INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team modbio

Modèles Informatiques en Biologie Moléculaire

Lorraine

THEME 2A

Table of contents

1. Team

Head of project-team
Alexander Bockmayr [Professor, University Henri Poincaré, Nancy 1]

Vice-head of project-team
Eric Domenjoud [CR CNRS]

Administrative assistant
Sophie Drouot [INRIA]

Staff member Universities
Yann Guermeur [MC UHP, until 9/2003]

Staff member CNRS
Yann Guermeur [CR, since 10/2003]

PhD students
Arnaud Courtois [UHP, cofinanced by the Région Lorraine]
Yannick Darcy [Allocataire MERT, since 10/2003]
Damien Eveillard [UHP, cofinanced by the Région Lorraine]

Post-doctoral fellows
Emmanuel Gothié [INRIA, until 9/2003]
Sandrine Peyrefitte [CNRS, since 10/2003, in collaboration with MAEM (UMR 7567)]
Nicolai Pisaruk [UHP, until 3/2003]

Technical staff
Abdelhalim Larhlimi [CNRS, from 08/2003 to 11/2003]

Student interns
Stéphanie Billaut [Maîtrise MGMC, 1/03 - 3/03]
Emmanuel Didiot [DEA Informatique, 2/03 - 7/03]
Sumit Jha [IIT Kanpur, 3/03 - 07/03]
Myriam Vezain [DESS EGOISt, Rouen]

2. Overall Objectives

The aim of the project MODBIO is to develop computational models for molecular and cell biology. We are interested in two types of problems:

- Determining the structure of biological macromolecules,
- Discovering and understanding the function of biological systems.

We approach these questions by combining techniques from constraint programming, discrete optimization, hybrid systems, and statistical learning theory.

2.1.1. Research themes

- Determination and analysis of macromolecular envelopes
- RNA structure and alternative splicing
- Predicting protein secondary structure
- Modeling biological systems with constraint programming

2.1.2. Scientific and industrial relations

- Participation in the "Génopole Strasbourg Alsace-Lorraine"
- Participation in the Bioinformatics project of the Région Lorraine
- Participation in the ACI project GENOTO3D
- Participation in the ARC INRIA "Process calculi and molecular networks"
- Participation in the LISCOS project of the European Community
- Various national and international collaborations
  - Laboratoire « Maturation des ARN et Enzymologie Moléculaire », MAEM, Nancy
  - Laboratoire de Cristallographie, LCM3B, Nancy
  - Institut de Biologie et Chimie des Protéines, IBCP, Lyon
  - Institute of Mathematical Problems in Biology, Russian Academy of Sciences
  - University of California, Irvine, USA

3. Scientific Foundations

3.1. Constraint Programming

Constraint programming [33] is a declarative programming language paradigm that appeared in the late 80's and has become more and more popular since then. A *constraint* is a logical formula containing variables that defines a relation to be satisfied by the values of these variables. For instance, the formula $x + y \leq 1$ expresses that the sum of the values of the variables x and y must be less than or equal to 1.

In *constraint programming*, the user programs with constraints, which means that he describes a problem with a set of constraints which may be connected by various *combinators* like conjunction, disjunction, or temporal operators (`always`). Each constraint gives a *partial* information about the state of the system under study. Constraint programming systems allow one to deduce new constraints from the given ones and to compute *solutions*, i.e., values for the variables which satisfy all constraints simultaneously.

One of the main goals of constraint programming is to develop programming languages which allow one to express constraint problems in a natural way, and to solve them efficiently.

3.1.1. Finite domain constraint programming and discrete optimization

In our work, we are first interested in constraint problems over finite domains. The domain of each variable (the set of values it may take) is then a finite set of natural numbers. Theory tells us that most constraint problems over finite domains are NP-hard, which means that there is little hope to solve them by algorithms polynomial in the size of the problem. In practice, these problems are handled by tree search methods which try successively different valuations of the variables until a solution is found. Because of the exponential number of possible combinations, it is crucial to reduce the search space as much as possible, i.e., to eliminate *a priori* as many valuations as possible.

There exist two generic methods to solve such problems. The first one is classical *integer linear optimization* which has been studied in mathematical programming and operations research for more than 40 years. Here, constraints are linear equations and inequalities over the integer numbers. In order to reduce the search space, one typically uses the linear relaxation of the constraint set. Equations and inequalities are first solved over the real numbers, which is much easier. Then the information obtained is used to prune the search tree.

The second method is *finite domain constraint programming* which arose in the last 15 years by combining ideas from declarative programming languages and constraint satisfaction techniques in artificial intelligence. In contrast to integer linear optimization one uses, in addition to simple arithmetic constraints,

more complex constraints, which are called *symbolic constraints*. For instance, the symbolic constraint `alldifferent`$(x_1, ..., x_n)$ expresses that the values of the variables $x_1, ..., x_n$ must be pairwise distinct. Such a constraint is difficult to express in a compact way using only linear equations and inequalities. Symbolic constraints are handled individually by specific filtering algorithms that reduce the domain of the variables. This information is propagated to other constraints which may further reduce the domains.

