REMC : Un algorithme de repliement des protéines dans le modèle HP

Auteur : Oussaren Mohamed

14 septembre 2022

Table des matières

Résumé

La prédiction de la structure 2D ou 3D d'une protéine à partir d'une séquence est un problème majeur. Une des solutions consiste à évaluer la fonction d'énergie. Nous avons pour cela implémenté une méthode de *replica exchange Monte-Carlo* (REMC) se basant sur une représentation de modèle HP [2] et la combinaison de mouvements bien identifiés. Nos résultats montrent que notre algorithme est capable d'identifier une conformation selon nos attentes.

1 Introduction

Un des problèmes majeurs du repliement ab initio des protéines est la prédiction du repliement natif à partir d'une séquence donnée. En effet, la résolution des structures de celles-ci pourrait fournir des infor-

mations inestimables dans le processus biologique mais actuellement que quelques centaines de milliers de structures de protéines résolues expérimentalement (environ 10%) couvrent l'ensemble des séquences protéiques humaines [4]. D'autant plus que les données massives obtenues par le séquençage haut débit s'appuient seulement sur des techniques et données expérimentales rendant le temps de recherche très couteux.

Par conséquent, l'utilisation d'outils informatiques est nécessaire pour permettre la prédiction des structures de protéines. Actuellement l'algorithme le plus connu pour la prédiction de protéines est AlphaFold2. Alphafold2 est un outil de modélisation de protéines 3D par prédiction. Il s'agit d'un algorithme de prédiction structurale en Machine Learning et a permis d'élargir le recouvrement des protéines résolues avec une prédiction fiable. Son principe est quant à lui plus complexe mais l'idée est la suivante : l'algorithme va chercher à créer une matrice de distance caractérisant la position de chaque acide aminée de la séquence cible et par la suite donné forme à une structure 3D de notre protéine [3].

Dans ce projet, nous sommes intéressés au développement d'une méthode qui est conceptuellement simple. Nous avons développé l'algorithme REMC (*Replica Exchange Monte-Carlo*)[1], une méthode se basant sur une représentation des résidus par des modèles HP de Dill [2] et la mise en place d'un système de mouvement pseudo-aléatoire. Cette méthode s'appuie sur une représentation 2D de la protéine mais reste assez couteuse en temps puisque sa prédiction ab initio de structure de protéines s'appuie une représentation interne très différente de celle d'Alphafold2.

2 Matériels et méthodes

L'objectif du REMC est de déterminer une conformation ayant une énergie minime. Cette méthode comprend deux étapes que nous décrivons : le changement transitoire des répliquats et l'évaluation des répliquats. Tout d'abord la séquence d'intérêt est récupéré avec laquelle nous générons n (un paramètre fourni lors de l'exécution du programme) répliquats avec des conformations pseudo-aléatoires. Ces répliquats vont subir des changements de conformation transitoire par le biais de la méthode de Monte-Carlo. Sur l'ensemble des nouvelles conformations candidats certaines subiront un échange que si la température d'un répliquats est voisin à une autre. Par ailleurs, l'objectif de cette méthode est de déterminer la conformation minimisant l'énergie. Ainsi sur un ensemble de conformations, il est possible

d'avoir plusieurs conformations minimisant l'énergie. À noter que chaque conformation est représenté par le modèle proposé par Dill [2] permettant de discriminer les résidus polaires (P) et hydrophobes (H), et de calculer l'énergie comme le nombre de résidus hydrophobes voisins non conjoints. Ainsi si deux résidus non voisins conjoints sont hydrophobes alors l'énergie baisse de -1.

Notre algorithme REMC est un programme interprété sur la version la plus récente de Python à savoir Python 3 et exploitable sur tout type de système d'exploitation. Le programme charge une séquence disponible sur un fichier FASTA correspondant à la structure de protéines que nous souhaitons prédire et va permettre de visualiser en 2D ou en semi-3D la conformation de la séquence. Le module python *matplotlib* est nécessaire à la visualisation tout comme le module *tqdm* ayant un peu moins d'importance. Un certain nombre d'options sont disponibles pour spécifier le type de sortie du programme souhaité.

