INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Équipe 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
Sophie Drouot [INRIA]

Personnel CNRS
Miki Hermann [CR, jusqu'au 31/3/2002]

Personnel Université
Yann Guermeur [Maître de Conférences, Université Henri Poincaré, Nancy 1]

Chercheurs doctorants
Arnaud Courtois [UHP, cofinancé région Lorraine]
Damien Eveillard [UHP, cofinancé région Lorraine]

Chercheurs post-doctorants
Emmanuel Gothié [UHP et INRIA]
Nicolai Pisaruk [UHP]

Chercheur invité INRIA
Vladimir Lunin [de 4/2002 à 7/2002]

Stagiaires
Damien Denis [DESS Informatique, Reims, 5/2002-9/2002]
Yasmine Khan [DEA Informatique, Nancy, 2/2002-7/2002]
Laïka Moussa [Maîtrise Informatique, Nancy, 7/2002-8/2002]
Deo Prakash Pandey [IIT Kanpur, Inde, 7/2002-8/2002]
Sébastien Vachenc [DESS RGTI, Nancy, 10/2002-12/2002]
Régis Vert [DEA Informatique, Nancy, 2/2002-9/2002]
Myriam Vezain [DESS EGOISt, Rouen, depuis 11/2002]

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 de l'apprentissage statistique.

2.1.1. Projets actuels

1. Détermination de la structure des macromolécules biologiques

   – Détermination et analyse des enveloppes macromoléculaires
   – Interactions entre ARN et ligands
   – Prédiction de la structure secondaire des protéines globulaires

2. Structures fonctionnelles
   - Motifs fonctionnels intergéniques
   - Analyse expérimentale des interactions ARN/protéine
3. Modélisation de la fonction de systèmes biologiques
   - Modélisation par programmation concurrente avec contraintes hybrides
   - Application à l'épissage alternatif

2.1.2. Principaux axes de recherche en informatique

- Programmation par contraintes ;
  - Contraintes sur les domaines finis et optimisation discrète ;
  - Programmation concurrente par contraintes.
- Apprentissage statistique et réseaux de neurones.

2.1.3. 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'ARC INRIA « Calculs de Processus et Biologie des Réseaux Moléculaires »
- Participation à l'action IMPG ;
- Participation au projet européen LISCOS ;
- Nombreuses collaborations nationales et internationales :
  - Laboratoire MAEM (Maturation des ARN et Enzymologie Moléculaire), Nancy ;
  - Laboratoire de Cristallographie LCM3B, Nancy ;
  - Institut des Problèmes Mathématiques en Biologie, Académie des Sciences, Poushchino, Russie ;
  - Institut de Biologie et de Chimie des Protéines, Univ. Claude Bernard, Lyon ;
  - Université de Californie, Irvine, Etats-Unis.

3. Fondements scientifiques

3.1. Programmation par contraintes

La programmation par contraintes est un nouveau paradigme de programmation qui a été proposé dans les années 80 et qui est de plus en plus utilisé depuis [44].

Une *contrainte* est une formule logique qui contient des variables et qui définit 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.

En *programmation par contraintes*, l'utilisateur programme avec des contraintes, c.-à-d. qu'il décrit un problème avec un ensemble de contraintes qui peuvent être liées avec différents *combinateurs* : conjonction, disjonction, opérateurs temporels (always), etc. Chaque contrainte donne une information *partielle* sur l'état du système étudié. Les logiciels de programmation par contraintes permettent de déduire de nouvelles contraintes à partir des contraintes données et de calculer des *solutions*, c.-à-d. des valeurs pour les variables qui satisfont simultanément l'ensemble des contraintes.

L'objectif général de la programmation par contraintes est de développer des langages de programmation dans lesquels on peut exprimer de manière naturelle des problèmes de contraintes et les résoudre efficacement.

3.1.1. Contraintes sur les domaines finis et optimisation discrète

Dans notre recherche, nous nous intéressons tout d'abord aux problèmes de contraintes sur les domaines finis. Le domaine de chaque variable (l'ensemble des valeurs qu'elle peut prendre) est alors un sous-ensemble fini des entiers naturels. La théorie nous enseigne que la plupart de ces problèmes sont NP-difficiles, ce qui signifie qu'il est très peu probable qu'on puisse résoudre ces problèmes par des algorithmes polynomiaux 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.-à-d. 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`$(x_1, ..., x_n)$ exprime que les variables $x_1, ..., x_n$ 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.

Un cadre unificateur pour la programmation entière et la programmation par contraintes sur les domaines finis est développé dans [1]. Une présentation systématique des techniques de résolution de contraintes sur les domaines numériques se trouve dans [2].

3.1.2. Programmation concurrente par contraintes

En programmation *concurrente* par contraintes, différents processus peuvent s'exécuter de façon concurrente [40]. Les interactions sont rendues possibles à l'aide d'un *pot de contraintes*, commun et accessible par tous les processus. Il contient toutes les contraintes du système connues à ce moment. Un processus peut *ajouter* une contrainte au pot (*tell*), ou *questionner* le pot (*ask*) pour savoir si une contrainte est actuellement déductible ; une action est alors décidée.

