De Azevedo Kévin
Duhamel Marine

Prédiction de la structure 3D du complexe barnase-barstar par docking moléculaire

Introduction

La prédiction des interactions protéine-protéine (ou protéine-peptide) est un domaine de recherche actif de la modélisation moléculaire, principalement à visée thérapeutique ou de recherche fondamentale. Les méthodes de modélisation d'interaction moléculaires sont fréquemment utilisées dans le domaine pharmaceutique afin de développer des molécules permettant d'interagir avec le site actif d'une protéine et ainsi de l'inhiber.

Le docking est une méthode de prédiction des interactions protéiques computationnelle, qui a pour but d'identifier la conformation optimale d'un ligand par rapport à son récepteur, permettant leur interaction physique. Elle est appliquée à des molécules dont la structure 3D a été prédite ou cristallisée, présentes dans la base de données PDB (Protein Data Bank). Cette méthode évalue la meilleure conformation d'un ligand par rapport à son récepteur, pour prédire la structure du complexe ligand-récepteur.

La position du récepteur est maintenue fixe pendant tous les calculs. Pour explorer la surface du récepteur, celle-ci est discrétisée selon deux angles (θ,φ). Chaque couple de valeurs d'angles correspond à une position du ligand par rapport au récepteur. Ainsi, la combinaison de deux angles permet d'obtenir un grand nombre de positions du ligand par rapport au récepteur. Pour chaque position du ligand, plusieurs orientations sont calculées. A nouveau, un couple de valeurs d'angles (θ',φ') est utilisé pour discrétiser la surface du ligand. Il y a alors autant de conformations ligand possibles que le produit du nombre de positions du ligand par le nombre d'orientations du ligand.

Les conformations du ligand par rapport au récepteur sont ensuite évaluées à l'aide d'une fonction de score, afin d'identifier la conformation réelle du ligand par rapport au récepteur. Les molécules sont généralement considérées comme rigides, pour faciliter les calculs. Il en résulte une difficulté importante à trouver la bonne solution lorsque l'interaction ligand-récepteur s'accompagne d'une modification conformationnelle de l'un ou des deux partenaires. Cependant les méthodes considérant les molécules rigides donnent de bons résultats dans le cas où les changements conformationnels sont faibles.

L'étape la plus cruciale des méthodes de docking est de trouver un critère efficace permettant de prédire avec la plus grande confiance possible la

structure du complexe : c'est le rôle de la fonction de score. Un début de définition de cette fonction de score peut être trouvée dans les modèles d'interactions moléculaires : en effet, les protéines se liant entre elles sont soumises à des forces moléculaires, certaines attractives comme les interactions faibles de Van der Waals ou répulsives entre molécules de charges égales. Cornell et al. (1) propose un tel modèle, prenant en compte les termes liés aux molécules (termes relatifs à la distance entre deux atomes ou leur angle de torsion) et des termes non liés (les interactions de Van der Waals et les charges des atomes) : ces termes sommés permettent d'obtenir l'énergie globale du complexe, qui tend à être minimisée dans les conformations stables, les plus probables à l'état naturel. Cependant, les acides aminés ont d'autres propriétés : certains présentent une polarité de charges, d'autres des propriétés hydrophobes ou hydrophiles, et d'autres encore possèdent des chaînes latérales pouvant causer de l'encombrement stérique. Ainsi, une fonction de score impliquant uniquement l'énergie de liaison devra être améliorée en tenant compte des autres propriétés physico-chimiques des résidus. De plus, ces critères doivent être pondérés selon l'importance qu'ils ont dans la formation du complexe.

Notre objectif est de développer une méthode pour évaluer les solutions de conformations de ligand par rapport à un récepteur, et de déterminer la structure réelle du complexe formé par leur interaction.

Matériel et méthodes

Matériel

Nous disposons d'un échantillon de 948 conformations de ligand contenues dans des fichiers pdb, et de la structure 3D du récepteur, également au format pdb. Les 948 conformations du ligand sont issues d'une première curation de 200 000 conformations possibles du ligand, réalisée avec une fonction de score qui nous est inconnue. Ces 200 000 conformations de ligand ont été obtenues à partir de 200 positions de ligand initiales et 1000 orientations pour chacune des positions.

Les données d'entrées sont donc des fichiers PDB de protéines. Seules certaines entrées dans ces fichiers nous intéressent : les chaînes protéiques, les acides aminés, les atomes et les coordonnées spatiales de ces atomes. Les fichiers d'entrée sont parsés à l'aide d'une fonction « *parserPDB* » qui range les données d'intérêt dans un dictionnaire.

