Rapport mi-parcours
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) de l'INRA, Jouy en Josas

2 Liste des participants au 1/4/05

- LORIA - MODBIO
  - Darcy, Yannick, Doctorant depuis le 01/10/03, allocation de recherche du ministère (MENRT), 100 %
  - Guermeur, Yann, CR CNRS, 50 %
  - Sur, Frédéric, Post-doctorant au CNRS depuis le 01/10/04, 50 %
- IBCP - Bioinformatique et RMN structurales
  - Benabdeslem, Khalid, Post-doctorant au CNRS depuis le 01/09/04, 100 %
  - Geourjon, Christophe, IR CNRS, 25 %
- LIF - BDAA
  - Denis, François, Professeur, 50%
  - Capponi, Cécile, Maître de conférences, 20%
  - Magnan, Christophe, Doctorant allocataire moniteur depuis le 01/09/04, 50%
  - Ralaivola, Liva, Maître de conférences, 20%
- IRISA - Symbiose
  - Coste, François, CR INRIA, 50%
  - Andonov, Rumen, Professeur Université de Rennes 1, 30%
  - Yanev, Nikola, Poste d'accueil CNRS pour un an, 80%
  - 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, 100%
  - Leroux, Aurélien, Doctorant INRIA, allocation Région Bretagne et INRIA, 20%
  - Mescam, Yoann, Doctorant INRIA, financement INRIA, 20%
  - Veber, Philippe, Doctorant INRIA depuis octobre 2004, bourse INRIA, 20%
  - Lahaye, Marie, Stagiaire Master 2 Informatique et Recherche, mars-juin 2005, 100%
  - Collet, Guillaume, Stagiaire Master 2 Informatique et Recherche, mars-juin 2005, 100%
  - Gruel Jeremy, Stagiaire Master 1 Bio-Informatique, avril-juin 2005, 100%

  - Wartelle, Nicolas, Stagiaire Master 1 Informatique, avril-juin 2005, 100%
- LIRMM - MAB
  - Bréhélin, Laurent, CR CNRS, 30%
  - Gascuel, Olivier, DR CNRS, 20%
  - Duprat, Elodie, Doctorante, 20%
- MIG
  - Martin, Juliette, Doctorante depuis le 01/10/02, allocation de recherche INRA (ASC), 100%
  - Taly, Jean-François, Doctorant depuis le 01/10/2003, allocation de recherche du ministère (MENRT), 100%
  - Marin, Antoine, CR INRA, 80%
  - Gibrat, Jean-Francois, DR INRA, 20%

3 Changements significatifs intervenus dans le projet

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 ; cependant, il a préféré rechercher un poste dans l'industrie plutôt que poursuivre une carrière de chercheur ou d'enseignant chercheur. Il ne fait donc plus partie du projet depuis janvier 2005.

Christophe Magnan est un nouveau doctorant (Allocataire-Moniteur) de l'équipe BDAA. 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 rejoint donc très naturellement le projet Genoto3D.

Liva Ralaivola a été recruté maître de conférences à l'université de Provence en septembre 04. Il a terminé en janvier 05 un séjour de près de 2 ans 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 bio-informatique : il a rejoint le projet Genoto3D depuis février 2005.

Frédéric Sur, Post-doctorant au CNRS, a rejoint le projet MODBIO le 01/10/04.

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

4 Résumé des principales avancées

L'identification des repliements pour les structures protéiques inconnues représente un problème fondamental en biologie structurale. Dans ce contexte, nous avons proposé une approche originale pour traiter les structures tridimensionnelles des protéines. Cette approche consiste (1) à générer une base de séquences significative pour l'apprentissage automatique et (2) 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, nous avons effectué une classification ascendante hiérarchique à partir de la matrice de dissimilarité extraite dans la première étape. A partir de cette classification, nous avons pu extraire des cœurs structuraux à tout niveau hiérarchique pour chaque famille de protéines. Finalement, un réseau de neurones a été conçu à partir de séquences modélisées par des matrices de covariance pour la reconnaissance de repliements. Par conséquent, grâce à l'extraction des cœurs structuraux, une base de méta séquences a été générée pour concevoir un système de reconnaissance puissant représentant un taux de réussite égal à 75 % sur 21 familles CATH. Ce travail a donné lieu à une publication nationale [11].

