UE: Langages de scripts – Python
Docking

Martin Larralde

Introduction

L'importance des protéines dans les systèmes biologiques eucaryotes n'est de nos jours plus à démontrer. Ces molécules, synthétisées dans les cellules à partir de l'information génétique contenue dans les molécules d'ADN, elles-mêmes rassemblées dans le noyau, sont très diverses dans leur structure, et incidemment dans leur fonction.

Quasiment toutes les protéines vont, entre le moment où elles sont synthétisées et celui où elles sont dégradées, s'associer avec d'autres molécules, qui peuvent être des métabolites, des hormones, ou encore d'autres protéines. Ces associations ont une grande importance biologique, qui peut aller de la perception de la douleur à la régulation du système digestif, en passant par le contrôle des divisions cellulaires.

La connaissance des interactions protéiques s'avère donc vitale pour la médecine, afin de mettre au point des médicaments efficaces, ou encore pour les biologistes qui souhaitent comprendre des mécanismes biologiques plus complexes.

La seule façon de déterminer une conformation native d'association de deux protéines fut pendant très longtemps de purifier le complexe, avant de le visualiser dans l'espace à l'aide de la cristallographie ou dans une moindre mesure de la RMN (résonance magnétique nucléaire).
Cependant, avec le gain en puissance des ordinateurs, et le développement de méthodes de calcul de plus en plus optimisées, il est possible de calculer des conformations uniquement *in silico.*

Deux méthodes sont largement utilisées : le calcul d'un *champ de force* et du *potentiel énergétique* d'une conformation spatiale d'un complexe, où l'on cherche ici à calculer les forces « réelles » s'exprimant sur les protéines, ainsi que la recherche par *potentiels statistiques*, où l'on cherche ici à trouver la conformation d'un complexe protéique en s'inspirant d'autres complexes déjà identifiés ; ces deux méthodes ne sont pas exclusives.

On s'intéresse ici aux méthodes de calcul de potentiel énergétique, utilisant une fonction de score prenant en compte une ou plusieurs forces physico-chimiques prenant part à l'interaction.

En particulier, on s'appliquera ici à utiliser des méthodes de prédiction d'association protéique pour étudier le complexe Barnase/Barstar, dont la structure a déjà été déterminée expérimentalement**[1]**.

Méthode

1. Fonction de score et potentiels

Une fonction de score est une fonction permettant d'assigner un score à une conformation prenant en compte deux protéines. Une pratique courante consiste à utiliser un ou plusieurs potentiels physico-chimiques comme fonction de score : une association sera d'autant meilleure qu'elle sera stable au niveau moléculaire, c'est à dire qu'elle aura un potentiel faible (celui-ci pouvant être négatif).

Il existe de nombreuses expressions mathématiques de potentiels physico-chimiques, chacune visant en général à modéliser une force.

Par exemple, le potentiel de Coulomb**[2]**:

(i) $$E_{Coulomb} = \sum_i \sum_j \frac{1}{4\pi\epsilon} \frac{q_i q_j}{d_{i,j}}$$

avec q_i *:* la charge du *i-ème* atome de la première protéine
q_j *:* la charge du *j-ème* atome de la seconde protéine
$d_{i,j}$ *:* la distance entre ces deux atomes
ε *:* la permittivité du milieu

permet de modéliser les forces électrostatiques entre tous les atomes de deux protéines. Un potentiel négatif signifiera que les deux protéines auront tendance à s'attirer, tandis qu'un potentiel positif signifiera que les deux protéines auront tendance à se repousser (selon la conformation proposée).

2. Première approche : champ de force de Cornell

En 1995, Cornell *et al.* proposent un champ de force**[3]** pour simuler la conformation spatiale d'une molécule organique, d'un acide nucléique, ou d'une protéine. Ce champ de force, étant supposé être utilisé pour une seule et unique molécule, comporte des termes pour analyser la torsion et la rotation des liaisons au sein de la molécule, la topologie des atomes, les forces électrostatiques et les interactions de Van der Waals.
Néanmoins, ce champ de force peut être adapté pour mesurer le potentiel énergétique d'une association protéine-protéine, en n'en considérant que les termes non-liés : interactions de Van der Waals et forces électrostatiques.

