INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Action MODBIO

Modèles Informatiques en Biologie Moléculaire

Lorraine

THÈME 2A

Table des matières

1 Composition de l'équipe

Responsable scientifique
Alexander Bockmayr [Professeur, Université Henri Poincaré, Nancy 1]
Responsable permanent
Eric Domenjoud [CR CNRS]
Assistante de projet
Christelle Etienne-Bergeret [INRIA, jusqu'au 30/4/2001]
Sophie Drouot [INRIA, à partir du 17/4/2001]
Personnel CNRS
Miki Hermann [CR]
Personnel Université
Yann Guermeur [Maître de Conférences, Université Henri Poincaré, Nancy 1]
Chercheurs doctorants
Arnaud Courtois [UHP, depuis le 1/11/2001]
Damien Eveillard [UHP, cofinancé Région-Lorraine]
Chercheurs post-doctorants
Emmanuel Gothié [UHP, depuis le 1/10/2001]
Nicolai Pisaruk [UHP]
Chercheur invité INRIA
Natasha Lunina [du 1/6/2001 au 30/6/2001]
Stagiaire
Arnaud Courtois [DEA]

2 Présentation et objectifs généraux

L'avant-projet MODBIO a été créé le 1er janvier 2001 par des anciens membres des projets PROTHEO et CORTEX. L'objectif de MODBIO est le développement de modèles informatiques pour la biologie moléculaire. Nous nous intéressons à deux types de problèmes :

- déterminer la structure de macromolécules biologiques ;
- comprendre leur fonction.

Notre approche est basée sur une combinaison de techniques de la programmation par contraintes, de l'optimisation discrète, des systèmes hybrides, et des réseaux de neurones.

Projets actuels

- Détermination et analyse des enveloppes macromoléculaires (partenaires : Laboratoire de Cristallographie LCM3B, Nancy ; Institut des Problèmes Mathématiques en Biologie, Académie des Sciences, Poushchino, Russie) ;
- Structure des ARN et épissage alternatif (partenaire : Laboratoire « Maturation des ARN et Enzymologie Moléculaire », Nancy) ;
- Prédiction de la structure secondaire des protéines (partenaires : Institut de Biologie et de Chimie des Protéines, Univ. Claude Bernard, Lyon ; Université de Californie, Irvine, Etats-Unis) ;

– Modélisation de systèmes biologiques en programmation par contraintes (partenaire : Institut Pasteur, Paris).

Principaux axes de recherche en informatique

– Programmation par contraintes ;
– Optimisation discrète ;
– Déduction automatique et théorie de la complexité ;
– Apprentissage statistique et réseaux de neurones.

Relations scientifiques et industrielles

– Participation au Génopole Strasbourg Alsace–Lorraine ;
– Participation au projet Bioinformatique du PRST Intelligence Logicielle de la Région Lorraine ;
– Participation à l'action IMPG ;
– Participation au projet européen LISCOS ;
– Nombreuses collaborations nationales et internationales.

3 Fondements scientifiques

3.1 Contraintes et optimisation discrète

Mots clés : Contrainte, optimisation discrète.

Participants : Alexander Bockmayr, Arnaud Courtois, Eric Domenjoud, Nicolai Pisaruk.

Beaucoup de questions scientifiques, industrielles ou techniques peuvent être formulées au moyen de *contraintes*. Par *contrainte*, on entend une formule mathématique portant des variables réelles et qui exprime une relation qui doit être satisfaite par les valeurs de ces variables. Par exemple la formule $x + y \leq 1$ exprime que la somme des valeurs des variables x et y doit être inférieure ou égale à 1. Une solution *faisable* d'un ensemble de contraintes est une valuation des variables qui satisfait toutes les contraintes simultanément. Dans l'exemple précédent, $x = 0$ et $y = 1$ est une telle solution. En général, une fonction de coût pour les solutions est donnée. Une solution *optimale* du problème est alors une solution faisable pour laquelle la valeur de cette fonction est minimale ou maximale.

En *programmation par contraintes*, il s'agit de développer des langages de programmation dans lesquels on peut exprimer de manière naturelle des problèmes de contraintes et les résoudre efficacement. Dans notre recherche, nous nous intéressons principalement aux problèmes de contraintes sur les domaines finis. Le domaine de chaque variable, c'est-à-dire l'ensemble des valeurs qu'elle peut prendre, est alors un sous-ensemble fini des entiers naturels. La théorie nous enseigne que la plupart des problèmes sont NP-difficiles, ce qui signifie qu'il est très peu probable qu'on puisse résoudre ces problèmes par des algorithmes polynômiaux en la taille des données. En pratique ces problèmes sont traités par des méthodes d'exploration d'un arbre

de recherche qui essaient successivement différentes valuations des variables jusqu'à ce qu'une solution soit trouvée. A cause du nombre exponentiel de combinaisons possibles, il est crucial de réduire au maximum l'espace de recherche, c'est-à-dire d'éliminer a priori le plus grand nombre de valuations.

Pour résoudre ces problèmes, il existe essentiellement deux méthodes. La première est l'*optimisation entière* classique comme elle est appliquée en mathématiques et en recherche opérationnelle depuis plus de 40 ans. Les contraintes sont des équations et des inéquations linéaires sur les entiers. Pour réduire l'espace de recherche, on considère souvent la *relaxation linéaire* de l'ensemble des contraintes. On résout les équations et les inéquations d'abord sur les réels, ce qui est beaucoup plus facile, puis on utilise cette information pour réduire le nombre des alternatives à énumérer.

