INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team symbiose

SYstèmes et Modèles BIOlogiques, BIOinformatique et SEquences

Rennes

THEME BIO

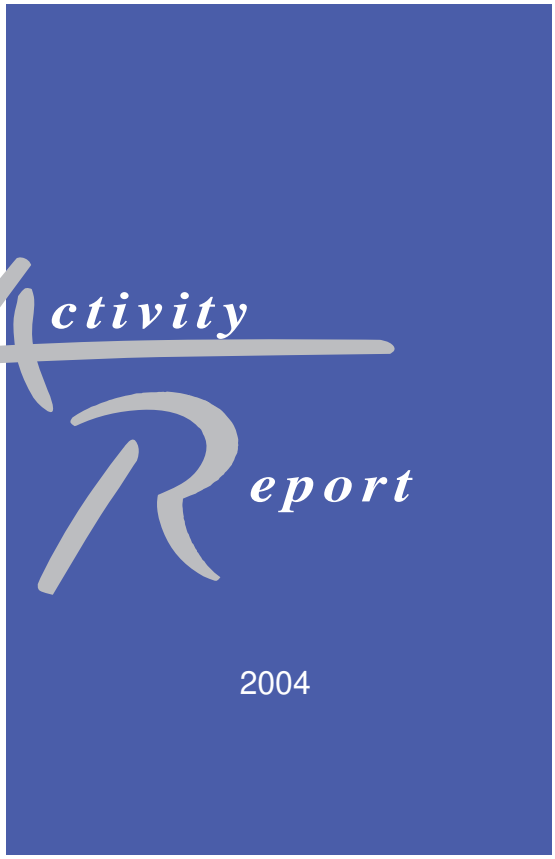

Table of contents

1. Team

The Symbiose project has been created in 2002. Its general purpose concerns 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. It is linked to Inria Research thema: Biological Systems.

Head of project
Jacques Nicolas [CR Inria]

Administrative assistant
Marie-Noëlle Georgault [AA Inria]

Inria staff members
François Coste [CR Inria]

CNRS staff members
Dominique Lavenier [DR CNRS]
Anne Siegel [CR CNRS]

Faculty members
Rumen Andonov [Prof., univ. Rennes 1 (since September 04, partner from univ. Valenciennes before)]
Catherine Belleannée [MC, univ. Rennes 1]
Michel Le Borgne [MC, univ. de 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)
Yves Bastide [Assistant professor, Ensar, Rennes]
Daniel Fredouille [Post-Doc, Aberdeen (since March 2004 teaching assistant in Rennes before)]
Frédéric Raimbault [MC, univ. Bretagne Sud (on secondment to CNRS until August 04)]
Stéphane Rubini [MC, univ. Bretagne Ouest]

Post-doctoral fellow
Xianyang Jiang [Post-doctoral fellow (Inria) since October 04]

Ph. D. students
Andre Floëter [Ph. D. student (cotutored Potsdam univ.)]
Mathieu Giraud [Ph. D. student (AMC)]
Stéphane Guyetant [Ph. D. student BDI CNRS/Région]
Goulven Kerbellec [Ph. D. student Inria/Région, since October 04]
Ingrid Jacquemin [Ph. D. student MENRT, teaching assistant since October 04]
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]
Philippe Veber [Ph. D. student Inria, since October 04]

Technical staff members
Patrick Durand [Senior research staff (Inria contract)]
Esther Kaboré [Project technical staff (Inria, Inria/Région contract genopole)]
Hugues Leroy [Engineer Inria, at 50%]
Michel Mac Wing [Project technical staff (Inria, ACI GénoGrid) until February 2004]
Emmanuelle Morin [Project technical staff (Inria, Inria contract genopole)]
Gregory Ranchy [Project technical staff (Inria) since August 04]
Elodie Retout [Project technical staff (Inra, at 20%, national program Inra/Sigenae)]

Anne-Sophie Valin [Project technical staff, (Inria, Inria contract genopole)]

Visiting scientist

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

Graduate student interns

Anurag Adarsh [Engineer, Kumpur, India /M. Le Borgne]
Romaric Gaudel [ENS Cachan Bretagne / F. Coste]
Goulven Kerbellec [DEA Génomique et Informatique / F. Coste]
Loubdna Khiar [DESA Diplômes d'Etudes Supérieures Approfondies, Ensa, Tanger, Maroc / D. Lavenier]
Ajith Jinjil [IBAB Engeneer, Bangalore, India /J. Nicolas]
Julien Ouy [DEA informatique /A. Siegel]
Ronan Trepos [DEA informatique / F. Coste]
Minh Quang Vo [DEA Génomique et Informatique / M. Le Borgne]

2. Overall Objectives

2.1. A project in Bioinformatics

We are interested in two types of data: sequences (DNA, 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 complex pattern filtering 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. Our approach is based on the definition of object oriented models of biological networks and the derivation of discrete or differential models for explaining and predicting (in a broad meaning) the behavior of the biological system 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, competing 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 [75]) :

- *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 consensual formats of recommandations 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, artifacts 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

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

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 [62], 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 [83] 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 [69]. 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 [91][92] ;
- 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 [53]. D. Searls proposed a more systematic approach based on logical grammars and a parser, Genlang [56]. 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 [46] or [65]. 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 [63]. 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 [74], 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 [45], using principles very close to those found in the work of M.-F. Sagot and A. Viari [89]. Another track explores variations on the search for cliques in a graph [72][48].

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 [58].

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 [90]. 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 [68]. 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 [98].

Boolean electronic circuits inspired one of the oldest dynamical models [70]: 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 [107]. They have proved to be good at studying dynamical properties such as stationary states or limit cycles [95], and allow the analysis and simulation of genetic networks of about 30 genes [108][106].

[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 [57] 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 [84]. 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

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

Mixing parallelism and genomics is both motivated by the large volume of data to handle and by the complexity of certain algorithms. First, there are data coming from intensive genome sequencing. Today (by the end of 2004), more than 220 genomes – including the human genome – are completely sequenced, and there exist more than 1000 other sequencing projects (see *Genomes onlines database*[5]). All these data are

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

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

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

The first algorithms for comparing genomic sequences have been developed in the seventies. They were essentially based on dynamic programming technics [81][94]. Then, with the increasing growth of data, faster algorithms have been designed to drastically speed-up the search. The Blast software [96] 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 [59] 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 [8], BISP [51], HSCAN [60] or BioScan [101], 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 [44]. 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 [75]. 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

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

All our developments are progressively made available within the bioinformatics platform of Ouest-Genopole. This platform has a strategy role in the genopole, offering access to various softwares and databases. We propose original tools for complex filtering of sequences. This includes Genofrag for PCR Scanning, STAN and ModelDesigner for pattern matching, and a set of pattern discovery algorithms. A first version of a graphical analyser for regulatory and metabolic networks is also available.

5.2. Bioinformatics Toolbox

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

The toolbox groups together accesses to standard tools (e.g. GCG package) and adapted softwares tailored to biologists needs collected in Ouest-genopole. One of the most recurrent demand is the possibility to make a Blast against a personal bank. This tool allows to perform a more relevant and faster search in this context. Access: [9]. The main activity concerns the generation of primers.

5.2.1. Specific primers

CAPS Tags. CAPS means Cleaved Amplified Polymorphic Sequence. The goal of this tool is to hightlight differences between two related sequences. First, we virtually digest the two sequences with Emboss restrict program, secondly we align them with Multalign. We display single enzyme cuts, taking into account the gaps appeared in the alignment. Differences are validated with the alignment, in this case a difference is a potential SNP. Access: [10].