Méthodes

Fonction de score

La fonction de score a été implémentée à partir des termes non liés de la fonction d'énergie décrit dans Cornell et al., JACS 1995 (1), notée Eij :

$$E_{ij} = \frac{A_{ij}}{R_{ij}^{8}} - \frac{B_{ij}}{R_{ij}^{6}} + f\frac{q_i q_j}{20R_{ij}}$$

où Aij et Bij sont des probabilité de transition entre les niveaux i et j tels que :

$$Aij = \varepsilon ij * Rij^{12} \qquad Bij = 2 * \varepsilon ij * Rij^{6}$$

où : $\varepsilon ij = \sqrt{(\varepsilon i * \varepsilon j)}$ est l'énergie d'interaction de Van Der Waals entre les atomes i et j et $Rij = \mathrm{ri} + \mathrm{rj}$ est la distance entre les atomes i et j (2).
f correspond à une constante égale à 332.0522 et q_i et q_j correspondent aux charges des atomes i et j respectivement. Cette fonction de score a été implémentée dans « *compEner* ».

Le but est de trouver l'interaction ligand-récepteur qui minimisera l'énergie du système, donc la fonction de score. Cette énergie est calculée pour toutes les positions i des 948 conformations de ligand données avec toutes les positions j du récepteur.

Dans un second temps, nous avons tenté d'améliorer cette fonction de score. Nous avons pris le parti de conserver la première étape de scoring, afin de filtrer les 100 meilleures conformations, soit environ 10 % des solutions ligand de départ. Ces 100 conformations ligand ont ensuite été analysées selon l'hydrophobicité de l'interface prédite avec le récepteur. Pour chaque solution ligand potentielle, l'interface avec le récepteur a été déterminée et le nombre de contacts hydrophobe/hydrophile entre un résidu du ligand et un résidu du récepteur appartenant à l'interface a été calculé. Le nombre de contacts hydrophobe/hydrophile a été ramené au nombre de résidus de l'interface, pour en faire une proportion. Cette seconde étape de scoring a été implémentée dans « *computeInterfaceScore* ». L'idée d'utiliser le critère d'hydrophobicité nous a été inspiré par l'algorithme GLIDE (3).

Afin d'identifier la solution ligand qui minimise la fonction d'énergie et maximise le nombre de contacts hydrophobes/hydrophiles, nous avons multiplié le score d'énergie par 1 – (proportion hydrophobes/hydrophiles), afin de chercher la conformation ligand qui minimise la nouvelle fonction de score « *interfaceHydrophobicity*. »

Ces deux étapes de scoring retournent un classement des scores obtenus dans un fichier texte. Pour y parvenir, les scores obtenus sont stockés dans une liste et associée à une liste d'entiers de 1 à 948, correspondant au numéro de fichier : ces listes seront générés par la fonction « *parseDirectory* ». Les listes ont l'avantage d'être ordonnées, ce qui permet d'obtenir facilement le classement souhaité. Une structure de tableau contenant le score et le numéro de fichier correspondant est réalisée grâce à la fonction « *scoreList* ». Puis, un fichier pdb du complexe prédit par la méthode de scoring est généré par la fonction « *writePDB* ». Cette fonction prend en entrée le dictionnaire du ligand solution et le dictionnaire du récepteur pour éditer un fichier nommé « complexe_predit_score1.pdb » le fichier pdb du complexe prédit.

Evaluation de la solution

La solution ligand est évaluée selon deux critères. D'abord, le nombre de résidus commun entre son interface avec le récepteur et l'interface du complexe natif. Puis, le calcul du RMSD (Root Mean Square Deviation), qui prend en compte la déviation structurale entre deux structures protéiques alignées. Le RMSD se révèle être un bon outil d'évaluation lorsque les protéines comparées ont des tailles similaires, ce qui est le cas ici. Le calcul du RMSD est

effectué sur le complexe entier, le ligand, et les résidus de l'interface.