La deuxième méthode est la *programmation par contraintes sur les domaines finis*, qui a émergé en informatique durant ces 15 dernières années. A l'opposé de l'optimisation entière, on utilise, en plus des contraintes arithmétiques simples, des contraintes complexes, dites *contraintes symboliques*. Par exemple, la contrainte symbolique `alldifferent`$([x_1, \ldots, x_n])$ exprime que les variables $x_1, \ldots, x_n$ doivent prendre des valeurs distinctes 2 à 2. Une telle contrainte est difficile à exprimer au moyen d'équations et d'inéquations. On résout les contraintes symboliques séparément par des algorithmiques spécifiques qui réduisent le domaine des variables. Cette information est propagée aux autres contraintes qui, à leur tour, réduisent le domaine des variables.

3.2 Complexité et algorithmes d'approximation

Mots clés : Complexité, algorithme d'approximation.

Participants : Alexander Bockmayr, Miki Hermann.

Les problèmes d'optimisation les plus naturels, qui incluent aussi le domaine de la bioinformatique, sont généralement NP-difficiles. C'est pourquoi, sous l'hypothèse courante adoptée par la majorité des chercheurs en algorithmique que P $\neq$ NP, leur résolution exacte demande un temps de calcul extrêmement long. La recherche d'approximabilité de ces problèmes par les algorithmes polynômiaux devient à ce titre un sujet plus qu'indispensable, qu'il s'agisse de la recherche théorique ou des applications industrielles.

Les problèmes d'optimisation NP-difficiles exhibent un éventail de comportements, allant de l'approximabilité à n'importe quel degré, passant par l'approximabilité constante, jusqu'à essentiellement prohibant toute approximation. Malgré cette diversité, quelques principes communs sont à la base de la construction des algorithmes d'approximation. Même si les problèmes d'optimisation NP-difficiles ne présentent pas de « crampons » pour trouver le résultat optimal efficacement, ils peuvent néanmoins offrir des points d'appuis pour trouver facilement des solutions presque optimales. Donc, au niveau général, le processus de création d'algorithmes d'approximation diffère peu de l'écriture d'un algorithme exact. Il demande toujours de révéler la structure principale du problème et ensuite de trouver les techniques algorithmiques pour les exploiter. Typiquement, cette structure se présente comme trop compliquée et, par conséquent, on utilise souvent des techniques algorithmiques généralisant et élargissant des techniques puissantes développées au cours de l'étude des algorithmes exacts.

Notre but est d'appliquer ces techniques dans le cadre des problèmes algorithmiques et des problèmes d'optimisation en bioinformatique.

3.3 Apprentissage statistique

Mots clés : Apprentissage statistique, statistique non paramétrique, réseaux de neurones, machines à vecteurs support (SVM).

Participants : Damien Eveillard, Yann Guermeur.

La théorie de l'apprentissage statistique est un domaine de la statistique inférentielle dont les fondements ont été posés par V.N. Vapnik à la fin des années 60. L'objet de cette théorie est de déterminer les conditions sous lesquelles il est possible d'apprendre à partir de données empiriques (obtenues par échantillonnage aléatoire simple). L'apprentissage se conçoit comme un problème de sélection de modèle, consistant à déterminer, dans une famille de fonctions donnée, de cardinalité ordinairement infinie, une fonction permettant d'obtenir les meilleures performances possibles sur un problème donné. Le problème en question peut relever de l'analyse discriminante, de l'approximation de fonctions (régression) ou de l'estimation de densité.

Cette théorie étudie particulièrement deux principes inductifs. Le premier, nommé principe de minimisation empirique du risque, consiste à minimiser l'erreur en apprentissage. Dans le cas des petits échantillons, on substitue à ce principe le principe de minimisation structurelle du risque, consistant à minimiser une borne sur l'espérance du risque (erreur en généralisation). Ce dernier principe est en particulier mis en œuvre dans les algorithmes d'apprentissage des machines à vecteurs support (SVM)[Vap98], qui obtiennent actuellement les meilleures performances sur de nombreuses tâches relevant des principaux domaines de la reconnaissance des formes.

4 Domaines d'applications

4.1 Panorama

Résumé : *Le domaine d'applications privilégié de l'équipe est la biologie moléculaire. Dans le même temps nous continuons à nous intéresser à des applications plus classiques de nos techniques dans le domaine de la recherche opérationnelle.*

4.2 Biologie Moléculaire

Mots clés : Biologie moléculaire, ADN, ARN, protéine, séquence, structure, fonction.

Participants : Alexander Bockmayr, Eric Domenjoud, Damien Eveillard, Emmanuel Gothié, Yann Guermeur, Miki Hermann.

La biologie moléculaire concerne l'étude de trois types de molécules biologiques : l'ADN, l'ARN et les protéines. Chacune de ces molécules peut être considérée comme une chaîne

[Vap98] V. Vapnik, *Statistical learning theory*, John Wiley & Sons, Inc., N.Y., 1998.

de caractères sur un alphabet fini. Ainsi, l'ADN et l'ARN sont des acides nucléiques basés respectivement sur les nucléotides A,C,G,T, et A,C,G,U tandis que les protéines sont des séquences d'acides aminés. Il existe 20 acides aminés qui constituent donc un alphabet de 20 lettres.

Le passage ADN $\longrightarrow$ ARN $\longrightarrow$ protéine se fait par un processus constitué de deux étapes : la transcription et la traduction. La transcription conduit, à partir de la séquence ADN double brin, à la formation d'un ARN pré-messager (pré-ARNm) simple brin, elle est suivie par un phénomène d'épissage alternatif qui conduit à la formation de l'ARN messager mature (ARNm), par élimination des introns et concaténation des exons restant.

