# Identification de la structure 3D de l'interaction de deux protéines par Docking

*BEN YEDDER Hassene & POULAIN Nicolas*

## Introduction

Lorsque l'on a connaissance de l'interaction entre deux protéines, il est intéressant de connaitre les parties des molécules participant à l'interaction. Le développement d'une méthode in silico permettant de prédire la surface d'interaction est un domaine de recherche intéressant. L'objectif du projet est d'évaluer des méthodes de prédiction de structures d'interfaces entre deux protéines, en les appliquant au complexe Barnase-Barnstar. La méthode de scoring des solutions évaluées pour l'interaction est basée sur les termes non liés de la fonction d'énergie proposée par Cornell (Cornell et al. 1995). Les résultats sont évalués par rapport à la conformation connue du complexe à l'aide de l'indice RMSD. L'ajout de la proportion de résidus hydrophobes dans les interfaces protéiques prédites sera ensuite testé dans le but d'améliorer la prédiction.

## Matériel et méthodes

### Matériel

Nous avons à notre disposition un jeu de données composée de fichiers pdb à analyser. Rappelons qu'un fichier pdb englobe toutes les informations sur la composition d'une protéine, sa conformation par la position de chaque résidu ainsi que sa chaine.

Les données se composent d'une conformation unique du récepteur natif pour laquelle, nous allons chercher la meilleure conformation de ligand parmi 948 conformations issues d'une curation de 200000 conformations possibles.

Nous disposons aussi de la conformation du ligand natif afin de pouvoir la comparer à la meilleure solution trouvée et en tirer des conclusions constructives.

### Méthodes

#### Fonction de Score de Cornell et al.

Pour évaluer la qualité des différentes conformations proposées, nous nous sommes basés sur le score proposé dans Cornell et al. JACS 1995. Étant donné qu'il s'agit dans tous les cas des deux mêmes protéines nous utilisons uniquement les termes non liés de la fonction d'énergie décrite dans ainsi que les paramètres associés afin d'évaluer énergétiquement toutes les solutions possibles :

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

avec f = 332.0522

$q_i$ = charge de i – idem pour j

Nature des variables de l'équation :
Aij et Bij sont les probabilités de transition entre i et j suivants deux équations distinctes :

- $Aij = \varepsilon ij * Rij^{8}$
- $Bij = 2 * \varepsilon ij * Rij^{6.}$

*et* $\varepsilon ij = \sqrt{(\varepsilon i * \varepsilon j)}$ est l'énergie d'interaction de Van Der Waals entre les atomes i et j

*Rij* est la distance entre les atomes i et j représenté dans « distancePoints » dans

« ComputeTools »

Ce calcul de score est implémenté dans « ScoreCornell »

Le but ici est de donner une liste de score de toutes les conformations du ligand contre le récepteur natif. Ainsi, nous pouvons retrouver les 100 meilleurs scores et en choisir la meilleure conformation ligand que l'on va comparer au ligand natif dans la suite.

### RMSD

Dans le but d'évaluer la proximité de la solution trouvée avec la conformation réelle, nous avons calculé le RMSD entre le ligand natif et la meilleure position du ligand vis-à-vis du récepteur. Rappelons la Définition du RMSD :
Le RMSD rend compte de la déviation structurale entre deux structures protéiques alignées. Dans notre cas, les structures ont toutes été déjà alignées sur la structure de départ. Plus le RMSD est petit, plus les structures des protéines alignées se ressemblent.

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

Le RMSD est un très bon moyen d'évaluation de l'interface de contact obtenu et facilite le choix et nous dirigera vers PYMOL dans la suite pour bien observé l'alignement entre ligands et le complexe prédit entre le récepteur et son meilleur ligand.

### Hydrophobicité de l'interface

Le calcul de Cornell a permis de déterminer les meilleures conformations du ligand, les 100 meilleures conformations pour cet indice ont été sélectionnées. Les interfaces des interactions entre

protéines sont généralement majoritairement hydrophobes (Tsai 1997), ce qui améliore la force de liaison entre les protéines.