Degenerate primers. A way to look for new genes is to use degenerate primers. Data are a set of protein sequences, from different species, with the same biological function. We align this set of sequences with Multalign. We extract from the calculated consensus sequence longest fragments with few ambiguous amino acids. After manual validation of one or several fragments, we degenerate each fragment from the 3' end. We have developed a module, working with degenerate alphabet and codon usage tables, who reverse translate protein sequences in nucleic sequences, computing and bounding a degeneration cost. Access: [11].

Microsatellite primers. Microsatellites are shorty repeated sequences that are primers markers in genome mapping. Data are a set of nucleic sequences in Fasta format. We use Sputnik to find microsatellites of chosen

[9] http://genouest.org/Services/
[10] http://genouest.org/Services/Amorces/CAPS.php
[11] http://genouest.org/Services/Amorces/genAm1.php

length in these sequences. Then we try to design PCR primers in the sequences containing a microsatellite with primer3. Access: [12].

5.2.2. GenoFrag

The goal of GenoFrag is to deal with Whole Genome PCR Scanning (WGPS), a means for analyzing bacterial genome plasticity. This software is developed for the design of optimized primers for Long-Range PCR on whole genomes. GenoFrag initially seeks all the potential primers on a chromosome. Then it calculates the best distribution of the primer pairs, thanks to combinatorial optimization algorithms. It was tested on *Staphylococcus aureus* strains but can be used for other bacterial or viral species [14][21]. A graphical interface is present on the Ouest-genopole bioinformatics platform server [13]. 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. Correspondant: Dominique Lavenier.

5.3. Tools for Databases

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

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

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

We have also set up the BioArray Software Environnement (BASE) which is a comprehensive database server to manage the massive amounts of data generated by microarray analysis. Access: http://idefix.univ-rennes1.fr:8080/www-base.

5.4. Pattern matching

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

Four pattern matching algorithms are available on the bioinformatics platform server. Two of them allow complex requests on complete genomes, STAN (Suffix Tree ANalyser) and WAPAM (Weighted Automata PAttern Matching).

STAN is based on a suffix tree data structure, and the patterns are represented in the form of a grammar. WAPAM is a tool to parse for proteic patterns expressed by weighted automata. Proteic databanks like Swiss-Prot or TrEMBL can be parsed too.

In both cases, the input patterns can be more complex than the usual regular patterns with the PROSITE syntax. Errors (substitutions and indels) are allowed. The biologist scientists are thus able to define precise signatures of biological functions.

[12] http://genouest.org/Services/Amorces/microSat.php
[13] http://idefix.univ-rennes1.fr:8080/Serveur-GPO/Services/GenoFrag/index.php?phpLang=en

The implementation programming languages are OCaml, C, Prolog, Python, PHP and JavaScript. The platform is available for all laboratories in OUEST-genopole. It should be opened soon to a larger public. Access: [14].

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

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

The entire platform is written using Java-based technologies: Java 2 Standard Edition for the client modules, and Java 2 Enterprise Edition for the server. The ModelDesigner server runs under an Open-Source Java Application Server (Tomcat from the Apache Software Foundation). ModelDesigner server module uses a proprietary Sicstus Prolog server.

ModelDesigner enters the beta-tests process within the Symbiose team and could be available for all laboratories in OUEST-genopole on 2005.

5.5. Pattern discovery

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

A Web platform grouping six pattern discovery algorithms is available. Access: [15]. It allows a more reliable and faster pattern discovery process 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 and validation of results, we propose a a toolbox with various modules: pattern matching in public databanks, visualization, statistical analysis, filtering.

The implementation programming languages are Python, PHP and JavaScript.

The platform is available for all laboratories in OUEST-genopole. It should be opened soon to a larger public.

5.6. Metabolic and genetic networks

Participants: Michel Le Borgne [correspondant], Anne Siegel, Anurag Adarsh.

For supporting our research in modeling and analysis of dynamical systems of biological metabolic and genetic interactions, a software has been developed and is still under development. This software, 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.

Models are specified using a declarative language. The user has to enumerate 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

[14] http://idefix.univ-rennes1.fr:8080/PatternMatching/
[15] http://idefix.univ-rennes1.fr:8080/PatternDiscovery/

and a graph generator builds a graph representation of the biological networks. Various tools allow for a flexible display of subnetworks of interactions. Types of biological interaction include signal pathways and gene expressions. The explanatory part of the software has been improved and allows one to extract from DNA-chips data the subset that do not fit with the model.

The generation of simulation programs for Matlab is a new feature of GARMEN. Informations are added to the description of the biochemical network from which GARMEN generates a Matlab program implementing the rate laws of the reactions and the stoechiometric matrix [41]. This is the first step in the realization of a software platform allowing for the study of a biological network from various point of view. Following the same idea, the generation of SIGNAL code (a language developed at IRISA) in order to model the combinatorics of gene/signal networks is under study.

Technical improvements have also been made on GARMEN in order to increase its portability, an important issue for a future distribution. This software is available on demand to the correspondant.

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/her model with respect to a set of sequences and to find new candidates in public sequence data banks.

The second situation concerns a biologist wishing a model for his family of interest. Our purpose is then to infer a model from sequences. More specifically, the goal of our research is to prove in the context of molecular biology the tractability of recognition and discovery of languages representing complex signatures.

6.1.1. Analysis by logical grammars

Participants: Jacques Nicolas, Catherine Belleannée, Patrick Durand, Mathieu Giraud, Emmanuelle Morin, Gregory Ranchy, Élodie Retout, Sébastien Tempel, Anne-Sophie Valin, Raoul Vorc'h.

We study the modeling of sequences with logical grammars in the line of Searls' work. We have specified a language, Logol, including string variables and constraints on strings.

Our objective is to make this logical grammar formalism accessible to the biologist, so that with minimum training he can design and test his own models. Because the biologist is usually not familiar with grammars, this supposes the design of graphical models which are translated in terms of logical grammars. We began the conception of an additional graphical interface to show parsing results inside the initial model, that is, to show graphically how the model matches each sequence [55]. These works also require to adapt expressiveness to biological specificities – to deal with helix structures for instance. Consequently, the design of models for DNA, RNA or proteins needs specific expressions even if the underlying analyzer remains the same. A first implementation of the graphical environement of Logol has been made available this year and is described in Section 5.5

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 suffix trees. This leads to two restricted tools for analysis, STAN and WAPAM, able of treating Prosite expressions and elementary repetitions with substitution costs. These are also described in Section 5.5. A first application has been the discovery of a gene in the mouse genome [38].

The software STAN (Suffix Tree ANalyzer) was used on the whole genome Arabidopsis thaliana (collaboration with UMR 6553) [6] for a systematical analysis of a family of transposons: the AtREP3 helitrons [39]. We proposed a new definition of domain and found twenty five internal domains in AtREP3. A survey of the complete genome of Arabidopsis thaliana was carried out to identify the biological function of all domains [39]. Some domains are characterized as minisatellite [93] or some associated domains as MITE-like

[49]. Some others domains possess a helitronic structure. The additional sequences that were found in some AtREP3 have been identified as a helitronic sequence (AtREP2 or AtREP20 [39]) or a MITE-like element. These experiments showed that AtREP3 is the target of nesting events (insertion/deletion) involving other transposable elements [33].