Dans la seconde étape, l'ARNm est traduit en protéine selon le code génétique qui associe chaque triplet de nucléotides à un acide aminé.

Les macromolécules biologiques ne sont pas seulement des séquences de nucléotides ou d'acides aminés. Il s'agit en réalité d'objets tridimensionnels complexes. L'ADN est structuré sous forme de structure en double hélice, tandis que les ARN et protéines adoptent des structures tridimensionnelles déterminées par les séquences sous-jacentes. La prédiction de structures tridimensionnelles à partir de la séquence primaire est l'un des problèmes majeurs en bioinformatique.

L'ARN est une chaîne de nucléotides simple brin dans laquelle un nucléotide d'une partie de la molécule peut s'associer avec un nucléotide complémentaire situé à un autre endroit de la molécule. Il en résulte une conformation moléculaire. La structure secondaire indique l'appariement des nucléotides. Elle peut être représentée par un graphe. La structure tridimensionnelle de l'ARN dépend du nombre et du type des appariements. A cette structure va être associée la fonction de l'ARN. Il y a donc des relations très étroites entre la structure, la fonction et la séquence des ARN.

Les protéines possèdent plusieurs niveaux de structures. L'enchaînement des différents acides aminés constitue la structure primaire. La structure secondaire correspond ensuite à l'agencement spatial de la protéine. Elle se caractérise par trois types d'éléments : les hélices α, les feuillets β, et les structures non-hélice et non-feuillet, nommées boucles. Une protéine peut posséder un ou plusieurs domaines protéiques qui sont des combinaisons d'éléments de structures secondaires avec quelques fonctions spécifiques. Un site actif d'une protéine est une zone d'interaction potentielle avec une molécule externe. On retrouve ainsi, et de la même manière que précédemment, des relations entre la structure, la fonction et la séquence protéique.

4.3 Cristallographie

Mots clés : Cristallographie, macromolécule, phasage.

Participants : Alexander Bockmayr, Eric Domenjoud.

L'analyse par rayons X constitue l'outil principal pour établir la structure tridimensionnelle des macromolécules biologiques. La détermination d'une structure en cristallographie comporte plusieurs étapes :

- purification et cristallisation de l'objet à étudier (protéine, ADN, ARN, virus, ou grand complexe de macromolécules) ;

- expérimentation par rayons X (généralement au moyen d'un synchrotron) ; collecte des données (jusqu'à un million d'observations indépendantes) et traitement primaire ;
- résolution du problème inverse de la théorie de la diffraction pour trouver la distribution de densité électronique dans l'objet étudié et l'interpréter en termes d'atomes.

Un problème clef de l'analyse de structure par rayons X est le problème du *phasage*. L'expérimentation permet de mesurer seulement la magnitude des coefficients de Fourier complexes de la distribution de densité électronique, mais pas leur phase. Une partie de l'information est donc perdue et doit être reconstruite par d'autres moyens.

4.4 Recherche opérationnelle

Mots clés : Recherche opérationnelle.

Participants : Alexander Bockmayr, Eric Domenjoud, Nicolai Pisaruk.

La recherche opérationnelle est un domaine d'application classique pour les techniques de résolution de contraintes et d'optimisation combinatoire. Dans le cadre des systèmes d'aide à la décision, on étudie des problèmes d'optimisation tels que la planification de la production, la répartition de ressources, ou encore des problèmes de transport. Suite à notre participation au projet européen LISCOS (Large-scale Integrated Supply Chain Optimisation Software) nous nous intéressons en particulier à des problèmes d'optimisation de la chaîne logistique.

5 Résultats nouveaux

5.1 Programmation entière et problème du phasage en cristallographie

Mots clés : Cristallographie, phasage, programmation entière.

Participants : Alexander Bockmayr, Eric Domenjoud.

L'objectif de ces travaux est le développement de nouvelles méthodes de détermination directe des images macromoléculaires cristallographiques à basse résolution sur la base des données de diffraction de rayons X par les cristaux[LLP+00]. Pour la première fois, nous appliquons des méthodes de résolution de contraintes et d'optimisation discrète à des problèmes de cristallographie macromoléculaire.

En collaboration avec le Laboratoire de Cristallographie LCM3B de l'Université Henri Poincaré, Nancy 1 (A. Urzhumtsev), et l'Institut de Problèmes Mathématiques en Biologie de l'Académie des Sciences de la Russie (V. Lunin) nous avons commencé à développer différents modèles de programmation entière pour le problème du phasage en radiocristallographie. En particulier, nous avons réussi à représenter des informations provenant des expériences ainsi que des propriétés cristallographiques de base (par exemple, l'atomicité de la structure et la connectivité des images) par des inégalités linéaires en variables 0-1. Les premiers résultats de cette approche sont décrits dans [20].

[LLP+00] V. Y. Lunin, N. L. Lunina, T. E. Petrova, T. P. Skovorada, A. G. Urzhumtsev, A. D. Podjarny, « Low resolution ab-initio phasing. Problems and advances », *Acta Cryst. D56*, 2000, p. 1223 – 1232.

5.2 Alignement de structures secondaires

Mots clés : Alignement, ARN, structure secondaire.

Participants : Alexander Bockmayr, Miki Hermann.