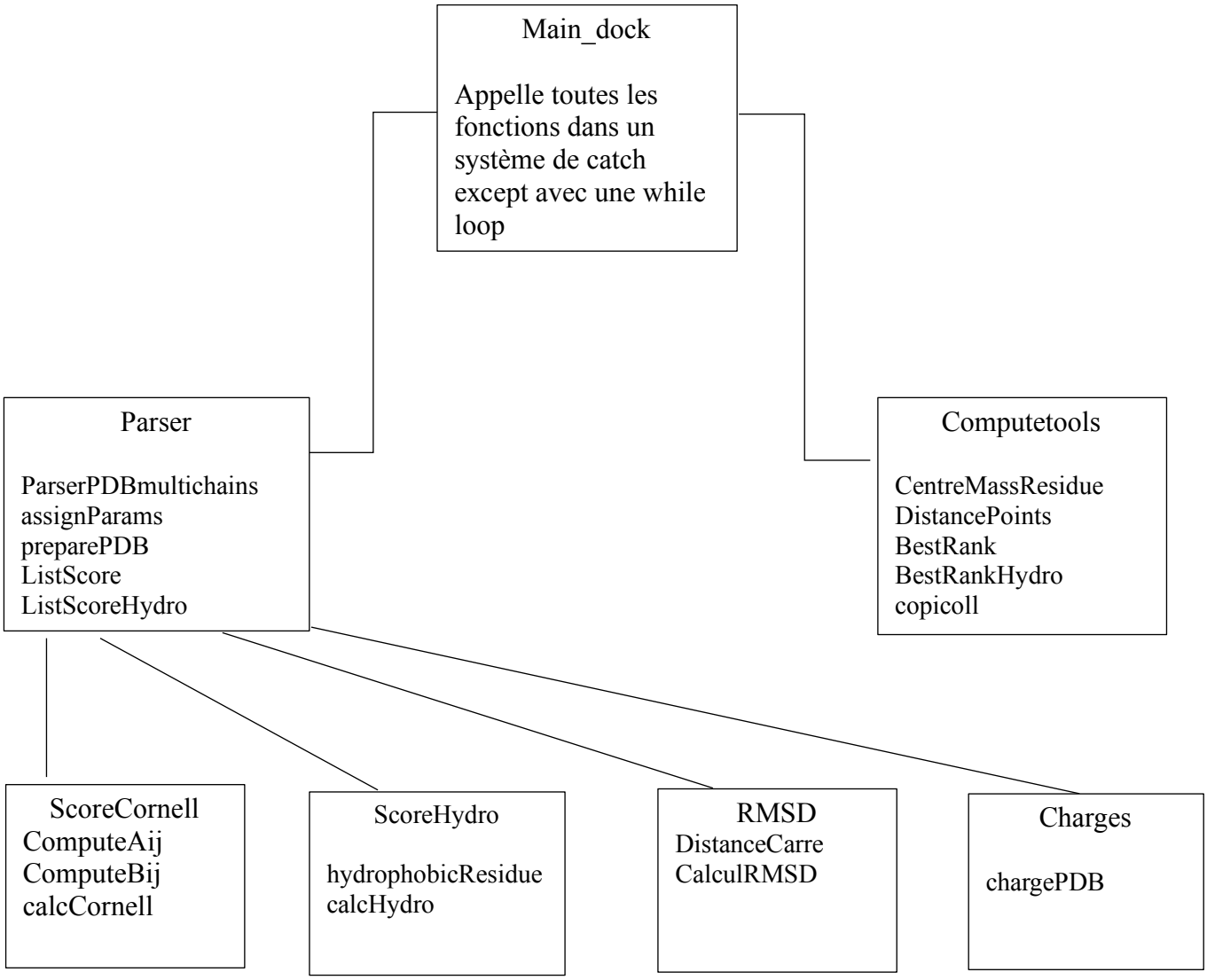
Afin d'affiner la sélection de la conformation optimale, nous avons choisi d'intégrer au calcul de score un indice d'hydrophobicité de l'interface. Les résidus distants de moins de 5 Angstrom de l'autre protéine sont considérés comme faisant partie de l'interface. Après avoir déterminé les résidus participant à l'interface, la proportion de résidus hydrophobes est calculée et utilisé comme facteur du score de Cornell obtenu précédemment :

$$ScoreCornell \times \frac{(Proportion_{Hydrophobe} \times nb_{interface} + Proportion_{Hydrophobe\_inverse} \times nb_{interface\_inverse})}{nb_{interface} + nb_{interface\_inverse}}$$

Cette méthode nous aidera dans l'évaluation de la meilleure solution, de l'hydrophobicité de l'interface mais encore nous allons pouvoir commenter la composition de l'interface Récepteur-Ligand et en tirer des conclusions.