La programmation concurrente par contraintes *hybrides*, `Hybrid cc`[32][33], extension de la programmation concurrente par contraintes, permet de modéliser et de simuler le devenir temporel de *systèmes hybrides*, c.-à-d. des systèmes ayant pour caractéristique de pouvoir changer d'état de façon discrète ou continue [43]. Les contraintes en `Hybrid cc` peuvent être algébriques ou sous forme d'équations différentielles. Les changements d'état peuvent être spécifiés en utilisant les combinateurs de la programmation concurrente par contraintes et de la logique *default*.

3.2. Apprentissage statistique

La théorie de l'apprentissage statistique [46] 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 celui 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), qui obtiennent actuellement les meilleures performances sur de nombreuses tâches relevant des principaux domaines de la reconnaissance des formes.

Les SVM sont des modèles connexionnistes conçus pour effectuer des tâches de discrimination (calcul de dichotomies), d'approximation de fonctions ou d'estimation de densité. Elles ont été introduites relativement récemment [24][28] comme extensions non linéaires de l'hyperplan de marge maximale [45]. Leur principale qualité est de permettre une bonne généralisation dans le cas des petits échantillons. [46][26][31][29]

4. Domaines d'application

4.1. Panorama

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

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

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

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 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* qui conduit à la formation de l'ARN messager mature (ARNm), par élimination des parties non codantes (*introns*) et concaténation des parties codantes (*exons*).

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.

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 *brins* β, et les structures non-hélice et non-brin, nommées *apériodiques*. La structure tertiaire correspond au repliement global de la protéine et comprend les coordonnées 3D de tous ses atomes. 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.

L'objectif final de la biologie moléculaire est de comprendre la fonction des macromolécules biologiques au niveau de la vie de la cellule. Cette fonction résulte de *l'interaction* entre différentes macromolécules et dépend de leurs structures. Le défi est trouver un chemin partant de la séquence, passant par la structure, pour finalement appréhender la fonction.

4.3. Cristallographie

Participants : Alexander Bockmayr, Eric Domenjoud.

Mots clés : *Cristallographie*, *macromolécule*, *phasage*.

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

Participants : Alexander Bockmayr, Eric Domenjoud, Nicolai Pisaruk.

Mots clés : *Recherche opérationnelle*.

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. Logiciels

5.1. M-SVM1

Participant : Yann Guermeur.

Nous avons rendu disponible le logiciel de la M-SVM nommée M-SVM1 dans [6] à l'adresse suivante : http://www.loria.fr/~guermeur/. La diffusion s'effectue avec une licence GNU GPL. L'algorithme mis en œuvre pour résoudre le problème de programmation quadratique correspondant à l'apprentissage est l'algorithme de Frank et Wolfe. Le logiciel résout donc une série de programmes linéaires. La conception modulaire de l'application permet à l'utilisateur d'employer pour ce faire, outre la méthode proposée par défaut, le « solveur » de son choix.

6. Résultats nouveaux

6.1. Programmation entière et problème du phasage en cristallographie

Participants : Alexander Bockmayr, Eric Domenjoud.

Mots clés : *Cristallographie*, *phasage*, *programmation entière*.

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 [38]. 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 montré que le problème de phasage en radiocristallographie peut être traduit en un problème de programmation entière en variables 0-1 [8][7]. L'idée de base est de binariser les magnitudes et les phases des coefficients de Fourier de la densité électronique. Cela permet de remplacer les équations pour les facteurs de structure par un système d'inégalités linéaires en variables 0-1. Nous avons évalué notre approche sur une structure de protéine connue. La limite actuelle est la taille de la grille tridimensionnelle utilisée dans la discrétisation. Même si la grille est très petite, nous pouvons obtenir des informations utiles qui peuvent servir comme point de départ pour une amélioration des phases par d'autres techniques.

6.2. Interaction ARN/Ligand

Participants : Alexander Bockmayr, Eric Domenjoud.

Mots clés : *ARN*, *ligand*, *contraintes*.

Nous nous intéressons, en collaboration avec Fabrice Leclerc du laboratoire MAEM, au problème de la synthèse de ligands qui vont interagir avec la molécule d'ARN.

Etant donnés deux sites de fixation dans la molécule d'ARN, nous recherchons une chaîne d'atomes qui relient ces deux sites. Cette chaîne, qui constitue le squelette du ligand, doit en outre d'une part minimiser l'énergie de Van der Waals de la molécule et d'autre part, respecter des contraintes géométriques telles que le fait que le ligand ne traverse pas la molécule d'ARN à laquelle il se fixe. Les longueurs des liaisons et les angles qu'elles forment dépendent du type des atomes (carbone, oxygène, azote, ...) et du type des liaisons inter-atomiques (simple, double).

