Proposition de recherche

Algorithmes d'apprentissage automatique à grande échelle pour l'élaboration de nouveaux composés pharmaceutiques

par : Alexandre Drouin

Directeur : François Laviolette Codirecteur : Mario Marchand

---

1 Introduction

Les interactions protéine-protéine interviennent dans la plupart les processus biologiques et cellulaires [14, 1]. Une protéine est constituée d'une séquence ordonnée de plusieurs acides aminés. C'est la combinaison des acides aminés d'une séquence qui détermine les propriétés biologiques d'une protéine.

Une intéraction entre protéines survient lorsque certains acides aminés d'une protéine se lient à quelques acides aminés d'une autre protéine. Ceux-ci forment alors un site de liaison qui est le point central de leur intéraction. Une fois cette liaison effectuée, on assiste à une cascade d'évènements pouvant inhiber ou déclencher certains processus biologiques. Le présent projet de recherche portera sur les interactions protéine-protéine impliquées dans le déclenchement de la réponse immunitaire. Plus précisément, nous nous intéresserons au développement médicaments servant à déclencher une réponse immunitaire, comme les vaccins.

À cette fin, nous nous intéresserons au déclenchement artificiel de la réponse immunitaire à l'aide d'outils biologiques comme les peptides. Un peptide est une petite biomolécule composée d'une courte séquence d'acides aminés pouvant servir de composant de base dans le développement de nouveaux médicaments. Par ailleurs, les peptides s'avèrent des composants de choix d'une part pour leur non-toxicité et leur fabrication en laboratoire, mais également parce qu'ils minimisent les interactions entre médicaments [2].

Des techniques expérimentales, comme la chimie combinatoire [15], permettent de générer un ensemble de peptides et de déterminer ceux qui se lient ou non à une protéine. Toutefois, le choix du meilleur peptide à utiliser pour inhiber une protéine requiert une recherche complexe. Par exemple, dans le cas où nous recherchons un peptide composé d'une séquence de 9 acides aminés, puisqu'il existe 20 acides aminés différents, l'ensemble de tous les peptides possibles de taille 9 est de l'ordre 20^9. Il n'est donc pas efficace de valider le potentiel de liaison de tous les peptides avec une protéine par l'entremise d'expériences in vitro.

De ce fait, la prédiction de la liaison peptide-protéine est un problème ayant fait l'objet de plusieurs études dans le domaine de la biologie computationnelle et ayant engendré le développement de nombreuses méthodes de prédiction basées sur l'apprentissage statisique (GS Kernel [5], NetMHCIIpan [10], NetMHCpan [9], MultiRTA [3], RTA [4], etc.). Ces méthodes se regroupent généralement en deux catégories : les méthodes de classification et les méthodes de régression.

Les méthodes de classification permettent de prédire, sous forme d'une réponse binaire, si un peptide et une protéine se lient. En revanche, les méthodes de régression permettent de prédire quantitativement l'affinité de liaison d'un peptide et d'une protéine. Ces dernières permettent donc de distinguer les complexes ayant une forte affinité de liaison de ceux ayant une affinité de liaison plus faible.

En somme, les méthodes de prédiction d'affinité de liaison peptide-protéine sont des outils de choix pour les chercheurs en biologie. Elles peuvent leur fournir des pistes pour cibler les composants à utiliser dans le développement de nouveaux composés pharmaceutiques, tel que les vaccins et les médicaments.

2 Problématique

Il existe plusieurs approches différentes d'apprentissage supervisé, tel que les réseaux de neurones, les méthodes à noyaux, les arbres de décision et plusieurs autres. Le principe de base de ces méthodes est le même : entraîner un prédicteur sur un ensemble d'exemples et l'utiliser pour inférer la réponse à de nouveaux exemples. En apprentissage automatique et en biologie computationnelle, les méthodes à noyau sont largement utilisées. Par exemple, Laurent Jacob et Jean-Philippe Vert[6], ont appliqué des méthodes à noyaux au problème de prédiction d'affinité de liaison peptide-protéine et ont obtenu des résultats surpassant ceux issus de l'état de l'art.