WAPAM and pattern discovery softwares were used on the dog and rat genomes (collaboration with UMR 6061). The olfactory receptors (OR) are genes devoted to the recognition of particular molecular substances. Biologists previously known 639 ORs located inside a 1.5x assembly [87]. In 2003, a 7x shotgun was conducted on the dog, but the first draft of the new assembly was only published in August 2004. In order to prevent the biologists for waiting the complete assembly, we developed a method which aims to directly analyze the sequenced runs. A pattern discovery step allowed to discover relevant patterns for OR and a very small subset of the runs was selected with the WAPAM tool, which keeps the sequences presenting the patterns expressed by weighted automata [27]. This subset was assembled using the CAP3 software [64], then further processed and cleaned. More than 400 new ORs were discovered and are further investigated by the biologists. This method allowed to spare the global assembly time while producing more sensitive results. Perspectives concern the conception of a tailored assembling algorithm.

6.1.2. Grammatical Inference

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

Our work concerns as well fundamental results or applied results in the field of bioinformatics.

6.1.2.1. Fundamental results

Searching for smallest consistent deterministic automata. We have proposed a new heuristic in the state merging algorithmic framework for the classical (NP complete) problem of the search of the smallest consistent deterministic finite state automata. This heuristic is based on minimizing the risk involved in making merges and allows a dramatic improvement in the size of the resulting automata. Whereas state of the art algorithms are mainly greedy algorithms, we have also studied backtrack search strategies in order to extend the visited search space. One of the result is a new heuristic limitation of the set of candidates after a backtrack, allowing to introduce more diversity in the search [20].

Introduction of background knowledge. 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 a better interpretation of the automaton to 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. It enables to exclude a (possibly infinite) set of strings from the inferred language.or to specify language including the target language; the second one reflects more semantic contraints. The typing semantics is integrated as constraints on the structure of the inferred automata. Our work as been to enable to consider non trivial typing functions, extending both existing formalism [73] and the understanding of the notion of typing.

Margin criteria. The introduction of classifiers automata [54] allows to shift from a characterization paradigm to a discrimination one. In that case, the size of the automata may be a wrong criterion with respect to the applications. We have proposed two new criteria, related to boosting and to support vector machines margins, for the inference of classifiers automata [99].

Complexity of context-free grammar inference. Context-free grammar inference is a challenging task along the way of improving the expressiveness of learned models. In order to compare different grammatical inference algorithms and to gain insight into the current state-of-the-art of context-free grammatical inference algorithms, we have designed a new measure of the complexity of inferring context-free grammars, used to rank target grammars in the Omphalos competition [31].

6.1.2.2. Characterization of genomic sequences

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.

Localization of covarying sites We focus on the localization of interacting pairs of sites. We assume that if a mutation occurs in one of the sites, either this mutation preserve the physico-chemical characteristics necessary to the interaction, or another mutation takes place in the other site to re-establish these proprieties. This process leads to statistically measurable correlations between sites and we proposed a new measure to estimate these correlations. We based our approach on the MoDEL algorithm [7] which searches a best candidate with a metaheuristic approach through all the space of possible motifs. We extended this algorithm to allow multiple sites to be localized, getting similar performances to those of the Gibbs motif sampler. This method shows good results on artificial sequences constrained to exhivit "natural" features [37][16]. A cooperation with Pasteur Institute is starting to test our approach on real biological sequences, focused on the protein aggregation problem.

Inference of automata on protein sequences We have developed our work on similarity based heuristics based on protein subsequences to learn characteristic automata of protein families. It proceeds from positive examples and can take advantage of the presence of negative examples (sequences not belonging to the family). A new generalization procedure based on the detection of important physico-chemical properties of the amino acids has been added. The application of this method has allowed us to clearly separate two subfamilies of the MIP proteins family (asserted by cross validation), even when increasing the precision of the models (automata)[25]. We are developing a more fundamental approach for taking into account the physico-chemical properties of amino acids. In our new merging scheme, working on transitions instead of states, we introduce a partial order on the alphabet corresponding to a lattice structure built on the set of amino acids. We propose a new score taking into account the dependence between amino acids and their context for the comparison of protein sequences. This score allows to characterize protein sequences containing "gapped" patterns. A ph-D thesis will be shortly available on this subject.

Learning Control Languages Motivated by the issue of predicting cysteins bonds within proteins, we have studied the application of the control language framework proposed by Takada [97]. 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 regular grammatical inference. Preliminary experiments have been made, studying various classical regular inference algorithms [67][66]. Our main conclusion is that one must design very carefully the description of instances. We have thus considered using the logic inductive programming approach to characterize the neighborhood of cysteins involved in a disulfide bond. We use Progol [80] on a sequence of windows extracted around the cysteins. The system generalizes these examples and produces an explicit, general rule, which can be used to identify future examples.

6.2. Gene expression data: analyzing data and modeling interactions

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

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

6.2.1. Classification

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

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

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, an additional contiguity constraint must be satisfied between merged

clusters. In collaboration with K. Bachar (ESSCA, Angers), we have set up a new hierarchical classification algorithm CAHCVR (Classification Ascendante Hiérarchique sous Contraintes utilisant les Voisins Réciproques). Two types of criteria for class association have been studied from both theoretical and experimental points of view, the classical inertia variation (Ward criterion) and a parametrized family of the criteria of the LLA. We have established that the computing complexity of our algorithm is linear (in average) with respect to the size of the object set, providing an accurate analysis and proof of this linearity. The slope of the linear increase is lowered by adopting an original strategy of multiple aggregation, instead of a binary one. We have assessed our approach in the framework of quick ascendant hierarchical classification method [23][34].

Quality of association rules in Data Mining. One fundamental objective in Data Mining consists in defining rule-relevant measures. Relative to a rule (implication) A->B, such implication index (measure) evaluates in a certain way the propensity of B, knowing A. A non symmetrical nature is required for this index. The LLA approach provides fruitful probabilistic indices for measuring the rule interest. However, a local definition depending solely on the rule to be evaluated becomes non discriminant for large data bases. In these conditions we propose a discriminant extension of the probabilistic indices obtained with respect to a set of potential interest rules. This work has been performed in collaboration with J. Azé of the LRI laboratory (Univ. Paris Sud) [28].

Clustering binary attribute data Binary attributes allow to describe many application domains. There exist a great number of similarity measures for such data in literature but most of them are 'biased' and are unfit to be used directly in clustering algorithms. We have developed a similarity index that has been derived starting from the the 'observation profiles' representation of data akin to the one used in correspondence analysis, and is based on the concept of correlation. We have also defined the corresponding cluster joining criterion and it has led to an elegant and efficient clustering method [32]. It has been used for the identification and characterization of common DNA sequences in a family (AtREP3) of transposons of *Arabidopsis thaliana* [39].

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 Ph-D 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; we have looked for a characterization of significant thresholds for metabolite variables, based upon a global analysis of Decision Forests. We have then applied the same kind of technics to discover combinatorial relations between metabolite variables. The method was able to confirm existing knowledge and to suggest interesting and new relationships. A Ph-D thesis will be shortly avalaible on this subject [15].

6.2.2. Modeling genetic networks inside metabolic or signaling pathways

Participants: Michel Le Borgne, Anne Siegel, Yves Bastide, Julien Ouy, Minh Quang Vo.

