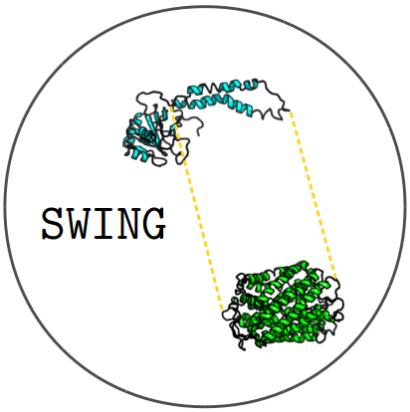

SWING : Sampling With INteroloG

DE AZEVEDO, Kévin
DUHAMEL, Marine
YAO, Hua-Ting

January 2, 2018

Version : 1.0

Contents

Introduction 1

Partie 1 Sampling

I Matériel et méthodes 2

1 *InterEvolAlign* 2

2 *Pymol* 3

3 Clustering 3

4 Rotations du ligand 3

5 Algorithme 4

II Résultats 4

1 Capacité de *SWING* à générer des conformations natives 4

2 Dépendance de *SWING* au regard de la qualité des interologues 6

III Discussion 6

Partie 2 Scoring

I Résultats 8

II Discussion 9

Conclusion et perspectives 10

Bibliography 11

Annexe

A Analyse sampling partie 1 : Capacité de *SWING* à générer des conformations natives 13

Introduction

Les interactions entre protéines ont un rôle de régulation majeur dans les fonctions cellulaires. Leur étude est essentielle afin de comprendre les processus biologiques, les mécanismes d'interaction, et dans le développement des médicaments qui ont vocation à inhiber ces interactions. La détermination expérimentale de la structure de ces interactions est parfois compliquée, longue et surtout très coûteuse. Par conséquent, la modélisation computationnelle des interactions protéine-protéine est une alternative fréquemment employée dans la recherche pharmacologique comme dans la recherche fondamentale, où elle représente un certain défi.

Ce défi réside dans deux difficultés principales. D'abord, la nécessité de produire des complexes putatifs d'interaction entre les deux protéines. En effet, la méthode la plus pratiquée pour modéliser les interactions protéine-protéine consiste à fixer l'une des protéines, appelée récepteur, et de générer des positions de l'autre protéine, appelée ligand, autour de ce récepteur. C'est ce que l'on appelle l'échantillonnage (ou sampling en anglais). La seconde difficulté réside dans l'évaluation (l'étape dit de scoring en anglais) des conformations ligand générée précédemment. Cette évaluation doit rendre compte de la qualité de la prédiction. Elle se base sur des critères physiques, géométriques, mais peut également considérer l'aspect évolutif de l'interaction, notamment par sa conservation dans le vivant. Une bonne méthode d'évaluation doit être capable d'identifier les prédictions les plus vraissemblables. Ces deux grandes étapes que sont l'échantillonnage et l'évaluation consistuent les deux étapes essentielles à un programme de modélisation d'interactions protéiques.

Le projet Meet-U vise à rassembler des étudiants d'universités d'Ile-de-France sur un projet de bioinformatique commun. Cette année, la thématique abordée est la modélisation d'interaction entre protéine. Les étudiants se sont réunis par équipe pour produire soit un programme d'échantillonage, soit un programme d'évaluation, puis ont intégré le programme complémentaire d'une autre équipe pour parvenir à une méthode de modélisation complète. Notre équipe a réalisé un programme d'échantillonnage.

Il existe plusieurs techniques pour déterminer des conformations ligand[Huang 2014]. La recherche exhaustive, qui utilise une grille et génère toutes les positions et rotations possibles du ligand autour du récepteur. Cette méthode est à priori idéale mais néanmois très coûteuse en temps de calcul. Une autre méthode, appelée corrélation de FFT (Fast Fourrier Transform)[Katchalski-Katzir et al. 1992], fonctionne globalement selon le même principe mais avec une puissance de calcul supérieure. Enfin, la recherche directe dans l'espace cartésien, moins efficace que la FFT, permet cependant l'introduction d'une certaine flexibilité dans la recherche de conformations ligand. Elle fournit notamment la possiblité d'apporter de l'information biologique lors du processus d'échantillonnage. C'est cette dernière méthode que nous avons employée dans notre projet, en exploitant les informations évolutives préexistantes sur les membres de l'interaction à modéliser.