Le potentiel total est alors obtenu en sommant le potentiel de Coulomb (décrit par **(i)**) et le potentiel de Lennard Jones**[4]** (décrit ci-dessous) :

(ii) $$E_{Lennard\,Jones} = \sum_i \sum_j \frac{A_{i,j}}{R_{i,j}^{12}} - \frac{B_{i,j}}{R_{i,j}^{6}} = \sum_i \sum_j \epsilon_{i,j} \left(\frac{R_{i,j}^{*}}{R_{i,j}} \right)^{12} - 2\,\epsilon_{i,j} \left(\frac{R_{i,j}^{*}}{R_{i,j}} \right)^{6}$$

avec $R_{i,j}$: la distance entre le *i-eme* atome de la première protéine et le *j-ème* atome de la seconde protéine
$R^*_{i,j}$: la distance de Van der Waals optimale pour une paire d'atomes de même nature que le *i-ème* atome de la première protéine et le *j-ème* atome de la seconde protéine
$\varepsilon_{i,j}$: la profondeur du puits de potentiel de Van der Waals pour la paire d'atomes considérée.

3. Première amélioration : Écrantage du champ électrique

Au sein d'un fluide contenant des porteurs de charges électriques mobiles, on constate un effet appelé *écrantage*, qui consiste en une atténuation du champ électrique. Le cytosol, qui contient de nombreux ions (Na^+, K^+, Ca^{2+}, etc.) et molécules organiques chargées (glutamate, arginine, etc.) est assez sensible à cet effet physique. Une expression du potentiel électrostatique prenant en compte l'écrantage a été proposée par Mehler *et* Eichele**[5]** :

(iii) $$E_{ScreenedCoulomb} = \sum_i \sum_j \frac{q_i q_j}{\epsilon(d_{i,j}) d_{i,j}}$$
$$\text{et } \epsilon(d_{i,j}) = A + \frac{B}{(1 + k * e^{-l*B*d_{i,j}})} \text{ où } B = \epsilon_{H_2O} - A$$

avec $d_{i,j}$: la distance entre le *i-eme* atome de la première protéine et le *j-ème* atome de la seconde protéine
q_i : la charge du *i-ème* atome de la première protéine
q_j : la charge du *j-ème* atome de la seconde protéine-p
ε_{H2O} : la permittivité du milieu (le cytosol dans la plupart des cas)
A, k, l : des constantes d'intégration (données par Mehler *et al.*)

4. Seconde amélioration : liaisons H

Les liaisons hydrogène jouent également un rôle important dans les liaisons biologiques, en particulier au sein des interactions protéine-protéine. En 2002, Fabiola *et al.* ont proposé une expression d'un potentiel prenant en compte ces liaisons H**[6]**:

(iv) $$E_{Fabiola} = \sum_D \sum_A \left[\left(\frac{\sigma}{R_{D-A}} \right)^6 - \left(\frac{\sigma}{R_{D-A}} \right)^4 \right] \cos^m(min(\theta - \theta_{low}, \theta - \theta_{high})) SW(R_{D-A})$$

avec $R_{D\text{-}A}$: la distance entre le donneur et l'accepteur d'électrons
θ_{low}, θ_{high} : les angles entre le donneur, l'accepteur et l'atome en alpha de l'accepteur pour lesquels la liaison est la plus forte
m : la puissance de coupure (usuellement m=4).
θ : l'angle entre le donneur, l'accepteur et l'atome en alpha de l'accepteur pour le couple donneur/accepteur considéré
SW : la fonction de switch définie dans Brunger *et al.*
$\sigma = R_0\sqrt{2}/3$ où R_0 est la distance pour laquelle le potentiel est minimum

D'après le travail des auteurs sur les données structurales de la PDB, θ_{low}=115°, θ_{high} = 155°, et R_0 = 2.9 Å.

Implémentation

Un module Python a été développé afin de répondre aux problématiques de l'énoncé : *dockerasmus.* Ce module permet d'importer des conformations protéiques depuis des fichiers de la PDB et d'attribuer un score à des structures de complexes.

