Rapport final
GENOTO3D

1 Liste des équipes impliquées

1. Projet MODBIO, LORIA, UMR 7503, Nancy
2. Laboratoire de Bioinformatique et RMN Structurales (LBRS) de l'IBCP, UMR 5086, Lyon
3. Equipe "Bases de Données et Apprentissage Automatique (BDAA)", LIF, UMR 6166, Marseille
4. Projet Symbiose, IRISA, UMR 6074, Rennes
5. Equipe MAB, LIRMM, UMR 5506, Montpellier
6. Unité Mathématique Informatique et Génome (MIG), INRA, Jouy-en-Josas

2 Liste des participants au 01/10/06

- LORIA - MODBIO
  - Darcy, Yannick, Doctorant depuis octobre 2003, allocation de recherche du ministère (MENRT), 100 %
  - Guermeur, Yann, CR CNRS, 50 %
- IBCP - Bioinformatique et RMN structurales
  - Benabdeslem, Khalid, Post-doctorant au CNRS à l'IBCP de septembre 2004 à février 2006, puis à l'INRIA Lorraine, de mars à mai 2006, 100 %
  - Geourjon, Christophe, IR CNRS, 25%
  - Sapay, Nicolas, Doctorant BDI de septembre 2002 à janvier 2006, 15%
- LIF - BDAA
  - Denis, François, Professeur, 50%
  - Capponi, Cécile, Maître de conférences, 20%
  - Ralaivola, Liva, Maître de conférences, 20%
  - Magnan, Christophe, Doctorant depuis septembre 2004, allocataire moniteur, 50%.
- IRISA - Symbiose
  - Coste, François, CR INRIA, 50%
  - Andonov, Rumen, Professeur, 50%
  - Yanev, Nikola, Professeur en détachement au CNRS de juin 2005 à juin 2006, 30%
  - Nicolas, Jacques, CR INRIA, 20%
  - Kerbellec, Goulven, Doctorant INRIA depuis octobre 2004, allocation de recherche Région Bretagne, 50%
  - Jacquemin, Ingrid, Doctorante, allocation de recherche MENRT (soutenance de thèse en décembre 2005), puis ATER, 50%
  - Veber, Philippe, Doctorant INRIA depuis octobre 2004, bourse INRIA, 30%
  - Leroux, Aurélien, Doctorant jusqu'à juin 2005 (soutenance de thèse), allocation de recherche Région Bretagne et INRIA, 20%
  - Mescam, Yoann, Doctorant jusqu'à juillet 2005, financement INRIA, 10%

- LIRMM - MAB
  - Bréhélin, Laurent, CR CNRS, 30%
  - Gascuel, Olivier, DR CNRS, 20%
  - Duprat, Elodie, Doctorante, 20%
- MIG
  - Martin, Juliette, Doctorante d'octobre 2002 à novembre 2005 (soutenance de thèse), allocation de recherche INRA (ASC), 100%
  - Taly, Jean-François, Doctorant depuis octobre 2003, allocation de recherche MENRT, 100%
  - Collet, Guillaume, Stagiaire M2, Bioinformatique, de janvier à juin 2006, 100%
  - Marin, Antoine, CR INRA, 80%
  - Zimmermann, Karel, Maître de conférences, 10%
  - Gibrat, Jean-François, DR INRA, 20%

3 Changements significatifs intervenus dans le projet

- Khalid Benabdeslem a été nommé Maître de conférences à l'Université Lyon 1 en septembre 2006. Il effectue sa recherche dans l'équipe COMAD du laboratoire PRISMa.
- François Denis a été accueilli en délégation dans le projet MODBIO de l'INRIA Lorraine durant l'année universitaire 2004-2005.
- Yann Esposito figurait dans le projet initial à hauteur de 30% ; il a soutenu sa thèse en décembre 2004 et obtenu son inscription sur les listes de qualification, à la suite de quoi, il a préféré rechercher un poste dans l'industrie plutôt que poursuivre une carrière de chercheur ou d'enseignant-chercheur. Il ne fait plus partie du projet depuis janvier 2005.
- Juliette Martin figurait dans le projet initial à hauteur de 100%. Elle a soutenu sa thèse en novembre 2005. Elle est maintenant Post-doctorante dans l'Équipe de Bioinformatique Génomique et Moléculaire (EBGM), INSERM U726.
- Christophe Magnan est Doctorant (Allocataire-Moniteur) au LIF depuis septembre 2004. Encadré par François Denis et Cécile Capponi, son sujet de thèse porte sur l'apprentissage semi-supervisé avec application à la biologie et en particulier, à la prédiction de la structure tridimensionnelle des protéines. Il a donc très naturellement rejoint le projet GENOTO3D.
- Liva Ralaivola a été nommé Maître de conférences à l'université de Provence en septembre 2004. Il a terminé en janvier 2005 un séjour d'un an à l'Université de Californie à Irvine (UCI) dans le laboratoire du Professeur Pierre Baldi. Ses travaux portent sur l'apprentissage automatique et ses applications à la bioinformatique : il a rejoint le projet GENOTO3D depuis février 2005.
- Frédéric Sur, Post-doctorant au CNRS, a rejoint le projet MODBIO d'octobre 2004 à août 2005.

À la suite des échanges des premières journées, il est apparu nécessaire de réfléchir à la définition et à l'acquisition de cœurs de structures. Cette thématique a été ajoutée au projet initial.