Les protéines partageant un ancêtre commun et une fonction similaire sont appelées orthologues. Lorsque des paires de protéines orthologues interagissent fonctionnellement chez différentes espèces, celles-ci sont appelées protéines interologues[Walhout et al. 2000]. La base de données interEvol[Faure et al. 2012] recense les interologues structuraux à partir de la PDB (Protein Data Bank). L'outil interEvolAlign associé[Faure et al. 2012] permet d'interroger cette base de données pour identifier les protéines interologues à deux partenaires protéiques. Ainsi, pour deux protéines que l'on sait interagir, il est possible de rechercher des interologues dont la structure est connue. SWING utilise interAlign pour identifier des interologues. Les protéines ligand et récepteur sont ensuite alignées sur l'interface interologue, générant ainsi

une position initiale pour l'échantillonnage. Des rotations du ligand sur lui-même et par rapport au récepteur permettent d'obtenir des conformations ligand.

Les conformations générées par SWING couvrent un espace conformationnel réduit par rapport à un échantillonnage naïf, offrant la possibilité de produire plus de conformations dans un espace d'interaction plus probable, compte tenu des informations évolutives.

Dans une première partie la capacité de SWING a générer des conformations natives (avec un RMSD ligand ou lRMSD $\leq$ à 5$\mathring{A}$) sera analysée, selon divers paramètres. D'abord, le nombre de conformations ligand à générer pour obtenir des conformations natives. Puis le degré de rotation, tiré au hasard dans un intervalle défini en paramètre. Nous examinons également la dépendance de ces performances à la qualité des interologues, en terme d'identité par rapport au complexe à prédire.

Dans une seconde partie, deux complexes seront évalués par deux méthodes, celles de l'équipe 1 (MeetDockOne) et de l'équipe 11 (DeNovo).

Partie I Sampling

I Matériel et méthodes

1 *InterEvolAlign*

L'information structurale des interologues a été extraite du site web *InterEvol* [Faure G. et al, 2012]. Ce site consiste en une base de données de complexes présentant des interologues connus, constituée de cette manière : à partir des complexes de la PDB, des groupes de complexes "redondants", partageant plus de 70% d'identité de séquences sur 70% de leur longueur, ont été définis par alignements profils-profils et superposition structurelle. Dans chacun de ces groupes, le complexe de meilleur résolution était défini comme "représentatif" du groupe. Pour pouvoir trouver des interologues éloignés en terme d'identité de séquence, ces groupes de complexes redondants ont été clusterisés au niveau de la super-famille via HHProfile et Matras, permettant de trouver des intérologues avec moins de 70% d'identité de séquences entre eux. La dernière mise à jour de la base de données recense 13 126 complexes de ce type : c'est donc une base de donnée extensive permettant de trouver des intéractions entre protéines qui ne seraient pas visibles par alignement de séquences classiques.

SWING utilise l'outil d'alignement du site *InterEvolAlign*. Les deux fichiers PDB donnés en entrée du programme sont traduits paren séquences FASTA qui font l'objet d'une requête d'alignement envoyée sur le serveur du site. Par défaut, le programme ne changera pas les paramètres par défaut de l'alignement. Par conséquent, une requête classique se déroulera de cette façon : les deux séquences seront blastées deux fois (via Psi-BLAST) contre une banque de données de génomes entiers et après chaque itération, les séquences sont réalignées via MUSCLE et la séquence de plus basse e-value est gardée pour chaque espèce, ce qui a pour conséquence que l'alignement final ne contient qu'une seule séquence pour chaque espèce. Pour finir, les alignements multiples sont utilisés pour déterminer leurs interologues dans la

base de données *InterEvol* via HHSearch.

De nombreux paramètres (e-value minimale, nombres d'interologues maximums etc.) peuvent être modifiés pour ajuster la requête. Dans cette optique, sont fournis avec le programme deux fichiers, blast.conf et parameters.conf, dont le contenu est un dictionnaire python classique dont les clés sont les paramètres et les valeurs correspondent à la valeur des paramètres. Changer la valeur contenue dans ces fichiers aura pour conséquence de changer celles de la requête transmise à *InterEvolALign*

2 *Pymol*

Les positions initiales ont été déterminées par un script *Pymol*. Le récepteur interologue est d'abord aligné sur le récepteur d'intérêt par la fonction *cealign*. Puis, le ligand d'intérêt est aligné sur le ligand interologue avec cette même fonction. Le ligand aligné est enregistré comme position initiale potentielle.

3 Clustering

La redondance des positions initiales a été supprimée par clustering hiérarchique ascendant sur la base de la distance entre les centres géométriques des positions initiales. Les positions initiales ont été progressivement regroupée en fonction d'une distance minimale. A chaque itération, la distance minimale entre deux membres de deux groupes distincts est évaluée. Si cette distance est inférieure ou égale à $5\mathring{A}$ (valeur fixée empiriquement), les deux groupes fusionnent.

4 Rotations du ligand

La recherche randomisée d'angles de rotation du ligand sur lui-même est par rapport au récepteur à partir des positions initiales ainsi générées permet des orientation du ligand plus probables. La rotation du ligand consiste la rotation extrinsèque et la rotation intrinsèque, paramétré par 2 et 3 angles d'Euler, respectivement.