Ce module est disponible sous la licence GPLv3, et peut-être téléchargé via *pip* (l'outil d'installation des packages sous Python), ou depuis le dépôt git à l'adresse suivante: https://gitlab.com/althonos/dockerasmus. La documentation complète est disponible en ligne à l'adresse suivante : https://dockerasmus.readthedocs.io/.

dockerasmus fonctionne indépendamment avec plusieurs librairies de calcul algébrique : ***theano, mxnet, tensorflow.*** Seul le module ***numpy*** est une dépendance (ainsi que *six* pour la compatibilité Python2 / Python 3), mais il est conseillé d'utiliser *theano* pour un maximum de performance dans les calculs, ainsi que *scipy* pour le calcul des matrices de distance.

À noter :
- la fonction de coupure pour le potentiel de Fabiola n'a pas été implémentée car sa spécification est introuvable
- pour tous les potentiels prenant en compte la constante diélectrique du milieu, la valeur utilisée est 65.0, ce qui correspond aux travaux de Cossins *et al.* sur le milieu cytosolique**[7]**.

dockerasmus est testé par Gitlab-CI contre différentes versions de Python (2.7, 3.4, 3.5, 3.6), avec un coverage de 91 % environ.

A la racine du dépôt git se trouve un répertoire *scripts* qui contient les deux scripts demandés dans l'énoncé du projet. Ceux-ci nécessitent *docopt, progressbar2,* ainsi que les dépendances de *dockerasmus* pour fonctionner. L'aide peut-être affichée à l'aide du paramètre `–help`.

Résultats

En terme de RMSD, la meilleure structure identifiée parmi les 1000 structures pré-générées est **1BRS_A_1BRS_B_allatom_28_DP.pdb**, avec un RMSD sur le ligand de *1.427292*, un RMSD sur le complexe entier de *0.9553977*, et un RMSD sur l'interface de *1.4342136787.* Les structures natives utilisées sont *Rec_natif_DP.pdb* et *Ligand_natif_DP.pdb.*

1. Fonction de score de Cornell

La fonction de score de Cornell (utilisant le potentiel de Lennard Jones et le potentiel de Coulomb uniquement) identifie la structure native comme étant la meilleure structure, avec un score de *-84.943.* La conformation native est la seule conformation ayant un score négatif parmi les conformations proposées.

Cependant, la conformation avec le score de Cornell le plus faible parmi les structures considérées (2.7) a un RMSD assez élevé vis-à-vis de la conformation native (environ 30 pour le ligand seul, 34 pour l'interface seule). La conformation la plus proche de la conformation native ne se retrouve qu'en 8^{e} position.

Conformation	Score	RMSD/ligand	RMSD/complexe
(native)	-84.943	0.0	0.0
593	**2.7177743944**	30.801579070	20.6178894561
370	177.49558832	24.817023491	16.6119615431
482	258.12052560	48.366119176	32.375200516
338	319.01419375	27.157972576	18.1789406042
523	334.97253447	36.069998895	25.7815899671
869	347.80684670	33.381771222	22.3450124864
498	357.84234253	53.181894512	35.5987730247
28	394.61450949	**1.4272923899**	**0.955397664843**
375	400.26106178	18.088324291	12.1079205009
90	440.21507413	49.206290030	32.9375921306

Table 1 : Score et RMSD des 10 meilleurs conformations d'après la fonction de score de Cornell *et al*.

2. Fonction de score améliorée

On propose ici d'utiliser ici une fonction de score n'utilisant que le potentiel de Fabiola. En effet, les autres potentiels tendent à surestimer l'importance de la proximité des deux protéines, tandis que le potentiel de Fabiola introduit une contrainte liée à l'alignement des donneurs et des accepteurs d'électrons.