Biologist teams we work with are concerned with biological systems that are regulated by genetic process. First, the lipid metabolism in the liver of chicken is studied at the Animal Genetics Lab. (Inra, Rennes) in order to understand the genetic origin of fatting state. Second, the signaling of TGF-beta in liver cancer (a molecule with a major influence on the expansion of the fibrosis) is studied in the U456 Lab. (Inserm Rennes). In both systems, datasets provide informations on the simultaneous state of hundreds of molecules. If the system is first modeled thanks to a formal organization of the avalaible knowledge, then a dynamical model shall be built for each system. Finally, the comparison between the predictions of the model and the experimental datasets allows to validate the model or explain the data and to propose new relevant experimentations.

Formal modeling of biological system with genetic regulations. A first step to understand the dynamics of a biological system is to have a precise model of the reactions involved in it. Our models are based on the object approach, which is closest to the description of biological phenomena in the literature. However databases are usually centered around products involved in a mechanism, whereas our dynamical purpose implies to forget products and focus on interaction between products. The content of specialized biological literature related to regulation of metabolic and signaling process have been carefully studied. It was realized that two different types of interactions exist. The first class is related to behavioral relations between two molecules, such as the action of the increase of the quantity of a protein on the quantity of another protein. Such an interaction

is qualitative and usually give no information about the underlying process. The second class of interactions provides informations on the mechanism of action of a product on another one.

A database was created from specialized literature informations related to this classification of interactions (behavioral versus mechanistic). Its purpose is not to be a comprehensive database of interactions but a repository of knowledge. 250 papers were read about the regulation of lipid metabolism in liver and 1900 genetic interactions were extracted from these papers [52]; about 100 papers were read on the signaling of TGF-β, providing 350 interactions about this system [100].

A computer model is then developed 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 development of a language to describe models. First versions of such data model and language were defined this year.

Graphical models. The derivative of a graph model has been done and some simple analysis, based on graph theory, already implemented. This work has led to the development of a software tool, GARMeN (Graphical Analyzer for Regulation and Metabolic Networks) (see Section 5.6). A special attention was devoted to the analysis of DNA-chips datasets in relation with literature models. A method was developed to extract automatically datasets that appear to be new or in contradiction with the literature model, as well as data that indicate that a pathway was preferred to another pathway in the system. The method is to build an influence graph related to the literature model [82]. It was tested on a small model (about 15 genes) and we plan to validate it on the whole model obtained from the literature.

Differential models. 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. More precisely, to understand the influence of genetic regulations on lipid metabolism, a very simplified differential model for this system allows to reduce the 100 products initial model to a 7 variables model. Then, the literature knowledge is interpreted in terms of differential equations. A qualitative study of stationary states of this system states that genetic regulations allow a continuous shift of the stationary states of the metabolism so that the system appear to have no multistationarity.

Discrete models. Differential models do not fit properly with the known behavior of some process such as the signaling network in liver cancer; indeed, signaling deals more with a discrete process of information transfer than with a continuous variation of fluxes. A model specified with the help of the SIGNAL language, was derived from the graph model described above. We used an interpretation of the graph model as a set of conditions implying the absence or presence of signal molecules. Some dynamical properties of this model where exhibited, as the existence of competing paths explaining opposite effects of TGF-beta. The subnetwork driving these opposite actions was also analyzed and the role of various molecule in context determination was underlined. We will also compare properties derived from the model with experimental data, in particular those coming from DNA-chips.

6.3. Parallelism

Participants: Rumen Andonov, Dominique Lavenier, Frédéric Raimbault, Stéphane Guyetant, Mathieu Giraud, Xianyang Jiang, Hugues Leroy, Michel Mac Wing, Stéphane Rubini, Nicolas Yanev.

The parallelism axis mainly focuses on two activities: (1) the design of specialized parallel machines for scanning genomic banks in relation with axis 6.1 (2) the parallelization of protein threading algorithms, and their deployment on a grid.

6.3.1. Specialized architectures for scanning genomic banks

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

Genomic databases are growing exponentially. As an example, GenBank, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research 2004 Jan 1;32(1):23-6) contained 38G bases and 32M sequences in this version. GenBank is daily scrutinized by thousands of researchers.

BLAST [42][43] has steadily become the reference software for exploring genomic banks. Large databases can be quickly and easily screened to detect similarity with a query sequence. This type of algorithm, and many other algorithms such as PATTERNHUNTER [76] or CHAOS [47], proceed in two steps: first they seek for anchors, then they extend them into alignments. The load balancing between this two tasks depends on the quality of the anchors. Since the alignment extension can be time consuming, the goal is to limit the number of hits by providing anchors of good quality, i.e. reflecting a good probability of generating a significant alignment.

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

6.3.1.1. RDISK project: filtering genomic banks with reconfigurable disks

The central idea of the RDISK project is to directly filter 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 idea of attaching computation capabilities near the disk for providing on-the-fly data filtering is not new. SmartDisk [79], Active Disk [40] or IDISK [71] are examples of such investigation. All of them are motivated by a major trend: hard disk controllers are designed with an increasing amount of general purpose processing power and on-chip memory. Thus, filtering the data by pushing computation closer to the storage system is becoming an attractive solution for providing reduction in data movement through the I/O system.

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 [61], 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 and it is likely to have 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. The complete system is thus made of a front-end computer connected to a bunch of hard disks coupled to reconfigurable processing and interconnected through a local network – in our case an Ethernet network. Depending of the type of the query, an adequate hardware filter is first downloaded to the FPGA component before scanning the banks. The filtering occurs locally and results are send back to the front-end computer for further post-processing.

In 2004, a 48-node system as been assembled and successfully tested. Genomic applications range from similarity search to complex motif extractions based on regular expression. Performances of this low-cost parallel system depends on the applications, but they can be compared to a PC-cluster of tenth of nodes. As an example, when performing complex motif extraction, the RDISK system has shown performances equivalent to a 192 PC cluster [26][36][11].

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

Compared to the previous project, the ReMiX project goes one step further by addressing the data access problem. As we previously mentioned, scanning genomic banks is bounded by the time to read data from the disks. The idea, here, is not to duplicate disk accesses, but to propose a hardware mechanism allowing fast

random accesses to Gbytes of data. In that way, indexing techniques to access only a fraction of the bank become highly efficient.

In the ReMiX architecture, hard drives are replaced by FLASH memories whose access time are 2 or 3 orders of magnitude lower. In the same way, data bandwidth is increased by accessing simultaneously a large number of FLASH memories. Banks are *flashed* into the memories each time a new version occurs. As in the RDISK project, data are process on-the-fly by reconfigurable hardware directly connected to the memory.

This new memory architecture is far from being a simple memory extension: the reconfigurable index memory does not fit in the addressing space of the processor but it is indirectly accessed by specific queries and the reconfigurable index memory does not hold any cache hierarchy, and therefore memory accesses do not have to worry about the data locality.

In 2004, we have set up a Java programming environment based on a framework philosophy to develop and test indexing techniques. The software run on a parallel system (a cluster of PCs). It allows the users to rapidly test indexing hypotheses and to execute them on huge volume of data.

We also designed a 64-Gbytes FLASH memory PCI board and a 8 board system (512 Gbytes) is expected to be assembled soon (spring 2005). The FLASH memory components remain pin-to-pin compatible across generations, allowing our system to evolve with technology and needs: FLASH density, as genomic banks, is doubling every 12/18 months.

6.3.2. Protein 3D Structure Prediction via Threading

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

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 previous section showe the most straightforward way to infer a homology relationship between two proteins, by comparing their amino acid sequences. However, at large evolutionary distance, it can be difficult to detect any sequence similarity.