4 Résumé des principales avancées

4.1 Introduction

Le but du projet GENOTO3D est d'appliquer les techniques de l'apprentissage automatique, symbolique ou numérique, au problème consistant à déterminer la structure tridimensionnelle (3D) des protéines à partir de leur séquence en acides aminés.

Résoudre ce problème est très important dans le cadre des nombreux projets de génomique en cours à l'heure actuelle (séquençage de génomes, etc.). En effet, une étape cruciale de l'analyse des données génomiques correspond à la prédiction de la fonction du produit des gènes, i.e., les protéines. Il est bien connu que la structure 3D adoptée par une protéine joue un rôle prépondérant dans sa fonction.

Il existe deux grandes catégories de méthodes permettant d'obtenir de l'information sur la structure 3D des protéines. La première catégorie est, par essence, comparative. On utilise des structures 3D connues pour bâtir les modèles des protéines dont on ne connaît que la séquence en acides aminés. Cette approche est fondée sur la notion évolutive d'homologie. La deuxième catégorie, qui comprend les méthodes *ab initio* ou *de novo*, cherche à prédire la structure 3D, ou des caractéristiques de la structure 3D, sans l'aide d'une structure déjà connue.

4.2 Techniques d'apprentissage pour les méthodes de recherche d'homologie

La construction d'un modèle de la structure 3D peut être plus ou moins difficile selon l'information dont on dispose. Dans le cadre de GENOTO3D nous nous sommes intéressés aux méthodes de reconnaissance de repliements. Ces techniques sont une généralisation des méthodes d'alignement de séquences dans le cas où, les protéines étant des homologues lointains, les séquences ont beaucoup divergé et sont très différentes. Nous nous trouvons donc dans un cas où l'obtention d'un modèle est difficile. Les structures 3D de protéines homologues étant bien mieux conservées que leurs séquences, ces techniques utilisent ces structures pour repérer des homologues lointains qui ne peuvent pas être mis en évidence par des techniques de comparaisons de séquences comme PSI-BLAST. Les développements effectués dans ce cadre concernent plus particulièrement les membres de l'IBCP, de MIG et du projet Symbiose.

L'IBCP s'est intéressé à l'extraction des cœurs de structures et à la reconnaissance des repliements de protéines. L'identification des repliements pour les structures protéiques inconnues représente un véritable problème en biologie structurale. Dans ce contexte, l'IBCP a proposé une approche originale pour traiter les structures tridimensionnelles des protéines. Cette approche consiste d'une part à assembler une base de séquences significative pour l'apprentissage automatique et d'autre part à concevoir un modèle de reconnaissance de repliements de protéines à partir de leurs structures primaires. Dans un premier temps, la méthode consiste à effectuer un alignement structural entre chaque paire de structures appartenant à une famille de protéines donnée. Ensuite, une classification ascendante hiérarchique est effectuée à partir de la matrice de dissimilarité extraite dans la première étape. À partir de cette classification, il est possible d'extraire des cœurs de structures à tout niveau hiérarchique pour chaque famille de protéines (22). Finalement, un réseau de neurones de type perceptron multi-couche prédit la nature des repliements à partir des séquences modélisées par des matrices de covariance (23; 21). Par conséquent, grâce à l'extraction des cœurs de

structures, une base de méta-séquences a été produite pour concevoir un système de reconnaissance puissant dont le taux de reconnaissance est de 72% sur une centaines de familles de repliements CATH (8). Le taux de reconnaissance a été porté à 86% en substituant au réseau de neurones la DSVM décrite dans la section 4.4.

L'IRISA s'est aussi intéressé à la définition du cœur caractéristique d'une famille de protéines. L'approche employée consiste en l'apprentissage d'automates caractéristiques d'une famille de protéines et vise à modéliser de façon automatique un ensemble de séquences protéiques sous une forme syntaxique. Dans le cadre de ce projet, le but est de constituer automatiquement une meilleure bibliothèque de cœurs pour FROST à partir d'ensembles de séquences partageant un même repliement. Un algorithme original d'apprentissage d'automates sur les protéines a été proposé (26; 27). Le prototype initial a été repensé et réimplémenté pour gagner en modularité et efficacité. Le programme est à présent disponible pour des expérimentations à plus grande échelle (travail en cours) et publication.

L'unité MIG dispose d'une méthode de reconnaissance de repliements nommée FROST [Marin et al., *Proteins*, **49** :493 (2002)]. Au cours des 3 ans écoulés, dans le cadre du projet GENOTO3D et en collaboration avec R. Andonov et N. Yanev de l'IRISA et S. Balev de l'université du Havre, les performances de la méthode qui, à l'origine, était très coûteuse en temps calcul, ont été considérablement améliorées. Un nouvel algorithme a été proposé pour résoudre la phase d'optimisation, qui est une étape importante dans l'approche FROST. Cet algorithme est basé sur un modèle linéaire en variables booléennes (6; 5; 19) qui a été efficacement résolu par la relaxation lagrangienne (42; 18). Une version parallèle du logiciel FROST a de plus été développée (38; 9). De nouveaux types d'alignement (semi-global et flexible (25; 24)) ont également été développés. Ces travaux d'amélioration des performances de la méthode de reconnaissance de repliements FROST ainsi que de développement de nouvelles méthodes d'alignement sont très significatifs. Ils permettent à FROST d'être plus versatile et capable de s'adapter aux problèmes rencontrés lors de l'analyse de séquences protéiques issues des projets de génomique. L'ensemble de ces travaux, pour le moment, n'a fait l'objet que de communications à des congrès ou des rapports internes et nous allons nous consacrer à les publier sérieusement.