Posons R, le centre géométrique du récepteur et L, celui du ligand. En prenant un repère cartésien sur R, la rotation du ligand autour du récepteur est paramétrée par ϕ et θ, où ϕ est l'angle de rotation autour du vecteur $\overrightarrow{O_z}$ et θ est celui autour du vecteur $\overrightarrow{O_z} \wedge \overrightarrow{RL}$. En mettrant le centre du repère sur L, la rotation intrinsèque du ligand est décrite par α, β et γ autour des vecteurs $\overrightarrow{O_z} \wedge (\overrightarrow{O_Z} \wedge \overrightarrow{RL})$, $\overrightarrow{O_z} \wedge \overrightarrow{RL}$ et $\overrightarrow{RL}$, respectivement.

L'espace de recherche est réduit autour des positions initiales, fixées par alignement des structures interologues. Selon l'hypothèse de conservation de l'interface, la véritable structure du complexe doit être relativement proche de cette position initiale.

Pour générer des conformations proches de la position initiale, la valeur d'angle de rotation doit être proche de 0. L'échantillonnage des cinq angles d'Euler suit donc une loi gausienne de moyenne 0 et d'écarte type $\frac{1}{3} * angle_{lim}$ ($angle_{lim} = 7,5°$, par défault).

Dans l'implémentation du programme, le quaternion est préféré aux angles d'Euler, réduisant ainsi le temps de calcul. Un quaternion q s'écrit sous la forme $q = a + bi + cj + dk$ où a, b, c, et d sont les nombres réels, i, j, et k des nombres imaginaires. La partie imaginaire $bi + cj + dk$ se comporte comme un vecteur $\overrightarrow{v}$, et la partie réelle a se comporte comme un scalaire. Cette décomposition est recommandée pour l'utilisation des quaternions en géométrie.

Soit $\overrightarrow{v}$ un vecteur de l'espace de dimension 3, $\overrightarrow{u}$ le vecteur unitaire, α un nombre

réel et $q_\alpha = \cos(\frac{\alpha}{2}) + \overrightarrow{u}\sin(\frac{\alpha}{2})$. Alors, $q_\alpha \overrightarrow{v} q_\alpha^{-1}$ est le vecteur résultant de la rotation du vecteur $\overrightarrow{v}$ de l'angle α selon le vecteur unitaire $\overrightarrow{u}$. Le calcul des rotations se résume ainsi au calcul $q_{int}\overrightarrow{LA_i}q_{int}^{-1}$ où $q_{int} = q_\alpha q_\beta q_\gamma$ et A est l'ensemble des atomes du ligand pour la rotation intrinsèque et $q_{ext}\overrightarrow{RL}q_{ext}^{-1}$ où $q_{ext} = q_\theta q_\phi$ pour la rotation extrinsèque du ligand.

5 Algorithme

Algorithme 1 : SWING

Données : Fichiers .pdb ligand et récepteur

```
Téléchargement des interologues depuis InterEvol;
Génération des positions initiales par PyMol;
Clustering des positions initiales;
pour chaque Position initiale faire
    Rotations intrinsèque du ligand (α, β, γ);
    Rotations extrinsèques du ligand (θ, φ);
    si Distance minimale entre deux atomes des partenaires < 5Å alors
        Reculer le ligand de 25Å;
    Minimisation par MaxDo
fin
```

II Résultats

Deux analyses principales ont été réalisées sur le sampling SWING. Premièrement, une évaluation de la capacité de SWING a générer des conformations natives (avec un RMSD ligand ou lRMSD $\leq$ à $5\mathring{A}$). L'obtention de conformations natives a été examinée selon trois conditions. D'abord, le nombre de conformations ligand minimum pour obtenir des conformations natives. Ensuite, le degré de rotation du ligand. Le degré de rotation est tiré aléatoirement dans un intervalle fixé par l'utilisateur, et définit par défaut à $[-7,5°,7,5°]$ (valeur fixée empiriquement). Enfin, la modalité de tirage, uniforme ou gaussienne. Le tirage uniforme consiste en une probabilité de tirage équivalente sur tout l'intervalle de rotation. Le tirage gaussien favorise les valeurs d'angles proches de la moyenne, ici 0ř, avec une probabilité de tirage décroissante pour les valeurs extrèmes. Dans ce mode, 95% des rotations sont choisies dans l'intervalle définit.

Dans un second temps, les performances de l'échantillonnage seront évaluées en fonction de la qualité des interologues. L'intuition est que plus les interologues seront proches, en terme d'identité, du complexe à prédire, plus il sera aisé d'obtenir une bonne prédiction du complexe.