La problématique de l'alignement des séquences et des structures constitue l'un des thèmes principaux de la bioinformatique. Dans le cas de la structure primaire, la solution est trouvée facilement à l'aide d'un algorithme de programmation dynamique, ce qui permet de résoudre le problème d'alignement des structures primaires en temps polynômial. Pour les ARN, la situation se complique radicalement dans le cas d'alignement des structures secondaires[LMZ01]. Il faut étudier notamment le problème d'alignement selon le classement topologique de la structure secondaire. L'alignement des séquences, où l'une ne possède aucune structure secondaire et la structure secondaire de l'autre ne présente pas de croisement d'arêtes, peut être décidé en temps polynômial aussi. Par contre, si la structure secondaire d'une des séquences présente des croisements, le problème de décision devient NP-complet et le problème d'optimisation correspondant, où le but est de minimiser le coût des opérations de changement de structure, est APX-complet. Ceci veut dire qu'on connaît un algorithme qui produit toujours une solution qui est au plus deux fois plus coûteuse que le minimum et en même temps nous ne pouvons espérer construire qu'un algorithme d'approximation qui produit des solutions qui diffèrent par rapport au minimum par un facteur constant, mais pas mieux, sinon P = NP. La complexité de l'alignement des séquences sans croisement de leurs structures secondaires est inconnue. Ce problème de décision est inclus dans la classe NP, mais on ne sait pas s'il est NP-complet ou s'il peut être résolu en temps polynômial. Alexander Bockmayr et Miki Hermann étudient actuellement cette problématique, en particulier (1) la complexité de différentes variantes d'alignement des structures secondaires sans croisement et (2) la possibilité d'améliorer la constante d'approximation pour l'algorithme d'approximation du problème de minimisation cité ci-dessus.

5.3 Principe de minimisation structurelle du risque pour les systèmes de discrimination multi-classes

Mots clés : Apprentissage statistique, SVM.

Participant : Yann Guermeur.

Si le taux de convergence du risque empirique vers l'espérance du risque est bien étudié dans le cas des modèles calculant des dichotomies, ou des modèles de régression, il n'en est pas de même dans le cas des systèmes discriminants à catégories multiples. Afin de combler cette lacune, nous avons poursuivi notre collaboration avec l'équipe d'Hélène Paugam-Moisy, à l'Institut de Sciences Cognitives (ISC) de Lyon, collaboration portant sur l'étude des lois fortes des grands nombres uniformes. Nous avons ainsi dérivé de nouvelles bornes de convergence uniforme pour certains types de familles de fonctions « Glivenko-Cantelli ». Ceci nous a permis

[LMZ01] G.-H. Lin, B. Ma, K. Zhang, « Edit distance between two RNA structures », *in : Computational molecular biology, RECOMB'01, Montreal*, ACM, p. 211 – 220, 2001.

d'étendre la théorie des SVM multi-classes (M-SVM) [19, 13]. Nos modèles ont trouvé une première application en identification de thèmes, dans le cadre d'une collaboration avec le projet PAROLE. L'application principale en bioinformatique est décrite dans la sous-section suivante.

5.4 Prédiction de la structure secondaire des protéines

Mots clés : Apprentissage statistique, structure secondaire des protéines.

Participant : Yann Guermeur.

Connaître la structure tridimensionnelle d'une protéine est essentiel pour en inférer la fonction. Prédire cette *structure tertiaire* à partir de la séquence d'acides aminés (*structure primaire*) demeure l'un des défis majeurs en biologie structurale. Une approche de ce problème consiste à prédire dans un premier temps la structure secondaire et à utiliser ensuite le résultat obtenu pour effectuer des calculs *ab initio* ou mettre en œuvre des algorithmes de *threading*[JH00]. Considéré du point de vue de la reconnaissance des formes, il s'agit d'un problème de discrimination consistant à associer à chaque résidu (acide aminé) d'une chaîne polypeptidique son état conformationnel (hélice, brin ou apériodique). Notre activité concernant la prédiction de la structure secondaire des protéines globulaires s'organise essentiellement autour de deux collaborations. Depuis plusieurs années, nous travaillons avec l'équipe de bioinformatique dirigée par Gilbert Deléage, à l'Institut de Biologie et Chimie des Protéines (IBCP) de Lyon. Nos recherches en cours portent sur l'exploitation de différents types d'informations, venant compléter l'information de séquence (structure primaire ou alignements multiples), afin d'améliorer les performances de systèmes de prédiction existants. Dans cette optique, nous avons commencé à développer une nouvelle version de la méthode de prédiction SOPM (self-optimized prediction method)[GD94]. Depuis août 2000, nous travaillons sur le même problème avec l'équipe de Pierre Baldi, à l'Université d'Irvine, en Californie. L'approche adoptée est cependant radicalement différente. Il s'agit ici de combiner plusieurs modules (BRNN) de l'un des systèmes de prédiction actuellement les plus performants, SSpro[BBF+99], au moyen de nos machines à vecteurs support. Cette étude s'inscrit donc dans la continuité de nos travaux sur la combinaison de modèles. Les premiers résultats expérimentaux, obtenus en combinant trois des méthodes de prédiction les plus utilisées au moyen de M-SVM, sont exposés dans [13]. Ils sont venus confirmer la théorie, qui prédisait que les SVMs devaient se révéler particulièrement efficaces pour cette tâche [19][13]. Des résultats initiaux concernant la combinaison de BRNN peuvent également être trouvés dans [17].

[JH00] D. JONES, C. HADLEY, « Threading methods for protein structure prediction », *in : Bioinformatics: Sequence, structure and databanks*, D. Higgins et W. Taylor (éditeurs), Oxford Univ. Press, 2000, p. 1–13.

