SWING : Sampling With INteroloG

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

January 2, 2018

1 Introduction

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 protéines. La modélisation computationnelle des interactions protéiques est essentielle à la compréhension des processus biologiques, des mécanismes d'interaction, et dans le domaine pharmacologique, la détermination expérimentale de ces complexes étant compliquée.

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, nommé *SWING* pour *Sampling With INteroloG*.

Dans ce document, la stratégie d'échantillonnage par *SWING* sera détaillée, ainsi que ses points forts et limites, illustrés par une évaluation succinte des performances de *SWING*. Dans une seconde partie, nous présenterons les stratégies de scoring choisies pour évaluer les conformations ligand générées par *SWING*.

2 Stratégie du sampling SWING

2.1 Principe

SWING utilise une méthode de recherche directe dans l'espace cartésien. Afin de réduire l'espace d'échantillonage, *SWING* exploite les informations évolutives pré-existantes 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 interologuesWalhout et al. [2000]. La base de données interEvolFaure et al. [2012] recense les interologues structuraux à partir de la PDB (Protein Data Bank). L'outil interEvolAlign associéFaure 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. Afin d'accélérer le processus de rotation du ligand, les quaternions ont été préférés aux angles d'Euler. Ce calcul se résume ainsi à $q_{int}\overrightarrow{LA_i}q_{int}^{-1}$ où $q_{int} = q_\alpha q_\beta q_\gamma$ et A est l'ensemble des atomes du ligand pour la rotation intrinsèque (par rapport à lui-même) et $q_{ext}\overrightarrow{RL}q_{ext}^{-1}$ où $q_{ext} = q_\theta q_\phi$ pour la rotation extrinsèque (par rapport au récepteur) du ligand.

Les conformations ligand obtenues ont ensuite été minimisées par *MaxDo*.

Le but de l'échantillonnage par *SWING* est de couvrir 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.

2.2 Avantages et limites

Le principal avantage de *SWING* est de fournir des conformations ligand à partir d'une position initiale cohérente avec les informations évolutives disponibles. Ainsi, pour le complexe *1AY7*, nous

avons pu obtenir 126 conformations natives (avec un RMSD ligand ou lRMSD $\leq$ à $5\mathring{A}$) avec le meilleur à $1,3\mathring{A}$, et 20 conformations natives pour le complexe *1PPJ_ CP*, avec le meilleur à $1,6\mathring{A}$. Les paramètres fixée étaient les paramètres par défaut et $n = 1000$, après minimisation.

L'inconvénient majeur est la dépendance à la disponibilité d'interologues structuraux dans la *Protein Data Bank* (*PDB*). Parmi le jeu de données à notre disposition, 33 des 44 complexes possédaient au moins un interologue, nous permettant ainsi, dans la majorité des cas, d'utiliser *SWING*.

D'autre part, nous avons observé la capacité de *SWING* à échantilloner des conformations ligand proches de la solution native en fonction du pourcentage d'identité (le plus faible pourcentage d'identité entre l'identité du ligand homologue et le récepteur homologue) (1). Ainsi, même pour des faibles valeur d'identité, *SWING* parvient à échantillonner des positions proches de la solution native, même si les fortes valeur d'identité assurent un meilleur résultats avec moins de variabilité.

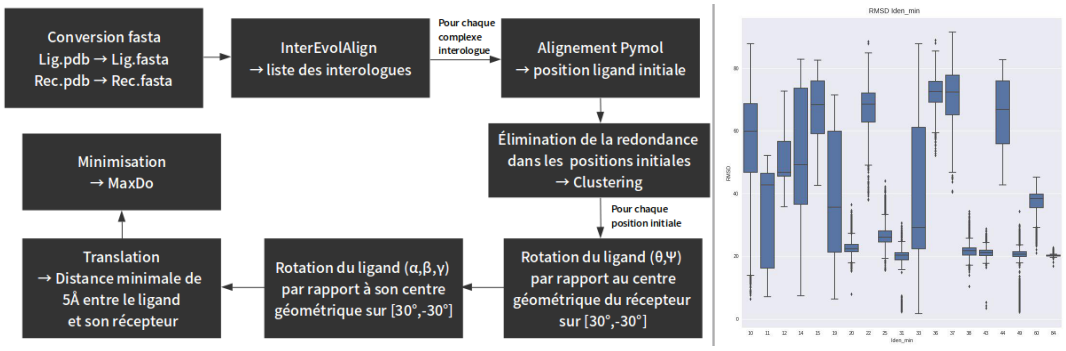

Figure 1: Workflow du programme *SWING* - RMSD ligand en fonction de l'identité minimale entre le ligand homologue et le récepteur homologue. Des faibles valeurs de RMSD peuvent être obtenues avec des interologues relativement éloignés.

2.3 Stratégie des programmes de scoring choisis

Deux stratégies ont été retenues pour scorer les conformations de *SWING* : celle de l'Equipe 1 (*MeetDockOne*) et celle de l'Equipe 11 (*DeNovo*). 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 intéressant dans un soucis de reproductibilité des résultats.

L'intégration des méthodes de scoring après *SWING* a été fait via un pipeline SnakeMake [Köster and Rahmann 2012]. Le pipeline est fournit avec un fichier de configuration (config.json) permettant à l'utilisateur de choisir quelle méthode de scoring adopter, et aussi de changer de nombreux paramètres des méthodes sampling ou scoring. Le fichier "README.md" fourni avec le pipeline explique en détail son utilisation.

References

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.

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.

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.