En complément des travaux précédents, l'unité MIG s'est également intéressée à l'étude de la qualité des modèles 3D obtenus. Quelle que soit la technique utilisée, modélisation par homologie, reconnaisance de repliements, méthodes *de novo*, on doit bâtir un modèle de la structure 3D de la protéine étudiée. On peut alors s'interroger sur la qualité du modèle obtenu. J-F Taly a développé, au cours de sa thèse, des méthodes pour analyser la qualité des modèles 3D obtenus (40; 20) basés sur un nouveau jeu de paramètres et des simulations de dynamique moléculaire.

4.3 Techniques d'apprentissage pour les méthodes *ab initio* et *de novo*

Le problème de prédire la structure 3D d'une protéine, dans toute sa généralité, par ces techniques est très complexe. Dans le projet GENOTO3D différentes équipes se sont attaquées à certaines composantes de ce problème qui, si elles sont résolues d'une manière satisfaisante, peuvent faire avancer l'état de l'art dans le domaine.

Le premier type de problème a consisté à prédire les interactions distantes dans les protéines, en particulier les ponts disulfures. Les équipes du LIF, du LIRMM et de l'IRISA ont étudié ce problème à l'aide de différentes techniques d'apprentissage.

Les ponts disulfures, liaisons covalentes entre deux cystéines oxydées d'une protéine, forment un élément de son repliement et leur prédiction fournirait une aide certaine dans la prédiction de la structure 3D de la protéine. Ce problème se décompose en fait en deux sous-problèmes que sont la prédiction de l'état, oxydé ou non, d'une cystéine, et la prédiction des appariements entre les cystéines oxydées.

La connaissance de l'état des cystéines est de plus un facteur important puisqu'elles sont fréquemment impliquées dans les sites actifs ou donnent de l'information sur la localisation cellulaire des protéines. En utilisant la programmation logique inductive (PLI), l'équipe Symbiose de l'IRISA s'est attachée à extraire et à valider des règles explicites de prédiction à partir des données biologiques. Le taux de prédiction obtenu atteint les 90% avec seulement 13 règles (35; 34; 2).

Concernant le problème de la prédiction des appariements entre les cystéines oxydées, une question ouverte est de savoir si ces ponts contraignent le repliement ou si le repliement détermine les ponts. Une première étape pour attaquer ce problème consiste à tenter de prédire ces interactions à partir de l'information contenue localement dans des fenêtres de tailles réduites centrées sur chaque cystéine (contextes locaux), éventuellement enrichie d'information évolutionnaire.

Dans ce cadre, l'équipe MAB du LIRMM s'est attachée à la recherche d'une modélisation probabiliste fine des interactions intervenant dans les contextes locaux aux cystéines. L'idée sous-jacente de ce travail, fondé sur les "cartes de contact" des interactions issues de la PDB, est que si ces interactions sont effectivement contraignantes, on doit pouvoir observer un biais de composition entre les acides aminés les plus proches physiquement.

Une piste différente suivie par l'équipe BDAA du LIF a été la recherche d'une modélisation dans le cadre de l'apprentissage semi-supervisé asymétrique, qui peut être replacé dans le cadre plus général de l'apprentissage supervisé avec bruit de classification. Ces travaux ont donné lieu à des résultats théoriques et appliqués, publiés dans (13; 31; 36; 39). Une autre modélisation plus approfondie est en cours d'élaboration : elle permettra d'étudier la présence d'information locale lors d'interactions distantes (non limitées aux ponts disulfures), des résultats théoriques nous assurant que l'existence d'une fonction d'affinité entre contextes locaux doit pouvoir être mise en évidence. Les algorithmes sous-jacents sont en cours d'implémentation au sein d'une plate-forme Java et leur application aux données biologiques fait l'objet du travail de thèse de Christophe Magnan.

D'autres problèmes ont été abordés par l'équipe MAB du LIRMM, l'IBCP, MODBIO et l'unité MIG.

L'équipe MAB du LIRMM a travaillé sur une extension des modèles de Markov cachés profils (profile hidden Markov models, PHMM). Ces modèles constituent un outil de base pour la modélisation de familles de protéines et l'inférence d'alignements multiples. Les PHMM classiques sont définis par une structure linéaire dépourvue de cycle. Bien qu'ils soient bien adaptés à la modélisation de la grande majorité des familles de protéines connues, ils conviennent mal pour la modélisation des protéines composées de motifs répétés dont le nombre de répétitions varie d'une protéine à l'autre au sein d'une même famille. Nous avons donc proposé une extension des PHMM spécialement adaptée à ces familles grâce à l'introduction de cycles, et l'avons appliquée à une famille particulière des protéines d'*Arabidopsis thaliana* (41).