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), des règles explicites de prédiction ont pu être obtenues et validées sur des données biologiques. Le taux de prédiction obtenu atteint les 90% avec seulement 13 règles [16]. 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 cette optique, l'équipe BDAA du LIF a proposé une modélisation qui a conduit à adopter un cadre d'apprentissage semi-supervisé asymétrique, à mettre au point des variantes de l'algorithme naïf de Bayes adapté à ce cadre, à montrer la validité théorique de l'algorithme, à étudier ses performances sur des données artificielles et à l'appliquer aux données biologiques. Cette première étude a donné lieu à une publication nationale [17], d'autres soumissions à des conférences internationales étant en cours. Dans la même optique, l'équipe MAB du LIRMM s'est attachée à la recherche d'une modélisation probabiliste plus fine que celle utilisée dans l'algorithme naïf de Bayes pour modéliser les interactions intervenant dans les contextes locaux aux cystéines. L'idée sous-jacente de ce travail fondé sur les "cartes de contact" des interactions est que si ces interactions sont effectivement contraignantes, on doit pouvoir observer un biais de composition entre les acide aminés les plus proches physiquement. Suite à ces études pratiques, l'équipe BDAA du LIF a proposé un cadre formel permettant d'étudier la présence d'information locale dans une telle interaction. Des résultats théoriques montrent que sous certaines conditions peu restrictives, 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 et leur applications aux données biologiques doit faire l'objet du travail des mois qui viennent.

L'équipe BDAA du LIF a également poursuivi une étude visant à appliquer les méthodes de Boosting en prenant les algorithmes d'alignement de séquences comme classifieurs faibles de base. Ce travail a été appliqué à un problème biologique (la prédiction de protéines partenaires de transporteurs ABC) autre que celui étudié dans l'ACI Genoto3d, mais les algorithmes d'apprentissage sous-jacents sont conçus pour pouvoir

être appliqués à des familles de séquences mal conservées et ont donc un spectre d'application plus large que celle qui a été réalisée. Ce travail a donné lieu à une publication internationale [7] et à deux publications nationales [12,13]. Un problème analogue a également été traité par l'équipe MAB du LIRMM. Il vise à prédire la liaison ou l'absence de liaison des protéines de la super famille MHC avec la beta2-microglobuline en combinant classifieur Bayésien naïf et alignement multiple [15].

Concernant l'apprentissage d'automates pour la caractérisation de protéines, une nouvelle approche basée sur la fusion de fragments significativement similaires a été proposée [14]. L'implémentation du programme est en cours de finalisation et le programme sera bientôt appliqué à la caractérisation de cœurs structuraux.
En ce qui concerne le problème de repliement de protéines dit "protein threading problem" nous avons proposé la technique "cost-splitting" pour sa résolution. Cette technique avancée d'optimisation combinatoire permet de donner une nouvelle formulation du dual lagrangien associé au problème. Le dual est résolu par un algorithme de complexité polynomiale. L'accélération que nous avons obtenue par rapport aux modèles MIP, le nôtre, ainsi que celui-ci du logiciel RAPTOR, est entre 100 et 250 pour des instances intéressantes. Grâce à ces résultats, le calcul des distributions, une des tâches les plus dures pour le "protein threading", peut maintenant être considéré comme une opération de routine (article soumis à WABI'05).