Nous avons récemment étudié le problème de prédiction de l'affinité de liaison entre molécules de complexe majeur d'histocompatibilité de classe II (MHC-II)[16] et peptides. Les molécules de MHC-II sont présentes dans les cellules des organismes vivants et participent au déclanchement de la réponse immunitaire. Puisque celles-ci peuvent être représentées par la séquence de la protéine qui les constitue, il s'agit d'un problème de prédiction d'affinité de liaison peptide-protéine.

Dans le but de résoudre ce problème, nous avons entraîné un algorithme de régression de ridge à noyau [13] sur des ensembles de données portant sur l'intéraction MHC-II-peptide et nous avons comparé nos résultats à ceux des méthodes NetMHCIIpan [10], RTA [4] et MultiRTA [3]. Les résultats obtenus surpassent significativement ceux de ces méthodes issues de l'état de l'art [5].

Bien que ces résultats soient encourageant, ils ne couvrent qu'une faible partie des molécules de MHC-II. Pour cette raison, nous comptons utiliser la régression de ridge à noyau pour entraîner un prédicteur sur un plus grand volume de données. Ceci nous permettra de développer un outil plus précis, permettant la prédiction de l'affinité de liaison entre une molécule de MHC-II quelconque et un peptide. Les données utilisées pour entraîner ce prédicteur proviendront de bases de données comme l'Immune Epitope Database (IEDB) [11], qui renferme de l'information sur plus de 220562 complexes MHC-peptide.

Tel que mentionné précedemment, les méthodes basées sur l'apprentissage automatique arrivent à bien performer pour prédire l'affinité de liaison peptide-protéine. Toutefois, plusieurs études portant sur la liaison MHC-peptide ont démontré que l'apprentissage multi-tâche peut significativement augmenter les performances des prédicteurs [6, 16, 5, 3, 9, 10]. Pour plus d'information sur l'apprentissage multi-tâche, le lecteur intéressé peut consulter [6]. Ce type d'apprentissage apprend plusieurs problèmes d'apprentissage à la fois et permet de partager l'information entre les problèmes similaires. Dans le cas de la prédiction des affinités de liaison MHC-II-peptide, un problème d'apprentissage consiste en la prédiction des peptides se liant à une molécule de MHC-II. Donc, ce type d'apprentissage s'avère un atout pour les molécules pour lesquelles peu de données ou aucunes données expérimentales sont disponibles. Toutefois, l'utilisation de telles approches augmente considérablement le nombre d'exemples considérés par le prédicteur et leur utilisation est restreinte par la puissance de calcul des ordinateurs actuels.

En plus du grand nombre d'exemples à considérer, la complexité algorithmique de la plupart des méthodes à noyau est de l'ordre de $\mathcal{O}(n^3)$ où n est le nombre d'exemples d'apprentissage. Ces méthodes sont aussi gourmandes au niveau de la mémoire, puisqu'elles nécessitent le calcul d'une matrice de similarité entre les exemples occupant un espace mémoire de l'ordre de $\mathcal{O}(n^2)$.

En outre, un autre problème vient du fait que pour effectuer une prédiction, les méthodes à noyaux doivent évaluer la valeur de la fonction de noyau entre le nouvel exemple et tous les exemples d'apprentissage. Ainsi, l'efficacité algorithmique de ces méthodes à l'étape de prédiction se détériore en fonction du nombre d'exemples d'entraînement. Pour palier à ce problème, l'utilisation de prédicteur creux, c'est-à-dire de prédicteurs qui accordent un poids nul à certains exemples dans la prédiction, pourrait être intéressante. Ceci permettrait d'omettre l'évaluation de la fonction de noyau pour plusieurs exemples d'entraînement, accélérant ainsi les prédictions.

En somme, avec l'augmentation de la taille des ensembles de données, l'utilisation des méthodes à noyaux sera limitée si les algorithmes existants ne sont pas adaptés pour faire face aux grands volumes de données.