En utilisant le même type de technique basée sur les modèles de Markov cachés (HMM), l'unité MIG s'est intéressée à la prédiction de la structure locale d'une chaîne polypep-

tidique dans le but d'extraire des fragments pertinents pour la modélisation *de novo*. Les méthodes *de novo* sont fondées sur l'assemblage judicieux de fragments de structure 3D extraits des bases de données de structures. Au cours de son travail de thèse (3) J. Martin s'est consacrée à la première étape de ces techniques qui vise à définir le plus précisément possible les fragments à assembler grâce à la prédiction de la structure locale de la chaîne polypeptidique. Le plus souvent cette prédiction concerne les structures secondaires régulières, hélices alpha et feuillets bêta et l'ensemble des structures complémentaires regroupées sous le terme d'apériodique (coil) (16). Ces dernières structures, qui représentent environ 50% des résidus, ne fournissent aucune information précise sur la structure locale de la chaîne polypeptidique. J. Martin a donc proposé une méthode basée sur des HMM (14; 37; 15) qui permet de prédire, non seulement les structures secondaires classiques mais également les zones d'angles phi et psi dans le diagramme de Ramachandran. Cette méthode permet de prédire correctement la localisation de 77% des résidus dans les 3 grandes zones du diagramme de Ramachandran.

Une autre contribution de l'équipe MAB du LIRMM a été, en collaboration avec l'équipe de Marie-Paule Lefranc à l'Institut de Génétique Humaine, le développement d'une méthode de prédiction automatique de la capacité des protéines de la superfamille du complexe majeur d'histocompatibilité (MHC) à se lier à la beta2-microglobuline (B2M). Ces protéines assurent une fonction essentielle au sein du système immunitaire, mais seule une partie d'entre elles a la propriété de pouvoir se lier de manière non-covalente à la B2M. La structure tridimensionnelle est suspectée d'être en partie responsable de cette capacité. Nous avons donc proposé une méthode de prédiction automatique de la liaison (ou de l'absence de liaison) à la B2M des protéines de la superfamille du MHC. La méthode proposée combine un classifieur Bayésien naïf et des alignements multiples. Elle est composée de deux étapes : un ensemble de descripteurs binaires discriminants (associant une position dans l'alignement et un groupe d'acides aminés) est tout d'abord extrait des données ; les fréquences de ces descripteurs sont ensuite estimées conditionnellement aux classes que l'on cherche à séparer, pour construire le classifieur. Cette approche a été appliquée à un jeu de données composé de 807 séquences alignées, correspondant aux allèles de 47 gènes de la superfamille du MHC. 18 descripteurs sont sélectionnés pour leur capacité à discriminer les protéines selon qu'elles se lient ou non à la B2M. L'analyse structurale des protéines du jeu de données montre que ces descripteurs correspondent à des sites potentiels de contact à la B2M ou à des sites impliqués dans le maintien d'une conformation favorable au contact (10).

Enfin, l'IBCP et l'équipe MODBIO se sont intéressés aux protéines membranaires. Les protéines membranaires représentent environ 25% des séquences codantes dans les génomes. On distingue deux catégories de protéines membranaires : les protéines polytopiques, traversant une ou plusieurs fois la membrane, et les protéines monotopiques, interagissant avec un seul côté de la membrane. Dans ce cas, l'ancrage membranaire peut être constitué d'une ou plusieurs hélices amphipathiques localisées à l'interface de la membrane, parallèlement au plan de la bicouche phospholipidique. Ce mode d'ancrage membranaire, aussi appelé ancrage "IPM" pour "in-plane membrane anchor", est mal compris et aucune méthode de prédiction n'est disponible. Dans ce contexte, l'objectif de l'IBCP était le développement d'une méthode bioinformatique de prédiction des ancrages IPM. Il a abouti à l'élaboration de la méthode AmphipaSeeK (4; 17), permettant de prédire les ancrages IPM dans les séquences protéiques. Cette méthode est basée sur un lot de 21 protéines membra-

naires monotopiques possédant un ancrage IPM caractérisé expérimentalement. Elle utilise une machine à vecteurs support (SVM) comme système de classification. Le noyau de cette machine, qui exploite le contenu d'une fenêtre d'analyse déterminant un contexte local autour du résidu à classer, prend en compte à la fois la nature des substitutions et l'importance relative, pour la prédiction, des différentes positions de la fenêtre. AmphipaSeeK s'est avérée être une méthode très spécifique. Ce développement méthodologique est le fruit d'une étroite collaboration entre le Laboratoire de Bioinformatique et RMN structurales (IBCP, Lyon) et le projet MODBIO.

4.4 Avancées fondamentales en apprentissage

L'équipe BDAA du LIF s'est intéressée à l'inférence d'automates probabilistes. Les automates probabilistes (PA) sont des objets formels permettant de modéliser des distributions de probabilités sur des ensembles de mots, appelées aussi langages stochastiques. Ils ont la même expressivité que les HMM et peuvent donc être utilisés dans les mêmes conditions, en particulier en bioinformatique. Pouvoir inférer un automate probabiliste à partir de données séquentielles est un problème clé dans nombre d'applications. La principale contribution du LIF au problème étudié ici a consisté à dégager une notion essentielle pour l'inférence, celle de *langages résiduels* d'un langage stochastique, à replacer le problème de l'inférence d'automates probabilistes dans le cadre plus général des langages stochastiques rationnels définissables au moyen d'automates à multiplicités, ce qui a permis à la fois d'obtenir des résultats théoriques non triviaux et de définir un algorithme d'inférence de langages stochastiques rationnels, performant et utilisable en pratique. Ces travaux ont été publiés dans (1; 29; 30; 33; 43). Une autre partie des travaux de l'équipe a concerné l'étude de l'apprenabilité de certaines classes de concepts (dont, notamment, les hyperplans séparateurs) lorsque différents types de bruits de classification affectent les données. Dans ce contexte, l'équipe BDAA a montré l'égalité de l'ensemble des classes de concepts apprenables avec bruit de classification uniforme et de l'ensemble des classes de concepts apprenables avec bruit de classification constant par morceaux (39). Ces résultats théoriques motivent l'utilisation de classifieurs linéaires (à noyaux) pour le problème de la prédiction de ponts disulfures lorsque celui-ci est (naturellement) formulé comme un problème d'apprentissage avec bruit de classification. Des travaux concernant l'identifiabilité de classifieurs naïfs de Bayes dans un contexte bruité viennent également légitimer l'usage de ce type de modèle pour le problème évoqué (31).

