Jean-François Gibrat,
Mathématiques et Informatique Appliquées
du Génome à l'Environnement,
INRA – Unité 1404.
Domaine de Vilvert, bât. 233,
78350 Jouy-en-Josas
Tél. : +33 (0)1 34 65 28 97
Email : jean-francois.gibrat@inra.fr

Jouy-en-Josas, le 10 novembre 2017

Thèse de doctorat de l'Université Paris-Saclay
Spécialité : biologie

Rapport sur la thèse présentée par David Mignon et intitulée « Computational Protein Design : un outil pour l'ingénierie des protéines et la biologie synthétique »

Description générale du travail de thèse

Le travail de thèse présenté par D. Mignon s'inscrit dans le cadre général de la conception de protéines par ordinateur. Ce travail utilise Proteus, un logiciel développé depuis de nombreuses années dans le laboratoire d'accueil et qui permet de générer des séquences polypeptidiques compatibles avec un repliement donné a priori. Plus précisément, le travail de thèse porte sur l'analyse de composants fondamentaux de Proteus : les différents algorithmes utilisés pour explorer l'espace de recherche et les énergies de référence de l'état dénaturé qui constituent un élément essentiel de la fonction d'énergie.
Le manuscrit de thèse est constitué de 5 chapitres. Le premier présente la conception assistée par ordinateur de protéines et décrit ses principaux éléments, fonctions d'énergie de la protéine et du solvant et méthodes d'exploration de l'espace de recherche. Le deuxième chapitre se focalise sur les caractéristiques de Proteus et fournit une brève description des outils d'analyse qui vont être utilisés dans la suite de la thèse. Le troisième chapitre présente différents développements algorithmiques apportés à Proteus, en particulier la parallélisation de l'algorithme d'échange de répliques. Les deux derniers chapitres présentent le travail de thèse proprement dit : le quatrième chapitre compare trois méthodes stochastiques pour explorer l'espace de recherche et le cinquième décrit la conception assistée par ordinateur du domaine protéique PDZ. Je reviendrai en détail sur ces deux chapitres dans la suite de ce rapport.

Remarques générales concernant la thèse

Remarques concernant la forme de la thèse

Le manuscrit est agréable à lire. Il est bien structuré et illustré par de nombreuses figures.
J'ai apprécié, dans la section bibliographie, la mention de la page où est citée la référence.

Les deux premiers chapitres fournissent un rappel bref de la conception de protéines assistée par ordinateur en général et de sa mise en œuvre dans le logiciel Proteus qui est très utile pour la suite du document. Bien que les deux derniers chapitres aient leur propre conclusion, je regrette un peu le manque d'un chapitre général de conclusion où D. Mignon aurait pu faire le bilan de son travail de thèse et le mettre en perspective.

Remarques concernant le fond de la thèse

Le travail de thèse porte sur l'analyse de points techniques du logiciel Proteus ce qui est toujours un peu « ingrat » à exposer. Une partie importante de ce travail a sans doute consisté en développements algorithmiques (ce qui n'est que partiellement reflété dans le bref chapitre 3 du manuscrit). Dans le manuscrit, D. Mignon se focalise plus sur les aspects techniques de son travail que sur les caractéristiques du modèle qu'il utilise et les approximations faites pour rendre les calculs possibles (par exemple, je ne trouve pas que l'interprétation physique de Proteus soit si transparente que ça – cf. le dernier paragraphe p. 89).

Ce travail a donné lieu à cinq publications, dont deux en premier auteur. L'article de 2016 où D. Mignon est premier auteur correspond sans équivoque au chapitre 4 du manuscrit. La relation entre l'article de 2017 et le chapitre 5 est un peu moins claire. Dans ce chapitre, D. Mignon s'en tient à l'aspect méthodologique (comparaison avec Rosetta, comparaison de différents modèles de solvant) sans insister, si l'on en juge par le titre de l'article, sur l'application qui en est faite (cette dernière étant sans doute la contribution de N. Panel).

Travail de thèse

Comparaison d'algorithmes stochastiques d'exploration de l'espace de recherche

Le chapitre 4 reprend l'article correspondant (où les références semblent avoir été mises en adéquation avec celles du manuscrit, sauf dans la section « Concluding discussion »).

Dans cet article, D. Mignon s'intéresse à 3 méthodes d'exploration de l'espace de recherche (une heuristique, une méthode de Monte-Carlo et une méthode d'échange de répliques) afin d'analyser leurs qualité et efficacité en ce qui concerne l'échantillonnage des régions de basse énergie. La qualité est mesurée en comparant les séquences générées dans ces régions de basse énergie avec des séquences réelles collectées dans la base de données SCOP. L'efficacité est mesurée en termes de temps CPU et de mémoire vive utilisée ainsi que la capacité de ces méthodes à trouver le minimum global d'énergie (en comparant avec une méthode exacte basée sur un algorithme de type « branch & bound »). Enfin, D. Mignon examine la densité d'états au-dessus du minimum global en traçant le nombre d'états pour une énergie donnée et l'entropie moyenne par position des séquences générées pour une énergie donnée (d'une façon pas très cohérente, car l'entropie ne fait intervenir que les types de résidus alors que les états correspondent aux combinaisons : types de résidus, rotamères).