A unifying framework for integer linear programming and finite domain constraint programming is presented in [12]. A systematic presentation of constraint solving techniques on numerical domains may be found in [2].

3.1.2. Concurrent constraint programming

In *concurrent* constraint programming (cc) [31], different computation processes may run concurrently. Interaction is possible via the *constraint store*. The store contains all the constraints currently known about the system. A process may *tell* the store a new constraint, or *ask* the store whether some constraint is entailed by the information currently available, in which case further action is taken.

Hybrid concurrent constraint programming (`Hybrid cc`) [28] is an extension of concurrent constraint programming which allows one to model and to simulate the temporal evolution of *hybrid systems*, i.e., systems that exhibit both discrete and continuous state changes. Constraints in `Hybrid cc` may be both algebraic and differential equations. State changes can be specified using the combinators of concurrent constraint programming and default logic.

3.2. Statistical learning

Statistical learning theory [35] is one of the fields of inferential statistics the bases of which have been established by V.N. Vapnik in the late 1960s. The goal of this theory is to specify the conditions under which it is possible to « learn » from empirical data obtained by random sampling. Learning amounts to solving a problem of model selection. More precisely, given a problem characterized by a joint probability distribution on couples made up of observations and labels, and a set of functions, of cardinality ordinarily infinite, the goal is to find in the set a function with optimal performance. Problems may belong to one of the three following areas: pattern recognition (discriminant analysis), function approximation (regression) and density estimation.

This theory considers more specifically two inductive principles. The first one, named empirical risk minimization (ERM) principle, consists in minimizing the training error. If the sample is small, one substitutes to this the structural risk minimization (SRM) principle. It consists in minimizing an upper bound on the expected risk (generalization error), a bound sometimes called a guaranteed risk. This latter principle is implemented in the training algorithms of the support vector machines (SVMs), which currently constitute the state-of-the-art for numerous problems of pattern recognition.

SVMs are connectionist models conceived to compute indicator functions, to perform regression or to estimate densities. They have been introduced during the last decade by Vapnik and co-workers [27], as nonlinear extensions of the maximal margin hyperplane [34]. Their main advantage is that they can avoid overfitting in the case where the size of the sample is small [35][25].

4. Application Domains

4.1. Panorama

The main application area of our work is molecular biology. At the same same, we continue to apply our techniques to more classical problems in operations research.

4.2. Molecular biology

Participants: Alexander Bockmayr, Arnaud Courtois, Yannick Darcy, Eric Domenjoud, Damien Eveillard, Emmanuel Gothié, Yann Guermeur, Abdelhalim Larhlimi, Myriam Vezain.

Molecular biology is concerned with the study of three types of biological macromolecules: DNA, RNA, and proteins. Each of these molecules can initially be viewed as a string on a finite alphabet: DNA and RNA are nucleic acids made up of nucleotides A,C,G,T and A,C,G,U, respectively. Proteins are sequences of amino acids taken among twenty different amino acids. Thus they may be represented by an alphabet of 20 letters.

Molecular biology studies the information flow from DNA into RNA, and from RNA to proteins. In a first step, called *transcription*, a DNA string ("gene") is transcribed into messenger RNA (mRNA). In the second step, called *translation*, the mRNA is translated into a protein, where each triplet of nucleotides encodes one amino acid ("genetic code"). During transcription, an intermediate maturation step can occur, which happens mainly in eukaryotic cells. In the so-called *splicing* process, introns are removed from the premessenger RNA. The remaining exons are concatenated yielding the mature RNA molecule.

Biological macromolecules are not just sequences of nucleotides or amino acids. Actually, they are complex three-dimensional objects. DNA shows the famous double-helix structure. RNA and proteins fold into complex three-dimensional structures, which depend on the underlying sequence. RNA is a single-stranded chain of nucleotides. However, a nucleotide in one part of the molecule can base-pair with a nucleotide in another part, following the Crick-Watson complementarity rules. This results in a folding of the molecule. The *secondary structure* of RNA indicates which nucleotides pair to which. This information can be represented by a graph.

Proteins have several levels of structure. Above the primary sequence is the *secondary structure*, which involves three basic types: α*-helices*, β*-sheets*, and structure elements that are neither helices nor sheets, called *loops*. A *domain* of a protein is a combination of secondary structure elements with some specific function. It contains an *active site* where an interaction with an external molecule may happen. A protein may have one or several domains.

The ultimate goal of molecular biology is to understand the *function* of biological macromolecules in the life of the cell. Function results from the *interaction* between different macromolecules, and depends on their structure. The challenge is to make the leap from sequence, through structure, to understand about the function.

4.3. Crystallography

Participants: Alexander Bockmayr, Eric Domenjoud.