## Composition et explication du code :

Notre code est organisé suivant une hiérarchie assez simple et claire afin qu'une personne qui n'a pas de grandes notions de programmation tel qu'un biologiste puisse la comprendre et puis l'utiliser le plus facilement possible.
Notre programme va utiliser un parseur de fichiers pdb, différentes fonctions de scores expliquées plus hauts ainsi que des fonctions de création de listes et de tableaux et enfin des fonctions de création et écriture dans des fichiers pour obtenir les résultats.
Le digramme suivant va clarifier l'organisation du programme.




En résumé, Parser.py à partir des fichiers pdb, appelle les fonctions de scores « calcCornell », « calcHydro » et « CalculRMSD » et crée des listes de scores et des listes de fichiers correspondants. Grace à Computetools.py, nous pouvons filtrer les scores des meilleurs résultats, afficher les résultats dans des fichiers textes et enfin créer le fichier de la structure 3D de la meilleure solution dans un fichier pdb nommé « complexe_predit_score1.pdb ».

### Visualisation Pymol

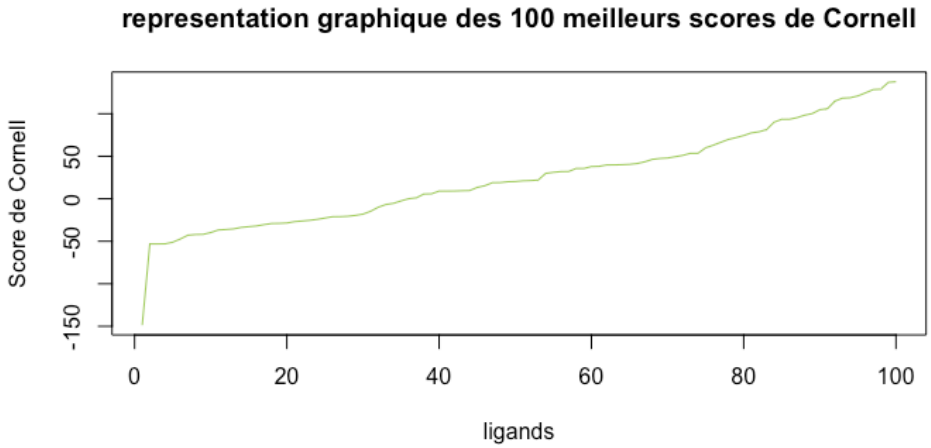
Les meilleures conformations obtenues pour les deux indices seront observées à l'aide de Pymol et comparées à la solution optimale.

# Résultats et Évaluation

## Résultats obtenus

Les 948 fichiers pdb représentant les différentes conformations du ligand ont été analysées et le score de Cornell a été calculé pour chacune des solutions proposées. Un premier fichier de score « score1.txt » est créé, contenant tous les scores dans l'ordre croissant, étant donné que l'on vise à minimiser la fonction d'énergie. La conformation ayant la meilleure valeur pour le score de Cornell est alors stockée dans le fichier « 1BRS_A_1BRS_B_allatom_28_DP.pdb » correspondant à la conformation du ligand n•28.

On observe un décrochage des valeurs obtenues pour les score de Cornell, entre le meilleur score et le suivant, c'est-à-dire un passage de -150 à -50.




Le fichier pdb « complexe_predict_score1.pdb » contient le complexe Récepteur-Ligand le plus probable selon le score de Cornell.
Une fois cette première étape faite, un deuxième fichier « score100.txt » contenant les cent meilleurs résultats est créé. Le calcul de la proportion de résidus hydrophobes a été calculé sur les 100 meilleurs conformations. Un troisième fichier est alors créé : « scorehydro.txt ». La meilleure conformation obtenue grâce à ce calcul est la même que celle reposant uniquement sur le score de Cornell. La meilleure solution est stockée dans un deuxième fichier pdb « complexe_predict_score2.pdb », qui est identique dans le cas de cette analyse.
Le fichier « RMSD.txt », contient les conformations correspondant aux 100 meilleurs scores de Cornell en plus du RMSD obtenu par comparaison avec la solution réelle. La meilleure solution obtenue pour les deux indices calculés présente la valeur de RMSD la plus basse et est donc la solution la plus proche de celle attendue parmi les conformations testées.

### Indices et Seuils

Dans le cas présent, la proportion de résidus hydrophobes ne modifie pas la conformation optimale obtenue à l'aide du score de Cornell uniquement. Cependant l'ordre des solutions suivantes est parfois modifié par l'ajout de cet indice.
Le seuil de 5 Å nous a permis de délimiter la surface de l'interface à peu près à 180 à 200 résidus dans l'interface, dont une cinquantaine d'hydrophobe pour le complexe prédit Recepteur_natif-Ligand28.