Nous avons en premier lieu développé un modèle correspondant à une situation *idéale* où le squelette du ligand ne comporte que des atomes de carbone avec des liaisons simples, en considérant que les liaisons inter-atomiques sont de longueur fixe et que les angles entre ces liaisons sont uniformes. Dans ce cas, les valeurs spécifiques des angles entre ces liaisons et des angles de torsion de la molécule permettent en première approximation d'obtenir un modèle du problème en variables 0/1. Nous pouvons alors appliquer à ce modèle des techniques d'optimisation en variables 0/1.

La réalité est toutefois plus complexe. Ni les longueurs des liaisons inter-atomique, ni les angles de la molécule (angles entre les liaisons et angles de torsions) ne sont homogènes. Tous ces paramètres dépendent du type des liaisons et peuvent subir de petites variations. Nous travaillons actuellement à un modèle permettant de rendre compte de cette complexité.

Une fois fixés les types des atomes du squelette et des liaisons entre ces atomes, les divers paramètres ne subissent plus que des variations faibles. Une première solution approchée est calculée puis affinée par une recherche locale.

Cette modélisation devra ensuite être enrichie en y intégrant notamment une modélisation de l'environnement. D'une part, le ligand ne doit pas traverser la molécule d'ARN et d'autre part, la géométrie de cette molécule elle-même peut subir de petites variations.

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

Participant : Yann Guermeur.

Mots clés : *Apprentissage statistique*, *SVM*.

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 continuer à combler cette lacune, nous avons poursuivi notre collaboration avec André Elisseeff et Dominique Zelus, collaboration portant sur l'étude des lois fortes des grands nombres uniformes et les mesures de capacité des familles de fonctions à valeurs vectorielles. Nous avons ainsi étendu les principaux résultats de la théorie des classifieurs « à grande marge » au cas des fonctions à valeurs dans $\mathbb{R}^Q$. Ceci nous a conduits à introduire une nouvelle dimension de Vapnik-Chervonenkis étendue, la M-fat-shattering dimension [18], qui généralise au cas multi-classe la fat-shattering dimension et introduit la notion de marge dans la dimension graphique. Dans [15], cette quantité a été bornée pour les modèles de régression linéaire multivariée, i.e. l'architecture partagée par l'ensemble des machines à vecteurs support multi-classes (M-SVM) publiées à ce jour. Ceci nous a permis de munir ces machines d'un cadre théorique unificateur. Ce cadre est exposé en détails dans [19]. Il est actuellement utilisé pour effectuer une étude comparative des différentes M-SVM. Les bornes qu'il fournit servent également de référence pour évaluer les avantages résultant de l'emploi de mesures de capacité empiriques.

6.4. Prédiction de la structure secondaire des protéines

Participants : Yann Guermeur, Régis Vert.

Mots clés : *Apprentissage statistique*, *structure secondaire des protéines*.

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 de la protéine. 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 repose essentiellement sur une collaboration avec l'équipe de Pierre Baldi, à l'Université d'Irvine, en Californie. Nous avons poursuivi l'étude de la combinaison des modules (BRNN) de la méthode de prédiction SSpro2 au moyen de M-SVM. Les résultats des expériences font apparaître un accroissement statistiquement significatif du taux de reconnaissance.

Parallèlement, le stage de DEA de Régis Vert [21], portant sur la conception et la mise en œuvre de machines à noyau dédiées au traitement de séquences biologiques, a trouvé son application dans le développement d'une méthode de prédiction exploitant directement les données présentées en entrée des BRNN, à savoir des profils d'alignement multiple issus de PSI-BLAST. Cette étude met en évidence l'importance du choix du noyau pour ce type de tâche.

6.5. Analyse des séquences interORF chez les levures - Application à la recherche des gènes d'ARN non codants (ARNnc)

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

Mots clés : *Séquence interORF*, *ARNnc*, *Génomes*, *SVM multi-classes*.

La somme considérable de données brutes extraites des programmes de séquençage nécessite de nouvelles techniques d'analyse. La première étape dans l'exploitation des génomes consiste à rechercher les régions codantes des protéines ORF (*Open Reading Frame*). Les séquences situées entre ces ORF, qui peuvent coder des ARN stables ou des ARN régulateurs, sont plus difficiles à étudier bien que très importantes. Le développement d'outils d'analyse appropriés à l'étude de ces régions est donc un challenge de premier ordre dans l'ère post-génomique.

Une étude portant sur des génomes de cellules eucaryotes de relatives petites tailles - les levures Hémiascomycètes - a été démarrée. En plus du génome complet de *Saccharomyces cerevisiae*, une étude récente - le projet Génolevures [30] - portant sur 13 espèces représentatives de la classe des Hémiascomycètes a donné lieu à une base de données conséquente [1] constituée de fragments d'ADN annotés par rapport au génome de *S. cerivisiae*. A partir de ces données nous avons extrait des banques de séquences interORF dont 17 correspondent aux chromosomes de *S. cerevisiae* et 13 aux génomes de Génolevures (59400 fragments). Parallèlement au développement de ces banques de données, les séquences relatives aux snoRNA connus pour *S. cerevisiae* ont été recherchées dans des bases de données publiques (Eddy Lab snoRNA Database [2] & Dmitry A. Samarsky and Maurille J. Fournier Database [3]) et environ 70 séquences ont été obtenues.