X-ray structure analysis is the main tool to establish the three-dimensional atomic structure of biological macromolecules and their complexes. The determination of a structure in X-ray crystallography passes through several stages:

- purification and crystallization of the object under study (a protein, DNA, RNA, virus, or a huge macromolecular complex, such as ribosome or lipoprotein particles);
- X-ray experiment (usually at synchrotron accelerators); data collection (up to a million of independent observations) and their primary processing;
- the solution of the inverse problem of the theory of diffraction to find the electron density distribution in the studied object and to interpret it in terms of atoms.

A key problem of X-ray structure analysis is the so-called *phase problem*. In an X-ray experiment, one can measure only the magnitudes of the complex Fourier coefficients of the electron density distribution under study, but not their phases. Half of the necessary information is therefore lost, and must be restored by other means.

4.4. Operations research

Participants: Alexander Bockmayr, Eric Domenjoud, Nicolai Pisaruk.

Operations research is a classical application area for techniques from constraint programming and discrete optimization. Typical problems include production planning, scheduling, resource allocation, or transportation. Due to our participation in the European project LISCOS (Large-scale Integrated Supply Chain Optimisation Software), we are particularly interested in supply chain optimisation problems.

5. Software

5.1. M-SVM: Multi-class Support Vector Machine

Participant: Yann Guermeur [correspondent].

We have improved the software of the general purpose M-SVM (Multi-class Support Vector Machine), releasing a new version available from the webpage « kernel machines », at http://www.kernel-machines.org. It is now more efficient in terms of running time and memory requirements, it has more functionalities and is easier to use. Furthermore we have made available at http://www.loria.fr/~guermeur/ a version of the machine devoted to protein secondary structure prediction. It corresponds to the approach described in [15] (see also Section 6.3). The system architecture has been designed in such a way that the user can easily adapt the software to other problems of biological sequence processing.

5.2. KOALAB 1.0: KOupled Algorithmic and Learning Approach for Biological sequences

Participants: Damien Eveillard [correspondent], Yann Guermeur, Abdelhalim Larhlimi.

Recent studies show that techniques based on M-SVMs (Multi-class Support Vector Machines) may improve the capability to discover biological motifs (see Sect. 6.5 and 6.4). However, using a M-SVM remains difficult for the non-specialist. To overcome this problem, we have developed a new tool for biologists, KOALAB 1.0, which provides a user-friendly interface to the M-SVM technology. KOALAB has been designed to lead the biologist in the discovery of biological motifs in an unknown genome. The main features of KOALAB 1.0 are as follows. First, the software gives access to the theoretical work of the MODBIO project. Using web technology, the user can apply the latest M-SVM techniques developed in the team, without having to be concerned about the technical details. KOALAB 1.0 can be installed on a local or remote web server. Thus, only a web browser is needed to use it. Second, KOALAB 1.0 provides a practical graphical interface to handle the output of M-SVMs. The interface incorporates a variety of tools, among them data analysis techniques. Combining these with a M-SVM is an appropriate way to detect both known and unknown biological motifs. Finally, a new version of GRAPPE, the motif finding algorithm developed by the ADAGE project-team, has been incorporated in KOALAB 1.0. By combining algorithmic and a statistical learning methods, KOALAB 1.0 may be particularly useful in exploring genomes that are not well-known yet, e.g. the genome of HIV-1. A new version, KOALAB 1.1, currently in test, will soon be available from our website under the GNU GPL licence.

6. New Results

6.1. Integer programming and the phase problem in crystallography

Participants: Alexander Bockmayr, Eric Domenjoud.

Key words: *Integer programming*, *crystallography*, *phase problem*.

The first stage of revealing the structure of a biological macromolecule by methods of X-ray crystallography is to find the function that represents the electron density distribution in the crystal of the studied object. This

function is periodical in the three space directions and may be represented by a three-dimensional Fourier series. In a standard X-ray diffraction experiment, one can only determine the magnitudes of the complex Fourier coefficients, but not their phases. The problem of restoring the phase values is called the phase problem of X-ray crystallography [9][17].

In collaboration with Vladimir Lunin (Laboratory of Macromolecular Crystallography, IMPB, RAS) and Alexandre Urzhumtsev (LCM3B, University Nancy 1), we have shown that the phase problem in X-ray crystallography can be modeled as a binary integer program [10]. The basic idea is to introduce a three-dimensional grid and to binarize the magnitudes and phases at the grid points. The integer programming problem may still have numerous solutions, but a special averaging technique allows a single, crystallographically meaningful solution to be derived. Our work this year has been focusing on reformulating the initial model and on increasing the grid size [13].

6.2. Structural risk minimization inductive principle for multi-class discriminant analysis

Participant: Yann Guermeur.

Key words: *Statistical learning theory*, *support vector machine*.