3 Objectifs

Dans le cadre de ce projet de recherche, nous proposons d'explorer deux types de méthodes à noyaux qui seraient adaptées pour l'apprentissage à grande échelle : les méthodes d'approximation matricielles et les méthodes d'ensembles.

Les méthodes d'approximation matricelles se basent sur l'approximation de la matrice de similarité entre les exemples. Elles permettent d'adapter les méthodes à noyaux aux grands volumes de données en calculant une approximation de la matrice de similarité de manière plus efficace. La méthode de Nyström [12] est une méthode appartenant à cette catégorie qui a récemment fait l'objet de plusieurs études [12, 7].

Les méthodes d'ensembles se basent sur la séparation d'un problème d'apprentissage à grande échelle en plusieurs problèmes de plus petite échelle. Typiquement, lorsqu'un calcul exigeant computationnellement doit être effectué, il y a une étape de réduction, c'est-à-dire où le problème initial est divisé en plusieurs sous-problèmes moins exigeants et une étape de combinaison, où le résultats des sous-problèmes sont recombinés pour fournir une solution au problème initial. Parmi ces méthodes, on compte la méthode Ensemble Nyström [7] qui construit plusieurs approximations de la matrice de similarité en utilisant la méthode de Nyström et qui les recombine pour produire une meilleure approximation de cette matrice.

Finalement, nous appliquerons ces algorithmes d'apprentissage à grande échelle à la biologie computationnelle, notamment au problème d'énergie de liaison peptide-protéine, de façon à évaluer la performance de nos algorithmes et à permettre une plus grande utilisation des données disponibles. Ces méthodes seront tout aussi bénéfiques pour la recherche en biologie computationnelle que pour le domaine de l'apprentissage automatique en général.

4 Méthodologie

Pour commencer, Alexandre Drouin fera une revue de littérature plus détaillée au sujet de l'application de méthodes à noyau à la biologie computationnelle. Par la suite, il reproduira les résultats en appliquant des méthodes à noyaux aux ensembles de données tirées des articles

qu'il aura trouvés. Parmi ces ensembles de données, il conservera ceux de grande taille ($\geq$ 10000 exemples) pour leur appliquer des méthodes à grande échelle.

De plus, il fera une revue de littérature au sujet des méthodes d'apprentissage à grande échelle comme les techniques d'approximation de matrices de similarité [12, 7] et les méthodes d'ensembles [7, 8]. Il essaiera d'appliquer ces méthodes aux ensembles de données liés à la biologie qu'il aura trouvés et à certains ensembles de données extraits des travaux réalisés par les collègues de son laboratoire, le GRAAL..

Tout au long de son projet de recherche, il collaborera avec d'autres étudiants travaillant sur des projets de biologie computationnelle. L'expertise qu'il développera dans le domaine des méthodes à noyaux à grande échelle l'aidera sans doute à apporter une bonne contribution à ces projets.

5 Échéancier

À la session d'automne 2012, Alexandre a suivi deux des cours de sa scolarité obligatoire de maîtrise. Parallèlement, il fera une revue de littérature portant sur l'application des méthodes à noyau en biologie computationnelle. De plus, il travaillera sur l'écriture de deux articles en vue de leur soumission dans le Journal of Immunological Methods.

À la session d'hiver 2013, il suivra les 3 derniers cours de sa scolarité obligatoire, dont le cours d'introduction à la recherche en informatique. Dans le cadre de ce dernier, il aura à fournir des livrables en lien avec la rédaction de son mémoire de maîtrise.

Par la suite, aux sessions d'été 2013 et d'automne 2013, il s'affairera à la rédaction de son mémoire, qui sera le livrable final de son projet de maîtrise.