Khalid Benabdeslem, à l'IBCP, a développé une SVM multi-classe appelée *Dendogram based SVM* (DSVM). Elle combine une classification hiérarchique et des SVM bi-classes (7). Cette méthode procède en deux phases. La première consiste à concevoir une arborescence à partir des classes étudiées de manière ascendante, i.e. effectuer une classification automatique sur ces classes plutôt que sur les données. La deuxième consiste à associer à chaque nœud de l'arborescence obtenue une SVM bi-classe. La SVM associée à un nœud donné a pour rôle de discriminer entre les deux sous-ensembles représentés par ce nœud. Le processus est ensuite répété de manière descendante depuis la racine jusqu'aux feuilles qui représentent les classes en question. Cette méthode présente deux avantages : celui de décomposer un problème multi-classe en un ensemble de problèmes bi-classes et celui de produire une trace de reconnaissance dans un chemin à partir de la racine de l'arborescence jusqu'au niveau des feuilles.

En parallèle, les membres du projet MODBIO ont étudié la mise en œuvre de SVM multi-classes standard (12) pour développer des modules d'une méthode de prédiction *ab initio* de la structure tertiaire. Ce travail s'est orienté suivant deux axes. D'une part, trois méthodes permettant de fixer les valeurs des hyperparamètres (paramètres du noyau et constante de marge douce) ont été proposées. Ces méthodes de sélection de modèles s'appuient sur des bornes sur le risque (probabilité d'erreur), bornes nommées risques garantis, et des bornes sur l'estimateur empirique du risque. Pour majorer les nombres de couverture des SVM multi-classes, le résultat reliant ces nombres aux nombres d'entropie de l'opérateur d'évaluation a été étendu au cas multi-classe (32). Toute la théorie de Vapnik-Chervonenkis (théorie VC) des systèmes discriminants multi-classes à grande marge (44) a également été développée autour de la notion de psi-dimensions à marge. Enfin, une extension multi-classe d'une borne "leave-one-out" sur le risque empirique, la borne "rayon-marge", a été proposée dans (28). D'autre part, la poursuite des travaux sur la spécification de noyaux dédiés au traitement de séquences biologiques exposés dans (11), a conduit au développement d'un noyau fondé sur un pair-HMM. Ce noyau est actuellement en cours d'évaluation sur le problème de la prédiction des ponts disulfures et de la structure secondaire.

4.5 Conclusions et perspectives

En soumettant le projet GENOTO3D à l'ACI "Masses de Données", notre objectif était double. Il s'agissait d'une part d'exploiter les ressources offertes par l'apprentissage automatique afin de faire progresser la prédiction de la structure tertiaire des protéines, à travers les deux grandes familles de méthodes que sont les méthodes de recherche d'homologie et les méthodes *ab initio* et *de novo*. Il s'agissait également de profiter de ces travaux pour dégager de nouveaux sujets de recherche amont en apprentissage, sujets relevant en particulier du traitement de données séquentielles, et initier l'étude de ces sujets. Nous souhaitions donc à la fois apporter des solutions opérationnelles à un problème de biologie structurale aussi important que difficile, et faire progresser la théorie et les algorithmes de l'apprentissage automatique.

Les principales avancées effectuées en prédiction par homologie et threading concernent la définition, la modélisation et l'identification des cœurs de structures, ainsi que le développement de la méthode de prédiction FROST. À notre connaissance, aucun groupe dans le monde ne dispose de l'équivalent en termes de performance et de flexibilité de l'algorithme d'alignement optimal séquence/structure qui est utilisé par FROST. Dans le domaine de la prédiction *ab initio*, les travaux se sont concentrés sur deux sous-problèmes, la prédiction des ponts disulfures et de la structure secondaire. En ce qui concerne les aspects fondamentaux de l'apprentissage, nos contributions concernent principalement l'inférence d'automates, ainsi que la conception et l'étude des propriétés de SVM multi-classes.

Au-delà de ces résultats scientifiques, le projet GENOTO3D aura été bénéfique à plus d'un titre, en permettant en particulier d'établir ou renforcer des collaborations entre les différentes équipes partenaires. Les contacts internationaux jouent également un rôle important. Ces collaborations, qui se poursuivront après la fin du projet, permettront en particulier de progresser dans le travail d'intégration des réalisations logicielles qui constituait la dernière étape de notre programme prévisionnel. Il ne s'agit plus à présent que d'une question de temps.

