Une opération de docking croisé à grande échelle pour la détection de partenaires protéiques potentiels.

Sophie Sacquin-Mora, Richard Lavery
Laboratoire de Biochimie Théorique, UPR 9080
Institut de Biologie Physico-Chimique
13 rue Pierre et Marie Curie
75005 Paris

Ladislas Trojan, Alessandra Carbone
Equipe de Génomique Analytique, INSERM U511
Université Pierre et Marie Curie
91 Bd de l'Hôpital
75013 Paris

Introduction

Ce travail, qui fait partie des trois projets sélectionnés dans le cadre du programme DECRYPTON (mis en place par l'Association Franaise contre les Myopathies, le CNRS et IBM) [1] pour l'année 2005, entend réaliser une étude à grande échelle des interactions protéine-protéine afin de mieux comprendre leur spécificité. En effet, les phénomènes d'association protéique jouent un rôle majeur dans le fonctionnement cellulaire, et leur compréhension représente un problème fondamental de la biologie, où des avancées auraient des conséquences immédiates dans le domaine de l'ingénierie protéique. **Notre objectif est donc de combiner des approches bioinformatiques (mises au point dans le groupe de Génomique Analytique) et de modélisation moléculaire (développées au Laboratoire de Biochimie Théorique), afin de pouvoir localiser les sites d'interaction à la surface des protéines et identifier les partenaires potentiels d'une protéine donnée** au sein d'une base de données comprenant des milliers structures (du type Protein Data Bank [2]). Ce projet a été développé en collaboration avec l'équipe de J.-M. Chesneaux du Laboratoire d'Informatique de Paris 6 [3] afin de mettre à profit leur compétence en matière de parallélisation d'algorithmes.

Résultats

Dans un premier temps nous avons développé un algorithme de Docking, permettant de rechercher les géométries d'interaction optimales entre deux partenaires protéiques. À cet effet nous avons repris la description réduite des protéines mise au point par l'équipe de M. Zacharias à Brême [4] où chaque acide aminé est représenté par un à trois pseudo-atomes selon la taille de sa chaîne latérale. Dans le cadre de cette représentation, les interactions protéine-protéine sont décrites via un potentiel simple qui comprend un terme de type van der Waals (reflétant les propriétés physico-chimiques des différents acides-aminés) et un terme de type Coulombien (pour les interactions électrostatiques).

Pour un couple récepteur (protéine fixe)/ligand (protéine mobile) donné, l'algorithme de docking génère un ensemble de positions de départ pour lesquelles l'énergie d'interaction va être minimisée en jouant sur quatre degrés de liberté : La distance ligand-

récepteur et les trois angles d'orientation du ligand. **Un calcul complet nous permet alors d'établir une carte de la surface énergétique du récepteur pour un ligand donné et notamment de localiser les zones d'interaction favorables à la surface du récepteur.** Du fait du très grand nombre de positions de départ du ligand nécessaires pour explorer correctement la surface d'interaction (entre 100000 et 500000 points selon la taille du récepteur), l'algorithme a été spécialement développé pour permettre la mise en place de calculs parallèles et l'exploitation de la grille de calcul universitaire DECRYPTHON, et donc réduire grandement les temps de calcul nécessaires.
Cet algorithme va être exploité dans le cadre d'une expérience de "Docking Croisé" sur une base de données de 89 complexes protéiques [5]. Le processus de docking sera appliqué non seulement aux partenaires protéiques connus, mais aussi à la totalité des paires de protéines possibles, qu'il s'agisse de partenaires identifiés expérimentalement ou non. Nous allons donc nous intéresser pour la première fois à des protéines qui, *en principe*, n'interagissent pas ensemble, ce qui nous permettra d'établir une base de "decoys" (faux positifs) de bonne qualité et qui pourront être exploités dans l'élaboration de potentiels d'interaction. L'ensemble de ces calculs (soit près de 25000 opérations de docking) sera distribué sur le World Community Grid [6] une grille d'internautes constituée d'ordinateurs personnels et mise en place avec IBM.