3 Résultats

Nous présentons ici différents résultats de notre algorithme. Dans un premier temps nous avons cherché à visualiser la structure 2D de la protéine p01013 notamment les régions hydrophobes et non hydrophobes qui théoriquement devraient être enfouies à l'intérieur et à l'extérieur de la protéine respectivement.

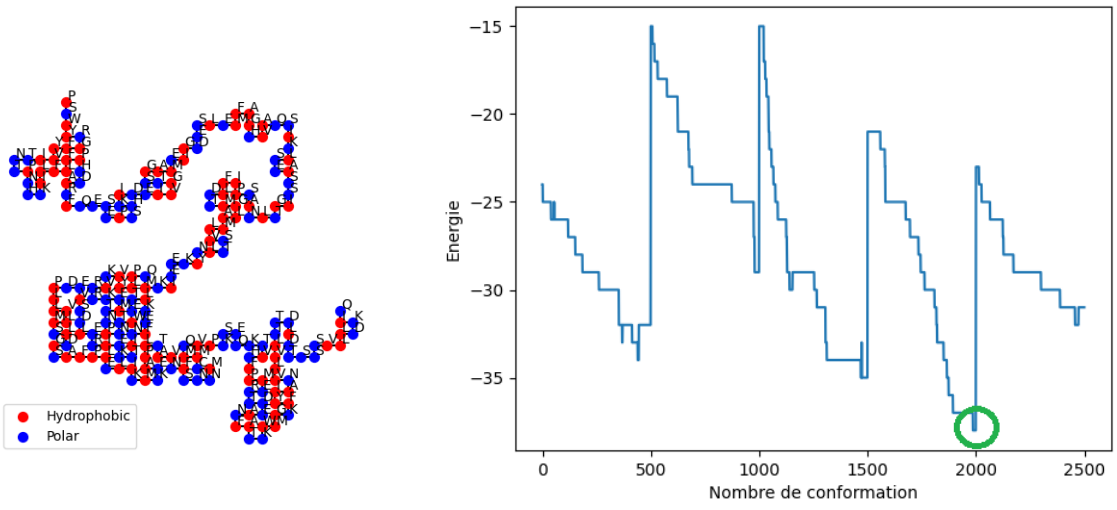

Figure 1 : Représentation de la structure 2D de la p01013 et de la fonction d'énergie obtenue après simulation

Nous avons fourni comme paramètres les valeurs suivantes : énergie optimale = -30, un nombre d'itérations

= 500, une température minimum et maximum à 220 et 250 respectivement et un nombre de répliquats = 5. D'après la figure ci-dessus, dans l'ensemble des conformations générées, un minimum global a pu être détecté à savoir la 2000ème conformation générée avec une énergie d'environ -40, c'est ce qui est représenté en vert.

4 Conclusion

Pour conclure, nous avons développé l'algorithme REMC permettant de prédire la structure 2D d'une protéine donnée. Les résultats présentés sont encourageants et montrent que notre algorithme est capable d'identifier une conformation dont le niveau d'énergie est minimisé. Des optimisations peuvent être apportées pour permettre d'effectuer des calculs sur des séquences potentiellement plus longues puisque l'algorithme effectue des calculs sur un temps polynomiales.

Références

[1] Thachuk C, Shmygelska A, Hoos HH. A replica exchange Monte Carlo algorithm for protein folding in the HP model. BMC Bioinformatics. 2007 Sep17 ;8 :342. PubMed PMID : 17875212 ; PubMed Central PMCID : PMC2071922.

[2] Dill KA. Theory for the folding and stability of globular proteins. Biochemistry. 1985 Mar 12 ;24(6) :1501-9. doi : 10.1021/bi00327a032. PMID : 3986190.

[3] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https ://doi.org/10.1038/s41586-021-03819-2

[4]] Tunyasuvunakool, K., Adler, J., Wu, Z. et al. Highly accurate protein structure prediction for the human proteome. Nature 596, 590–596 (2021). https ://doi.org/10.1038/s41586-021-03828-1