[GD94] C. GEOURJON, G. DELÉAGE, « SOPM: a self-optimized method for protein secondary structure prediction », *Protein Engineering 7*, 2, 1994, p. 157–164.

[BBF+99] P. BALDI, S. BRUNAK, P. FRASCONI, G. SODA, G. POLLASTRI, « Exploiting the past and the future in protein secondary structure prediction », *Bioinformatics 15*, 11, 1999, p. 937–946.

5.5 Epissage alternatif

Mots clés : Epissage alternatif, SELEX, structure des ARN.

Participants : Damien Eveillard [en collaboration avec l'UMR CNRS 7567 MAEM], Yann Guermeur.

L'épissage alternatif intervient au niveau de la maturation des ARN messagers. Il est contrôlé par différents activateurs et inhibiteurs comme les protéines SR[LZK98]. Dans le cadre d'une thèse sous la direction de Christiane Branlant (MAEM) et Alexander Bockmayr nous cherchons à identifier les sites de fixation des protéines SR grâce à des résultats expérimentaux SELEX[LCS01]. Ce type d'expérience récupère par sélection *in vitro* des séquences nucléiques qui possèdent une forte affinité pour une protéine donnée [TG90]. Les expériences SELEX possèdent un caractère stochastique développé par des études théoriques [VHPBDG98].

L'exploitation des données SELEX se fait souvent par un alignement multiple de type local[DA00,BG98]. Le caractère stochastique du SELEX pouvant générer un biais, nous avons mis en place une analyse statistique pour tester la stabilité des résultats face à une perte de séquences. Il apparaît qu'une suppression aléatoire d'une ou deux séquences puisse modifier de manière significative le résultat de l'alignement multiple. Notre étude met ainsi en évidence que des techniques performantes d'alignement multiple ne sont pas toujours appropriées aux expériences SELEX. Les conditions d'utilisation de l'algorithme d'alignement doivent suffisamment prendre en compte la variabilité biologique qui dans notre cas est très importante. Une de nos optiques futures est de mieux caractériser cette variabilité biologique en incorporant des informations supplémentaires concernant la structure secondaire des ARN. L'autre de nos perspectives est d'intégrer les différentes connaissances au sein d'un modèle de régulation post-transcriptionnelle qui représenterait les différentes modulations des protéines SR dans l'épissage alternatif. Dans ce cadre, nous envisageons une collaboration avec Hidde de Jong du projet HELIX de l'INRIA Rhône-Alpes.

[LZK98] H. Liu, M. Zhang, A. R. Krainer, « Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins », *Genes & Development 12*, 1998, p. 211 – 221.

[LCS01] F. Lejeune, Y. Cavaloc, J. Stevenin, « Alternative Splicing of Intron 3 of the Serine/Arginine-rich Protein 9G8 Gene », *Journal of Biological Chemistry 276*, 11, 2001, p. 7850 – 7858.

[TG90] C. Tuerk, L. Gold, « Systematic evolution of ligands by exponential enrichment : RNA ligands to bacteriophage T4 DNA polymerase », *Science 249*, 1990, p. 505 – 510.

[VHPBDG98] B. Vant-Hull, A. Payano-Baez, R. Davis, L. Gold, « The mathematics of SELEX against complex targets », *Journal of Molecular Biology 278*, 1998, p. 579 – 597.

[DA00] L. Duret, S. Abdeddaim, « Multiple alignements for structural, functional, or phylogenetic analyses of homologous sequences », *in : Bioinformatics: Sequence, structure and databanks*, D. Higgins et W. Taylor (éditeurs), Oxford Univ. Press, 2000, p. 51 –76.

[BG98] T. L. Bailey, M. Gribskov, « Methods and statistics for combining motif match scores », *Journal of Computational Biology 5*, 1998, p. 211 – 221.

5.6 Analyse des séquences intergéniques chez les levures

Mots clés : Séquence intergénique.

Participant : Emmanuel Gothié [en collaboration avec l'UMR CNRS 7567 MAEM].

La somme considérable de données brutes extraites des programmes de séquençage nécessite de nouvelles techniques d'analyse. La plupart des outils actuels sont orientés vers les séquences codantes, mais il existe de très nombreux ARN non-codants (séquences intergéniques) possédant des rôles variés (structural, régulateur, catalytique, etc.). Le développement d'outils d'analyse permettant la détermination de tels ARN dans les génomes nouvellement séquencés, mais aussi la découverte de nouveaux ARN de régulation aux fonctions inconnues, est donc nécessaire. L'existence de génomes disponibles pour différents membres d'une même famille permet d'envisager une étude par comparaison de séquences. Cette approche permet de mettre en évidence des séquences essentielles conservées entres les différents génomes apparentés (sélectionnées par pression de sélection).

Une étude portant sur des génomes de cellules eucaryotes de relative petites tailles - les levures Hémiascomycètes - a été démarrée. Actuellement, seul le génome de *Saccharomyces cerevisiae* est séquencé complètement, mais une étude récente[Gén00] portant sur 13 espèces représentatives de la classe des Hémiascomycètes a donné naissance à une base de données conséquente[1] constituée de fragments d'ADN annotés par rapport au génome de *S. cerivisiae*. A partir de ces bases de données nous avons extrait une banque de séquences intergéniques qui sert de point de départ à l'étude. Les données expérimentales décrites dans la littérature telles que le contexte, ou la structure des ARN seront intégrées au sein d'algorithmes de recherche afin d'améliorer la mise en évidence de ce type d'ARN. Par exemple, en ce qui concerne la structure des ARN de régulation, des éléments structuraux définissant leur classe et leur fonction sont déjà décrits (exemple des boîtes C/D ou H/ACA dans les snoRNAs). La découverte de courtes zones très conservées dans des séquences possédant un haut score d'alignement entre différents génomes devrait aussi permettre de caractériser de nouvelles structures clefs. Ces dernières pourront s'avérer être autant de signatures d'ARN non-codants importants.