5 Réalisations obtenues dans le cadre du projet

La méthode AmphipaSeeK est disponible via une interface Web à l'adresse : http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_amphipaseek.html.

Le logiciel Protomata Learner d'apprentissage d'automates sur les protéines sera diffusé à partir du site http://www.irisa.fr/symbiose/coste/ conjointement à la publication de l'article correspondant.

Une version du logiciel de SVM multi-classes développé dans le projet MODBIO a été dédiée au traitement des données protéiques. Diffusée sous licence GPL à l'adresse suivante : http://www.loria.fr/~guermeur/Dev2.tar.gz, elle a fait l'objet d'un dépôt à l'APP sous le numéro IDDN : IDDN.FR.001.170014.000.R.P.2005.000.10000.

Le logiciel KAKSI d'assignation des structures secondaires est mis à disposition sur le site Web de l'unité MIG http://migale.jouy.inra.fr/mig/mig_fr/servlog/kaksi/ sous licence GPL.

Le logiciel de reconnaissance de repliements FROST est disponible sur le site http://genome.jouy.inra.fr/frost sous licence GPL.

Le logiciel de prédiction des structures secondaires et des zones d'angles phi et psi du diagramme de Ramachandran sera distribué sous licence GPL après publication de l'article soumis (voir le site de l'unité MIG http://migale.jouy.inra.fr/mig).

Un serveur Web d'extraction des cœurs de structures des protéines a été développé http://pig-pbil.ibcp.fr/cgi-bin/asce/asce.

Une première version de DSVM pour la discrimination multi-classe est disponible dans http://www710.univ-lyon1.fr/~kbenabde/.

6 Réunions et Conférences organisées dans le cadre du projet

Les informations relatives aux réunions propres au projet sont disponibles sur le site de GENOTO3D, à l'adresse suivante : http://www.loria.fr/~guermeur/ACIMD/.

Les pages des différentes réunions contiennent des liens sur les supports des présentations qui y ont été faites. La liste est la suivante :

1. Réunion initiale du projet le 23/11/03 à l'IBCP
2. Réunion le 22/06/04 au LORIA
3. Réunion le 22/11/04 au LIP6, à Paris (présentation de l'avancement des travaux devant Patrick Gallinari, Professeur à l'université Paris 6 et rapporteur du projet)
4. Réunion le 24/02/05 à l'Unité MIG

5. Réunion le 30/05/05 à Nice, sur le thème des problèmes d'appariement dans les protéines
6. Réunion finale du projet le 12/10/06 à l'IRISA

7 Soutiens obtenus en liaison avec ce projet

7.1 Postes chercheurs

La thèse intitulée "Apprentissage automatique de motifs structuraux" qu'effectue Goulven Kerbellec depuis octobre 2004 est financée par une bourse Renouvellement des Compétences de la Région Bretagne.

Nikola Yanev a bénéficié d'un détachement CNRS d'un an à l'IRISA - Symbiose de juin 2005 à juin 2006.

7.2 Postes ingénieurs

7.3 Contrats nationaux

- Les mutations structurales avec leurs conséquences sur le phénotype des pathologies humaines. Association Française contres les Myopathies (AFM), Programme DECRYPTHON, 2005-2007, 100 000 euros. O. Poch, G. Deléage, E. Bettler et C. Geourjon (coordinateurs : O. Poch et G. Deléage).

- Détermination de la structure 3D des protéines par spectrométrie de masse et bioinformatique structurale. Région Rhône-Alpes, Programme EMERGENCE, 2005-2008, 110 000 euros. E. Forest, J. Martinez, G. Deléage et C. Geourjon (coordinateur : C. Geourjon).

Le projet MODBIO est partenaire de l'opération "Modélisation de la protéine FAK (Focal Adhesion Kinase) en vue de l'identification de molécules anti-métastases". Ce projet est une opération du thème "Bioinformatique et applications à la génomique" du "PRST Intelligence Logicielle" pour les années 2005 à 2007. Il est également financé par l'ANR, dans le cadre du projet "Tyrosines kinases de la famille FAK. Bases structurales de la régulation et de la localisation intracellulaire", retenu par l'ANR non-thématique pour les années 2006 à 2008.

L'équipe Symbiose coordonne le projet MODULOME : Deciphering and modelling the structural organization of genomes, de 2006 à 2009, financé par l'ANR pour un montant de 160 000 euros (coordinateur Jacques Nicolas).

7.4 Contrats européens

7.5 Contrats internationaux hors CEE

7.6 Contrats industriels

7.7 Contacts internationaux dans le cadre de ce projet

Pierre Baldi, Professeur à l'Université de Californie à Irvine (UCI) USA et directeur de l'Institute for Genomics and Bioinformatics de l'UCI, était invité à la réunion du projet du 30 mai 2005 à Nice, où il a tenu un séminaire intitulé "Protein structure prediction from SCRATCH using machine learning, evolutionary information, graph matching, and physical constraints" sur les récentes avancées, notamment en apprentissage automatique, pour la prédiction *de novo* de la structure des protéines. Il a effectué un séjour d'un mois, en juin 2005, dans l'équipe BDAA du LIF.

Giuseppe Lancia, Professeur associé au Département de Mathématiques et d'Informatique de l'Université d'Udine, en Italie, a donné un séminaire à l'occasion de la réunion du projet à Nice, puis a effectué un séjour d'un mois dans l'équipe Symbiose en février 2006, en bénéficiant d'un financement IFSIC.