A côté de l'approche qui a permis de générer, pour le cas de Génolevures, des bases de données interORF par similitude au génome de référence *S. cerevisiae* (alignement par Blast), nous avons choisi d'utiliser les SVM (*Support Vector Machine*) [46] afin d'obtenir, par une autre approche, les séquences interORF. En effet, en utilisant comme référence le génome de *S. cerevisiae* et en analysant directement les séquences RST (Random Sequence Tag) des différents génomes dans leur intégralité, nous espérons obtenir des bases de données pertinentes. Le sujet traité s'exprimant comme un problème de discrimination à catégories multiples, nous utiliserons des SVM multi-classes (M-SVM), qui ont été conçues et implémentées dans l'équipe [6], au lieu des SVM (bi-classes) classiques.

L'apprentissage se fera sur la base de données de *S. cerevisiae* (14-15 séquences de chromosomes). Le test sera réalisé sur les séquences restantes. La validation de cette étape, avec éventuellement mise au point, par ajout d'autres paramètres (signaux de début et de fin d'ORF etc.) permettra de l'appliquer sur la prédiction de séquences interORF. Ainsi, ce test sera appliqué sur les bases de données brutes des RST du projet Génolevures en vue de l'obtention d'une base de données d'interORF.

La comparaison des différentes bases de données obtenues (par similitude à *S. cerevisiae* ou par discrimination) nous permettra de juger de l'intérêt de ces modèles de reconnaissance de formes dans l'exploitation des bases de données génomiques. Si l'analyse par discrimination s'avère plus efficace ou complémentaire, ce type d'analyse constituera un outil généralisable intéressant. La seconde application directe des M-SVM sera la recherche de séquences d'intérêt - snoRNA à boîtes C/D ou H/ACA - sur les bases de données interORF de levures (13 génomes).

Le projet a fait l'objet d'un poster présenté en octobre 2002 lors de la conférence « 4e Rencontre SifrARN » à Nancy [14].

6.6. Traitement de données SELEX

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

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

Les interactions ARN-Protéines jouent un rôle fondamental dans le fonctionnement cellulaire. Des travaux expérimentaux mettent en évidence les motifs nucléiques qui interviennent dans ces interactions. Les expériences SELEX permettent ce type de résultats [42]. Elles caractérisent les ligands potentiels d'une protéine dans une banque d'oligonucléotides générée aléatoirement par un processus de sélection *in vitro*. Cette méthode permet

[1] http://www.genoscope.cns.fr/
[2] http://rna.wustl.edu/snoRNAdb/Sc/Sc-snos-bysno.html
[3] http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_DataBase.html

plus précisément de sélectionner les motifs nucléiques qui auront pour fonction de se lier à la protéine donnée. Les analyses classiques des résultats expérimentaux mettent ensuite en évidence les séquences consensus des motifs ainsi sélectionnés [22]. Il apparaît cependant que les différents résultats obtenus avec cette approche ne sont pas consistants.

En collaboration avec le laboratoire « Maturation des ARN et Enzymologie Moléculaire » (MAEM) de Nancy (Dir. Christiane Branlant), nous étudions alors ce problème sous une approche différente en intégrant la variabilité biologique propre aux interactions ARN-Protéine. Notre première démarche est d'évaluer la justification analytique et biologique des consensus. En effet, après traitement statistique, il apparaît que les consensus obtenus sont peu stables sur certaines protéines. C'est pourquoi, nous proposons dans un deuxième temps, une nouvelle méthode analytique pour interpréter les résultats de SELEX. Cette approche s'appuie sur l'utilisation d'outils d'analyse de données qui partitionnent les données SELEX en ensembles homogènes. Cette partition expérimentale permet ainsi de générer de nouvelles séquences consensus qui sont cette fois-ci stables. Ces deux approches sont développées dans [13][12]. Notre démarche analytique permet de conserver la variabilité biologique que possèdent certaines interactions ARN-Protéines comme celles illustrées par les protéines SR. Fort de ces premiers résultats, nous avons confronté cette démarche à des données expérimentales dont les données nucléiques sont naturellement réparties suivant un critère structural. Une fois encore la classification obtenue converge vers celle déterminée expérimentalement suivant la structure secondaire des ARN. Les résultats préliminaires confortent alors la démarche de classification des données SELEX. Ainsi, nous travaillons actuellement à l'étude théorique d'une méthode issue des outils connexionnistes dédiés au traitement expérimental des acides nucléiques. Cette démarche reposerait sur l'emploi des *string kernel* [37] afin de comparer les séquences entre elles. Des outils de classification connexionnistes tels que les *SVC* [23] et les *Kernel-PCA* [41] seraient alors appropriés dans un cadre théorique validé. Cette démarche issue du connexionnisme permettra d'obtenir un outil puissant de par sa capacité de généralisation des informations de diverses natures issues des expériences SELEX. Les motifs ainsi caractérisés seront alors représentés par leur séquences mais aussi par d'autres critères biologiques.