Les premiers calculs effectués sur un ensemble réduit de cinq complexes protéiques montrent l'efficacité du programme. Dans chaque cas les cartes énergétiques établies par minimisations multiples mettent en évidence un puit de potentiel au niveau de la position cristallographique du ligand par rapport au récepteur. De plus, l'algorithme permet de retrouver pour chacun des complexes une conformation où les atomes du ligand présentent un écart quadratique moyen par rapport à leur position cristallographique inférieur à 3 Å(voir la figure 1). Néanmoins, certains "faux" complexes peuvent présenter des énergies d'interactions ou des interfaces comparables à celles des complexes expérimentaux, ce qui souligne le problème que pose actuellement l'identification des partenaires spécifiques au sein d'une large base de données protéique.
Dans une seconde phase d'analyse des données issus des calculs de docking croisé, nous avons mis au point un indice d'association qui tient compte à la fois de l'énergie d'interaction obtenue lors d'un calcul de docking et des résidus présents au niveau de l'interface protéique résultante. Pour un couple protéique donné, cet indice est d'autant plus important que l'interaction entre les deux protéines est favorable. Pour toutes les

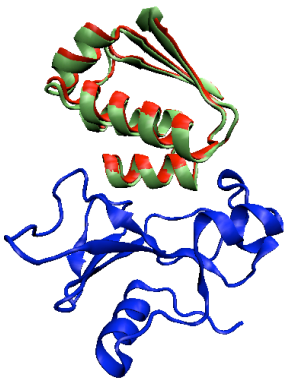

FIG. 1 – *Complexe protéique barnase (en bleu)/barstar, avec en rouge la position cristallographique du ligand et en vert sa position prédite par l'algorithme de docking.*

protéines de notre base réduite, le partenaire protéique présentant l'indice d'association maximal correspond systématiquement au partenaire expérimental (voir la figure 2). **Pour la première fois, notre algorithme permet donc de déterminer comment deux protéines vont pouvoir s'associer, mais aussi quelles sont les protéines au sein d'une base de données qui sont susceptibles d'interagir pour former un complexe spécifique.**

Parallèlement à ces travaux, l'équipe d'A. Carbone a mis au point un programme de détection des sites fonctionnels dans les protéines, appelé Joint Evolutionary Trees (JET), exploitant la méthode "Evolutionary Trace" [7]. À partir d'un ensemble d'arbres phylogénétiques construits pour une famille de protéines donnée, cette approche permet d'extraire des résidus "trace" de la séquence en acides aminés, ces résidus correspondant à des positions conservées dans les différentes branches des arbres. Des études préalables ont montré que les résidus trace ainsi obtenus forment des clusters dans la structure tridimensionnelle de la protéine et sont localisés au niveau des sites fonctionnels ou des interfaces macromoléculaires (voir la figure 3). **Cette méthode peut donc nous apporter des informations concernant les sites d'interaction à partir de la seule structure primaire d'une protéine et sans aucune donnée sur ses partenaires potentiels.**

