Compte rendu de l'avancement des travaux - 3e année de thèse

David Mignon

29 octobre 2015

Sujet de thèse : Computational protein design (CPD) : un outil pour l'ingénierie des protéines et la biologie synthétique.

Le CPD est en développement au sein de notre laboratoire depuis déjà quelques années, avec plusieurs succès à son actif. Ce sont ces résultats prometteurs qui fondent notre motivation à aller de l'avant en améliorant nos outils, en enrichissant nos programmes pour progresser encore dans nos résultats. En particulier, mon travail se concentre sur le problème de l'exploration de l'espace des séquences d'acides aminés et son contrôle.

Cette année, une série d'évaluations de la qualité de nos séquences calculées a été réalisée, ainsi qu'une étude sur le voisinage du minimum global de notre fonction d'énergie pour quelques-uns des systèmes déjà étudiés.

Pour évaluer nos résultats, nous avons utilisé le programme toulbar2 d'une équipe de l'Université de Toulouse (UMR792) qui propose un algorithme de recherche de type Dead End Elimination (DEE). Cet algorithme est dit «exact» , c'est-à-dire qu'il arrive au minimum global, lorsqu'il se termine, en éliminant successivement les configurations ne pouvant pas appartenir à ce minimum. Cette méthode fonctionne bien si l'espace de recherche n'est pas trop grand. En deuxième année,nous avions défini un ensemble de près de mille de systèmes biologiques en fixant plusieurs acides aminés dans neuf protéines de trois domaines différents,afin de travailler avec des espaces de recherche de petite taille. Les tests montrent que :

Le programme toulbar2 aboutit presque toujours (dans le temps imparti de 24h) pour les systèmes avec au plus dix positions actives (positions dans la séquence où l'acide aminé n'est pas fixé). L'heuristique de proteus trouve quasiment systématiquement le minimum global lorsque nous le connaissons et donne presque toujours le meilleur résultat pour les systèmes avec au plus trente positions actives. Les méthodes stochastiques sont presque toujours à moins d'une kcal/mol des meilleurs, et ne sont significativement moins bonnes que sur un petit nombre de cas.

Pour les tests avec toutes les positions actives, c'est une méthode stochastique qui domine le Monte-Carlo avec échange de répliques configuré avec huit marcheurs est des températures comprises entre 3 et à 0,1.

Une étude détaillée de la topologie au voisinage du minimum global de notre fonction d'énergie a été effectuée pour quelques-uns des systèmes ou les méthodes donnent les résultats différents.

D'autres aspects que l'énergie sont intéressants pour juger de la qualité des séquences. L'un des plus importants est la similarité avec les séquences naturelles ayant une structure 3D proche.Pour cela nous avons utilisé la classification de structure de protéine SCOP, couplé avec la base de données de modèles de Markov cachés Superfamily qui fait correspondre un modèle à chaque domaine SCOP.Nos résultats sont bons, avec pour huit de nos neuf protéines, 100% des meilleures séquences reconnues comme structure du domaine SCOP de notre séquence de départ.

L'identité des séquences est également utilisée avec les matrices de coûts de substitution d'acide aminé BLOSUM et la base de données de Pfam qui propose des alignements multiples de familles de séquences naturelles.Ensuite, des comparaisons d'entropie par position, pour mesurer la diversité de nos séquences, ont été effectuées.

Comme nos algorithmes Monte-Carlo produisent en théorie des ensembles de séquences qui suivent une distribution de Poisson-Botzman, nous avons examiné les densités d'états produites par notre programme.

Un article qui résume les résultats de ces deux dernières années est en cours de rédaction.

Plusieurs améliorations ont été apportées au programme proteus. Dans le Monte-Carlo,le critère d'acceptation de Hasting a été perfectionné pour prendre au compte les différences de nombre de positionnement de la chaîne latérale,selon l'acide aminé.Un nouveau mode de sélection des séquences à imprimer, basé sur la meilleure énergie courante, a été introduit. Il existe maintenant, un système de probabilité de sélection qui permet une pondération à chaque position du taux de mutations et de changements de rotamères. Enfin,le programme propose maintenant le nombre d'occurrence d'un même état au cours de la trajectoire.