Nous étudions actuellement la possibilité d'optimiser cette étude par une approche connexionniste. L'utilisation de SVM permettrait en effet d'incorporer des descripteurs supplémentaires représentants de la structure secondaire des ARN. Après adaptation des SVM au problème, nous envisageons d'utiliser cette même méthode sur la reconnaissance de motifs fonctionnels du génome HIV, travail initié au cours du stage de maîtrise de Sébastien Vachenc. Cette dernière démarche *in silico* permettra d'établir une cartographie fonctionnelle du génome HIV. Une recherche préliminaire permet dans un premier temps de structurer le génome par les signaux connus nécessaires à l'épissage. Il est ensuite possible sur cette base de rechercher les motifs plus fins de fixation des protéines SR. Cette dernière extension connexionniste généralisera à des génomes mal connus l'expertise fonctionnelle expérimentale acquise grâce aux expériences SELEX.

6.7. Modélisation de systèmes biologiques

Participants : Alexander Bockmayr, Arnaud Courtois, Damien Eveillard, Yasmine Khan.

Mots clés : *Programmation par contraintes*, *système hybride*, *biologie des systèmes*.

Actuellement les projets génome, transcriptome ou protéome, dont le but est de déterminer complètement tous les gènes, ARN ou protéines d'un organisme donné, produisent une quantité de données qui augmente exponentiellement. Le défi principal consiste maintenant à exploiter toutes ces données, et à comprendre comment les différents composants d'un système biologique (*i.e.* les gènes, ARN, protéines, *etc.*) interagissent pour former des fonctions biologiques complexes. La *biologie des systèmes* est un nouveau domaine de recherche qui vise à la compréhension des systèmes biologiques à différents niveaux [35]. Alors que la biologie conventionnelle examine de façon isolée un gène ou une protéine, la biologie des systèmes étudie des interactions complexes à différents niveaux d'information biologique - ADN génomique, ARNm, protéines, voies d'information et réseaux moléculaires - afin de comprendre comment ils fonctionnent ensemble.

Nous avons commencé à étudier l'utilisation de la programmation par contraintes pour la modélisation et la simulation de systèmes biologiques. Comme un système biologique peut être décrit par un ensemble de

processus concurrents, gérés par des lois continues et/ou discrètes, nous avons utilisé comme point de départ le langage `Hybrid cc` [32][33].

Dans [10][11], nous montrons que le langage `Hybrid cc` est bien approprié pour la modélisation de systèmes biologiques. Il existe une correspondance naturelle entre les possibilités de `Hybrid cc` et des phénomènes à modéliser en biologie. En particulier, la programmation par contraintes semble bien adaptée pour exploiter au mieux des informations *partielles* sur la dynamique d'un système, qui correspondent à l'état actuel des connaissances en biologie. Dans un stage de DEA [20], `Hybrid cc` a été comparé avec d'autres formalismes, notamment le π-calcul et les réseaux qualitatifs.

Dans une seconde phase, nous comptons nous servir de `Hybrid cc` pour l'étude de phénomènes biologiques nouveaux. Une partie du travail passé et actuel consiste à étudier un modèle biologique complexe, dans le but de développer le formalisme `Hybrid cc`. Il s'agit de modéliser une carte d'interactions de protéines [36] intervenant dans le cycle cellulaire eucaryote. Plus précisément, nous nous sommes intéressés à un sous-réseau, la *cyclin box*, qui est à l'étude conjointement avec les autres membres de l'action de recherche coopérative INRIA « Calculs de processus et biologie des réseaux moléculaires ». Jusqu'à présent, plus de 200 équations génériques ont été écrites pour constituer ce sous-réseau complexe. Deux buts importants de la modélisation sont de procéder à l'analyse structurelle ou logique du réseau, ainsi que la recherche ou la validation de propriétés d'intérêt biologique.

6.8. Régulation de l'épissage alternatif

Participants : Alexander Bockmayr, Damien Eveillard [en collaboration avec l'UMR CNRS 7567 MAEM], Deo Prakash Pandey, Myriam Vezain.

Mots clés : *Epissage alternatif*, *modélisation*, *validation qualitative*.