One of the alternative strategies consists of development of fold recognition (also called threading) methods. These methods are essentially based on two 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 and that we know a sizable fraction of all existing folds. The *protein threading problem* (PTP) 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 a NP-complete challenging optimization problem.

Currently we follow two complementary axis in this research domain. The first proposes that Mixed Integer Programming (MIP) models are very successful for solving the PTP problem [102][103] [12]. It significantly outperforms the state of the art in the domain. In addition, these results show that in practice the problem can be easier than in theory. 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. Our second research axis consists of deriving an efficient parallel algorithm with rare communications based on the splitting strategies [104][85] [19].

Next step will be a complete integration of our MIP model in the fold recognition package (FROST) developed by the researchers from Inra/MIG [78][77], and its deployment on a grid (GenoGRID project, see Section 8.2).

6.4. Other contributions

6.4.1. Comparative genomics of bacteria using LR-PCR

Participants: Rumen Andonov, Dominique Lavenier, Nicolas Yanev.

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

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 goal is to cover the genome of a reference strain with overlapping segments of nearly identical size, constrained by starting and ending-primers. Primers are short synthetic oligonucleotides used in PCR. 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), *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 [14]. Thus, the problem of segmenting a complete bacterial genome is reduced to split each domain into segments of nearly identical size. Along a domain, there are specific positions (i.e. small 25 DNA character string) corresponding to all possible primer sites. The overlapping segments can only start and end at these positions. If we assume that a solution is made of a list of N segments, and that each segment can take only P different positions, then the number of possibilities equals P^N. Finding the best one when N is large is clearly a combinatorial problem (N>100).

We have explored various approaches for solving this problem. Given a domain, i.e. a DNA sequence ranging from a few 100 Kpb to a few Mbp, together with all potential primer positions, we need to cover it with a sequence of overlapping segments of nearly identical size. Two cases have been considered. In the first one we search for a sequence of overlapping segments, each one of size in the interval $[\underline{L}, \overline{L}]$ and as close as possible to a *given* ideal size L. In the second case L is considered as *unknown* and we look for L^*, $\underline{L} \leq L^* \leq \overline{L}$, such that the best segmentation with respect to it is of minimal error. In both cases, we solved the problem by dedicated graph algorithms (see [21] for details), allowing a short computation time (1-2 minutes).

This research is an active collaboration with Y. Leloir and N. Ben Zacour from the Inra Ensar UMR 1055 microbiology laboratory, Rennes (see [14]). Implementation of the two algorithms have been performed and packaged into the GenoFrag software (see Section 5.2).

6.4.2. *Iterated morphisms*

Participant: Anne Siegel.

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

Iterated morphisms of the free monoid are very simple combinatorial objects which produce infinite sequences by replacing iteratively letters with words [86]. It naturally generates a minimal symbolic dynamical system that have many arithmetical, geometrical and dynamical properties. The Fibonacci morphism $\sigma(1) = 12$, $\sigma(2) = 1$ (related to the golden ratio) provided a good illustration of these different properties and the relations between them. In [22], relationships between this morphism, the addition of the golden ratio mod 1 and the multiplication of the golden ratio mod 1 are detailed.

In some specific case (unimodular morphism of Pisot type), iterated morphisms can be understood in a geometrical framework, thanks to the construction of a Rauzy fractal [17], that is, a self-similar compact subset of the Euclidean space [50]. A fundamental question about dynamical systems associated with an iterated morphism is whether they have a pure discrete spectrum. In [18], an algorithm and effective necessary and sufficient condition is given to answer this question.

From another point of view, Rauzy fractals may generate several classes of tilings. These tilings should be seen as generalizations of Thurston's tilings associated to a β-numeration system. In cite [30][29], we explain why the question of pure discrete spectrum is deeply related to properties of tilings of Rauzy fractals.

From the point of view of arithmetics, it is well known that real numbers with a purely periodic decimal expansion are the rationals having, when reduced, a denominator coprime with 10. This results extends to

beta-expansions with a Pisot base beta which is not necessarily a unit: real numbers having a purely periodic expansion in such a base are characterized thanks to Rauzy fractals [35].

In [13], 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.

7. Contracts and Grants with Industry

7.1.1. Defensins project

Participants: Yves Bastide, Yoann Mescam, Jacques Nicolas, Grégory Ranchy.

In collaboration with Germh laboratory, Symbiose team scanned human genome to search for new antimicrobial peptides known as *defensins*. These antimicrobial peptides have therapeutic applications. The Suffix Tree ANalyzer software (STAN) found 40 new defensins in the human genome. Symbiose and Innova Proteomics, a Rennes based biotechnology company, entered a collaboration to select and to validate the best predicted defensins. Currently, 13 putative defensins have been synthetized and tested. Rennes1 University has protected all putative defensins by submitting a new patent to the E.P.O.

8. Other Grants and Activities

8.1. Regional initiatives

8.1.1. OUEST-genopole

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

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

8.1.2. Bioinformatics Platform

Participants: Esther Kaboré, Hugues Leroy, 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*.

The bioinformatics platform is under the responsability of the Inria Symbiose project, and propose a complete set of tools and databases for biologists and bioinformaticians. The web site is http://genouest.org/. This platform received the national RIO label in december 2003.

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

8.1.3. Sigenae

Participants: Elodie Retout, Jacques Nicolas.

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

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

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 data extraction and analysis, and to contribute to the setting up of the new SIGENAE website.

This year, as part of the SIGENAE team, E. Retout was fastened to the various missions of development: tools to publish private sequences to public databases and to assess sequence quality in sequencing; tools related to data annotations; choosing a CMS (Contents Management System) to set up the new SIGENAE website.

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).
- Working group GiGn: "Grid for Genomics" of IMPG (D. Lavenier).
- National *contract Interface de la numération*, funded by the French ministry of research (Ministry Grant (ACI) Mathematical Interfaces program.
- National contracts GENOTO3D, ReMiX, GenoGRID, RDISK, MathResoGen, VICANNE. These contracts are detailed heraafter.

8.2.1. Project GENOTO3D

Participants: François Coste, Jacques Nicolas, Rumen Andonov, Ingrid Jacquemin, Aurélien Leroux, Yoann Mescam.

The goal of GENOTO3D is to develop and integrate machine learning approaches for the protein tertiary structure prediction task. It is a great challenge both for the difficulty of the task and for its applications in many fields (biology, genetics, drug design, etc.). An increasing number of structures are available in the Protein Data Bank PDB[16] which may be used by programs to predict the structure of a query protein sequence. The GENOTO3D project proposes to use numerical and symbolic machine learning approaches to predict long-term dependencies - which are still badly exploited by the classical prediction methods - and a divide-and-conquer strategy to integrate the different prediction levels in a single model.

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

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

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

Indexing is a well-known technique that accelerates searches within large volumes of data such as the ones needed by applications related to genomics. Very large indexes (larger than the main memory capacities) need to be stored on the hard disk drives. In that case, the design of indexes is concerned with low level notions such

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

as pages, fill-factors, tracks, cylinders, etc and indirectly impacts the search algorithms that navigate within the index.

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

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

Experimentation on this platform will be carried out with three application domains where huge volume of data are manipulated: genomic bank search, content-based image retrieval, and text information retrieval in heterogeneous XML knowledge databases [88].

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.