### Évaluation du Score RSMD

Le RMSD est une mesure classique de comparaison de structures des protéines. Elle repose sur la distance entre deux atomes correspondant dans deux conformations et évalue la superposition de deux structures protéiques.
Dans notre cas, nous avons mesuré les différents RMSD entre le ligand natif et les cent conformations du ligand ayant obtenu le meilleur score de Cornell. Les valeurs obtenues sont très grandes pour la plupart des conformations, ce qui indique que certaines solutions présentant un bon score d'énergie sont très éloignées de la solution optimale. Néanmoins, pour la meilleure conformation nous avons obtenues une distance de 1140, ce qui indique une importante similitude avec la solution attendue.
En effet, une valeur très petite signifie que les deux conformations sont proches. Cette méthode est utilisée pour évaluer la qualité de notre algorithme de recherche de la meilleure solution.

### Visualisation sur PYMOL

Les résultats obtenus nous ont permis de bien visualiser et de comparer les rapport Recepteur-Ligand_natif, Recepeteur-Ligand28 et Ligand_natif-Ligand. La meilleure conformation a été alignée avec la conformation native de ce ligand.

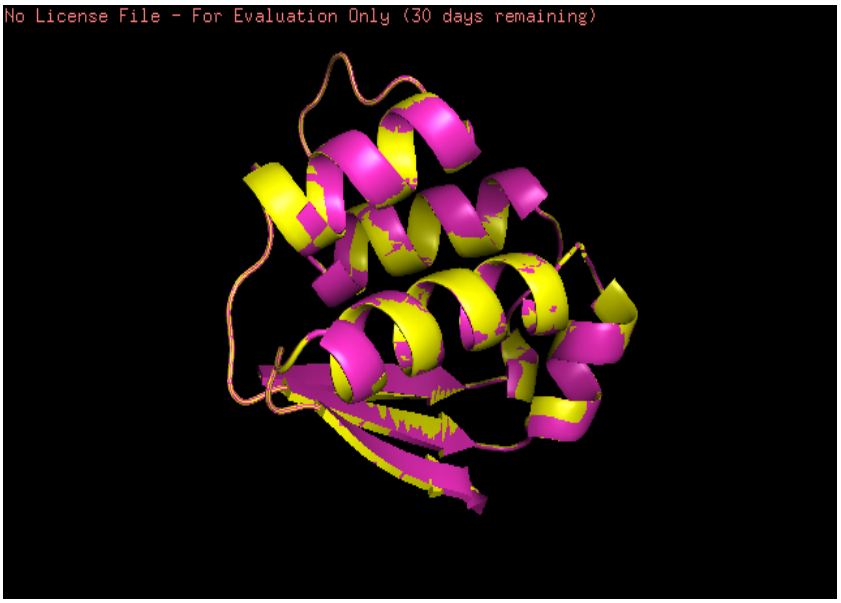



*Figure.1 : Alignement des deux Ligands 28(rose) et natif(jaune)..*

Cet Alignement montre un bon choix de conformation, si nous regardons bien le score d'alignement, nous sommes à 0.001 de probabilité de fautes. Donc les deux conformations sont très proches. Ceci peut être consolidé grâce au score de RMSD trouvé de 1140, le plus faible score obtenu pour l'évaluation ligand-ligand.

La solution obtenue et la conformation réelle du complexe ont ensuite été alignées cote à cote afin de visualiser leur similitude.