Nikola Yanev, Professeur à l'Université de Sofia, en Bulgarie, a rejoint le projet Symbiose pendant un an dans le cadre d'un détachement au CNRS. La collaboration qui s'est établie avec lui, collaboration qui a déjà donné lieu à de nombreuses publications, se poursuit actuellement.

8 Publications obtenues dans le cadre du projet

Thèses et manuscrits d'habilitation

[1] Y. ESPOSITO, *Contribution à l'inférence d'automates probabilistes*, PhD Thesis, Université de Provence, 2004.

[2] I. JACQUEMIN, *Découverte de motifs relationels en bioinformatique : application à la prédiction de ponts disulfures*, PhD Thesis, Université Rennes 1, december 2005.

[3] J. MARTIN, *Prédiction de la structure locale des protéines par des modèles de chaînes de Markov cachées*, Spécialité : Analyse des génomes et modélisation moléculaire, Université Paris 7, École doctorale B2M, novembre 2005.

[4] N. SAPAY, *Les Peptides d'ancrage à l'Interface Membranaire : Analyse structurales par RMN et dynamique moléculaire et développement d'une méthode de prédiction bioinformatique*, PhD Thesis, Université Claude Bernard Lyon 1, janvier 2006.

Journaux internationaux et chapitres de livres

[5] R. ANDONOV, S. BALEV, N. YANEV, "Protein Threading : From Mathematical Models to Parallel Implementations", *INFORMS Journal on Computing 16*, 4, 2004, p. 393–405, Special Issue on Computational Molecular Biology/Bioinformatics.

[6] R. ANDONOV, S. BALEV, N. YANEV, *High Performance alignment methods for protein threading*, Joh Wiley & Sons 2OO6 Wiley-Interscience, 2006, p. 427–457, in Parallel Computing for Bioinformatics and Computational Biology, edited by Prof. Albert Zomaya.

[7] K. BENABDESLEM, Y. BENNANI, "Dendogram based SVM for multi-class classification", *CIT*, 2006, à paraître.

[8] K. BENABDESLEM, G. DELÉAGE, C. GEOURJON, "Structural cores extraction for fold recognition improvement", *Bioinformatics*, 2006, Soumis.

[9] G. COLLET, V. POIRRIEZ, A. MARIN, J.-F. GIBRAT, R ANDONOV, *Grid for Bioinformatics and Computational Biology*, edition A Zomaya & E-G Talbi, John Wiley and Sons, 2007, ch. Protein threading on grid, en préparation.

[10] E. DUPRAT, M.-P. LEFRANC, O. GASCUEL, "A simple method to predict protein-binding from aligned sequences - application to MHC superfamily and β2-microglobulin", *Bioinformatics 22*, 4, 2006, p. 453–459.

[11] Y. GUERMEUR, A. LIFCHITZ, R. VERT, "A kernel for protein secondary structure prediction", *in : Kernel Methods in Computational Biology*, B. Schölkopf, K. Tsuda, and J.-P. Vert (editors), The MIT Press, 2004, p. 193–206.

[12] Y. GUERMEUR, "SVM multiclasses", *in : Support Vector Machines et autres méthodes à noyau*, S. Canu, C. Richard, M. Davy, and A. Rakotomamonjy (editors), Hermès, 2007, (à paraître).

[13] C. MAGNAN, "Asymmetrical Semi-Supervised Learning and Prediction of Disulfide Connectivity in Proteins", *RIA*, 2006, à paraître.

[14] J. MARTIN, J.-F. GIBRAT, F. RODOLPHE, "Choosing the optimal Hidden Markov Model for secondary-structure prediction.", *IEEE Intelligent Systems 20*, 2005, p. 19–25.

[15] J. MARTIN, J.-F. GIBRAT, F. RODOLPHE, "Analysis of an optimal hidden Markov model for secondary structure prediction.", soumis.

[16] J. MARTIN, G. LETELLIER, A. MARIN, J. TALY, A. DE BREVERN, J. GIBRAT, "Protein secondary structure assignment revisited : a detailed analysis of different assignment methods.", *BMC Struct Biol 5*, 2005, p. 17.

[17] N. SAPAY, Y. GUERMEUR, G. DELÉAGE, "Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier", *BMC Bioinformatics 7*, 255, 2006.

[18] N. YANEV, R. ANDONOV, P. VEBER, S. BALEV, "Lagrangian Approaches for a class of Matching Problems in Computational Biology", *Computers and Mathematics with Applications*, accepted, also available as RR INRIA No 5973, August 2006.

[19] N. YANEV, R. ANDONOV, "Parallel Divide&Conquer Approach for the Protein Threading Problem", *Concurrency and Computation : Practice and Experience 16*, 2004, p. 961–974.

[20] K. ZIMMERMANN, J.-F. TALY, A. MARIN, J.-F. GIBRAT, "Detecting periodicity and repetitions in biological sequences.", soumis.

Publications dans des conférences et ateliers de travail

[21] K. BENABDESLEM, G. DELÉAGE, C. GEOURJON, "A Neural Network System based on Structural Alignment and Clustering for Proteins Fold Recognition", *in : ECCB'05*, 2005.

[22] K. BENABDESLEM, G. DELÉAGE, C. GEOURJON, "Alignement structural et classification hiérarchique pour l'extraction des cœurs structuraux", *in : BIO-EGC'06*, p. 09–17, 2006.