Les travaux de l'unité MIG concernent 2 axes complémentaires mais distincts. Le premier a trait à des améliorations apportées à la méthode de reconnaissance de repliements FROST, développée dans l'unité et le second à des développements portant sur la modélisation *de novo* des protéines.
En ce qui concerne FROST, les améliorations portent sur la généralisation du modèle protéique employé et l'optimisation des techniques d'alignement séquence-structures tridimensionnelles (3D) utilisées.
A l'origine, FROST a été conçu pour permettre de comparer des séquences avec des structures de protéines globulaires. Pour cela, des descripteurs de la structure comme les hélices alpha, les feuillets bêta et l'enfouissement des acides aminés ont été utilisés avec succès. Des modifications ont été entreprises qui visent à généraliser la notion de conservation de la structure qui peut être définie par n'importe quel critère a priori, et aussi sur la notion d'état structural qui peut, lui aussi, être défini de manière totalement libre. La conséquence majeure de ces modifications est la possibilité de modéliser n'importe quel type de protéine, comme les protéines transmembranaires, mais également de tester différentes alternatives pour la définition des régions structurellement conservées.
En collaboration avec R. Andonov, S. Balev, N. Yanev et V. Poirriez, l'algorithme pour la résolution du problème d'alignement de la séquence sur les structures 3D a été considérablement amélioré. Ce problème, démontré NP-difficile, constitue l'étape limitante des techniques de reconnaissance de repliements. Nous disposons maintenant d'une série d'algorithmes pour effectuer ces alignements dont le plus performant permet d'explorer des espaces dont le nombre de solutions est de l'ordre de 10^{77} en quelques minutes. Cet algorithme a de plus été parallélisé pour pouvoir être exécuté sur une grappe d'ordinateurs.
Les méthodes de modélisation *de novo* ont pour but de déterminer la structure 3D d'une protéine uniquement à partir de sa séquence en acides aminés. Une première étape dans ce processus consiste à prédire 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). Ces dernières structures, qui représentent environ 50% des résidus, ne fournissent aucune information précise sur la structure locale de la chaîne polypeptidique. Nous avons donc développé une méthode basée sur des modèles de Markov cachés qui permet de prédire les zones d'angles phi et psi du diagramme de Ramachandran caractéristiques des protéines. Cette méthode permet de prédire correctement la localisation de 77% des résidus dans les 3 grandes zones du diagramme de Ramachandran. En parallèle, nous avons également développé un nouveau jeu de paramètres permettant d'estimer la qualité des modèles obtenus par modélisation moléculaire.

L'étude de la mise en œuvre de SVM multi-classes [3] pour développer des modules d'une méthode de prédiction *ab initio* de la structure tertiaire s'est orientée suivant deux axes. Tout d'abord, nous avons proposé un algorithme permettant de fixer automatiquement les valeurs des hyperparamètres [8]. Cette méthode de sélection de modèle s'appuie sur une borne de convergence uniforme du risque. Son évaluation est en cours en prédiction de la structure secondaire et en prédiction des ponts disulfures. Poursuivant nos travaux sur la spécification de noyaux dédiés au traitement de séquences biologiques [1], nous développons également un noyau fondé sur un pair-HMM. Il devrait permettre de mesurer de manière plus précise les similitudes entre segments protéiques en prenant en compte plus finement les phénomènes de l'évolution biologique, en particulier les insertions/délétions.

5 Réalisations obtenues dans le cadre du projet

Une version du logiciel de M-SVM développé dans le projet MODBIO a été dédiée au traitement des données protéiques. Elle a fait l'objet d'un dépôt à l'APP sous le numéro IDDN suivant : 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 Web de l'unité MIG `http://genome.jouy.inra.fr/frost/` sous licence GPL.

Un serveur Web d'extraction des cœurs structuraux des protéines a été développé à l'IBCP (`http://pig-devel.ibcp.fr/cgi-bin/asce/asce.py`). Il est actuellement en phase de validation en interne du laboratoire et devrait être mis en ligne officiellement début juin.

6 Réunions et Conférences organisées dans le cadre du projet

Réunions propres au 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

Plusieurs orateurs extérieurs participeront à la réunion du 30 mai. Cette réunion sera en particulier l'occasion pour le projet de renforcer et d'étendre sa collaboration avec Pierre Baldi, invité par François Denis.

7 Soutiens obtenus en liaison avec ce projet

7.1 Postes chercheurs

Pour l'année universitaire 2004-2005, François Denis, Professeur à l'Université de Provence, a obtenu un accueil en délégation dans le projet MODBIO de l'INRIA Lorraine.

Khalid Benabdeslem, Post-doctorant au CNRS, a rejoint le laboratoire de Bioinformatique et RMN structurales de l'IBCP le 01/09/04.

La thèse intitulée "Conception, mise en œuvre et évaluation de machines à noyau dédiées au traitement de séquences biologiques" qu'effectue depuis le 01/10/03 Yannick Darcy, sous la direction d'Alexander Bockmayr et Yann Guermeur, est financée par une allocation de recherche ministérielle accordée par l'ACI "Masses de Données".

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.

7.2 Postes ingénieurs

7.3 Contrats nationaux

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" financé par le thème bioinformatique du PRST "Intelligence Logicielle" du CPER lorrain. Cette opération, qui débute en 2005, durera jusqu'en 2007.