5.7 Modélisation de systèmes biologiques

Mots clés : Programmation par contraintes, système hybride, biologie systémique.

Participants : Alexander Bockmayr, Arnaud Courtois.

Nous avons commencé à étudier l'utilisation de la programmation par contraintes pour la modélisation et la simulation de systèmes biologiques.

Dans le cadre d'un stage de DEA [18], nous avons vérifié l'hypothèse selon laquelle un langage de programmation concurrente par contraintes hybrides est à même de reproduire les différentes techniques de modélisation employées en biologie systémique. Tenant compte du fait

[1] `http://www.genoscope.cns.fr/`

[Gén00] « Special Issue: Génolevures », *FEBS Letters 487*, 1, 2000, `http://cbi.labri.u-bordeaux.fr/Genolevures/biblio.php3`.

qu'un modèle biologique peut être décrit comme un ensemble de processus concurrents, gérés par des lois continues et/ou discrètes, il a été fait usage du langage `Hybrid cc`[GJSB94,GJS98] afin de capturer ces phénomènes hybrides [14]. Dans la suite de ces travaux, nous partirons de l'hypothèse que les *systèmes hybrides* sont satisfaisants pour capturer les différents aspects des systèmes biologiques et leur devenir temporel. L'objectif final est de développer un langage basé sur les contraintes, dédié au domaine de la biologie moléculaire. Il offrirait des opérateurs puissants et la possibilité de modéliser et simuler des systèmes biologiques de façon déclarative. Le choix concret des phénomènes biologiques à modéliser ne sera fait qu'avec l'aide de nos partenaires dans le domaine de la biologie.

Un premier contact dans ce champ d'étude s'est matérialisé sous la forme d'une coopération avec l'équipe de Magali Roux-Rouquié à l'Institut Pasteur de Paris. Il s'agit de créer un modèle visant à modéliser les différents mécanismes inhérents au cycle cellulaire. Un but premier est d'analyser le rôle d'une protéine, BRG1, soupçonnée d'être fondamentale vis-à-vis du point de contrôle entre les phases G1 et S du cycle cellulaire. L'intérêt informatique y est clair, car cette étude est particulièrement exhaustive par rapport à la connaissance biologique actuelle, ce qui permettra de tirer des enseignements sur la fabrication d'un modèle complexe : faisabilité, possibilités de développement incrémental, raffinements, qualité du paramétrage.

5.8 Résolution de contraintes numériques

Mots clés : Contrainte numérique, équation, inégalité.

Participant : Alexander Bockmayr.

En collaboration avec Volker Weispfenning (Université de Passau, Allemagne), nous avons donné dans [12] une présentation systématique des techniques de résolution de contraintes sur les domaines numériques. Nous traitons les contraintes linéaires sur les corps, les contraintes linéaires diophantiennes, des contraintes non-linéaires sur les domaines continus, et les contraintes non-linéaires diophantiennes.

5.9 Plans de coupe et complexité de la clôture élémentaire

Mots clés : Programmation entière, plan de coupe, complexité.

Participant : Alexander Bockmayr.

La clôture élémentaire P' d'un polyèdre P est l'intersection de P avec tous ses plans de coupe de type Gomory-Chvátal[Sch86]. P' est un polyèdre rationnel si P est rationnel. Toutes les bornes connues sur le nombre d'inégalités nécessaires pour définir P' étaient exponentielles, même en dimension fixe. En collaboration avec Fritz Eisenbrand (Max-Planck-Institut f. In-

[GJSB94] V. Gupta, R. Jagadeesan, V. A. Saraswat, D. G. Bobrow, « Programming in Hybrid Constraint Languages », *in: Hybrid Systems*, p. 226–251, 1994.

[GJS98] V. Gupta, R. Jagadeesan, V. A. Saraswat, « Computing with Continuous Change », *Science of Computer Programming 30*, 1-2, 1998, p. 3–49.

[Sch86] A. Schrijver, *Theory of Linear and Integer Programming*, 1986.

formatik, Allemagne) nous avons montré dans [11] que le nombre d'inégalités nécessaires pour définir P' est polynômial en dimension fixe.

5.10 Collaboration de résolveurs pour la résolution des problèmes d'équilibrage

Mots clés : Programmation entière, programmation par contraintes, collaboration de résolveurs, équilibrage.

Participants : Alexander Bockmayr, Nicolai Pisaruk.

Le problème d'équilibrage consiste à répartir le travail nécessaire pour fabriquer un ensemble de produits sur différents postes d'une ligne d'assemblage. Pour résoudre ces problèmes, nous développons dans [16] une approche hybride qui combine des techniques de la programmation linéaire en nombres entiers et de la programmation par contraintes sur les domaines finis.

5.11 Une contrainte globale pour les problèmes de flot

Mots clés : Programmation par contraintes, contrainte globale, problème de flot.

Participants : Alexander Bockmayr, Nicolai Pisaruk.

Les problèmes de flot jouent un rôle important en mathématiques et en informatique. Ils ont de nombreuses applications, par exemple dans le domaine des transports, en télécommunication ou pour l'optimisation de la chaîne logistique. Beaucoup de modèles de flot classiques peuvent être résolus de manière très efficace, ils admettent des solutions entières, et donnent un langage de modélisation plus naturel que par exemple la programmation linéaire.