Les interfaces sont calculées à l'aide de la fonction « *computeInterface* », qui détermine les résidus appartenant à l'interface entre le ligand et le récepteur selon deux paramètre : le seuil, qui correspond à la distance maximale entre les résidus du ligand et du récepteur, et le mode de calcul de distance, soit un calcul de distance entre les atomes, soit un calcul de distance entre les centres de masse des résidus. Par défaut, le seuil est fixé à deux, qui est légèrement supérieur au rayon de van der Waals de l'atome de souffre (1.80Å), qui est lui-même supérieur au rayon de van der Waals des atomes de carbone, d'hydrogène, d'oxygène et d'azote, qui sont les principaux constituants des molécules organiques telles que les protéines. Le mode par défaut est le calcul de distance entre les centres de masse des résidus, car il semble plus précis.

Le nombre de résidus communs aux interfaces du complexe prédit et du complexe natif est calculé à l'aide de la fonction « *compareInterface* » qui compare les résidus appartenant à l'interface du complexe prédit à ceux de l'interface du complexe natif.

Le RMSD est calculé à l'aide de la formule suivante :

$$RMSD=\sqrt{(\frac{1}{N}\Sigma(i=1 à i=N)\delta i^2)}$$

où N est le nombre de paires d'atomes utilisés dans le calcul et δi la distance spatiale séparant les atomes i des deux molécules.

Dans notre programme, il est déterminé par la fonction « *computeRMSD* » qui prend en argument les dictionnaires du ligand (ou du complexe) prédit et natif, ainsi que la liste des atomes sur laquelle est calculée le RMSD. Par défaut, les atomes concernés sont le carbone alpha, l'azote, le carbone et l'oxygène de la chaîne principale. Mais l'utilisateur peut spécifier les atomes souhaités en les nommant, ou en utilisant l'argument ALL pour tous les utiliser dans le calcul du RMSD.