La conclusion de l'article est que globalement l'heuristique et la méthode d'échange de répliques fournissent des résultats très similaires en termes de qualité des séquences générées (qui sont un peu moins diverses que les séquences réelles collectées dans SCOP) et d'efficacité d'exploration de l'espace de recherche.

À plusieurs endroits dans le manuscrit, D. Mignon semble confondre l'énergie d'un état avec l'énergie libre (p. 27 « *Pour le CPD l'énergie qui est pertinente est celle qui prend en compte l'énergie interne de la protéine, mais aussi son environnement aqueux moyen. Il s'agit donc d'une énergie libre de Gibbs* » ; p. 62, si $E_u(S)$ est une énergie libre, elle ne doit pas apparaître

comme telle dans l'exponentielle de l'équation 4.3 p. 62). Il y a aussi quelques inexactitudes (par exemple la probabilité d'acceptation, Éq. 4.8 p. 64, est juste $\min(1, \frac{n_f(t)}{n_f(t')} e^{-\beta \Delta E_M})$. La différence d'énergie, ΔE_M, n'intervient directement que si la probabilité de génération des conformations est symétrique).
D. Mignon s'intéresse aux états de faible énergie, mais n'indique jamais à partir de quand il considère qu'une séquence est incompatible avec le repliement étudié. Est-ce que les 10000 séquences/conformations qu'il retient pour ses comparaisons avec les séquences de SCOP ont toutes une énergie compatible avec le repliement ? Il serait intéressant de connaître l'énergie de la séquence native et des séquences de SCOP du repliement utilisé dans le modèle de Proteus.
Je me demande aussi pourquoi D. Mignon ne caractérise pas les trajectoires de Monte-Carlo effectuées (calcul de l'énergie moyenne de la trajectoire, de la variance éventuellement de l'autocorrélation). Si l'énergie moyenne et la variance sont telles que la zone d'énergie explorée est nettement supérieure au minimum global d'énergie, il y a peu de chances d'obtenir des séquences de faible énergie (de plus, cela permet de s'assurer qu'il y a bien un recouvrement entre les zones explorées par les différentes répliques).
La conclusion de ce chapitre laisse un peu le lecteur sur sa faim. Ne pourrait-on pas envisager de coupler l'heuristique et la méthode d'échange de répliques ? Le Monte-Carlo permet de faire un échantillonnage préférentiel (importance sampling), c'est-à-dire d'explorer en priorité les zones d'énergie favorables à une température donnée. La méthode d'échange de répliques permet à l'algorithme de Monte-Carlo d'éviter de rester bloqué dans un minimum local. On dispose donc d'une méthode efficace d'exploration de l'espace de recherche. Pour explorer les zones de faible énergie se trouvant sous l'énergie moyenne d'une trajectoire, on peut ensuite utiliser l'heuristique (qui est un moyen de minimiser l'énergie en suivant la plus forte pente dans un espace non continu).

Conception assistée par ordinateur du domaine protéique PDZ

Dans le dernier chapitre de sa thèse, D. Mignon, s'intéresse dans un premier temps à l'optimisation des énergies de référence de l'état dénaturé. Ces énergies sont des paramètres empiriques et D. Mignon les optimise en maximisant la vraisemblance d'un ensemble de séquences appartenant à un domaine protéique particulier. Cela favorise la génération de séquences par Proteus possédant des fréquences de résidus proches des fréquences observées dans les séquences appartenant au domaine choisi, ici le domaine PDZ. Dans un second temps, ces paramètres optimisés sont utilisés avec deux versions différentes d'un modèle de solvant (le modèle de « Born généralisé ») plus élaboré que celui utilisé dans le chapitre 4, afin de générer des séquences compatibles avec trois protéines du domaine PDZ. Les séquences ainsi générées sont comparées avec celles obtenues avec Rosetta. Les résultats montrent que la version la plus élaborée du modèle de solvant « Born généralisé » (appelé FDB) fournit des résultats plus satisfaisants que la version NEA. FDB fournit des résultats similaires en qualité à ceux de Rosetta. Cependant, les séquences générées avec Proteus et Rosetta, comparées aux séquences réelles des domaines PDZ correspondants, montrent une diversité moindre (mesurée selon l'entropie des séquences générées). Il est frappant, quand on compare les figures montrant les scores de similarité des séquences de ces deux méthodes avec ceux des séquences de SCOP entre elles, de voir que la distribution des scores de ces dernières est bien plus large que celles obtenues avec Proteus et Rosetta.

D. Mignon conclut ce chapitre en récapitulant les résultats obtenus et en présentant les limitations inhérentes à l'outil qu'il utilise.

Conclusion

D. Mignon a fourni un travail très conséquent visant à améliorer l'outil de conception assistée par ordinateur de protéines du laboratoire. Il s'agit d'un travail très technique s'intéressant aux composants fondamentaux de cet outil, la méthode d'exploration de l'espace des configurations et les paramètres empiriques d'énergie utilisés. Il a bien valorisé le travail effectué avec 5 publications, dont 2 en premier auteur.
En conséquence, j'émets un avis très favorable pour la soutenance de thèse de D. Mignon.

Fait à Jouy-en-Josas le 10 novembre 2017.

J-F Gibrat