If the rate of convergence of the empirical risk to the expected risk has been extensively studied in the case of sets of indicator functions computing dichotomies or polychotomies, as well as sets of real-valued functions computing dichotomies, so far, the case of large margin multi-class discriminant models has seldom been studied independently. This is all the more unsatisfactory that the study of this case, which generalizes the former ones, is of central importance, for instance to develop the theory of multi-class kernel machines such as M-SVMs. Furthermore, one cannot tackle it by extending in a straightforward way the results derived in the bi-class case. To bridge this gap, we have continued our collaboration with André Elisseeff and Dominique Zelus on the derivation of guaranteed risks for large margin multi-class discriminant models. Our last results, exposed in [14], highlight the central importance played by scale-sensitive extensions of the Ψ-dimensions to study the capacity, and thus the generalization capabilities, of these models. Once more, we focused on the special case of M-SVMs, for which it is possible, for instance, to bound tightly the scale-sensitive Natarajan dimension.

6.3. Protein structure prediction

Participants: Yannick Darcy, Yann Guermeur, Sumit Jha.

Key words: *Statistical learning*, *protein secondary structure*, *disulfide bridges*.

Knowing the three-dimensional structure of a protein can greatly help to infer its function. Predicting this *tertiary structure* from the sequence of amino acids (or *primary structure*), remains one of the central open problems in structural biology. This is the subject of the « GENOTO3D » project that we coordinate. This year, our main efforts have been concentrated on two intermediate problems: the prediction of disulfide bridges and the prediction of the secondary structure. The prediction of disulfide bridges has been one of the two main applications selected by the working group « Apprentissage et Séquences » of the « Action Spécifique Apprentissage et Bioinformatique ». In this field, our contribution has mainly consisted in the implementation of a SVM with a specifically designed kernel, taking into account high-level knowledge sources provided, among others, by Olivier Gascuel. When integrated in an appropriate way, such predictors as the parity of the number of cysteins in the chain, or the location of the protein in the cell, have proved once more their relevance for the task. This study was the subject of Sumit Jha's internship. Since October, it has been resumed by Yannick Darcy. As for the secondary structure prediction, we continued our collaboration with Pierre Baldi's research group on the combination of prediction methods [16]. This gave birth to a new version of SSpro which should be available online in the months to come. Furthermore, we developed a M-SVM devoted to the prediction directly from the primary structure, or the profile of multiple alignments [15] (see also [21]).

6.4. Search for non-coding RNA genes

Participants: Emmanuel Gothié [in collaboration with the UMR CNRS 7567 MAEM], Sandrine Peyrefitte [in collaboration with the UMR CNRS 7567 MAEM], Yann Guermeur.

Key words: *InterORF sequences*, *ncRNA*, *support vector machine*, *pattern discovery*.

Genome sequence data are the object of various in silico analyses in order to provide accurate automatic annotation or to search for functional subsequences: ORF, regulatory elements, etc. The vast majority of current analysis tools have been designed for research on protein encoding sequences. At the same time, increasing evidence on the number and functions of non-coding RNAs (ncRNA) requires a systematic research and dedicated tools for this purpose. Our goal is to come up with a new tool for the identification of genes encoding for untranslated but functional RNA within the genomes.

Our research is based on statistical learning, more precisely the M-SVM (Multiclass Support Vector Machines) developed in our group. After a training phase on well-documented genomes (for example, yeast or Drosophila etc.), these techniques can be applied to newly released sequence data that still have to be annotated.

Small nucleolar RNAs (snoRNA, involved in two types of post-transcriptionnal modifications of the ribosomal RNA (rRNA)) were selected as templates among non coding small RNAs to develop and assess our approach. In a first stage, the biological model organism baking yeast *Saccharomyces cerevisiae* allowed testing the new tool before possible generalization to other genomes. The system is able to retrieve the whole set of snoRNA (56 sequences) among the total intergenic sequences of this yeast. Additional positive signals are also present, which suggests that other snoRNA could also be discovered with this approach. Sharper analysis of these signals, as well as experimental checks will allow us to validate these potential candidates. In order to test whether the method may be generalized, the approach was applied to intergenic sequence data sets (provided by S. Muller, MAEM) from three entirely sequenced Archaea genomes belonging to the Pyrococcus genus: *P. abyssi*, *P. furiosus* and *P. horikoshii*. The analysis showed that by learning on the sequence data from a single genome, it was possible to retrieve sRNA in the two other related genomes. This confirmed the results obtained independently by comparative analysis. The tests carried out so far on *Saccharomyces cerevisiae* and different Archaea genomes are promising and suggest that the SVM-based technique may be a good tool for identifying ncRNA genes [23].

6.5. SELEX data processing

Participants: Stéphanie Billaut, Damien Eveillard [in collaboration with the UMR CNRS 7567 MAEM], Yann Guermeur, Abdelhalim Larhlimi.

Key words: *Statistical learning theory*, *support vector machine*, *SELEX*, *pattern discovery*.

RNA-protein interactions play an important role in the cell cycle. Recent work shows the importance of nucleic motifs in these interactions. SELEX experiments [32] can automatically characterize the potential ligands for a given target protein, starting from a random oligonucleotidic database. As shown in [4], processing SELEX data is a non-trivial task. The biological motifs generally cannot be directly identified from the experimental database. In particular, this holds for the binding sites of SR proteins, a protein family that is important in the regulation of the alternative splicing process, see Section 6.7.