Hiérarchie du programme

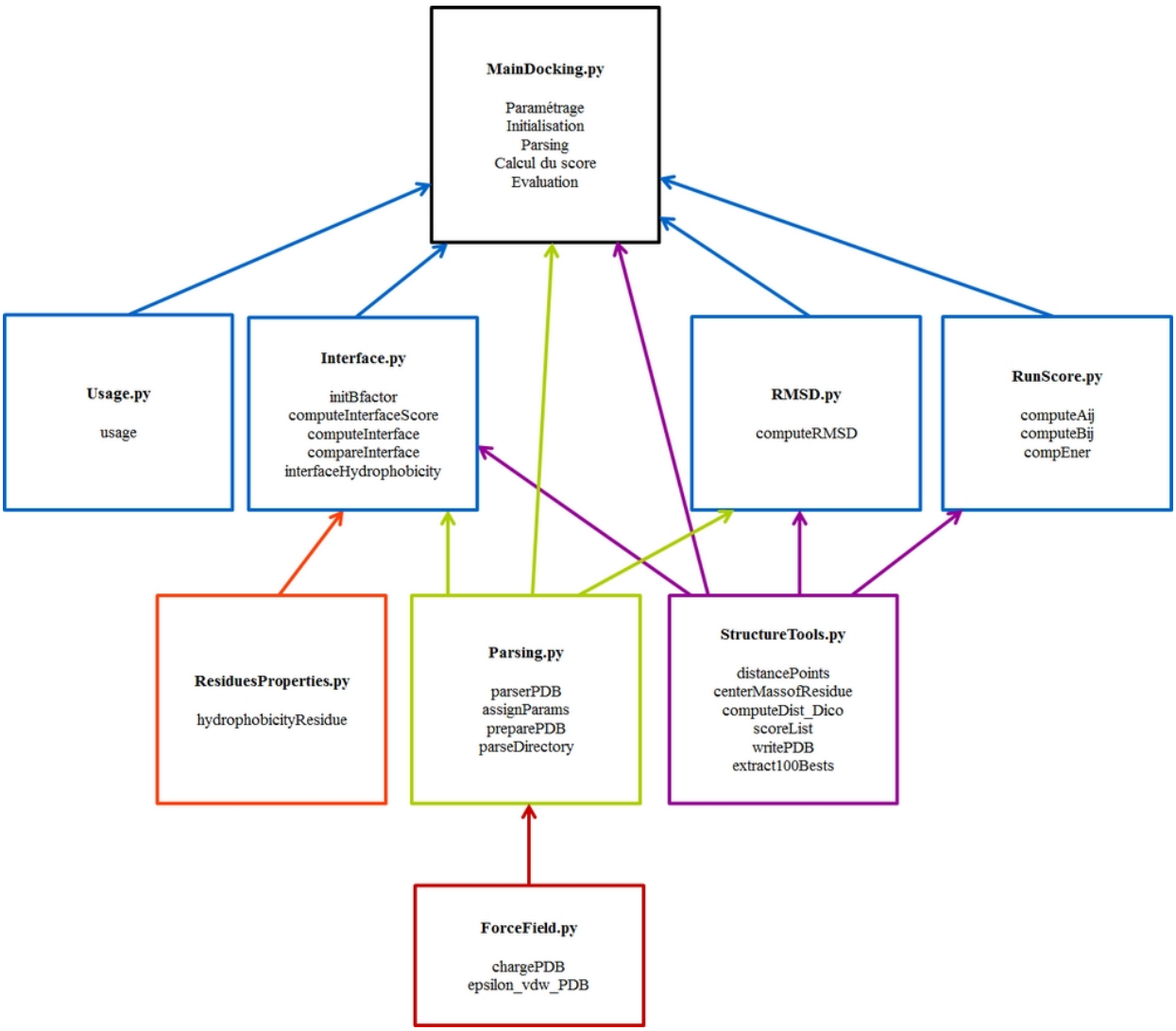

Figure 1 : Représentation de la modularité du programme de docking. Le programme pointé par la flèche utilise le programme a la base de la flèche.

Résultats

L'application notre programme à l'ensemble des 948 conformations de ligands génère deux fichiers textes qui se trouvent dans le répertoire « scoring_Cornell ». Le premier est intitulé « scoring1.txt » et correspond au classement des scores dans l'ordre décroissant. En effet, le choix de la meilleure solution se fait en minimisant la fonction d'énergie, les solutions sont donc ordonnées du score le plus faible au score le plus élevé. Le meilleur score est donné par le fichier ligand n°28 : celui-ci est automatiquement utilisé pour calculer la structure prédite dont la sortie est un fichier PDB nommé « complexe_predit_score1.pdb ». A partir des 100 meilleurs scores de « scoring1.txt » est créé un second classement de scores « scoring2.txt » où sont appliquées au score les modifications dues à la proportion de contact hydrophobes/hydrophiles de l'interface. Le meilleur score est encore une fois donné par le fichier n°28 qui est utilisé pour donner la structure prédite sous forme du fichier PDB « complexe_predit_score2.pdb ».

Interface et seuil

Les classements donnés par « scoring1.txt » et « scoring2.txt » sont identiques : il n'y pas de modifications apportée par notre fonction calculant les contacts hydrophobes/hydrophiles de l'interface. Au vu de la manière dont la fonction calculant ce pourcentage est codée (fonction « *computeInterfaceScore* »), il est possible qu'un seuil trop bas engendre un pourcentage de 0. Nous avons donc exécuter notre programme avec deux valeurs de distance seuil : une de 2 Å et une de 5 Å. Dans les deux cas, les classements restent identiques et la structure prédite inchangée. En revanche, le calcul des résidus présent dans l'interface donne des résultats intéressants :

	Seuil (Å)	Résidus dans l'interface
Complexe natif	5	160
	2	16
Complexe prédit	5	182
	2	30

Table 1 : Résidus trouvés dans l'interface

Ces résultats sont à mettre en relief avec le nombre de contacts natifs identifiés selon le seuil fixé. Avec le seuil de 2Å, le nombre de contacts en commun trouvé par notre programme est de 11 tandis qu'il est de 49 pour le seuil de 5Å. Bien que la taille de l'interface augmente avec le seuil, et par conséquent augmente également le nombre de contacts natifs dans l'interface du complexe prédit, cette augmentation n'est pas proportionnelle. En effet, la proportion de contacts natifs ramenés au nombre de résidus dans l'interface native est plus élevée avec un seuil de 2Å qu'un seuil de 5Å. Il semble donc qu'il y ait un effet de « plateau » à partir duquel l'augmentation du seuil de détermination de l'interface ne permet pas d'améliorer la détermination de contacts communs.

Par la suite, les résultats présentés ont été obtenus avec un seuil de 5Å.