7.4 Contrats européens

7.5 Contrats internationaux hors CEE

7.6 Contrats industriels

7.7 Contacts internationaux dans le cadre de ce projet

8 Publications obtenues dans le cadre du projet

– Chapitres de livres

[1] Y. Guermeur, A. Lifchitz et R. Vert (2004). A kernel for protein secondary structure prediction. Dans *Kernel Methods in Computational Biology*, édité par B. Schölkopf, K. Tsuda et J.-P. Vert, The MIT Press, Chapitre 9, 193-206.

– Journaux internationaux

[2] R. Andonov, S. Balev et N. Yanev (2004). Protein threading problem : From mathematical models to parallel implementations. *INFORMS Journal of Computing, special issue on computational molecular biology/bioinformatics*, **Vol. 16**, N°4, 393-405.

[3] Y. Guermeur, A. Elisseeff et D. Zelus (2005). A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers. *Applied Stochastic Models in Business and Industry (ASMBI)*, **Vol. 20**, N°2, 199-214.

[4] J. Martin, J.-F. Gibrat et F. Rodolphe. Choosing an optimal hidden Markov model for protein secondary structure prediction. *IEEE Intelligent System, Special Issue on Data Mining for Bioinformatics*, (à paraître).

[5] J. Martin, G. Letellier, A. Marin, J.-F. Taly, A.G. de Brevern et J.-F. Gibrat. Protein secondary structure assignment revisited : a detailed analysis of different assignment methods, (soumis).

[6] N. Yanev et R. Andonov (2004). Parallel divide & conquer approach for the protein threading problem. *Concurrency and Computation : Practice and Experience*, **Vol. 16**, 961-974.

– Conférences internationales

[7] C. Capponi, G. Fichant, Y. Quentin et F. Denis (2005). Classification of domains with boosted blast. Actes de *Applied Stochastic Models and Data Analysis (ASMDA'05)*, 136-144.

[8] Y. Guermeur, M. Maumy et F. Sur (2005). Model selection for multi-class SVMs. Actes de l'*ASMDA'05*, 507-516.

[9] J. Martin, J.-F. Gibrat et F. Rodolphe (2005). Hidden Markov model for protein secondary structure. Actes de l'*ASMDA'05*, 180-187.

[10] V. Poirriez, A. Marin, R. Andonov et J.-F. Gibrat (2005). FROST : Revisited and distributed, Actes du *Fourth IEEE International Workshop on High Performance Computational Biology (HiCOMB'05).*

– Journaux nationaux

– Conférences nationales

[11] K. Benabdeslem, G. Deléage et C. Geourjon (2005). Core's extraction based neural network model for proteins fold recognition. Actes des *Journées*

Ouvertes Biologie Informatique Mathématiques (JOBIM'05), (à paraître).

[12] C. Capponi, G. Fichant F. Denis et Y. Quentin (2005). Boosting BLAST for classifying proteins. Actes de *JOBIM'05*, (à paraître).

[13] C. Capponi, G. Fichant et Y. Quentin (2005). Classification of domains with boosted blast. Actes de la *Conférence d'Apprentissage (CAp'05)*, (à paraître).

[14] F. Coste, G. Kerbellec, B. Idmont, D. Fredouille et C. Delamarche (2004). Apprentissage d'automates par fusions de paires de fragments significativement similaires et premières expérimentations sur les protéines MIP. Actes de *JOBIM'04*.

[15] E. Duprat, M.-P. Lefranc et O. Gascuel (2005). Prédire l'interaction des protéines de la superfamille du MHC avec la beta2-microglobuline en combinant classifieur Bayésien naïf et alignement multiple IMGT. Actes de *JOBIM'05*, (à paraître).

[16] I. Jacquemin et J. Nicolas (2005). Modélisation de cystéines oxydées à l'aide de la programmation logique inductive. Actes de *JOBIM'05*, (à paraître).

[17] C. Magnan (2005). Apprentissage semi-supervisé asymétrique et estimations d'affinités locales dans les protéines. Actes de *CAp'05*, (à paraître).

[18] J. Martin, J.-F. Taly, J.-F. Gibrat et F. Rodolphe (2005). Choice of the optimal hidden Markov model for secondary structure. Actes de *JOBIM'05*, (à paraître).

[19] J.-F. Taly, J. Martin, A. Marin et J.-F. Gibrat (2005). Définition de mesures décrivant l'environnement local des acides aminés pour une application à l'évaluation des modèles structuraux. Actes de *JOBIM'05*, (à paraître).