To overcome this problem, we have developed this year a new method to localise a protein binding motif, based on statistical learning. We optimised a kernel method (M-SVM), dedicated to the recognition of SR motifs. Our machine was trained on experimental SELEX data. To analyse the M-SVM results biologically, a graphical interface on top of the M-SVM results was developed. Using data analysis in addition to the graphics interpretation, we can now predict SR binding sites in the HIV-1 genome.

6.6. Modeling biological systems

Participants: Alexander Bockmayr, Arnaud Courtois, Damien Eveillard.

Key words: *Constraint programming*, *hybrid systems*, *modeling*, *system biology*.

Systems biology is a new area in biology that aims at achieving a systems-level understanding of biological systems. While current genome projects provide a huge amount of data on genes or proteins, lots of research is still necessary to understand how the different parts of a biological system interact in order to perform complex biological functions. Computational models that help to analyze, explain or predict the behavior of biological systems play a crucial role in systems biology. While traditional biology examines single genes or proteins in isolation, system biology simultaneously studies the complex interaction of many levels of biological information - genomic DNA, mRNA, proteins, informational pathways and networks - to understand how they work together.

In [1] and [22], we have introduced hybrid concurrent constraint programming (`hcc`) [28] as a promising alternative to existing modeling and simulation approaches in systems biology. Hybrid cc is a declarative compositional programming language with a well-defined semantics. It allows one to model and simulate the dynamics of hybrid systems, which exhibit both discrete and continuous change. We show that Hybrid cc can be used naturally to model a variety of biological phenomena, such as reaching thresholds, kinetics, gene interaction or biological pathways.

We are currently using this approach to model alternative splicing regulation in HIV-1, see Section 6.7.

6.7. Regulation of alternative splicing

Participants: Alexander Bockmayr, Arnaud Courtois, Damien Eveillard [in collaboration with the UMR CNRS 7567 MAEM], Myriam Vezain.

Key words: *Modeling*, *hybrid system*, *constraint programming*, *alternative splicing*, *HIV-1*.

Alternative splicing is a key process in post-transcriptional regulation, by which several kinds of mature RNA can be obtained from the same premessenger RNA. The resulting combinatorial complexity contributes to biological diversity, especially in the case of the human immunodefficiency virus HIV-1. In collaboration with our partners from the Laboratory MAEM in Nancy, we have started to model the alternative splicing regulation in HIV-1.

Molecular biology studies the information flow in the transcription of DNA into RNA, and the translation of RNA in proteins. The DNA molecule first yields a premessenger RNA molecule. The premessenger RNA can be decomposed in exons and introns. During splicing, introns are removed. Only some exons are kept for the messenger or mature RNA (mRNA). The regulation of the splicing process depends on donor and acceptor sites. The donor site is located at the end of an exon A, the acceptor site at the beginning of exon B. Together, they define the intron to be excised from the premessenger RNA.

In the HIV-1 case, splicing regulation is a complex phenomenon involving 4 donor sites and 8 acceptor sites, which may yield 40 mature messenger RNAs [30]. This combinatorial complexity is achieved by regulating the selection of the acceptor site. Recent biological studies show that SR (Serine ARginine rich) proteins play a crucial role in this regulation.

In a first step, we have developed a model for the SR regulation of the A3 acceptor site in HIV-1 [20][22]. The qualitative behavior of the model depends on the values of the reaction kinetic parameters. Experimental results available to us validate this first approach in the equilibrium phase.

In a second step, we have developed a constraint programming approach that allows us to integrate different single-site models into a multi-site model of alternative splicing regulation. The model described in [20][22], involves the acceptor sites A3, A4, and A5. We are using an hybrid automaton implemented in `Hybrid cc` to model the selection of a particular acceptor site.

In a third step, we have developed a new single-site model of the acceptor site A7, which is again based on differential equations [24]. We are currently integrating a multi-site model of alternative splicing regulation [19] based on the sites A3, A4, A5, and A7 into a model of the full HIV-1 life cycle proposed in [29]. In parallel, we perform a theoretical analysis of the different multi-site models. In particular, this allows us to determine the possible qualitative behaviours depending on the parameter values, which may be characterized by simple mathematical formulas.

The ultimate goal is to obtain a model that can be validated qualitatively both on the scale of a single splicing site and on the scale of the whole HIV-1, in order to represent the global effect of alternative splicing in the HIV-1 cycle.

6.8. Disjunction of polytopes

Participants: Alexander Bockmayr, Nicolai Pisaruk.

Key words: *integer programming*, *cardinality constraints*, *disjunction*, *separation*.