Évaluation de la solution

Interprétation du RMSD

Le calcul du RMSD entre le ligand solution et le ligand connu donne une valeur de 0.96 : cette valeur est satisfaisante car faible, et permet d'être confiants dans la solution ligand identifiée. En revanche, le RMSD calculé sur le complexe et les résidus de l'interface vaut respectivement 49.75 et 44.15 : ces valeurs sont beaucoup trop élevées pour être satisfaisantes. Notre choix de ligand paraissant satisfaisant, l'hypothèse d'un problème au niveau du récepteur peut être émise. L'initiative a donc été prise de calculer le RMSD via la fonction « *computeRMSD* » du fichier récepteur fourni « Rec_natif_DP.pdb » et la chaîne B du complexe natif tel qu'il nous l'a été fourni soit « cplx_natif.pdb ». Le RMSD entre les deux récepteurs supposés identiques

vaut 39.45 : il y a donc une importante différence de structure entre le récepteur natif et le récepteur une fois en contact avec son ligand, expliquant les grandes valeurs de RMSD trouvées pour le complexe et l'interface. L'hypothèse d'un changement de conformation lors de l'interaction aurait pu expliquer ce résultat. Cependant, il semblerait que l'hypothèse de corps rigide faite pour le docking est valable pour l'interaction Barnase-Barstar (4).

Visualisation à l'aide de PyMOL

PyMOL est un logiciel de visualisation de structures protéiques que nous avons utilisé pour évaluer la pertinence de nos résultats. Dans les images qui suivent, un code couleur a été adopté :
- Pour le complexe natif donné, la chaîne B (correspondant au récepteur) est de couleur rouge, la chaîne D (correspondant au ligand) est de couleur verte, et l'interface est en jaune.
- Pour le complexe prédit, la chaîne B est en orange, la chaîne D en bleu et l'interface en blanc.

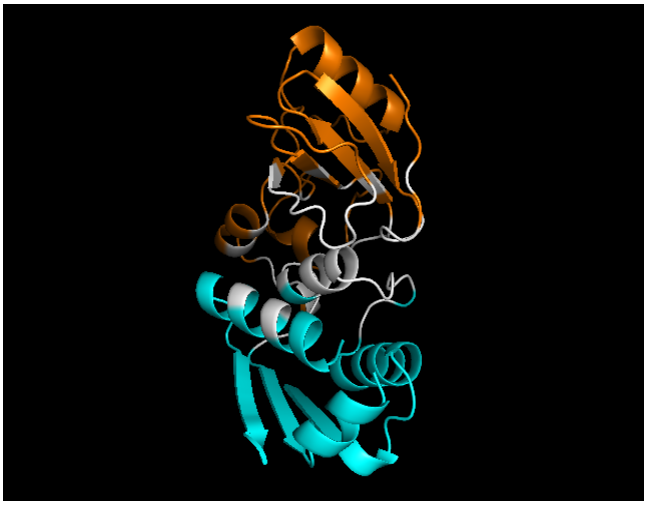

Figure 2 : (gauche) Structure 3D du complexe natif (représentation sous forme «Cartoon»)
Figure 3 : (droite) Structure 3D du complexe prédit (représentation sous forme «Cartoon»)

Cette visualisation révèle une potentielle erreur dans le calcul des résidus communs aux interfaces du complexe prédit et natif. En effet, il y a très peu de différences entre les deux interfaces lorsque le seuil de distance est fixé à 5Å. Or l'algorithme n'a identifié que 49 résidus communs aux deux interfaces, s respectivement 160 et 184 résidus à l'interface pour le complexe natif et le complexe prédit.

Figure 4 : Alignement des deux séquences par PyMOL

L'alignement des deux séquences montre que les deux séquences sont presque parfaitement superposées, confirmant la structure prédite par notre algorithme.

Calcul des contacts hydrophiles/hydrophobes

Notre initiative d'utiliser la proportion de contact hydrophiles/hydrophobes comme moyen de modifier le score n'est pas fructueuse car elle ne modifie pas les scores calculés par la fonction d'énergie de Cornell et. Al (1). Techniquement, cela peut signifier qu'il n'y a pas de contacts hydrophiles/hydrophobes dans l'interface. Or, une vue latérale du complexe prédit avec l'interface en blanc et les acides aminés hydrophobes en magenta montre que l'interface n'est pas dominé par des acides aminés soit hydrophobes soit hydrophiles. Il doit donc exister des contacts hydrophobes/hydrophiles au niveau de l'interface du complexe.

Par conséquent nous soupçonnons un défaut de programmation non détecté qui empêcherait de détecter les contacts hydrophiles/hydrophobes.