Un autre domaine d'application au cœur de nos intérêts biologiques, est le rôle des protéines SR durant l'épissage alternatif. Ce processus biologique intervient dans la maturation des ARN. De récentes études expérimentales [39] insistent sur l'importance des protéines SR dans la régulation globale de l'épissage alternatif. Ce dernier phénomène prend concrètement toute son ampleur au cours du cycle de vie de HIV, justifiant ainsi l'enjeu de la modélisation de l'épissage. Néanmoins, les modèles existants de HIV [34] ne tiennent pas compte de la complexité de la régulation de l'épissage alternatif par les protéines SR. Nous proposons de compenser cette lacune en modélisant le rôle des protéines SR au cours de l'épissage alternatif. Cette approche est possible grâce à l'utilisation de données expérimentales fournies par une collaboration avec le laboratoire MAEM. De nombreux formalismes sont envisageables pour modéliser cette régulation. Une réflexion concertée avec Hidde de Jong du projet HELIX (INRIA Rhône-Alpes) nous a permis d'isoler trois approches que nous désirons confronter afin de modéliser au mieux la problématique biologique. La première consiste à formuler un modèle discret en relation avec les résultats expérimentaux à notre disposition [48]. Un deuxième formalisme probabiliste [25] permet de représenter les variations qualitatives de l'épissage alternatif. Le troisième type de formalisme utilise des équations différentielles pour effectuer un modèle continu [17]. Cette dernière méthode finalise ainsi un gradient de formalismes couvrant les domaines discrets à continus adaptables à notre problématique biologique. Après validation qualitative, le modèle fonctionnel obtenu pourra s'inscrire en aval du modèle de structure fonctionnelle de HIV présentée précédemment en section 6.6. On sera alors en possession d'une étude globale de l'action des protéines SR sur l'épissage alternatif consistant en un modèle de structure interagissant sur un modèle fonctionnel.

Le modèle fonctionnel met en évidence la régulation de l'épissage qui peut se manifester par un rendement d'épissage. Nous nous sommes alors interessés à l'importance de ce paramètre dans le cycle de vie de HIV-1 tel qu'il est décrit dans [34]. Deo Prakash Pandey s'est attaqué à cette tâche en testant la flexibilité du comportement d'un modèle global face à des fluctuations du rendement d'épissage. Il apparaît que le critère de rendement possède une importance non négligeable dans les modèles validés jusqu'alors, confirmant ainsi l'intérêt d'étudier l'épissage du virus HIV-1.

C'est dans l'optique de perfectionner le modèle existant que Myriam Vezain, actuellement en stage de DESS, travaille sur la modélisation de l'épissage alternatif.

6.9. Disjonction de polytopes

Participants : Alexander Bockmayr, Nicolai Pisaruk.

Mots clés : *Programmation entière*, *disjonctions*, *plans de coupe*.

En collaboration avec E. Balas (CMU Pittsburg, Etats-Unis) et L. Wolsey (CORE, Belgique) nous avons caractérisé l'enveloppe convexe d'une union de polytopes par des inégalités linéaires [16]. Nous avons obtenu une description dans le même espace qui n'utilise pas de nouvelles variables. Si les polytopes sont monotones, nous pouvons caractériser de manière explicite les facettes de l'enveloppe convexe et donner un algorithme efficace pour la séparation. Nos résultats généralisent des travaux sur les règles de cardinalité [47] et sur les unions de polyèdres de matroides [27].

6.10. Contraintes symboliques pour la programmation entière

Participant : Alexander Bockmayr.

Mots clés : *Programmation entière*, *contraintes symboliques*.

Les contraintes symboliques sont une idée clef de la programmation par contraintes qui peut être introduite aussi en programmation entière [1]. En collaboration avec T. Kasper (SAP, Allemagne), E. Althaus et K. Mehlhorn (MPI Sarrebruck, Allemagne), M. Elf et M. Jünger (Univ. Cologne, Allemagne) nous avons participé à la conception et la réalisation d'un logiciel SCIL (Symbolic Constraints in Integer Linear Programming) qui permet de définir et d'utiliser des contraintes symboliques dans le cadre d'un algorithme branch-and-cut-and-price [9] (http://www.mpi-sb.mpg.de/SCIL/). Une bibliothèque de contraintes symboliques pour la programmation entière est en cours de construction.

7. Contrats industriels

7.1. LISCOS

Participants : Alexander Bockmayr, Nicolai Pisaruk.

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

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 en janvier 2000 et qui devrait se terminer en mars 2003, 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 sont basés sur une intégration de la programmation entière et de la programmation par contraintes sur les domaines finis.

8. Actions régionales, nationales et internationales

8.1. 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 et le laboratoire UMR CNRS 7567 « Maturation des ARN et Enzymologie Moléculaire » à l'Université Henri Poincaré, Nancy 1.

8.2. Actions nationales

Depuis février 2002 nous participons à l'Action de Recherche Coopérative INRIA « Algèbres de processus et réseaux de biologie moléculaire ». Nos partenaires sont le projet Contraintes de l'INRIA (F. Fages), la société Hybrigenics (V. Schächter), l'Institut Pasteur à Paris (M. Roux-Rouquié), le laboratoire PPS-CNRS (V. Danos).

Nous avons des contacts réguliers avec les projets HELIX (Rhône-Alpes), SYMBIOSE (Rennes) et COMORE (Sophia-Antipolis). En particulier, nous avons commencé une collaboration avec Hidde de Jong (HELIX) sur la modélisation de l'épissage alternatif.

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 » (en particulier avec le Laboratoire IBCP à Lyon) et « Bioinformatique Fonctionnelle des Systèmes de Régulations Génétiques » de l'Action Ministérielle IMPG : Informatique, Mathématique, Physique pour la Génomique.