A well-known result on unions of polyhedra in the same space gives an extended formulation in a higher-dimensional space whose projection is the convex hull of the union. In [11], we study the unions of polytopes in different spaces, giving a complete description of the convex hull without additional variables. When the underlying polytopes are monotone, the facets are described explicitly, generalizing results of Hong and Hooker on cardinality rules [36], and an efficient separation algorithm is proposed. These results are based on an explicit representation of the dominant of a polytope, and an algorithm for the separation problem for the dominant. For non-monotone polytopes, both the dominant and the union are characterized. We also give results on the unions of polymatroids both on disjoint ground sets and on the same ground set generalizing results of Conforti and Laurent [26].

6.9. Combining MIP and CP

Participants: Alexander Bockmayr, Nicolai Pisaruk.

Key words: *constraint programming*, *integer programming*, *cooperative solving*.

Mixed integer programming (MIP) and constraint programming (CP) are two complementary approaches for solving complex combinatorial optimization problems. Combining these two methods has been an important research topic during the last years. At the same time, it has also been crucial for solving various real-world industrial applications.

Continuing our previous work, we develop in [12] a unifying view of integer programming and finite domain constraint programming. We present the two modeling and solution approaches in a uniform framework, branch-and-infer. The goal of this framework is to clarify the relationship between the two techniques, and to indicate possible ways towards their integration. We illustrate the different concepts by examples from discrete tomography and supply chain optimization.

In [18], we develop a new hybrid MIP/CP solution approach in which CP is used for detecting infeasibilities and generating cuts within a branch-and-cut algorithm for MIP. Our framework applies to MIP problems augmented by monotone constraints that can be handled by CP. We illustrate our approach on a generic multiple machine scheduling problem, and present a number of computational experiments.

7. Contracts and Grants with Industry

7.1. LISCOS

Participants: Alexander Bockmayr, Nicolai Pisaruk.

The group has participated in the project LISCOS (Large-scale Integrated Supply Chain Optimisation Software Based upon Branch & Cut and Constraint Programming Methods) funded by the European Community. The partners of the project, which started in January 2000 and was completed in March 2003, were Barbot (P), BASF (D), CORE (B), COSYTEC (F), Dash (UK), DEIO (P), LORIA (F), PSA (F), Procter and Gamble (B). The main goal of LISCOS was to develop new software for modeling and solving supply chain optimization problems. This software is based on an integration of mixed-integer programming and finite domain constraint programming. The main results were presented during a Seminar on New Technology for Supply Chain Planning and Scheduling in March 2003 at Brussels (see http://www.dashoptimization.com/liscos.html).

8. Other Grants and Activities

8.1. Regional projects

We participate in the « Génopole Strasbourg Alsace-Lorraine » together with the laboratory MAEM (« Maturation des ARN et Enzymologie Moléculaire », UMR 7567 CNRS-UHP) in Nancy and the IGBMC in Strasbourg.

In the framework of the CPER Lorraine 2000-2006, we participate in the project « Bioinformatics and Applications to Genomics » of the PRST « Intelligence Logicielle ». Our partners here are the Laboratory of Crystallography LCM3B (UMR CNRS 7036) and the MAEM (UMR 7567) at the University Henri Poincaré, Nancy 1.

8.2. National projects

Since February 2002, we have been participating in the cooperative research action ARC CPBIO « Process calculi and Biology of Molecular Networks ». Our partners are the project team CONTRAINTES from INRIA Rocquencourt (F. Fages), the Genoscope (V. Schächter) and the laboratory PPS (V. Danos).

We have regular contacts with the INRIA project teams HELIX (Rhône-Alpes), SYMBIOSE (Rennes) and COMORE (Sophia-Antipolis). In particular, we are collaborating with Hidde de Jong (HELIX) in modeling the regulation of alternative splicing.

We are members of two « Actions Spécifiques » CNRS, entitled « Support Vector Machines (SVM) and Kernel Methods » and « Machine Learning and Bioinformatics ». In the context of the second project, we organize a working group entitled « Machine Learning and Sequences ».

Since September 2003, we are coordinating a project called GENOTO3D which is funded by the « Action Concertée Incitative » (ACI) « Masses de Données ». The aim of this project is to apply machine learning approaches to the prediction of the tertiary structure of globular proteins. Our partners are the IBCP in Lyon, the LIF in Marseille, the project team SYMBIOSE from IRISA, the LIRMM in Montpellier, and the MIG laboratory of INRA in Jouy-en-Josas.

8.3. European projects

In the framework of the Growth programme of the European Community, we have been participating in the project LISCOS (Large-scale Integrated Supply Chain Optimisation Software), Contract No. G1RD-CT-1999-000034.

8.4. International relations

Within the French-Russian Institute Liapunov, we have a joint project with the Institute for Mathematical Problems in Biology (IMPB) of the Russian Academy of Sciences in Pushchino (V. Y. Lunin).

We have been collaborating with researchers from Carnegie-Mellon University (E. Balas, John N. Hooker), the Center of Operations Research CORE in Louvain-la-Neuve (L. Wolsey), the Max Planck Institute for Computer Science in Saarbrücken (E. Althaus, K. Mehlhorn), SAP AG (T. Kasper), the University of California at Irvine (P. Baldi), IBM at Zurich (A. Elisseeff), and the Wiener laboratories in Rosario (D. Zelus).