Malgré cette importance, les systèmes de programmation par contraintes actuels n'offrent pas de support particulier pour les problèmes de flot. Dans [15], nous introduisons une nouvelle contrainte globale `flow` pour la modélisation et la résolution de problèmes de flot en programmation par contraintes. Cette contrainte est en cours d'intégration dans le système de programmation par contraintes CHIP de la société COSYTEC S.A.

6 Contrats industriels (nationaux, européens et internationaux)

6.1 LISCOS

Mots clés : Chaîne logistique, programmation entière, programmation par contraintes.

Participants : Alexander Bockmayr, Nicolai Pisaruk.

L'équipe participe au projet européen LISCOS (Large-scale Integrated Supply Chain Optimisation Software Based upon Branch & Cut and Constraint Programming Methods). Les partenaires de ce projet qui a commencé le 1/1/2000 et qui est prévu pour une durée de 3 années sont : Barbot (P), BASF (D), CORE (B), COSYTEC (F), Dash (UK), DEIO (P), LORIA (F), PSA (F), Procter and Gamble (B). L'objectif du projet est le développement de logiciels

pour la modélisation et la résolution de problèmes d'optimisation de la chaîne logistique. Ces logiciels seront basés sur une intégration de la programmation entière et de la programmation par contraintes sur les domaines finis.

7 Actions régionales, nationales et internationales

7.1 Actions régionales

Nous participons au Génopole Strasbourg Alsace-Lorraine avec comme partenaire le MAEM UMR 7567 à Nancy et l'IGBMC à Strasbourg.

Dans le cadre du CPER 2000-2006 pour la Région Lorraine, nous participons également au projet « Bioinformatique et Applications à la Génomique » du Pôle de Recherche Scientifique et Technologique « Intelligence Logicielle ». Nos partenaires ici sont le laboratoire de cristallographie LCM3B, UMR CNRS 7036, Equipe « Théorie, phasage » et le laboratoire UMR CNRS 7567 « Maturation des ARN et Enzymologie Moléculaire » à l'Université Henri Poincaré, Nancy 1.

7.2 Actions nationales

Nous participons au groupe de recherche du STIC–CNRS « Mathématiques des systèmes perceptifs et cognitifs » (MSPC), ainsi qu'aux groupes thématiques « Analyse systématique des structures tridimensionnelles et des interactions » et « Bioinformatique Fonctionnelle des Systèmes de Régulations Génétiques » de l'Action Ministérielle IMPG : Informatique, Mathématique, Physique pour la Génomique.

7.3 Actions européennes

Dans le cadre du programme « Growth » de la Commission Européenne, nous participons au projet de recherche LISCOS (Large-scale Integrated Supply Chain Optimisation Software), Contrat No. G1RD-CT-1999-000034.

Nous participons aussi aux groupe de travail ERCIM *Constraints* coordonné par K. Apt (CWI, Amsterdam).

7.4 Actions internationales

Nous entretenons des relations avec l'Institut pour les Problèmes Mathématiques en Biologie (IMPB) de l'Académie des Sciences de la Russie à Poushchino (Vladimir Y. Lunin), ainsi qu'avec l'Université Carnegie-Mellon à Pittsburgh (Egon Balas, John N. Hooker) et l'Université de Californie à Irvine (Pierre Baldi).

7.5 Visites et invitations de chercheurs

Natasha Lunina, Institut des Problèmes Mathématiques en Biologie, Académie des Sciences de la Russie, a travaillé un mois dans l'équipe sur la détermination des enveloppes macromoléculaires.

8 Diffusion de résultats

8.1 Animation de la communauté scientifique

Alexander Bockmayr est responsable de l'action « Bioinformatique » du LORIA et de l'INRIA Lorraine, responsable du projet « Bioinformatique et Applications à la Génomique » du PRST Intelligence Logicielle ; membre du conseil scientifique du PRST Intelligence Logicielle ; membre du comité de coordination de la bioinformatique des génopoles ; membre du *Steering Committee* du projet européen LISCOS ; Associate Editor de INFORMS J. Computing ; coodinateur de « Optimization Online », `http://www.optimization-online.org` ; membre des comités de programme de CPAIOR'2001 et de WFLP'2001 ; co-responsable de la filière « Algorithmique numérique et symbolique » du DEA Informatique ; membre du conseil de laboratoire du LORIA ; membre du comité de projets du LORIA et de l'INRIA Lorraine ; membre du conseil d'orientation scientifique du LORIA ; membre suppléant de la Commission de Spécialistes 27e section de l'Université Henri Poincaré, Nancy 1, et de l'Université de Metz.

Eric Domenjoud est membre de la section 07 du Comité National de la Recherche Scientifique.

Yann Guermeur est membre de la commission de choix de l'IUT de Saint-Dié des Vosges, correspondant pour le LORIA du groupe de recherche du STIC - CNRS « Mathématiques des systèmes perceptifs et cognitifs » (MSPC), membre du groupe thématique « Analyse systématique des structures tridimensionnelles et des interactions » de l'action IMPG, ainsi que du groupe de travail ESPRIT « Neural Networks and Computational Learning Theory » (NeuroCOLT2).

Miki Hermann est membre élu du Comité Scientifique du département STIC au CNRS.

8.2 Enseignement

Alexander Bockmayr et Yann Guermeur sont des enseignants-chercheurs à l'Université Henri Poincaré, Nancy 1. Ils assurent une partie de leur service avec des enseignements de bioinformatique (Maîtrise « Biologie Cellulaire et Physiologie » ; DESS « Ressources Génomiques et Traitements Informatiques »).