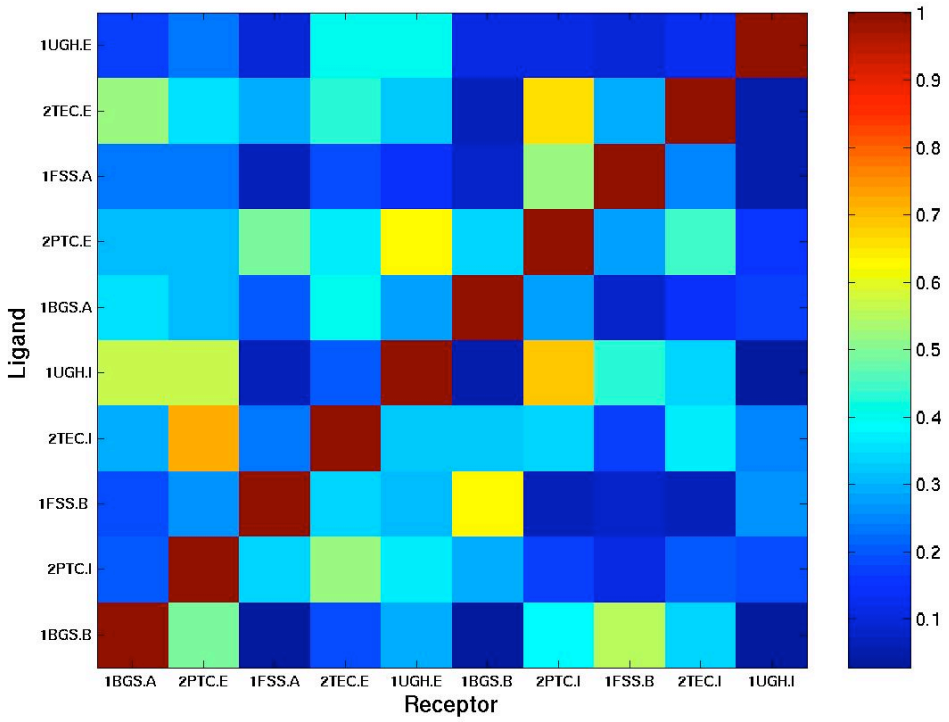

FIG. 2 – *Matrice de docking croisé obtenue pour une base réduite de cinq complexes protéique, soit dix protéine distinctes. Les protéines ont été ordonnées de manière à ce que les partenaires expérimentaux soit placés sur la diagonale, ceux ci présentent systématiquement le meilleur indice d'association.*

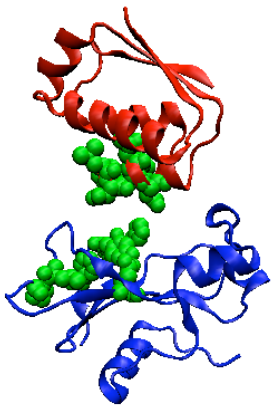

FIG. 3 – *Complexe protéique barnase (en bleu)/barstar(en rouge), les résidus d'interface détectés par ET sont représentés en vert.*

JET permet de prendre en compte des protéines de basse homologie (avec moins de 30% d'identité par rapport à la séquence de référence) lors de la constructions des arbres phylogénétiques et donc d'améliorer la robustesse des prédictions du programme.

Aprés avoir été testées séparément, les deux approches (bioinformatique et modélisation) seront appliquées aux protéines de la base de données de Mintseris et Col. [5]. **Les informations ainsi obtenues sur les interfaces macromoléculaires seront recoupées pour mettre au point une base de données des sites d'interaction protéiques. Les résultats des calculs de docking croisé effectués à grande échelle seront utilisés pour comparer interactions spécifiques (entre partenaires expérimentaux) et non-spécifiques. Les informations obtenues par ET permettront quant à elles de réduire le coût des calculs de docking en limitant l'exploration des surfaces protéiques aux sites d'interaction détectés prélablement** (ce qui représente une réduction des points de départ nécessaires par un facteur cent). Le gain réalisé en matière de temps de calculs rendra alors possible l'analyse de bases de données protéiques nettement plus large (comprenant plusieurs milliers de structures).
A terme, cette approche pluridisciplinaire sera appliquée à un ensemble de protéines connues pour leur implication dans les maladies neuromusculaires afin d'identifier des partenaires d'association potentiels au sein des bases de données protéiques.

[1] http ://www.decrypthon.fr
[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *The Protein Data Bank*, Nucleic Acids Research, **28**, 235 (2000).
[3] http ://www-anp.lip6.fr
[4] M. Zacharias, *Protein-protein docking with a reduced protein model accounting for side-chain flexibility*, Protein Science **12**, 1271 (2003).
[5] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, and Z. Weng, *Protein-protein docking benchmark 2.0 : An update*, Proteins **60**, 214 (2005).
[6] http ://www.worldcommunitygrid.org.
[7] O. Lichtarge and M. E. Sowa, *Evolutionary predictions of binding surfaces and interactions*, Curr. Opin. Structur. Biol. **12**, 21 (2002).