Conformation	Score	RMSD/ligand	RMSD/complexe
28	**-46.469232256292**	**1.42729238992**	**0.955397664843**
(native)	-44.913186056597	0.0	0.0
920	-41.583424307556	40.8970850852	27.3755958241
71	-40.745361938959	44.166728406	29.5642220758
332	-40.155402373234	17.3058185602	11.5841286322
674	-40.076726605149	35.2773347236	23.6138603798
299	-37.371071661085	42.3008104843	28.3152182713
53	-36.910476839803	41.7029669351	27.915035145
885	-36.890985954321	37.6895689597	25.22855613
416	-36.510812107930	32.5797629735	21.8081660673

Table 2 : Score et RMSD des 10 meilleurs conformations d'après la fonction de score utilisant le potentiel de Fabiola *et al.*

On remarque ici que la conformation native n'arrive qu'en second, mais que la conformation la plus proche de la conformation native est également celle qui a le score le plus bas. Les RMSD obtenus sont moins élevés en moyenne que pour la fonction de score de Cornell (19.2 contre 21.4 pour les ligands, 27.9 contre 32.9 pour le complexe). La fonction de score utilisant le potentiel de Fabiola semble donc mieux favoriser les résidus bien positionnés à l'interface, qui sont ceux pour lesquels les termes du calcul du potentiel sont non-négligeables.

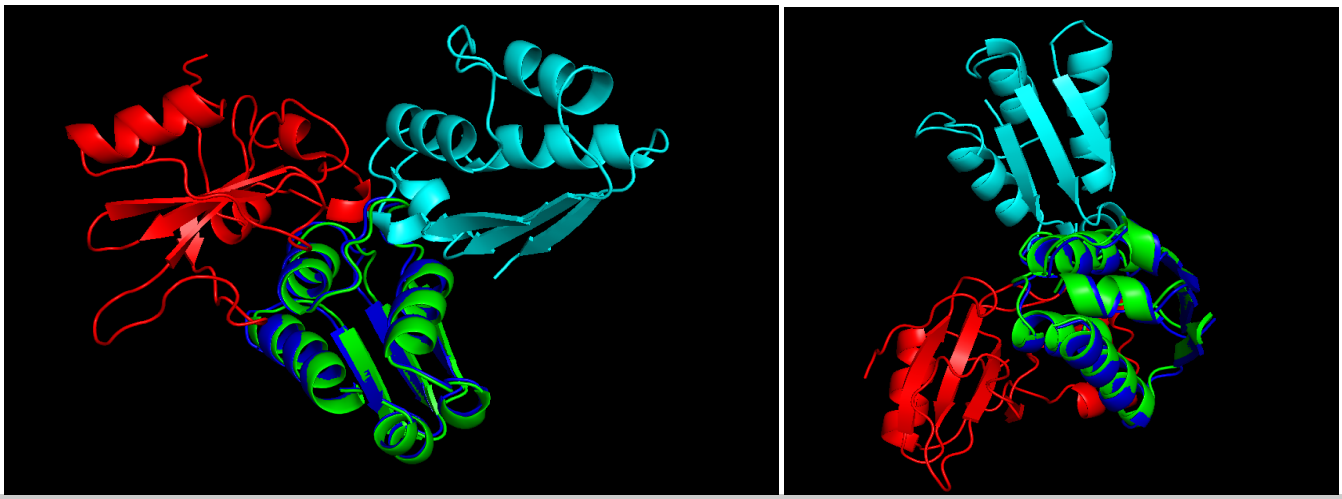

Figure 1 : Meilleures conformations obtenues avec *dockerasmus* (sous PyMol)
rouge : conformation native de la barnase
vert : conformation native de la barstar,
cyan : meilleure conformation de la barstar d'après Cornell
bleu : meilleure conformation de la barstar d'après Fabiola