GénoGRID aims at experimenting with a grid of parallel computers for time-consuming genomic computations. 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.
2. *A transparent use of the resources* The grid is composed of several parallel computers (nodes) geographically dispatched in the western part of France and including the main genomic banks and softwares. 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* Three applications have been selected as a first validation of the grid: 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. Third one is related to the detection of repeat sequences inside full genomes.

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.

8.2.4. Project MathResoGen: Mathematical models for networks dynamics

Participants: Michel Le Borgne, Anne Siegel, Yves Bastide.

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

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

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

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

8.2.5. Project VicAnne: animation of community of biological networks

Participants: Michel Le Borgne, Anne Siegel, Yves Bastide.

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

8.3. 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:

- BIOMIS, ENS Cachan Rennes (B. Lepioufle): Bio-chips for cell electroporation (D. Lavenier)
- Ecole Nationale de la Santé Publique, Rennes: Identification of pathogen vibrio cholerae strain (D. Lavenier).
- Ensar-Rennes (G. Douaire): Ascendant hierarchical classification applied to image segmentation (I.-C. Lerman).
- Ensar-Inra Rennes - Laboratoire de Génétique Animale : Analysis of gene regulation involved in the lipid metabolism (Y. Bastide, M. Le Borgne, J. Nicolas, A. Siegel).
- GURIH: Micro-environnent cellulaire moléculaire, Medecine faculty, Rennes: Characterization and modelization of the TNF (Tumor Necrosis Factor) ligands and receptors families (F. Coste, G. Kerbellec).
- Irmar, Rennes : Mathematical modeling of lipogenesis (A. Siegel, M. Le Borgne).
- Inra Rennes - Technologie Laitière - Microbiologie : Study of Staphylococcus aureus genome plasticity (R. Andonov, D. Lavenier).
- Inra Nantes, Unité de Recherche sur les Protéines végétales et leurs Interactions: Prediction of disulfur bounds (J. Nicolas, I. Jacquemin).

- Inserm U625 GERHM Rennes : Human defensins (J. NIcolas, G. Ranchy)
- Inserm U456 (Détoxication et réparation tissulaire). Study of gene regulations in TGF-beta signalling in liver cancer (M. Le Borgne, A. Siegel).
- UMR-CNRS 6061 - Génétique et Développement : Olfactive receptors of dog and rat (M. Giraud, E. Morin, J. Nicolas, E. Retout, A.-S. Valin).
- UMR 6553 - EcoBio : Arabidopsis thaliana transposons (J. Nicolas, S. Tempel).
- UMR 6197 Laboratoire de microbiologie des environnements extrêmes Brest: Study for genomic diversity of virus and hyperthermophil plasmids (J. Nicolas, P. Durand)
- UMR-CNRS 6026 (Equipe récepteurs et canaux membranaires) : Study of the structure of MIP proteins (F. Coste, G. Kerbellec).

8.4. National collaborations

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

- ADAGE, Loria, Nancy (L. Noé, G. Kucherov): Sequence indexation (M. Giraud)
- CEA, Saclay (N. Ventroux): Reconfigurable computing (D. Lavenier)
- ESSCA, Angers (K. Bachar) : Ascendant hierarchical classification applied to image segmentation (I.-C. Lerman)
- EURISE, Univ. J. Monnet, St Etienne (C. de la Higuera, C. Kermorvant): Introduction of background knowledge in automata inference (F. Coste).
- IMB, Dijon (E. Pécou): Mathematical modeling of lipogenesis (A. Siegel, M. Le Borgne).
- Inserm E0018 (CHU, Angers): GenoGRID, massive genomic banks comparison (D. Lavenier)
- LIRMM, Montpellier (V. Berthé) and IML, Marseille (P. Arnoux): Substitutive dynamical systems (A. Siegel).
- LRI, Univ. Paris Sud (J. Azé): Validation of a probabilistic association rule (I.-C. Lerman)
- MIG, Inra, Jouy en Josas (J.-F. Gibrat, A. Marin): Protein threading, GenoGRID (R. Andonov, D. Lavenier).
- UMR 5171 Ifremer/CNRS-GPIA Montpellier: Discovery of antimicrobial peptides in oysters (J. Nicolas)

8.5. International Collaborations

- University of Geneva (SIB). Motif discovery with metaheuristics (Y. Mescam, J. Nicolas, R. Andonov and F. Coste).
- University of Sofia (Bulgaria). Exchange research program RILA'2003 (PAI) managed by the French Ministry of Foreign Affairs [17]. The project focusses on the application of combinatorial optimisation techniques in two different domains, Protein Threading and automata inference for discovering signatures of a sequence. Both domains are rich in NP-hard problems and the goal of the project is to propose and to analyze new mathematical models allowing to accelerate the solution of these problems (R. Andonov, J. Nicolas, F. Coste and D. Lavenier).
- Postdam university. Learning in metabolic pathway. A co-tutored Ph-D thesis started in 2002 and is almost completed.
- Department of Computer Science & AI, University of Malta. Searching for smallest consistent deterministic automata (F. Coste).

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

8.6. Visiting scientists

The following scientists visited the Symbiose project.

- Prof. Nicolas Yanev (Sofia University, Bulgaria).
- R. Gras, D. Hernandez and P. Hernandez (Swiss Institute of Bioinformatics).
- John Abela and Sandro Spina (University of Malta)

9. Dissemination

9.1. Leadership within scientific community

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

The second meeting dealing with the Bioinformatics platform of OUEST-genopole held at Irisa, Rennes, on 18th November 2004. This year, emphasis is on the new services proposed (complex filtering schemes for sequences), and representative of other genopole's bioinformatics platforms. A round table on "Which complementarity services between platforms?" will be helded at the end of the meeting.

9.1.2. BioInfoOuest thematic-day conferences

The Symbiose project regularly organizes thematic-day conferences on bioinformatics subjects[18]. The public of this thematic-day is made of computer scientists as well as biologists. Usually, this public gathers 50 persons (with 50 % of biologists) coming from all western France. Four thematic-day conferences were organized during the year 2002-2003. 13 talks were given in this framework. The themes were Computing Grids (D. Lavenier, V. Breton, C. Blanchet, W. Saurin); Proteomics (C. Pinault, H. Hondermack); Genome Structure (A. Viari, R. Grossi, A. Lefebvre).

9.1.3. Symbiose Seminar

The Symbiose seminar is held on a weekly basis. 24 talks were given in this framework during the year 2003-2004. 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.

9.1.4. Conferences, meetings and tutorial organization

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

- FPL : International Conference on Field Programmable Logic and Applications (D. Lavenier).
- FPT : International Conference on Field Programmable Technology (D. Lavenier).
- ERSA : International Conference on Engineering of Reconfigurable Systems and Algorithms (D. Lavenier).
- EGC'2004: Extraction et Gestion des Connaissances, Clermont-Ferrand (I.-C. Lerman).
- Workskop β-numeration, generalized substitution and tilings, Marseille (A. Siegel).
- Workshop *Méthodes pour l'analyse des réseaux de régulation génétiques et métaboliques*, Dijon (A. Siegel).
- ICGI: International Colloquium Grammatical Inference 2004 (F. Coste, J. Nicolas, Steering Committee)
- JOBIM'2004, Montréal, Canada, (J. Nicolas, Steering Committee).
- Omphalos Context-Free Grammar Learning Competition for ICGI'04 (7th International Colloquium on Grammatical Inference) (F. Coste, organizer).