1 Capacité de *SWING* à générer des conformations natives

Des conformations ligand ont été générées pour le complexe *1AY7* afin d'évaluer les paramètres optimaux pour l'obtention de conformations natives. Ce complexe ne possède qu'un seul interologue, le complexe *2ZA4* . Etant donné le coût en terme de temps de calcul de la minimisation par *MaxDo*, l'influence du choix des paramètres a d'abord été observée avant minimisation (Figure 1.1).

Le tirage des angles selon une loi gaussienne centrée sur 0ř permet d'enrichir l'échantillon en conformations natives, et cela quelque soit la valeur de n, le nombre de conformations

ligand à échantillonner. C'est donc ce paramètre qui sera utilisé dans la suite de l'analyse.

Concernant le nombre de conformations, $n = 5000$ semble être un seuil au-delà duquel le nombre de conformations natives n'augmente plus avec n. Cependant, le temps nécessaire pour minimiser 5000 conformations est conséquent. En outre, le nombre de conformations n est a généré pour chaque position initiale. *1AY7* ne possède qu'une position initiale, mais certains complexes en comptent jusqu'à neuf, rendant le nombre de conformation ligand a générer trop important pour les ressources dont nous disposions. La valeur $n = 1000$ semble plus raisonnable, pour un résultat correct.

Après minimisation (Figure 1.2), la distribution des conformations ligand en fonction du RMSD pour $n = 500$ et $n = 1000$ est équivalente. Néanmoins, le complexe *1AY7* est relativement aisé à prédire. C'est donc le paramètre $n = 1000$ qui a été retenu pour la suite des analyses. Par ailleurs, l'enrichissement en conformations natives pour $n = 10000$ est moindre après minimisation, ce qui nous conforte dans le choix d'un seuil plus raisonnable en coût.

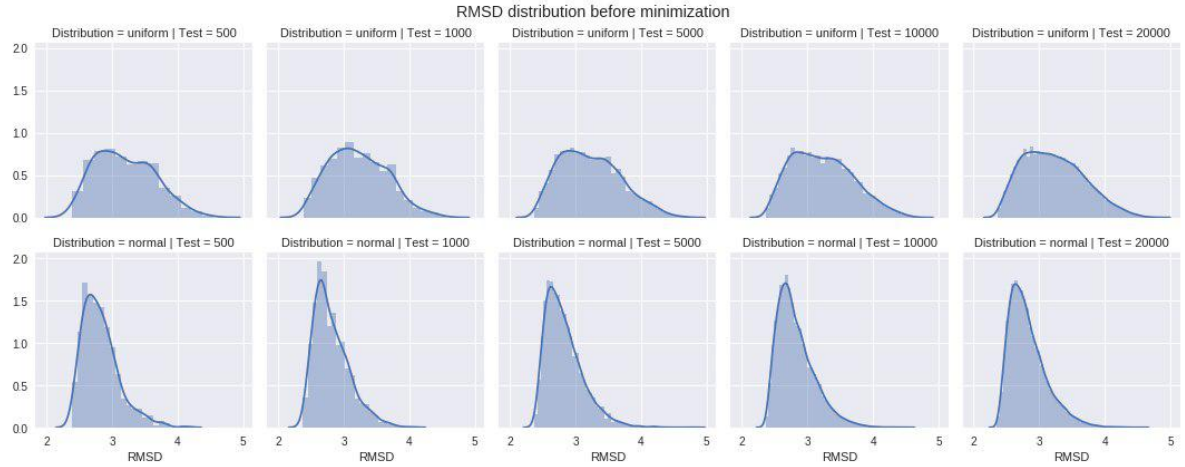

FIGURE 1.1: Densité de distribution en fonction du RMSD des conformations ligand du complexe *1AY7* avant minimisation. Le meilleur compromis entre le temps de calcul et la performance est la génération de n=1000 conformations ligand, avec une modalité de tirage gaussienne.

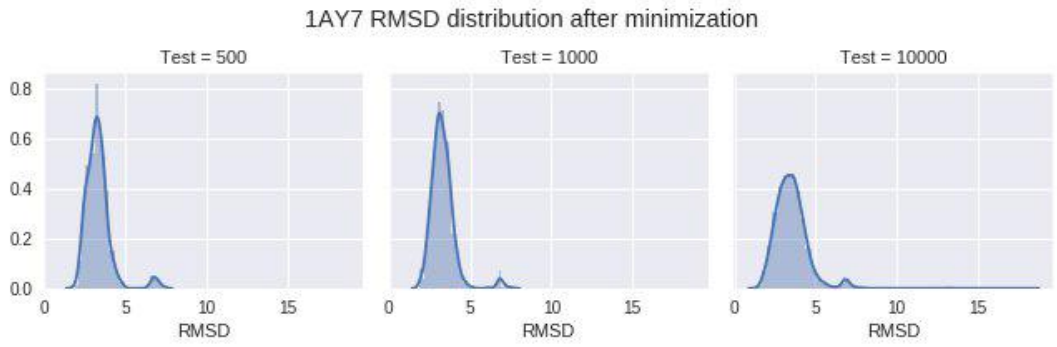

FIGURE 1.2: Densité de distribution en fonction du RMSD des conformations ligand du complexe *1AY7* après minimisation. Après minimisation, la génération de n=1000 conformations ligand et la modalité de tirage gaussienne sont toujours les paramètres optimaux.