Nous participons à deux Actions Spécifiques du CNRS : « Machines à Vecteur Support et Méthodes à Noyau » et « Apprentissage, fouille de données et bioinformatique ».

8.3. 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 au groupe de travail ERCIM *Constraints* coordonné par K. Apt (CWI, Amsterdam).

8.4. Relations internationales

Dans le cadre de l'Institut franco-russe Liapunov nous avons un projet commun avec l'Institut pour les Problèmes Mathématiques en Biologie (IMPB) de l'Académie des Sciences de la Russie à Poushchino (V. Y. Lunin).

Nous avons collaboré en 2002 avec l'Université Carnegie-Mellon à Pittsburgh (E. Balas, John N. Hooker), le Centre de Recherche Opérationnelle CORE à Louvain-la-Neuve (L. Wolsey), le MPI Informatique à Sarrebruck (E. Althaus, K. Mehlhorn), l'Université de Cologne (M. Elf, M. Jünger), la Société SAP (T. Kasper), l'Université de Californie à Irvine (P. Baldi), le MPI Cybernétique biologique à Tübingen (A. Elisseeff) et les laboratoires Wiener (D. Zelus).

8.5. Visites et invitations de chercheurs

Vladimir Lunin de l'Institut pour les Problèmes Mathématiques en Biologie de l'Académie des Sciences de la Russie, a travaillé quatre mois dans l'équipe sur la détermination des enveloppes macromoléculaires.

9. Diffusion des résultats

9.1. 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 JOBIM'2002 et CP'2002 ; 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 des conseils de l'UFR STMIA et de la Faculté des Sciences de l'Université

Henri Poincaré, Nancy 1, membre suppléant de la Commission de Spécialistes 27e section de l'Université Henri Poincaré, Nancy 1, et de l'Université de Metz.

Eric Domenjoud a été membre de la section 07 du Comité National de la Recherche Scientifique jusqu'en juillet 2002.

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).

9.2. Enseignement

Alexander Bockmayr et Yann Guermeur sont 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 RGTI « Ressources Génomiques et Traitements Informatiques »).

Damien Eveillard intervient dans le DESS RGTI ;

Arnaud Courtois est moniteur à l'Université Henri Poincaré, Nancy 1.

9.3. Participation à des colloques, séminaires, invitations

Alexander Bockmayr a donné une conférence invitée à l'Ecole Internationale sur l'Optimisation au Croisic en Mars 2002. Il a participé aussi à la préparation du workshop « Formal methods and biological reasoning » de la conférence internationale sur la biologie des systèmes ICSB'02 à Stockholm.

Yann Guermeur a présenté deux exposés, l'un dans le séminaire de l'équipe « Statistiques et Modélisation Stochastique » du Laboratoire de Modélisation et Calcul (Grenoble) et l'autre dans le séminaire du groupe de travail « Machines à Vecteur Support » de l'Université Paris Sud à Orsay.

10. Bibliographie

Bibliographie de référence

[1] A. BOCKMAYR, T. KASPER. *Branch-and-Infer : A unifying framework for integer and finite domain constraint programming.* in « INFORMS J. Computing », numéro 3, volume 10, 1998, pages 287 - 300.

[2] A. BOCKMAYR, V. WEISPFENNING. *Solving numerical constraints.* éditeurs A. ROBINSON, A. VORONKOV., in « Handbook of Automated Reasoning », volume 1, Elsevier, 2001, chapitre 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] 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 », série Lecture Notes in Computer Science, volume 976, Springer Verlag, pages 18-35, septembre, 1995.

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

Articles et chapitres de livre

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

[7] V. Y. LUNIN, A. URZHUMTSEV, A. BOCKMAYR. *Binary integer programming and its use for envelope determination.* in « CCP4 Newsletter on Protein Crystallography », numéro 40, mars, 2002.

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

Communications à des congrès, colloques, etc.

[9] E. ALTHAUS, A. BOCKMAYR, M. ELF, T. KASPER, M. JÜNGER, K. MEHLHORN. *SCIL - Symbolic Constraints in Integer Linear Programming.* in « 10th European Symposium on Algorithms - ESA'02, Rome, Italie », Springer, LNCS 2461, éditeurs R. H. MÖHRING, R. RAMAN., pages 75-87, septembre, 2002.

[10] A. BOCKMAYR, A. COURTOIS. *Modélisation de systèmes biologiques en programmation concurrente par contraintes hybrides.* in « Programmation en logique avec contraintes, JFPLC'02 », Hermès, pages 167 - 180, 2002.

[11] A. BOCKMAYR, A. COURTOIS. *Using hybrid concurrent constraint programming to model dynamic biological systems.* in « Logic Programming, ICLP'02 », Springer, LNCS 2041, pages 85 - 99, 2002.

[12] D. EVEILLARD, Y. GUERMEUR. *Statistical processing of SELEX results.* in « Intelligent Systems in Molecular Biology, ISMB03 », Edmonton, 2002, (Poster).