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

9.1.5. Journal board

I.-C. Lerman takes part in the editorial board of the following journals: *La Revue de Modulad*, *Mathématiques et Sciences Humaines, Mathematics and Social Sciences*, *RO-Operations Research*.

9.1.6. Miscellaneous administrative functions

- Jury of the Habilitation-thesis of R. Gras, December 2004, Rennes (R. Andonov, J. Nicolas).
- Chair of the jury of the Ph-D thesis of G. Youness, Paris 6 university, July 2004 (I.-C. Lerman). Chair of the jury of the Ph-D thesis of D. Eveillard, May 2004, Nancy (J. Nicolas). Rewiever of the Ph-D thesis of F. Boyer, July 2004, Grenoble (J. Nicolas). Jury of R. Groult, Rouen, June 2004 (D. Lavenier). Jury of the PhD thesis of S. Guyetant, december 2004 (D. Lavenier).
- Organization of Omphalos Context-Free Grammar Learning Competition for ICGI'04 (7th International Colloquium on Grammatical Inference).

9.2. Faculty teaching

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

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

1. DEA Génomique et Informatique. (F. Coste, D. Lavenier, J. Nicolas, B. Tallur)
2. DEA IFSIC. (F. Coste, H. Leroy)
3. DESS MITIC. (B. Tallur)
4. DESS Mathématiques. (B. Tallur)
5. DEA de l'université libanaise (H. Leroy)
6. Maîtrise Bio BCP et Biochimie (M. Le Borgne, S. Tempel)
7. Formation permanente Inra. (D. Lavenier)
8. Formation permanente Cnrs (D. Lavenier)
9. Specialized trainings: Genostar, Brest (P. Durand); Bioinformatique dans l'industrie, ISSB, Angers (P. Durand); Systèmes dynamiques, Bordeaux (A. Siegel)

9.3. Conference and workshop committees, invited conferences

9.3.1. Meetings

We attended the following meetings:

- 12ème Colloque Elément Transposables, Tours, France.
- CIAA 2004: Ninth International Conference on Implementation and Application of Automata, Kingston, Canada.
- 2nd EDAA Ph.D. Forum, Paris, France.
- EGC 2004: Extraction et Gestion de Connaissances, Clermont-Ferrand.
- HiComb 2004: Third IEEE International Workshop on High Performance Computational Biology, Santa Fe, USA.
- ICGI 2004 (International Colloquium on Grammatical Inference), Athens.
- IFCS: Meeting of the International Federation of Classification Societies, Chicago.
- ISMB/ECCB 2004, Glasgow.
- Jobim 04, Montréal.
- Journée des doctorants en architecture et compilation" du GDR ARP, Antibes.
- Number Theoretic Algorithms and Related Topics, Strobl, Austria.
- Aperiodic Order: Dynamical Systems, Combinatorics and Operators, Banff, Canada.
- 11es Rencontres de la Société Francophone de Classification, Bordeaux.
- WATA 2004: Weighted Automata: Theory and Applications, Dresden, Germany.
- Workskop *β-numeration, generalized substitution and tilings*, Marseille.
- Réunion *Méthodes pour l'analyse des réseaux de régulation génétiques et métaboliques*, Dijon.

9.3.2. Invited conferences

- Université des Sciences et Technologies de Lille (LIFL), 25 mai (I.-C. Lerman).
- CEA, 20 juillet (S. Guyetant).
- Loria, Nancy, équipe ADAGE, juin 2004 (M. Giraud).
- 1er carrefour Ouest genopole, Rennes, janvier 2004 (J. Nicolas, *Mise en évidence de nouvelles défensines par une approche bioinformatique*).
- 2èmes rencontres autour de la plate-forme bioinformatique, Rennes, novembre 2004 (J. Nicolas, *Filtrages complexes de séquences*).
- Ecole nationale polytechnique d'Alger, Alger, 5-6 décembre 2004 (H. Leroy, *Challenges in grid computing*).

9.3.3. Invitations

The Symbiose project supported the following scientific visits:

- Invited visiting scientist at SIB, Geneve, (10 days, J. Nicolas).
- Montpellier (LIRMM), 7 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, Crete Island, Greece", ACM Press, 2001, p. 153–162.

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

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

[4] F. COSTE, D. FREDOUILLE, C. KERMORVANT, C. DE LA HIGUERA. *Introducing Domain and Typing Bias in Automata Inference*, in "ICGI'04 (7th International Colloquium on Grammatical Inference)", 2004.

[5] 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", vol. 81, 1999.

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

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

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

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

[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", vol. 101, 1999, p. 77-81.

Doctoral dissertations and Habilitation theses

[11] S. GUYETANT. *Architecture parallèle reconfigurable pour le filtrage de banques de données non structurées ; application à la génomique.*, Ph. D. Thesis, IRISA, 2004.

Articles in referred journals and book chapters

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

[13] P. ARNOUX, V. BERTHÉ, A. SIEGEL. *Two-dimensional iterated morphisms and discrete planes*, in "Theoretical Computer Science", vol. 319, 2004, p. 145–176.

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

[15] A. FLOETER, J. NICOLAS, T. SCHAUB, J. SELBIG. *Threshold extraction in metabolite concentration data*, in "Bioinformatics", vol. 20, 2004, p. 1491-1494.

[16] R. GRAS, D. HERNANDEZ, P. HERNANDEZ, N. ZANGGER, Y. MESCAM, J. FREY, O. MARTIN, J. NICOLAS, R. APPEL. *Artificial Intelligence Methods and Tools for Systems Biology*, Springer Series Computational Biology vol 5, chap. Cooperative metaheuristics for exploring proteomic data, W. Dubitzky and F. Azuaje Eds, 2004.

[17] A. SIEGEL. *Fractals a la carte*, in "Tangente", vol. Hors Série 18, 2004, p. 80–84.

[18] A. SIEGEL. *Pure discrete spectrum dynamical system and periodic tiling associated with a substitution*, in "Annales de l'Institut Fourier", vol. 54, n° 2, 2004, p. 288-299.

[19] N. YANEV, R. ANDONOV. *Parallel Divide&Conquer Approach for the Protein Threading Problem*, in "Concurrency and Computation:Practice and Experience", vol. 16, 2004, p. 1–14.

Publications in Conferences and Workshops

[20] J. ABELA, F. COSTE, S. SPINA. *Mutually compatible and incompatible merges for the search of the smallest consistent DFA.*, in "ICGI'04 (7th International Colloquium on Grammatical Inference)", 2004.

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

[22] P. ARNOUX, A. SIEGEL. *Dynamique du nombre d'or*, in "Université d'été Sciences mathématiques et modélisation, Bordeaux", To be published by DESCO, 2004.

[23] K. BACHAR, I. LERMAN. *Fixing Parameters in the Constrained Hierarchical Classification Method: Application to Digital Image Segmentation*, in "Classification, Clustering and Data Mining Applications", D. BANKS, A. EDITORS (editors)., Springer, 2004, p. 85-94.

[24] F. COSTE, D. FREDOUILLE, C. KERMORVANT, C. DE LA HIGUERA. *Introducing Domain and Typing Bias in Automata Inference*, in "ICGI'04 (7th International Colloquium on Grammatical Inference)", 2004.

[25] F. COSTE, G. KERBELLEC, B. IDMONT, D. FREDOUILLE, C. DELAMARCHE. *Apprentissage d'automates par fusions de paires de fragments significativement similaires et premières expérimentations sur les protéines MIP.*, in "JOBIM'04, Montreal", 2004.