Le choix de l'angle a été évalué sur différents complexes (*1QDL 1PPJ_CP 1CGI 1DQJ*

1inl 3FN1 et 1AY7)(annexe 2.5 et table 1.1). Selon les complexes, différents comportements ont été observés, avant et après minimisation. Dans l'ensemble, nous avons recueilli peu de conformations natives. Le sampling avec les paramètres par défaut définis précédemment n'a pas permis d'obtenir de conformation native dans le cas de 1inl, quelques soient les conditions. Des conformations natives ont été obtenues pour *1DQJ* et *3FN1* seulement après minimisation. Pour les complexes *1PPJ_CP* et *1AY7*, le nombre de conformations natives diminue après minimisation. Ces deux derniers complexes sont également ceux qui ont le meilleur rendement en conformations natives.

Globalement, l'intervalle [−15°,15°] semble le meilleur pour obtenir des conformations natives. Cependant, presque aucun conformation native n'a été échantillonnée pour la majorité des complexes. Lors de la livraison du programme aux équipes de scoring, c'est la valeur 30ř qui était fixée par défaut. Cette valeur a été conservée dans l'analyse du scoring.

Angle	1AY7		1QDL		1PPJ_CP		1CGI		1DQJ		1inl		3FN1	
	0	1	0	1	0	1	0	1	0	1	0	1	0	1
7.5	351	305	0	1	2	0	0	0	0	0	0	0	0	1
15	292	239	1	0	59	22	0	2	0	1	0	0	0	2
30	192	126	0	0	57	20	1	0	0	0	0	0	0	0

TABLE 1.1: Nombre de conformations natives de chaque complexes avant (0) et après (1) la minimisation

2 Dépendance de *SWING* au regard de la qualité des interologues

Considérant le pourcentage d'identité des interologues, un fort pourcentage d'identité semble augmenter le nombre de conformation natives ou e tous cas diminuer le RMSD des conforamtions ligand (Figure 1.3). La représentation du RMSD ligand en fonction de l'identité de l'interologue (identité minimale entre l'interologue ligand et l'interologue récepteur) ne montre aucune tendance globale, si ce n'est les faibles RMSD pour les fortes identités ($\geq 38\%$ globalement).

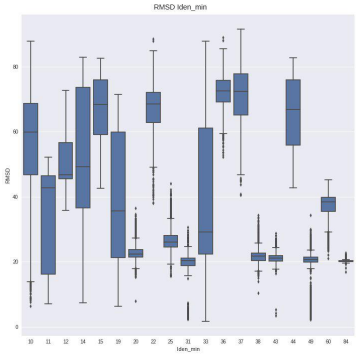

FIGURE 1.3: RMSD ligand en fonction de l'identité minimum entre le ligand et le récepteur interologue.

III Discussion

Les paramètres par défaut validés après cette analyse sont le tirage gaussien des angles de rotation, et le nombre de conformations $n = 1000$. L'intervalle des angles de rotation reste incertain, compte tenu des mauvais résultats pour les complexes évalués (Table). Une

raison probable quant à ces mauvais résultats est l'opération de recul du ligand lorsque la distance entre les partenaires est trop faible. Ce recul avait été ajouté pour éviter que les conformations ligand trop proches ne soient rejetées par le minimineur *MaxDo*. En effet, ces conformations auraient pu être correctes à une translation près. Ce recul pourrait diminuer le nombre de conformatives, si les conformations ayant subit la translation sont trop éloignées du récepteur pour être correctement minimisées. Supportant cette hypothèse, l'analyse sur un échantillonnage restreint à $n = 500$, sur le complexe *1AY7*, montre bien que Le recul du ligand diminue le nombre de conformations natives (Figure 1.4). Par extention, nous suposons que le retrait de cette contrainte augmenterait le nombre de conformations natives pour les autres complexes. Une augmentation du nombre de conformations générées pourrait également être un plus.