[13] D. EVEILLARD, Y. GUERMEUR. *Traitement statistique des résultats SELEX.* in « Journées Ouvertes Biologie, Informatique, Mathématiques, JOBIM'02 », éditeurs J. NICOLAS, C. THERMES., pages 277-283, St Malo, 2002.

[14] E. GOTHIÉ, Y. GUERMEUR, F. LECLERC, C. BRANLANT, A. BOCKMAYR. *Analyse des Séquences InterORF chez les Levures - Application à la Recherche des petits ARN.* in « Structure, Intégration, Fonction et Réactivité des ARN (sifrARN) », 2002, (Poster).

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

Rapports de recherche et publications internes

[16] E. BALAS, A. BOCKMAYR, N. PISARUK, L. WOLSEY. *On unions and dominants of polytopes.* Discussion Paper, numéro No. 2002/8, CORE, février, 2002, Also available as Management Science Report #MSRR-669, GSIA, Carnegie-Mellon-University.

[17] D. EVEILLARD, D. ROPERS, H. DE JONG, C. BRANLANT, A. BOCKMAYR. *Modeling the effects of SR proteins on alternative splicing.* rapport technique, numéro A02-R-117, LORIA, septembre, 2002.

[18] Y. GUERMEUR, A. ELISSEEFF, D. ZELUS. *Bounding the Capacity Measure of Multi-Class Discriminant Models.* rapport technique, numéro NC-TR-2002-123-R, NeuroCOLT2, 2002, (revised).

[19] Y. GUERMEUR. *A Simple Unifying Theory of Multi-Class Support Vector Machines.* rapport technique, numéro RR-4669, INRIA, 2002, http://www.inria.fr/rrrt/rr-4669.html.

[20] Y. KHAN. *Formalismes de modélisation en biologie moléculaire.* Rapport de DEA, Univ. Henri Poincaré, LORIA, juillet, 2002.

[21] R. VERT. *Conception et mise en œuvre de M-SVM dédiées au traitement de séquences biologiques.* Rapport de DEA, Univ. Henri Poincaré, LORIA, septembre, 2002.

Bibliographie générale

[22] T. L. BAILEY, M. GRIBSKOV. *Methods and statistics for combining motif match scores.* in « Journal of Computational Biology », volume 5, 1998, pages 211-221.

[23] A. BEN-HUR, D. HORN, H. T. SIEGELMANN, V. VAPNIK. *Support Vector Clustering.* in « Journal of machine learning research », volume 2, 2001, pages 125-137.

[24] B. BOSER, I. GUYON, V. VAPNIK. *A training algorithm for optimal margin classifiers.* in « COLT'92 », pages 144-152, 1992.

[25] éditeurs J. BOWER, H. BOLOURI., *Computational Modelling of Genetic and Biochemical Networks.* The MIT Press, 2000.

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

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

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

[29] N. CRISTIANINI, J. SHAWE-TAYLOR. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

[30] *Special Issue : Génolevures.* série FEBS Letters, numéro 1, volume 487, 2000, http://cbi.labri.u-bordeaux.fr/Genolevures/biblio.php3.

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

[32] V. GUPTA, R. JAGADEESAN, V. SARASWAT, D. G. BOBROW. *Programming in Hybrid Constraint Languages.* in « Hybrid Systems II », Springer, LNCS 999, pages 226-251, 1995.

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

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

[35] H. KITANO. *Systems Biology : A Brief Overview.* in « Science », volume 295, 2002, pages 1662-1664.

[36] K. KOHN. *Molecular interaction map of the mammalian cell cycle control and DNA repair systems.* in « Molecular Biology of the Cell », volume 10, 1999, pages 2703 - 2734.

[37] H. LODHI, C. SAUNDERS, J. SHAWE-TAYLOR, N. CHRISTIANINI, C. WATKINS. *Text Classification using String Kernels.* in « Journal of Machine Learning Research », volume 2, 2002, pages 419-444.

[38] 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 a dvances.* in « Acta Cryst. », volume D56, 2000, pages 1223 - 1232.

[39] M. NISSIM-RAFINIA, B. KEREM. *Splicing regulation as a potential genetic modifier.* in « Trends in genetics », numéro 3, volume 18, 2002, pages 123-127.

[40] V. A. SARASWAT. *Concurrent constraint programming.* série ACM Doctoral Dissertation Awards, MIT PRESS, 1993.

[41] B. SCHÖLKOPF, A. SMOLA, K.-R. MÜLLER. *Nonlinear Component Analysis as a Kernel Eigenvalue Problem.* in « Neural Computation », numéro 5, volume 10, 1998, pages 1299-1319.

[42] C. TUERK, L. GOLD. *Systematic evolution of ligands by exponential enrichment.* in « Science », 1990.

[43] A. VAN DER SCHAFT, H. SCHUMACHER. *An introduction to hybrid dynamical systems.* Springer, Lecture Notes in Control and Information Sciences, Vol. 251, 2000.

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

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

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

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

[48] C.-H. YUH, H. BOLOURI. *Genomic Cis-Regulatory logic : experimental and computational analysis of a sea urchin gene.* in « Science », volume 279, 1998, pages 1896-1902.