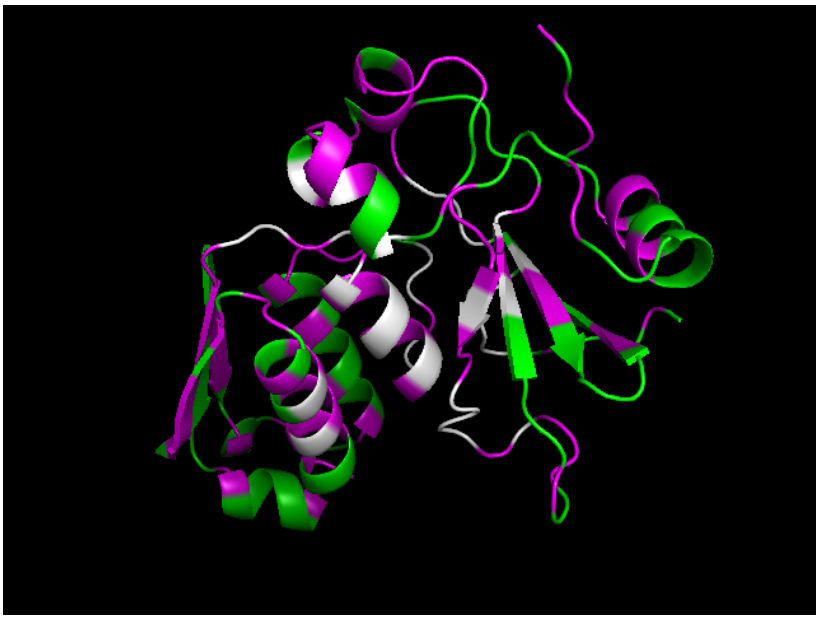

Figure 4 : Vue latérale de la structure prédite

Discussion

L'utilisation de la fonction de score d'énergie de Cornell et al. (1) a permis d'identifier une solution ligand, qui, au regard des valeurs de RMSD et de visualisation sous PyMOL semble relativement satisfaisante. L'implémentation du critère de pourcentage de contacts hydrophobes/hydrophiles à l'interface du ligand et du récepteur n'a pas amélioré le score de RMSD. Cependant ce résultat pourrait être une conséquence d'une erreur de codage. De plus, le RMSD du récepteur natif et du récepteur issu du complexe natif est très élevé, ce qui fausse nos valeurs de RMSD pour le complexe et l'interface. D'autres critères biochimiques pourraient être pris en compte dans la fonction de score, tels que l'encombrement stérique, la conservation des acides aminés à l'interface ou encore la contrainte de la présence de résidus clés à l'interface (5).

Nous avons pris le parti de ne pas tenir compte du rayon de giration des deux protéines, afin de ne pas limiter l'interaction aux extrémités les plus proéminentes du ligand et du récepteur. Cependant, le complexe prédit pourrait en conséquence résulter d'un enfouissement des deux molécules, ce qui n'est apparemment pas le cas ici.

L'efficacité de notre algorithme pourrait être améliorée. En effet, le calcul des scores d'énergie est très coûteux en temps. Un calcul matriciel serait préférable.

Conclusion

Pour le complexe barnase-barstar, la fonction de score de Cornell et al. (1) impliquant les termes non liés semble être suffisante pour déterminer la meilleure solution pour prédire le complexe, étant donné une valeur de RMSD très faible pour le ligand.

Cependant notre seconde étape de scoring a vraisemblablement échoué, remettant en cause cette conclusion. Le critère des contacts hydrophobes/hydrophiles nous semble pourtant pertinent car il renforce l'affinité du ligand pour son récepteur. D'autres critères sont pris en compte par différents algorithmes de docking (6) et pourraient largement enrichir notre programme.

Bibliographie

(1) Cornell et al., A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. Journal of the American Chemistry Society, 1995, vol 117, p 5179-5197.

(2) GROMIHA, Michael. Protein Bioinformatics : From Sequence to Function. Academic Press, 2011, 339 p.

(3) Friesner et al., Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein Ligand Complexes. Journal of Medicinal Chemistry, 2006, vol 49, p 6177-6196.

(4) Chris et al., Is the rigid-body assumption reasonnable ? Insights into the effects of dynamics on the electrostatic analysis of barnase-barstar. Journal of Non-Crystalline Solids, 2011, vol 357, p. 707-716.

(5) Jucovic et Hartley, Protein-protein interaction : A genetic selection for compensating mutations at the barnase-barstar interface. PNAS Biochemistry, 1996, vol 93, p. 2343-2347.

(6) Xu et al., Comparing sixteen scoring functions for predicting biological activities of ligand for protein targets. Journal of Molecular Graphics and Modelling, 2015, vol 57, p. 76-88.