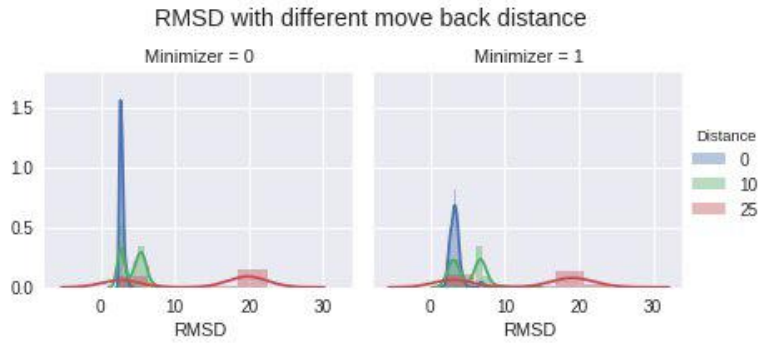

FIGURE 1.4: Densité de distribution en fonction du RMSD des conformations ligand du complexe *1AY7* avant et après minimisation, selon le recul imposée lorsque la distance entre les partenaires est inférieure à $5\mathring{A}$ avant minimisation. $n = 500$, tirage gaussien dans l'intervalle $[-7.5°,7.5°]$. Le recul du ligand diminue la quantité de conformations natives.

Des conformations proches de la solution native sont plus aisément obtenues lorsqu'au moins le ligand ou le récepteur interologue est évolutivement proche du ligand ou du récepteur du complexe, c'est à dire avec une forte identité. Cependant, ce critère ne semble par complétement limitant pour obtenir des conformations proches de la solution native (Figure 1.3).

Partie II Scoring

Les conformations fournies par *SWING* ont été évalués par deux méthodes de scoring : *MeetDockOne* (Equipe 1) et *De Novo* (Equipe 11). Les deux méthodes sont intéressantes du fait que l'approche machine learning permet la combinaison de plusieurs méthodes de scoring allant de la minimisation énergétique à la complémentarité de forme en passant par la conservation des contacts entre acides aminés. La méthode *SWING* était purement un sampling naïf couplé d'une information positionnelle basée sur l'évolution (via la recherche d'interologues), par conséquent, toutes ces approches de scoring sont en quelque sorte complémentaires de l'approche de sampling et non redondantes. De plus, l'utilisation d'Anaconda pour *MeetDockOne* est très intéressante dans un soucis de reproductibilité des résultats.

Le complexe sur lequel le docking complet (le sampling et chacune des méthodes de scoring) a été réalisé est 1AY7 (Barnase-Barstar), du fait des tests extensifs faits sur ce complexe lors de l'élaboration du sampling, sa petite taille ainsi que le fait qu'il soit un complexe classique utilisé pour tester des programmes de docking. Avec le scoring DeNovo, le complexe 1DQJ (anticorps anti-lysozymes) a également été utilisé, pour des raisons développées ci-après. Sont considérées uniquement dans la suite les conformations du "Top 10", c'est à dire les 10 conformations ayant obtenu les meilleurs scores.

I Résultats

Les conformations minimisées obtenues ont été évaluée via le programme *SWING* pour *1AY7*, un angle de rotation de 30ř et un nombre de conformations demandées de $n = 1000$. Pour *MeetDockOne*, a été testé la pertinence d'utiliser naccess ou msms, sachant que le dernier ne fonctionne que pour des petits complexes comme *1AY7*.

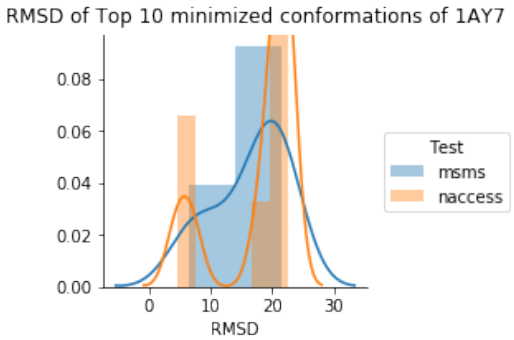

FIGURE 2.1: Densité de distribution en fonction du RMSD du Top 10 pour 1AY7 en fonction de l'utilisation de msms ou naccess

DeNovo fournit quant à lui deux scores : un basé sur l'énergie (en), et un autre sur la conservation des contacts entre acides aminés (stat). Par conséquent, il y a deux Top 10 différents dont les RMSD ont été calculés.

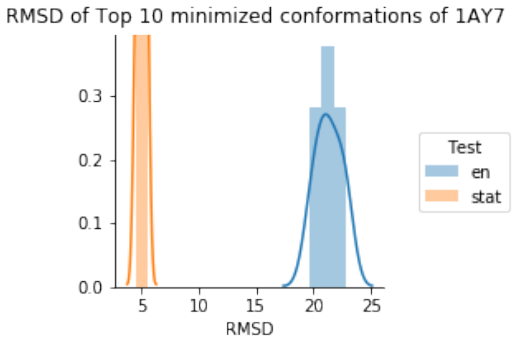

FIGURE 2.2: Densité de distribution en fonction du RMSD du Top 10 pour 1AY7 en fonction du score