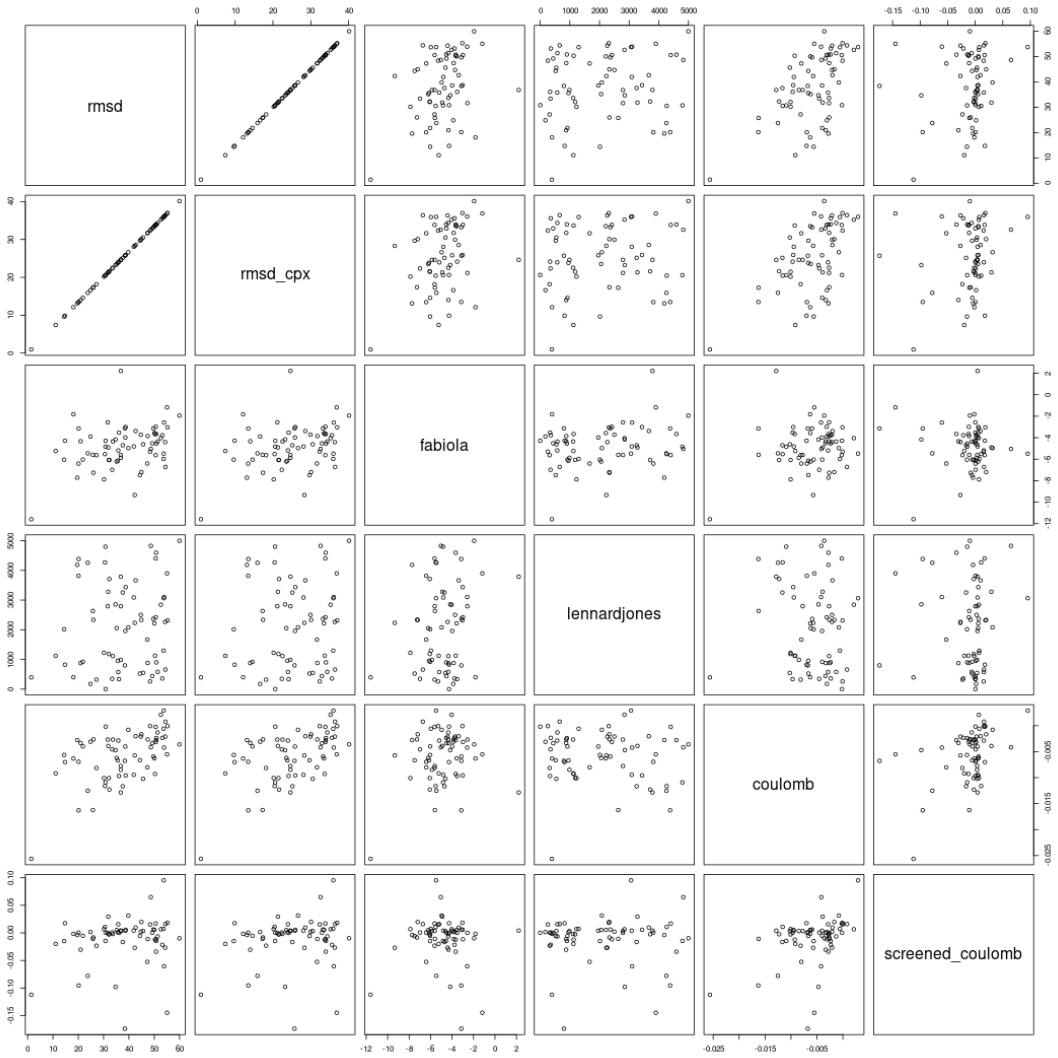

Figure 2: Graphe des paires des différents potentiels et des RMSD (ligand seul ou complexe) sur le jeu de données barnase/barstar.
Les points ayant un potentiel de Lennard Jones supérieur à 10000 ont été retirés pour avoir une représentation plus lisible.

On remarque finalement que le potentiel ayant une distribution la mieux corrélée avec le RMSD est le potentiel de Coulomb (corrélation de 30 % environ). Sous R, si on teste le modèle linéaire de Cornell (RMSD = Lennard Jones + Coulomb), le modèle n'a qu'un R^2 ajusté de 9 %. Ce R^2 ajusté monte à 11 % pour un modèle linéaire utilisant Fabiola et Coulomb (RMSD = Fabiola + Coulomb). On pourrait éventuellement utiliser les poids obtenus pour recalculer les scores avec les poids optimaux (ce qui n'améliorerait pas tant les résultats obtenus, mais pourrait être utile si l'on voulait travailler par homologie sur une autre structure d'interactions proche du complexe barnase-barstar).

Conclusion

Au terme de ce projet, la fonction de score développée a permis d'identifier la meilleure conformation du complexe barnase/barstar parmi un ensemble de conformations donné.