9. Dissemination

9.1. Serving the scientific community

Alexander Bockmayr is leading the research action « Bioinformatics » of LORIA and INRIA Lorraine, and the project « Bioinformatics and Applications to Genomics » of the PRST « Intelligence Logicielle ». He has been on the Scientific Board of the ACI IMPBio « Informatics, Mathematics, and Physics in Molecular Biology »

of the French Ministry of Research, and a member of the programme committees of CMSB'03, CPAIOR'03, and CP'03. He is an associate editor of INFORMS J. Computing and coordinator of « Optimization Online ».

Yann Guermeur has been a member of the program committees of CAP'03 and RFIA'04.

9.2. Teaching

Alexander Bockmayr is a professor of computer science at the University Henri Poincaré, Nancy 1.

Yann Guermeur is assistant professor in computer science at the University Henri Poincaré, Nancy 1.

Arnaud Courtois is a teaching assistant (« moniteur ») in computer science at the University Henri Poincaré, Nancy 1.

Damien Eveillard is teaching bioinformatics in the DESS RGTI.

9.3. Divers

Alexander Bockmayr has given invited lectures at the Spring School « Modelling and simulation of biological processes » at Dieppe, the Colloquium « Protein-protein interactions » at Palaiseau, the Portuguese Conference on Artificial Intelligence (EPIA'03), at ENS Cachan, INSA Rouen, and the University of Metz. He has been co-organizing the French-Cuban symposium on Bioinformatics at Havana in February 2003.

Yann Guermeur has been an invited speaker at the satellite workshop « Kernel Methods in Computational Biology » of RECOMB'03.

10. Bibliography

Major publications by the team in recent years

[1] A. BOCKMAYR, A. COURTOIS. *Using hybrid concurrent constraint programming to model dynamic biological systems.* in « 18th International Conference on Logic Programming, ICLP'02, Copenhagen », Springer, LNCS 2401, pages 85-99, 2002.

[2] A. BOCKMAYR, V. WEISPFENNING. *Solving numerical constraints.* A. ROBINSON, A. VORONKOV, editors, in « Handbook of Automated Reasoning », volume 1, Elsevier, 2001, chapter 12, pages 751-842.

[3] E. DOMENJOUD, C. KIRCHNER, J. ZHOU. *Generating feasible schedules for a pick-up and delivery problem.* in « Electronic Notes in Discrete Mathematics », volume 1, 1999.

[4] D. EVEILLARD, Y. GUERMEUR. *Traitement statistique des résultats SELEX.* in « JOBIM », J. NICOLAS, C. THERMES, editors, pages 277-283, St Malo, 2002.

[5] Y. GUERMEUR, A. ELISSEEFF, D. ZELUS. *Bound on the risk for M-SVMs.* in « Statistical Learning, Theory and Applications », pages 48–52, 2002.

[6] Y. GUERMEUR, C. GEOURJON, P. GALLINARI, G. DELÉAGE. *Improved performance in protein secondary structure prediction by inhomogeneous score combination.* in « Bioinformatics », number 5, volume 15, 1999, pages 413–421.

[7] Y. GUERMEUR. *Combining discriminant models with new multi-class SVMs.* in « Pattern Analysis and Applications », number 2, volume 5, 2002, pages 168–179.

[8] Y. GUERMEUR, H. PAUGAM-MOISY. *Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines.* M. SEBBAN, G. VENTURINI, editors, in « Apprentissage Automatique », Hermès, 1999, pages 109–138.

[9] V. Y. LUNIN, N. LUNINA, A. PODJARNY, A. BOCKMAYR, A. URZHUMTSEV. *Ab initio phasing starting from low resolution.* in « Z. Kristallogr. », number 12, volume 217, 2002, pages 668-685.

[10] V. Y. LUNIN, A. URZHUMTSEV, A. BOCKMAYR. *Direct phasing by binary integer programming.* in « Acta Crystallographica Section A », volume 58, 2002, pages 283-291.

Articles in referred journals and book chapters

[11] E. BALAS, A. BOCKMAYR, N. PISARUK, L. WOLSEY. *On unions and dominants of polytopes.* in « Mathematical Programming, Series A », 2003, To appear.

[12] A. BOCKMAYR, T. KASPER. *Branch-and-Infer : a Framework for Combining CP and IP.* M. MILANO, editor, in « Constraint and Integer Programming. Toward a Unified Methodology », series Operations Research/Computer Science Interfaces, volume 27, Kluwer, Oct, 2003, chapter 3, pages 59-87.

[13] A. BOCKMAYR, V. LUNIN, N. LUNINA, A. URZHUMTSEV. *Mathematical methods for the direct solution of the phase problem in X-ray structural analysis of biological macromolecules.* in « French-Russian A. M. Liapunov Institute for Applied Mathematics and Computer Science Transactions, Volume 4 », INRIA, Sep, 2003, pages 200-232.

