRIA, 2003, http://www.inria.fr/rrrt/rr-4907.html.

Miscellaneous

[42] A. FLOETER, J. KOPKA, T. SCHAUB, J. NICOLAS, J. SELBIG. *Finding combinatorial causal relationships in metabolite concentration data using decision tree heuristics.* ECCB'03 European Conference on Computational Biology (Poster), 2003.

[43] S. GUYÉTANT, D. LAVENIER. *Evaluation of anchoring scheme for fast DNA Sequence Alignment.* ECCB'2003 European Conference on Computational Biology (Poster), Paris, 2003.

[44] I. JACQUEMIN, J. NICOLAS. *Grammatical inference for disulfid bonds prediction within protein.* ECCB'03 European Conference on Computational Biology (Poster), 2003.

[45] H. LEROY. *Le projet GenoGrid.* Gign - Grilles pour la genomique- kick-off meeting, Lyon, 2003.

[46] J. NICOLAS. *Rapport de présentation de la bioinformatique pour Ouest-genopole.* Journees de bilan de la bioinformatique des genopoles, Lyon-Gerland, 2003.

Bibliography in notes

[47] A. ACHARYA, M. UYSAL, J. SALTZ. *Active Disks: Programming Model, Algorithms and Evaluation.* in « ASPLOS-VIII, San Jose, California », 1998.

[48] V. ADOUE. *Elaboration d'un logiciel d'explication de classes pour une classification de données génotypiques.* Stage de DESS CCI, Université de Rennes1, Irisa, 2003.

[49] F. ALMEIDA, R. ANDONOV, D. GONZALEZ, L. MORENO, V. POIRRIEZ, C. RODRIGUEZ. *Optimal tiling for the RNA base pairing problem.* in « SPAA'02: 14th ACM Symposium on Parallel Algorithms and Architectures », pages 173-182, Winnipeg, Canada, 2002.

[50] S. ALTSCHUL, W. GISH, W. MILLER, E. MYERS, D.J. LIPMAN. *Basic local alignment search tool.* in « J. Mol. Biol. », volume 215, 1990.

[51] S. ALTSCHUL, T. MADDEN, A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, D. LIPMAN. *Gapped Blast and PSI-Blast: a new generation of protein database search programs.* in « Nucleic Acids Research », number 17, volume 27, 1997.

[52] D. BENSON, I. KARSCH-MIZRACHI, D. LIPMAN, J. OSTELL, B. RAPP, D. WHEELER. *GenBank.* in « Nucleic Acids Research », number 1, volume 30, 2002.

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

[54] A. BRAZMA, I. JONASSEN, I. EIDHAMMER, D. GILBERT. *Efficient discovery of conserved patterns using a pattern graph.*. in « Cabios », number 13, 1997, pages 509-522.

[55] A. BRAZMA, I. JONASSEN, I. EIDHAMMER, D. GILBERT. *Approaches to the Automatic Discovery of Patterns in Biosequences.* in « Journal of Computational Biology », number 2, volume 5, 1998, pages 277-304.

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

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

[58] E. CHOW, T. HUNKAPILLER, J. PETERSON. *Biological Information Signal Processor.* in « ASAP », 1991, pages 144-160.

[59] J. COLLADO-VIDES. *A Transformational-Grammar Approach to the Study of The Regulation of Gene Expression.* in « J. Theor. Biol. », number 6, volume 13, 1989, pages 403-425.

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

[61] H. DE JONG, J. GEISELMANN, D. THIEFFRY. *On Growth, Form, and Computers.* Academic Pres, 2003, chapter Qualitative modeling and simulation of developmental regulatory networks, pages 109-143.

[62] H. DE JONG, M. PAGE. *Qualitative simulation of large and complex genetic regulatory systems.* in « Proceeding of the 14th European Conference on Artificial Intelligence, ECAI 2000 », IOS Press, W. HORN, editor, pages 141-145, Amsterdam, 2000.

[63] F. DENIS, A. LEMAY, A. TERLUTTE. *Learning Regular Languages using RFSA.* in « Proceedings of the 12th International Conference on Algorithmic Learning Theory, ALT'01 », pages 348-363, 2001.

[64] S. DONG, D. SEARLS. *Gene structure prediction by linguistic methods.* in « Genomics », volume 23, 1994, pages 540-551.

[65] N. FRIEDMAN, D. KOLLER. *Being Bayesian about Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks.* in « Machine Learning », volume 50, 2003, pages 95-126.

[66] E. GLEMET, J. CODANI. *LASSAP: a LArge Scale Sequence compArison Package,.* in « Cabios », number 2, volume 13, 1997, pages 137-143.

[67] 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.

[68] 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.

[69] T. HEAD. *Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviours.* in « Bull. Math. Biology », volume 49, 1987, pages 737-759.

[70] J. HENIKOFF, S. HENIKOFF. *BLOCKs database and its applications.* in « Methods Enzymol. », volume 266, 1996, pages 88-105.

[71] J. HUDAK, M. MCCLURE. *A comparative analysis of computational motif-detection methods.* in « Pacific Symposium of Biocomputing PSB 1999 », pages 138-139, 1999, http://www-smi.stanford.edu/projects/helix/psb99.