Ce résultat a motivé une analyse subsidiaire : les conformations non minimisées issues de *SWING* (qui se trouvent dans le dossier "Proteins" dans le dossier de sortie) ont été évalués via les deux méthodes (en utilisant cependant uniquement naccess pour *MeetDockOne*)

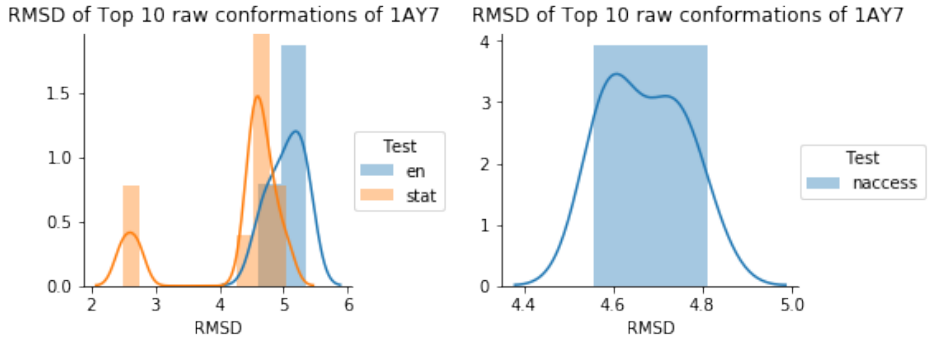

FIGURE 2.3: Densité de distribution en fonction du RMSD du Top 10 avant minimisation pour 1AY7

Enfin, le complexe *1DQJ* a également été évalué avec *DeNovo*, toujours avec $n = 1000$ conformations demandées en entrée et un angle de rotation de 30ř. Ce complexe a été choisit car il possède trois positions initiales au lieu de une seule pour *1AY7*. De nouveau, les conformations minimisées et non minimisées ont été évaluées.

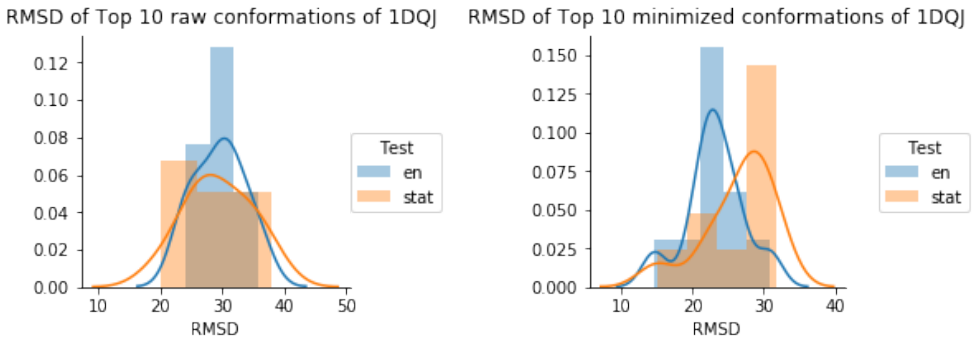

FIGURE 2.4: Densité de distribution en fonction du RMSD du Top 10 avant et après minimisation pour 1DQJ

II Discussion

Le scoring des conformations minimisées de *1AY7* par *MeetDockOne* ne fournit aucune conformation native (Figure 2.1), même s'il montre une meilleure performance de naccess par rapport à msms. Cela ne semble pas être un problème inhérent au programme quant à sa capacité à évaluer correctement et trouver la bonne conformation. En effet, l'évaluation des mêmes conformations par *DeNovo* (Figure 2.2) montre un résultat intéressant : le scoring énergétique ne donne aucune cnformation native tandis que le scoring statistique donne une distribution de RMSD peu éparse centrée autour de 5 ångström. Le problème doit donc résider dans le fait qu'un élément de partie sampling n'est pas adapté à un calcul énergétique.

L'évaluation des conformations non minimisées suggère que la minimisation impacte négativement le rendement en conformations natives pour *1AY7*(Figure 2.3), comme vu dans la partie I. L'explication est que le minimiseur utilisé, *MaxDo*, ne se prête pas à un calcul énergétique typiquement tout-atome, comme c'est le cas pour les méthodes de scoring considérées, car c'est un minimiseur dit "gros grains". En particulier, le scoring énergétique de *De Novo* tient compte des hydrogènes du complexe, qui ne sont pas présents dans les fichiers de sorties de MaxDo.

Le scoring de 1QDL via *DeNovo* (Figure **??**) donne d'autres informations : en effet, la qualité des conformations obtenues pour 1QDL est très faible en termes de nombre de natives. Le scoring des conformations minimisées ne donne pas le même profil que pour 1AY7, montrant que la qualité des conformations samplées reste importante pour le programme.