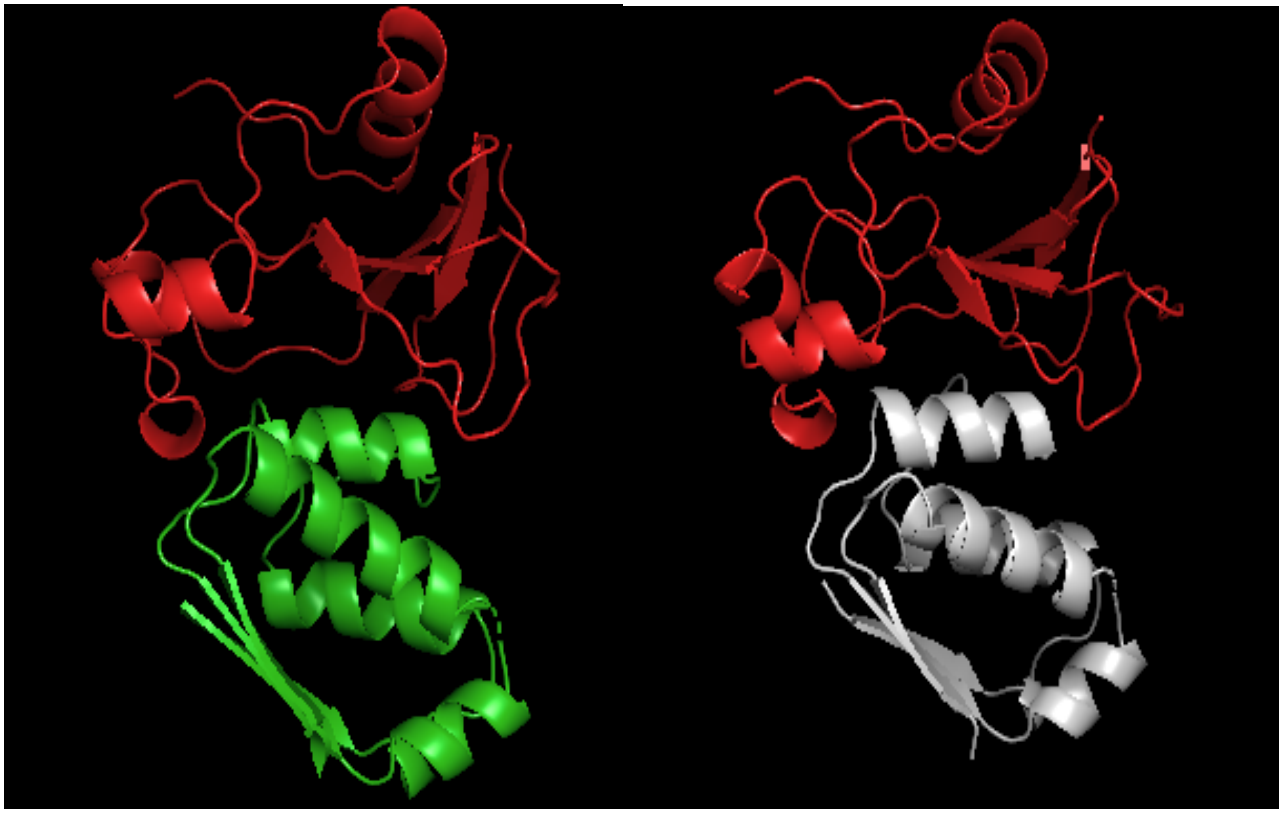


*Figure.2 : Représentation Cartoon du Recepteur-ligand_natif à gauche et Recepteur-ligand28 à droite. Le Récepteur est en rouge, le ligand natif en vert et le ligand28 en Blanc.*

Les deux complexes se ressemblent étroitement mais les conformations des deux ligands ne sont pas identiques et cela est visible grâce aux hélices α qui n'ont pas la même position entre le ligand natif et le ligand solution.
Pour pouvoir choisir entre les deux, les distances des résidus de l'interface ont été calculés à l'aide de PYMOL, La fonction est Compute => surface area => per residue. Les distances obtenues sont entre 0.5 et 0.9, pour l'interface du complexe natif.

Cependant, pour le complexe Rec_natif-ligand28, nous avons obtenu un maximum de 0.7 comme le montre la visualisation suivante :

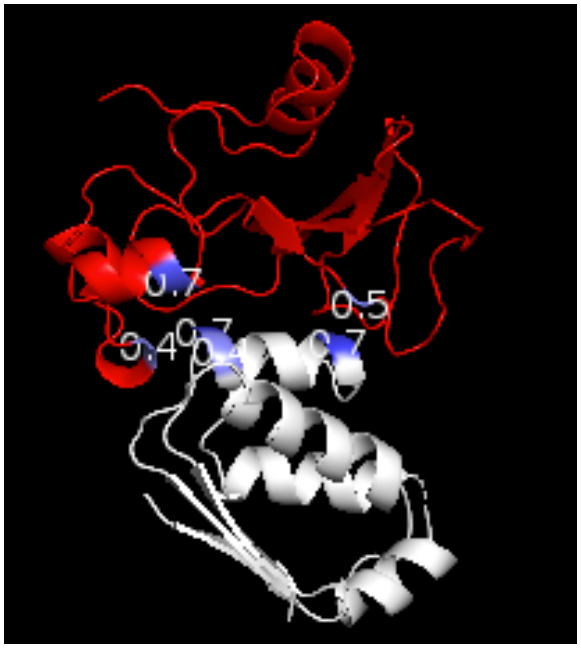



*Figure.3 : Mesure des distances de l'interface*
*Avec PYMOL*

Le complexe prédit en surface a été visualisé, en particulier le site d'interaction Ligand-Récepteur pour vérifier qu'il s'agisse d'une interface valable pour la conformation ligand 28 obtenue par Scoring.