6 Répartition des crédits de recherche

Session	**Nombre de crédits**	**Objectifs**
Automne 2012	7	Activité de recherche
Hiver 2013	7	Activité de recherche
Été 2013	8	Rédaction du mémoire
Automne 2013	8	Rédaction du mémoire

7 Ressources nécessaires

Le développement sera fait par Alexandre Drouin sur son ordinateur personnel. Il occupera un bureau au GRAAL situé au PLT-3908. Tous les calculs seront effectués sur les superordinateurs appartenant au consortium Calcul Canada.

Références

[1] Albert, R. Scale-free networks in cell biology. *Journal of cell science 118*, Pt 21 (Nov. 2005), 4947–57.

[2] Ayoub, M., and Scheidegger, D. Peptide drugs , overcoming the challenges , a growing business. *Chemistry Today 24*, 4 (2006), 46–48.

[3] Bordner, A., and Mittelmann, H. MultiRTA : A simple yet reliable method for predicting peptide binding affinities for multiple class II MHC allotypes. *BMC Bioinformatics 11*, 1 (2010), 482.

[4] Bordner, A. J., and Mittelmann, H. D. Prediction of the binding affinities of peptides to class II MHC using a regularized thermodynamic model. *BMC Bioinformatics 11*, 1 (2010), 41.

[5] Giguère, S., Marchand, M., Laviolette, F., Drouin, A., and Corbeil, J. Learning a peptide-protein binding affinity predictor with kernel ridge regression. 22.

[6] Jacob, L., and Vert, J.-P. Efficient peptide–mhc-i binding prediction for alleles with few known binders. *Bioinformatics 24*, 3 (2008), 358–366.

[7] Kumar, S., and Mohri, M. Ensemble nystrom method. *Neural Information Processing Systems* (2009), 1–9.

[8] Machart, P., Universit, A.-m., Peel, T., Universit, A.-m., Anthoine, S., Ralaivola, L., and Universit, A.-m. Stochastic Low-Rank Kernel Learning for Regression.

[9] Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Rø der, G., Peters, B., Sette, A., Lund, O., and Buus, S. r. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PloS one 2*, 8 (Jan. 2007), e796.

[10] Nielsen, M., Lundegaard, C., Blicher, T., Peters, B., Sette, A., Justesen, S., Buus, S. r., and Lund, O. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence : NetMHCIIpan. *PLoS computational biology 4*, 7 (Jan. 2008), e1000107.

[11] Peters, B., Sidney, J., Bourne, P., Bui, H.-H., Buus, S., Doh, G., Fleri, W., Kronenberg, M., Kubo, R., Lund, O., Nemazee, D., Ponomarenko, J. V., Sathiamurthy, M., Schoenberger, S., Stewart, S., Surko, P., Way, S., Wilson, S., and Sette, A. The immune epitope database and analysis resource : from vision to blueprint. *PLoS biology 3*, 3 (Mar. 2005), e91.

[12] Seeger, C. W., and Matthias. Using the Nyström Method to Speed Up Kernel Machines. In *Advances in Neural Information Processing Systems 13* (2001), MIT Press, pp. 682–688.

[13] Shawe-Taylor, J., and Cristianini, N. *Kernel Methods for Pattern Analysis*, illustrated edition ed. Cambridge University Press, June 2004.

[14] Toogood, P. L. Inhibition of protein-protein association by small molecules : approaches and progress. *Journal of medicinal chemistry 45*, 8 (May 2002), 1543–58.

[15] Wilson, S., and Czarnik, A. Combinatorial chemistry, synthesis and application. *European Journal of Medicinal Chemistry 32*, 11 (1997), 842–842.

[16] Zhang, L., Udaka, K., Mamitsuka, H., and Zhu, S. Toward more accurate pan-specific MHC-peptide binding prediction : a review of current methods and tools. *Briefings in bioinformatics* (Sept. 2011).

Signatures :

Directeur : _____ Date: _____

Codirecteur : _____ Date: _____

Étudiant : _____ Date: _____

Appréciation et signature des membres du cas

_____ Date: _____

_____ Date: _____

_____ Date: _____

Décision du Comité d'admission et de supervision

_____

_____

_____

_____

_____

_____

_____

_____

Date : _____ Signature du président : _____