Conclusion et perspectives

L'échantillonnage par *SWING* dans sa version actuelle possède quelques défauts. D'abord, la contrainte d'un recul du ligand dans le cas où les partenaires sont trop proches, diminue les conformations natives. Cette otpion sera prochainement retirée, afin de tester l'obtention de meilleures conformations ligand. D'autre part, le nombre de conformation $n = 1000$ n'est peut être pas suffisant pour certains complexes. Cependant, avec des ressources limitées, il est compliqué d'augmenter cette valeur, en considérant les complexes ayant un grand nombre de positions initiales. Néamoins, le tirage des angles selon une modalité gaussienne semble être un bon paramètre.

La faible dépendance de la performance de *SWING* à la qualité des interologues, en terme d'identité, est très encourageante pour la suite. Ce résultat suggère que de bonnes conformations ligand peuvent être obtenues même lorsque l'identité est faible. L'augmentation du nombre de conplexes échantillonné permettrait de confirmer ce résultat.

Pour la partie scoring, il a été déterminé que le minimiseur *MaxDo* n'était pas forcément adapté aux méthodes de scoring choisies. En revanche, un programme de sampling ne peut généralement pas se passer de minimiseur, car il permet d'éliminer les conformations contenant des clash stériques entre le ligand et le récepteur. Ces clash peuvent être observés dans certaines conformations du dossier "Proteins" (donc non minimisées) sous *Pymol*. Pour pouvoir continuer à utiliser *DeNovo* à son potentiel maximal, il serait nécessaire de se procurer ou programmer un minimiseur tout-atome. Cependant, l'algorithme de *MeetDockOne* gère les clash en calculant la complémentarité de forme des conformations. Une autre solution serait donc d'utiliser uniquement *MeetDockOne* avec *SWING* à l'état actuel. En outre, l'évaluation des conformations a été fait avec un angle de 30ř et un nombre de conformations générées de $n = 1000$ car antérieur à la détermination du nombre d'angles et de conformations optimales, il serait donc judicieux de refaire le scoring avec ses paramètres optimaux. Enfin, il est absolument nécessaire de scorer d'autres conformations, particulièrement des conformations de complexes particuliers tels que des homomères ou des macro-assemblages.

La perspective principale suite à ces résultats est la création d'un programme mixte, utilisant une alternance d'échantillonnage et d'évaluation. Ainsi, les meilleures conformations seraient récupérées à chaque itération, servant de position initiale pour générer de nouvelles conformations, en diminuant progressivement l'intervalle de rotation. Une autre amélioration serait l'ajout de rotation au niveau des chaînes latérales, afin d'obtenir des modèles hautement résolutifs.

Bibliography

Faure, G., Andreani, J., and Guerois, R. (2012). InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res.*, 40(Database issue):D847–856.

Huang, S.-Y. (2014). Search strategies and evaluation in proteinprotein docking: principles, advances and challenges. *Drug Discovery Today*, 19(8):1081 – 1096.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U.S.A.*, 89(6):2195–2199.

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, 287(5450):116–122.

Annexe

A Analyse sampling partie 1 : Capacité de *SWING* à générer des conformations natives

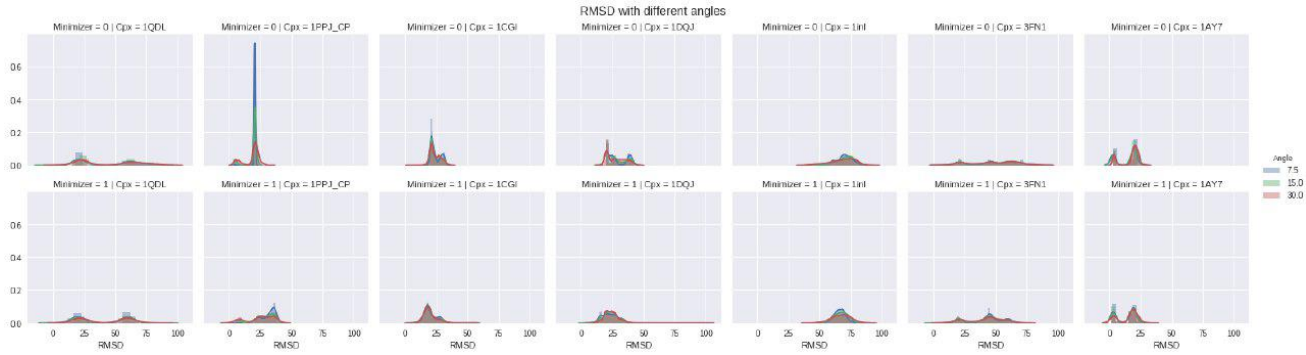

FIGURE 2.5: Densité de distribution en fonction du RMSD pour les complexes *1QDL 1PPJ_CP 1CGI 1DQJ 1inl et 3FN1*, avec des angles dans l'intervalle $[-7,5°,7,5°]$, $[-15°,15°]$ ou $[-30°,30°]$, tirage gaussien.