Compte rendu de l'avancement des travaux - 2ème année de thèse

David Mignon

24 octobre 2014

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 aminées et son contrôle.

Cette année un système d'exécution parallèle a été ajouté à notre programme de recherche de séquences (proteus). Il s'agit d'une parallélisation de type mémoire partagée effectue avec la bibliothèque openMP.

En plus d'une meilleure utilisation de la mémoire des ordinateurs, cela nous a permis d'ajouter un autre algorithme de recherche le Monte Carlo avec échange de répliques (MCER). Dans cet algorithme plusieurs marches Monte Carlo à température différentes s'exécutent en parallèle. De temps en temps deux températures consécutives sont échangées. Cette évolution du Monte Carlo conserve les propriétés importantes : Les échantillons produits tendent vers la distribution souhaitée (ici la distribution de Boltzmann) et respectent la condition de "balance détaillée". De plus, le MCER , avec ses températures les plus hautes, balaie mieux l'ensemble de l'espace de recherche. Et donc, il travaille pendant les périodes de recherche à basses températures sur un ensemble de régions plus variées.

Une série de comparaisons entre les algorithmes de proteus ont été effectuées. Ces comparaisons portent sur trois familles de protéines (SH2,SH3 et PDZ) avec dans chaque famille deux ou trois représentants dont la taille varie entre 57 et 109 acides aminés. Le jeu de données constitué a également été utilisé pour optimiser certain paramètre de MCER : distribution des températures, nombre de répliques, fréquences des échanges, etc. Les résultats montrent que le MCER avec un protocole de huit marcheurs et des températures comprises entre 3 et 0,2 est systématiquement meilleur en terme d'énergie ou équivalent que tous les autres protocoles testés (le MCER huit marcheurs avec des températures plus écartées, le MCER avec quatre marcheurs, le Monte-Carlo, l'optimisation itérative sur chaque position).

Toujours pour évaluer proteus, 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 trouve le minimum global 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. Nous avons alors fixé plusieurs acides aminés de nos protéines afin de diminuer la taille de l'espace de recherche. Si on laisse libre entre dix et vingts acides aminés selon les protéines alors toulbar2 donne un résultat en moins de 24 heures de calculs sans demander plus de 8 Go de mémoire vive. Des comparaisons sont en cours pour des espaces allant de zéro à vingts acides aminés libres. Les premiers résultats montrent que proteus trouve systématiquement le minimum global.

Un objectif important de notre projet est le passable d'un modèle à squelette de protéine fixe vers un modèle multi-squelette. Pour cet objectif, plusieurs étapes logicielles ont été franchies : Une nouvelle version de proteus a été écrite qui fusionne le travail de Karen Druart (doctorante dans l'équipe) sur le modèle multi-squelette avec le MCER. Notre programme de calcul d'énergie (Xplor) a été enrichi d'un système extensible de gestions des coordonnées des atomes. Les scripts ont été rendus plus modulaire en séparant des phases de préparation du modèle avec des phases de calcul de l'énergie. Ces extensions et cette modularité accrue vont permettre d'aborder la situation nettement plus complexe où le squelette de la protéine devient mobile.

Maintenant, il faut poursuivre la mise au point du proteus multi-squelette. Une protéine PDZ avec deux squelettes est à l'étude. L'expérience avec le toulbar2 nous conduit à penser qu'il serait possible d'assouplir les critères de sélections du DEE pour obtenir un nouvel algorithme prometteur. Pour l'étape suivante du squelette flexible, une collaboration avec "la maison de la simulation" au CEA et l'IDRIS débutée cette année permettra une parallélisation encore plus poussée du logiciel et donc le passage de nos programmes sur les super-ordinateurs.