[14] Y. GUERMEUR, A. ELISSEEFF, D. ZELUS. *A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers.* in « Applied Stochastic Models in Business and Industry », 2003, (to appear).

[15] Y. GUERMEUR, A. LIFCHITZ, R. VERT. *A kernel for protein secondary structure prediction.* B. SCHÖLKOPF, K. TSUDA, J.-P. VERT, editors, in « Kernel Methods in Computational Biology », The MIT Press, 2003, (to appear).

[16] Y. GUERMEUR, G. POLLASTRI, A. ELISSEEFF, D. ZELUS, H. PAUGAM-MOISY, P. BALDI. *Combining Protein Secondary Structure Prediction Models with Ensemble Methods of Optimal Complexity.* in « Neurocomputing », 2003, (in press).

Publications in Conferences and Workshops

[17] A. BOCKMAYR, V. LUNIN, A. URZHUMTSEV. *A binary integer programming approach for the determination of structures of biological macromolecules (Extended Abstract).* in « Moscow Conference on Computational Molecular Biology 2003 - MCCMB'03, Moscou, Russie », pages 44-45, July, 2003.

[18] A. BOCKMAYR, N. PISARUK. *Detecting Infeasibility and Generating Cuts for MIP using CP.* in « 5th International Workshop on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems - CPAIOR'03, Montréal, Canada », May, 2003.

[19] D. EVEILLARD, A. COURTOIS, A. BOCKMAYR. *Hybrid concurrent constraint programming : A well-suited formalism for modelling alternative splicing regulation (Abstract).* in « Modeling and Simulation of Biological

Regulatory Processes - ECCB Satellite Meeting, Paris, France », October, 2003.

[20] D. EVEILLARD, D. ROPERS, H. D. JONG, C. BRANLANT, A. BOCKMAYR. *Multiscale modeling of alternative splicing regulation.* in « Computational Methods in Systems Biology, CMSB'03 », volume 2602, Springer LNCS, C. PRIAMI, editor, pages 75–87, Rovereto, Italy, 2003.

Internal Reports

[21] E. DIDIOT. *Conception et mise en oeuvre de M-SVM dédiées au traitement de séquences biologiques.* Rapport de DEA, July, 2003.

[22] D. EVEILLARD, D. ROPERS, H. D. JONG, C. BRANLANT, A. BOCKMAYR. *A Multi-Site Constraint Programming Model of Alternative Splicing Regulation.* Research Report RR 4830, INRIA, May, 2003, http://www.inria.fr/rrrt/rr-4830.html.

[23] E. GOTHIÉ, Y. GUERMEUR, S. MULLER, C. BRANLANT, A. BOCKMAYR. *Recherche des gènes de petits ARN non codants.* Research Report RR-5057, INRIA, December, 2003, http://www.inria.fr/rrrt/rr-5057.html.

[24] M. VEZAIN. *Modélisation de la régulation de l'épissage alternatif au site A7 de HIV-1.* Rapport de Stage, DESS EGOISt, Université Rouen, June, 2003.

Bibliography in notes

[25] C. BURGES. *A tutorial on support vector machines for pattern recognition.* in « Data Mining and Knowledge Discovery », number 2, volume 2, June, 1998, pages 121–167.

[26] M. CONFORTI, M. LAURENT. *On the facial structure of independence system polyhedra.* in « Mathematics of Operations Research », volume 13, 1988, pages 543 - 555.

[27] C. CORTES, V. VAPNIK. *Support-Vector Networks.* in « Machine Learning », volume 20, 1995, pages 273–297.

[28] V. GUPTA, R. JAGADEESAN, V. SARASWAT. *Computing with Continuous Change.* in « Science of computer programming », number 1-2, volume 30, 1998, pages 3-49.

[29] B. J. HAMMOND. *Quantitative Study of the Control of HIV-1 Gene Expression.* in « J. Theor. Biol », volume 163, 1993, pages 199–221.

[30] D. PURCELL, M. MARTIN. *Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity.* in « J. Virol. », number 11, volume 67, 1993, pages 6365–6378.

[31] V. A. SARASWAT. *Concurrent constraint programming.* series ACM Doctoral Dissertation Awards, MIT Press, 1993.

[32] C. TUERK, L. GOLD. *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.* in « Science », volume 249, 1990, pages 505-510.

[33] P. VAN HENTENRYCK, V. SARASWAT. *Strategic directions in constraint programming.* in « ACM Computing Surveys », number 4, volume 28, 1996, pages 701 – 726.

[34] V. VAPNIK. *Estimation of Dependences Based on Empirical Data..* Springer-Verlag, N.Y., 1982.

[35] V. VAPNIK. *Statistical learning theory.* John Wiley & Sons, Inc., N.Y., 1998.

[36] H. YAN, J. N. HOOKER. *Tight representation of logical constraints as cardinality rules.* in « Mathematical Programming », volume 85, 1999, pages 363-377.