Miki Hermann a enseigné la théorie du codage et l'imagerie numérique aux étudiants de la 2^e année d'IUP GEII, spécialisation « Réseaux Numériques de Communication », à l'Université Henri Poincaré pendant les années universitaires 2000-2001 et 2001-2002. Il a également enseigné une partie du cours du DEA Informatique ANS3 (la partie codage) en 2000-2001 et une partie du cours du DEA Informatique ANS2 (la partie complexité) en 2001-2002.

8.3 Divers

Alexander Bockmayr a participé aux jury de thèse de Guy A. Narboni à l'ENS Cachan.

Miki Hermann a donné avec Nadia Creignou und Reinhard Pichler un tutoriel sur la complexité des problèmes de résolution de contraintes à la Conférence « Principles and Practice of Constraint Programming (CP'2001) » à Paphos en Chypre.

9 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] A. Bockmayr, F. Eisenbrand, « Cutting Planes and the Elementary Closure in Fixed Dimension », *Mathematics of Operations Research 26*, 2, 2001, p. 304–312.

[2] A. Bockmayr, T. Kasper, « Branch-and-Infer : A unifying framework for integer and finite domain constraint programming », *INFORMS J. Computing 10*, 3, 1998, p. 287 – 300.

[3] A. Bockmayr, V. Weispfenning, « Solving numerical constraints », *in : Handbook of Automated Reasoning*, A. Robinson et A. Voronkov (éditeurs), *1*, Elsevier, Amsterdam, 2001, ch. 12, p. 751–842.

[4] E. Domenjoud, C. Kirchner, J. Zhou, « Generating feasible schedules for a pick-up and delivery problem », *Electronic Notes in Discrete Mathematics 1*, 1999.

[5] E. Domenjoud, A. Tomás, « From Elliott-MacMahon to an Algorithm for General Linear Constraints on Natural », *in : Proceedings 1st International Conference on Principles and Practice of Constraint Programming, Cassis, Lecture Notes in Computer Science, 976*, Springer Verlag, p. 18–35, septembre 1995.

[6] A. Durand, M. Hermann, L. Juban, « On the Complexity of Recognizing the Hilbert Basis of a Linear Diophantine System », *Theoretical Computer Science*, 2002, à paraître.

[7] Y. Guermeur, C. Geourjon, P. Gallinari, G. Deléage, « Improved performance in protein secondary structure prediction by inhomogeneous score combination », *Bioinformatics 15*, 5, 1999, p. 413–421.

[8] Y. Guermeur, « Combining Discriminant Models with new Multi-Class SVMs », *Pattern Analysis and Applications Journal*, 2002, à paraître.

[9] M. Hermann, P. G. Kolaitis, « Computational Complexity of Simultaneous Elementary Matching Problems », *Journal of Automated Reasoning 23*, 2, août 1999, p. 107–136.

[10] M. Hermann, P. G. Kolaitis, « Unification Algorithms Cannot Be Combined in Polynomial Time », *Information and Computation 162*, 1-2, 2000, p. 24–42.

Articles et chapitres de livre

[11] A. Bockmayr, F. Eisenbrand, « Cutting Planes and the Elementary Closure in Fixed Dimension », *Mathematics of Operations Research 26*, 2, 2001, p. 304–312.

[12] A. Bockmayr, V. Weispfenning, « Solving numerical constraints », *in : Handbook of Automated Reasoning*, A. Robinson et A. Voronkov (éditeurs), *1*, Elsevier, Amsterdam, 2001, ch. 12, p. 751–842.

[13] Y. Guermeur, « Combining Discriminant Models with new Multi-Class SVMs », *Pattern Analysis and Applications Journal*, 2002, à paraître.

Communications à des congrès, colloques, etc.

[14] A. Bockmayr, A. Courtois, « Modeling biological systems in hybrid concurrent constraint programming (Poster) », *in : 2nd Int. Conf. Systems Biology, ICSB'01, Pasadena, CA*, 2001.

[15] A. Bockmayr, N. Pisaruk, A. Aggoun, « Network flow problems in constraint programming », *in : Principles and Practice of Constraint Programming, CP'2001, Paphos, Cyprus, LNAI, 2239*, Springer, p. 196 – 210, 2001.

[16] A. Bockmayr, N. Pisaruk, « Solving Assembly Line Balancing Problems by Combining IP and CP », *in : Sixth Annual Workshop of the ERCIM Working Groupon Constraints, Prague, Czech Republic*, juin 2001, `http://arXiv.org/abs/cs.DM/0106002`.

[17] Y. Guermeur, D. Zelus, « Combining Protein Secondary Structure Prediction Models with Ensemble Methods of Optimal Complexity », *in : JOBIM 2001, Toulouse, France*, p. 97–104, juin 2001.

Rapports de recherche et publications internes

[18] A. Courtois, « Modélisation de systèmes biologiques en programmation par contraintes », *Rapport de DEA*, Univ. Henri Poincaré, LORIA, juillet 2001.

[19] A. Elisseeff, Y. Guermeur, H. Paugam-Moisy, « Margin error and generalization capabilities of multi-class discriminant models », *Rapport de recherche n° NC-TR-99-051-R*, NeuroCOLT2, 1999, revisé en 2001.

[20] V. Y. Lunin, A. Urzhumtsev, A. Bockmayr, « Direct phasing by binary integer programming », *Rapport de recherche n° A01-R-307*, LORIA, octobre 2001.