[23] K. BENABDESLEM, G. DELÉAGE, C. GEOURJON, "Cores extraction based Neural Network Model for Proteins fold recognition", *in : Septièmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM)*, Lyon, juillet 2005.

[24] G. COLLET, A. MARIN, N. YANEV, R. ANDONOV, J.-F. GIBRAT, "Implémentation d'un algorithme d'alignement semi-global utilisant des paramètres non locaux pour la reconnaissance de repliement des protéines", *in : Proc. of the ROADEF conference*, Lille, Janvier 2006.

[25] G. COLLET, N. YANEV, A. MARIN, R. ANDONOV, J.-F. GIBRAT, "A flexible model for protein fold recognition", *in : Septièmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM)*, 2006.

[26] F. COSTE, G. KERBELLEC, "A Similar Fragments Merging Approach to Learn Automata on Proteins", *in : European Conference on Machine Learning (ECML-2005)*, J. Gama, R. Camacho, P. Brazdil, A. Jorge, L. Torgo (editors), *LNAI*, Springer, p. 522–529, Porto, Portugal, 2005.

[27] F. COSTE, G. KERBELLEC, "Learning Automata on Protein Sequences", *in : JOBIM'06*, Bordeaux, p. 199–210, 2006.

[28] Y. DARCY, E. MONFRINI, Y. GUERMEUR, "Borne "rayon-marge" sur l'erreur "leave-one-out" des SVM multi-classes", *in : JdS'06*, 2006.

[29] F. DENIS, Y. ESPOSITO, A. HABRARD, "Learning Rational stochastic languages", *in : Proc. of the 19th Annual Conference on Learning Theory*, *LNAI*, *4005*, p. 274–288, 2006.

[30] F. DENIS, Y. ESPOSITO, "Learning classes of Probabilistic Automata", *in : COLT 2004*, *LNAI*, 3120, p. 124–139, 2004.

[31] F. DENIS, C. MAGNAN, L. RALAIVOLA, "Efficient Learning of Naive Bayes Classifiers under Class-Conditional Classification Noise", *in : Proc. of the 23rd Int. Conf. on Machine Learning*, p. 265–272, 2006.

[32] Y. GUERMEUR, M. MAUMY, F. SUR, "Model selection for multi-class SVMs", *in : ASMDA'05*, p. 507–516, 2005.

[33] A. HABRARD, F. DENIS, Y. ESPOSITO, "Using Pseudo-Stochastic Rational Languages in Probabilistic Grammatical Inference", *in : Proc. 8th Int. Conf. on Grammatical Inference*, 2006.

[34] I. JACQUEMIN, J. NICOLAS, "Disulfide bonds prediction using inductive logic programming", *in : Workshop on Constraint Based Methods for Bioinformatics, WCB*, p. 56–65, Sitges, Spain, 2005.

[35] I. JACQUEMIN, J. NICOLAS, "Modélisation de cystéines oxydées à l'aide de la programmation logique inductive", *in : JOBIM 2005*, p. 331–340, Lyon, France, 2005.

[36] C. MAGNAN, "Apprentissage semi-supervisé asymétrique et estimations d'affinités locales dans les protéines", *in : Actes de CAP 05*, F. Denis (editor), PUG, p. 297–312, 2005.

[37] J. MARTIN, J.-F. GIBRAT, F. RODOLPHE, "Hidden Markov Model for protein secondary structure.", *in : Proc. of Applied Stochastic Models and Data Analysis*, J. J. . P. Lurca (editor), p. 180–187, Brest, May 2005.

[38] V. POIRRIEZ, A. MARIN, R ANDONOV, J.-F. GIBRAT, "FROST : Revisited and distributed.", *in : HICOMB on line proceedings*, J. J. . P. Lurca (editor), 4th IEEE International Workshop on High Performance Computational Biology, Denver, April 2005.

[39] L. RALAIVOLA, F. DENIS, C. MAGNAN, "CN=CPCN", *in : Proc. of the 23rd Int. Conf. on Machine Learning*, p. 721–728, 2006.

[40] J.-F. TALY, J. MARTIN, A. MARIN, J.-F. GIBRAT, "Définition de mesures décrivant l'environnement local des acides aminés pour une application à l'évaluation des modèles structuraux", *in : Sixièmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM)*, 2005.

[41] R. URICARU, E. RIVALS, L. BRÉHÉLIN, "Hidden Markov models for the detection of motifs repeats in protein sequences", *in : IPG'06*, 2006. (soumis).

[42] P. VEBER, N. YANEV, R. ANDONOV, V. POIRRIEZ, "Optimal protein threading by cost-splitting", *in : Algorithms in Bioinformatics-WABI 2005*, G. M. R. Casadio (editor), p. 365–375, October 3-6, 2005, Mallorca, Spain 2005. Lecture Notes in Bioinformatics, 3692.

Rapports de recherche

[43] F. DENIS, Y. ESPOSITO, "Rational stochastic languages", *research report*, LIF - Université de Provence, http ://hal.ccsd.cnrs.fr/ccsd-00019728, 2006.

[44] Y. GUERMEUR, "Large margin multi-category discriminant models and scale-sensitive Ψ-dimensions", *Research Report RR-5314*, INRIA, September 2004, (révisé en 2006).