Il convient de nuancer ce résultat car si, en effet, les deux fonctions ont permis l'identification d'une conformation du complexe, on constate que les scores obtenus avec les différents potentiels sont loin d'identifier les meilleures conformations en dehors de la conformation native : si l'on ne connaissait pas la conformation native, il serait impossible de savoir quel potentiel utiliser *a priori.* En ce sens, avoir connaissance d'une homologie de conformation au préalable représente une information très importante dans le cadre du docking, car on peut ainsi essayer d'assigner un poids relatif à l'importance de chacun des potentiels dans le calcul des scores : forces électrostatiques, interactions de Van der Waals, liaisons H, etc.

Enfin, ce projet ne fait pas usage de la notion d'ajustement induit, et suppose que les protéines restent dans leur conformation native lorsqu'elles interagissent entre elles. Ce modèle clef-serrure, suffisant pour certaines interactions, peut passer à côté des véritables solutions pour d'autres complexes (par exemple dans le cas de la multimérisation de HFBI liée au déplacement d'une épingle β**[8]**). Néanmoins, prendre en compte la torsion des protéines nécessiterait des calculs beaucoup plus complexes.

Perspectives

dockerasmus a été développé pour fonctionner de façon modulaire, pour le calcul des potentiels ou pour le calcul des scores. L'implémentation a permis de tester différentes librairies de calcul algébrique (*theano, numpy, tensorflow, mxnet*), mais certaines librairies n'ont pas été implémentées par faute de temps (*PyCUDA, PyTorch*).

De même, seuls quatre potentiels ont été implémentés : Lennard Jones, Coulomb, Coulomb écranté, et Fabiola. Le module gagnerait à être développé en ajoutant des potentiels supplémentaires pour permettre de raffiner les scores.

Enfin, on pourrait essayer d'améliorer les résultats obtenus en appliquant des transformations spatiales supplémentaires aux meilleures conformations obtenues pour tenter de prédire une conformation encore meilleure, à l'aide d'un moteur 3D simpliste (*dockerasmus.spatial*). Cela n'a pas été fait par manque de temps, mais il serait tout à fait possible d'implémenter un algorithme génétique utilisant une fonction de score quelconque pour obtenir une conformation de docking à partir de deux conformations natives, sans pré-génération de structures intermédiaires.

Bibliographie

[1] Buckle, Ashley M., Gideon Schreiber, and Alan R. Fersht. “Protein-Protein Recognition: Crystal Structural Analysis of a Barnase-Barstar Complex at 2.0-.ANG. Resolution.” *Biochemistry* 33, no. 30 (August 1994): 8878–89. doi:10.1021/bi00196a004.

[2] Coulomb, C. A. “Premier mémoire sur l’électricité et le magnétisme”. Histoire de l’Académie Royale des Sciences, 569-77 (1785).

[3] Cornell, Wendy D., Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.” *Journal of the American Chemical Society* 117, no. 19 (May 1, 1995): 5179–97. doi:10.1021/ja00124a002.

[4] Jones, J. E. “On the Determination of Molecular Fields. II. From the Equation of State of a Gas”. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 106, 463–477 (1924).

[5] Mehler, E. L., and G. Eichele. “Electrostatic Effects in Water-Accessible Regions of Proteins.” *Biochemistry* 23, no. 17 (August 1, 1984): 3887–91. doi:10.1021/bi00312a015.

[6] Fabiola, Felcy, Richard Bertram, Andrei Korostelev, and Michael S. Chapman. “An Improved Hydrogen Bond Potential: Impact on Medium Resolution Protein Structures.” *Protein Science : A Publication of the Protein Society* 11, no. 6 (June 2002): 1415–23.

[7] Cossins, Benjamin P., Matthew P. Jacobson, and Victor Guallar. “A New View of the Bacterial Cytosol Environment.” *PLoS Computational Biology* 7, no. 6 (June 9, 2011). doi:10.1371/journal.pcbi.1002066.

[8] Riccardi, Laura, and Paolo Mereghetti. “Induced Fit in Protein Multimerization: The HFBI Case.” Edited by Ozlem Keskin. *PLOS Computational Biology* 12, no. 11 (November 10, 2016): e1005202. doi:10.1371/journal.pcbi.1005202.