[72] B. IDMOND. *Mesures de similarité et d'entropie pour l'apprentissage d'automates classifieurs de protéines.* Rapport de DEA informatique, Université de Rennes 1, Irisa, 2003.

[73] N. JAMSHIDI, S. JEREMY, J. EDWARD, T. FAHLAND, G. CHURCH, B. PALSSON. *Dynamic simultion of the human red blood cell metabolic network..* in « Bioinformatics », volume 17, 2001, pages 286-287.

[74] L. KARI, G. PAUN, G. ROZENBERG, A. SALOMAA, S. YU. *DNA computing, Sticker systems and universality.* in « Acta Informatica », volume 35, 1998, pages 401-420.

[75] S. KAUFFMAN. *The large scale structure and dynamics of gene control circuits: an ensemble approach.* in « Journal of Theorical biology », volume 44, 1974, pages 167.

[76] K. KEETON, D. A. PATTERSON, J. M. HELLERSTEIN. *A Case for Intelligent Disks (IDISKs).* in « SIGMOD Record », number 3, volume 27, 1998.

[77] V. KEICH, A. PEVZNER. *Findind motifs in the twilight zone.* in « Proceedings of RECOMB02 », ACM Press, pages 195-203, Washington, USA, 2002.

[78] C. KERMORVANT, C. HIGUERA (DE LA). *Learning languages with help.* in « Grammatical Inference: Algorithms and Applications, ICGI'02 », 2002, pages 161-173.

[79] K. J. LANG, B. A. PEARLMUTTER, R. A. PRICE. *Results of the Abbadingo One DFA Learning Competition and a New Evidence-Driven State Merging Algorithm.* in « Lecture Notes in Computer Science », volume 1433, 1998, pages 1–12.

[80] 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 », volume 262, 1993, pages 208-214.

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

[82] B. MA, J. TROMP, M. LI. *PatternHunter: Faster And More Sensitive Homology Search.* in « Bioinformatics », number 3, volume 18, 2002.

[83] A. MARIN, J. POTHIER, K. ZIMMERMANN, J.-F. GIBRAT. *FROST: A Filter Based Recognition Method.* in « Proteins: Struct. Funct. Genet. », volume 49, 2002.

[84] A. MARIN, J. POTHIER, K. ZIMMERMANN, J.-F. GIBRAT. *Protein structure prediction: bioinfromatic approach.* I. Tsigelny Ed. International University Line, 2002, chapter chapter Protein threading statistics: an attempt to assess the significance of a fold assignment to a sequence.

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

[86] 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. », volume 48, 1970, pages 443-453.

[87] O. PAREIGE. *Analyse synthaxique de génomes.* Stage de DEA Informatique, Université de Rennes1, Irisa, 2003.

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

[89] E. PECOU. *Qualitative dynamics of metabolic pathways and their genetic regulation.* Technical report, number RR 341, Institut de Mathématiques de Bourgogne, 2003.

[90] J. PLEY, R. ANDONOV, J.-F. GIBRAT, A. MARIN, V. POIRRIEZ. *Parallélisation d'une méthode de reconnaissance de repliements de protéines.* JOBIM 2002 Journées ouvertes en biologie, informatique et mathématiques (Poster), St Malo, 2002.

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

[92] 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 », series 1075, Springer-Verlag, Berlin, D. S. HIRSCHBERG, E. W. MYERS, editors, pages 186-208, Laguna Beach, CA, 1996.

[93] Y. SAKAKIBARA. *Recent advances of grammatical inference.* in « Theoretical Computer Science », volume 185, 1997, pages 15-45.

[94] D. B. SEARLS. *String Variable Grammar: A Logic Grammar Formalism for the Biological Language of DNA.* in « Journal of Logic Programming », number 1/2, volume 24, 1995, pages 73-102.

[95] D. SEARLS. *Formal language theory and biological macromolecules.* in « Theoretical Computer Science », volume 47, 1999, pages 117-140.

[96] T. SMITH, M. WATERMAN. *Identification of common molecular subsequences.* in « J. Mol. Biol. », number 147, 198, pages 195-197.

[97] E. SNOUSSI. *Necessary conditions for multistationnarity and stable periodicity.* in « J. Biol. Syst. », volume 6, 1998, pages 1-23.

[98] D. STATES, W. GISH, S. ALTSCHUL. *Basic local alignment search tool,.* in « J. Mol. Biol. », volume 215, 1990, pages 403-410.

[99] B. STEURICH, G. BORN. *Siemens "System-on-Silicon" Integrated hard drive Hard-disk for Controller to Gigabyte HD Drives.* in « Siemens Tools Partners », 2000.

[100] Y. TAKADA. *Learning formal languages based on control sets.* in « Lecture notes in AI », 1994.

[101] 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.

[102] S. TEMPEL. *Etude systématique de la structure et de la répartition d'un nouvel élément transposable à l'étude d'un génome entier par analyse de signatures.* Stage de DEA Génomique et Informatique, Université de Rennes1, Irisa, 2003.

[103] 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 », pages 504-509, 1991.

[104] N. YANEV, R. ANDONOV. *The protein threading problem is in P ?.* RR, number 4577, Inria, 2002, http://www.inria.fr/rrrt/rr-4577.html.

[105] 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 », pages 38-45, 1995.