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

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

[28] I. LERMAN, J. AZE. *Indice probabiliste discriminant (de vraisemblance du lien) d'une règle d'Association en cas de "très grosses données*, in "Mesures de Qualité pour la Fouille des Données, RNTI-E-1, Cépadues, Toulouse", H. BRIAND, R.-E.-1. AL. EDITORS (editors)., 2004, p. 69-94.

[29] A. SIEGEL. *Coverings associated to a beta-shift and different conditions for tilings*, in "Number Theoretic Algorithms and Related Topics, Strobl, Austria", 2004.

[30] A. SIEGEL. *Spectral theory of substitutive systems: combinatorial conditions for pure discrete spectrum and tilings*, in "Aperiodic Order: Dynamical Systems, Combinatorics and Operators, Banff, Canada", 2004.

[31] B. STARKIE, F. COSTE, M. VAN ZAANEN. *The Omphalos Context-Free Grammar Learning Competition*, in "Proceedings of the International Colloquium on Grammatical Inference (ICGI); Athens, Greece", 2004, p. 16–27.

[32] B. TALLUR. *Binary data clustering*, in "International Federation of Classification Societies Conference, Chicago, USA", 2004.

[33] S. TEMPEL. *Genome-wide analysis of domain organization in higher plant helitronic structures*, in "12ème Colloque Eléments Transposables, Tours", 2004.

Internal Reports

[34] I. LERMAN, K. BACHAR. *Construction et Justification d'une Méthode de Classification Ascendante Hiérarchique Accélérée Fondée sur le critère de la Vraisemblance du Lien en Cas de Données de Contiguïté. Application en Imagerie Numérique*, Technical report, n° PI 1616, Irisa, 2004.

Miscellaneous

[35] V. BERTHÉ, A. SIEGEL. *Purely Periodic beta-Expansions in the Pisot Non-unit Case*, submitted to Journal of Number Theory, 2004.

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

[37] R. GRAS. *Structure des espaces de recherche, complexité des algorithmes d'optimisation combinatoire stochastique et applications à la bioinformatique*, Habilitation à Diriger les Recherches, Université de Rennes 1, décembre 2004, Habilitation à Diriger les Recherches, Université de Rennes 1.

[38] D. SCHAUSI, V. VALLET-ERDTMANN, C. TIFFOCHE, G. TILLY, M. GUERROIS, B. JÉGOU, M. THIEULANT, J. NICOLAS, S. TEMPEL. *Regulation Of An Intronic Promoter Of Rat Estrogen Receptor Alpha Gene. Targeting Of Promoter To The Pituitary In Transgenic Mice*, Conférence Découverte des génomes et expression des gènes, Paris mai 2003 (Poster), 2004.

[39] S. TEMPEL, I. COUÉE, J. NICOLAS, I.-C. LERMAN, A. E. AMRANI. *Domain organization in non-autonomous helitrons*, 2004.

Bibliography in notes

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

[41] A. ADARSH. *JMS: A GARMeN extension for simulating metabolic networks in matlab*, Stage d'ingénieur, IRISA, 2004.

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

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

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

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

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

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

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

[49] F. C., C. MOUCHES. *Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon*, in "Mol Biol Evol.", vol. 17, 2000, p. 730-7.

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

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

[52] J. CLUCHAGUE, P. LHOMME. *Constitution d'une base de donnée sur les interactions impliquées dans le métabolisme hépatique des lipides*, Technical report, Inra, 2004.

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

[54] F. COSTE. *Apprentissage d'automates classifieurs en inférence grammaticale*, Ph. D. Thesis, IRISA, Université de Rennes 1, janvier 2000.

[55] 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.

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

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

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

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

[60] 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.

[61] 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.

[62] T. HEAD. *Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviours*, in "Bull. Math. Biology", vol. 49, 1987, p. 737-759.

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

[64] X. HUANG, A. MADAN. *CAP3: A DNA Sequence Assembly Program.*, in "Genome Research", vol. 9, 1999, p. 868-877.

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

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

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

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

[69] L. KARI, G. PAUN, G. ROZENBERG, A. SALOMAA, S. YU. *DNA computing, Sticker systems and universality*, in "Acta Informatica", vol. 35, 1998, p. 401-420.

[70] S. KAUFFMAN. *The large scale structure and dynamics of gene control circuits: an ensemble approach*, in "Journal of Theorical biology", vol. 44, 1974, 167.

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

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

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

[74] C. E. LAWRENCE, S. F. ALTSCHUL, M. S. BOGUSKI, J. S. LIU, A. F. NEUWALD, J. C. WOOTTON. *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.*, in "Science", vol. 262, 1993, p. 208-214.

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

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

[77] A. MARIN, J. POTHIER, K. ZIMMERMANN, J.-F. GIBRAT. *FROST: A Filter Based Recognition Method*, in "Proteins: Struct. Funct. Genet.", vol. 49, 2002.

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

[79] 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.

[80] S. MUGGLETON. *Inverse Entailment and Progol*, in "New Generation Computing, Special issue on Inductive Logic Programming", vol. 13, nº 3-4, 1995, p. 245-286.

[81] S. NEEDLEMAN, C. WUNSCH. *A general method applicable to the search of similarities in the amino acid sequences of two protein,*, in "J. Mol. Biol.", vol. 48, 1970, p. 443-453.

[82] J. OUY. *Modélisation de réseaux biologiques et application pour l'analyse de données expérimentales*, Stage de DEA Informatique, IFSIC, Université de Rennes1, Irisa, 2004.

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

[84] E. PECOU. *Qualitative dynamics of metabolic pathways and their genetic regulation*, Technical report, nº RR 341, Institut de Mathématiques de Bourgogne, 2003.

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

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

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

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

[89] M.-F. SAGOT, A. VIARI. *A Double Combinatorial Approach to Discovering Patterns in Biological Sequences*, in "Proceedings of the7th Annual Symposium on Combinatorial Pattern Matching, Laguna Beach, CA", D. S. HIRSCHBERG, E. W. MYERS (editors)., 1075, Springer-Verlag, Berlin, 1996, p. 186-208.

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

[91] D. B. SEARLS. *String Variable Grammar: A Logic Grammar Formalism for the Biological Language of DNA*, in "Journal of Logic Programming", vol. 24, nº 1/2, 1995, p. 73-102.

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

[93] C. SLAMOVITS, S. ROSSI. *Satellite DNA: Agent of chromosomal evolution in mamals. A review*, in "J. Neotrop. Mammal", vol. 9, 2002, p. 297-308.

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

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

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

[97] Y. TAKADA. *Learning formal languages based on control sets*, in "Lecture notes in AI", 1994.

[98] 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.

[99] R. TREPOS. *Apprentissage d'automates pour la discrimination*, Stage de DEA Informatique, IFSIC, Université de Rennes1, Irisa, 2004.

[100] M.-Q. VO. *Modélisation des réseaux de signalisation du facteur de croissance TGF-beta*, Stage de DEA Génomique et Informatique, Université de Rennes1, Irisa, 2004.

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

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

[103] 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.

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

[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", 1995, p. 38-45.

[106] H. DE JONG. *Modeling and Simulation of genetic Regulatory Systems: a Literature Review*, in "Journal of Computational Biology", vol. 9 (1), 2002, p. 69-105.

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

[108] 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, Amsterdam", W. HORN

(editor)., IOS Press, 2000, p. 141-145.