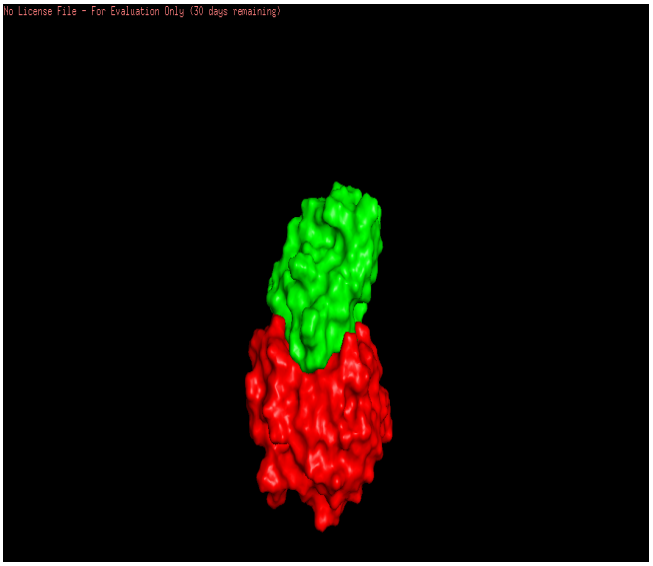



*Figure.4 : Complexe prédit en 3D (style surface)*
*Ligand en vert, Récepteur en Rouge*

## Discussion

Le calcul du score de Cornell a permis d'identifier la meilleure solution parmi les conformations proposées. Ce résultat est confirmé par l'observation des valeurs de RMSD des différentes conformations comparées. L'ajout d'un indice d'hydrophobicité de l'interface n'a pas modifié le résultat proposé, ce qui est attendu étant donné que le meilleur résultat était identifié lors de la première analyse. La différence importante entre la première et la deuxième solution en termes de score de Cornell peut expliquer le fait que la solution ne change pas. L'ordre des conformations a cependant été modifié parmi les moins bonnes valeurs de Cornell. On peut donc s'attendre à ce que cet indice trouve son utilité dans des situations où les scores observés soient moins distincts. Il serait donc intéressant de tester cette méthode pour identifier la structure de l'interaction entre deux protéines sur des complexes différents.

Si le score d'hydrophobicité modifiait les résultats, il serait intéressant de tester l'impact du seuil servant à déterminer l'interface. Ce qui pourrait notamment être fait en testant de meilleures solutions pour identifier l'interface au lieu de simplement tenir compte de la distance avec la protéine partenaire. Les résidus hydrophobes composent rarement la totalité des interfaces, au contraire, ils sont répartis sous forme de patches en frontière d'interface, l'algorithme proposé bénéficierait donc de l'amélioration du calcul du score d'hydrophobicité (Larsen 1998).

On s'attend à ce que la différence de score entre les solutions soit moins nette dans d'autres études. On propose d'ajouter d'autres indice permettant d'évaluer la qualité de l'interface prédite tels que la taille de l'interface ou la composition en certains acides aminés caractéristiques des interfaces entre protéines (Moreira 2007).

La visualisation de la solution optimale sur Pymol a permis de vérifier la cohérence de la solution prédite et sa proximité avec la solution attendue.

# Bibliographie

Cornell, Wendy D., et al. "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules." *Journal of the American Chemical Society* 117.19 (1995): 5179-5197.

Larsen, Teresa A., Arthur J. Olson, and David S. Goodsell. "Morphology of protein–protein interfaces." Structure 6.4 (1998): 421-427.

Moreira, Irina S., Pedro A. Fernandes, and Maria J. Ramos. "Hot spots—A review of the protein–protein interface determinant amino-acid residues." Proteins: Structure, Function, and Bioinformatics 68.4 (2007): 803-812.

Tsai, Chung-Jung, et al. "Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect." Protein Science 6.1 (1997): 53-64.