Projet INF442-3 **
Critical Assessment of protein Structure Prediction
Analyse de données biologiques

Sujet proposé par Amélie Héliou
amelie.heliou@polytechnique.edu

X2014

Introduction

CASP (Critical Assessment of protein Structure Prediction) est une expérience à l'échelle mondiale qui propose d'évaluer des résultats de prédiction de la structure et des interactions de protéines.

Obtenir de bons résultats à une expérience de CASP est un critère essentiel à la validation d'une méthode de prédiction.

Nous allons ici nous intéresser aux résultats de la 10ème expérience de CASP.

Entre avril et juillet 2012, des structures sur le point d'être résolues par des méthodes expérimentales ont été identifiées. Leurs séquences ont été rendues publiques alors que leurs structures 3D étaient encore conservées secrètes.

Les équipes travaillant à la mise en place de méthodes de prédiction ont alors pu travailler sur ces séquences. Les équipes prédisent des structures et les soumettent.

Chaque prédiction soumise a ensuite été comparée aux résultats expérimentaux pour être évaluée.

But

Votre objectif est d'analyser à votre tour les prédictions pour extraire les meilleures et comparer avec les résultats de l'expérience. Dans une seconde étape vous pourrez mettre en pratique un ou des algorithmes de regroupement vu en cours pour comparer les prédictions entre elles.

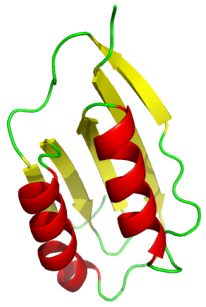

FIGURE 1 – Domaine 1 de la cible T0651

1 Lecture des fichiers de prédiction

Nous allons nous intéresser plus particulièrement au domaine 1 de la cible T0651 qui a eu 451 prédictions venant de 116 groupes.

Les fichiers sont disponibles ici :
`http://www.enseignement.polytechnique.fr/profs/informatique/Amelie.Heliou/projetINF442-3/index.html`

Les fichiers de prédictions sont au format pdb pour les atomes. `http://deposit.rcsb.org/adit/docs/pdb_atom_format.html`

Le domaine 1 correspond à tous les résidus de 1 à 95.

Pour extraire les informations d'un fichier texte vous pouvez lire le fichier ligne à ligne. Puis extraire des lignes commençant par « ATOM » les informations qui vous intéressent (identifiant, résidu, coordonnées, etc.).

1) Ecrire un programme qui récupère les coordonnées d'un certain type d'atomes dans un fichier (par exemple les carbones alpha « CA »).

2 Calcul d'une distance entre deux structures

Nous allons utiliser un critère pour trouver les dix meilleures prédictions selon ce critère.

Il existe de nombreux critères pour estimer la précision d'une prédiction.

Ici nous nous intéresserons à la dRMSD. La dRMSD est un score qui compare les distances entre paires d'atomes.

Pour gagner du temps nous ne considérons que le carbone alpha (CA dans le fichier pdb) de chaque résidu.

$$dRMSD(X^A, X^B) = \frac{1}{N \times N - 1} \sum_{i \neq j} (d(x_i^A, x_j^A) - d(x_i^B, x_j^B))^2$$

2) Ecrire un programme qui calcule la dRMSD entre deux structures.

3 Utilisation du calcul parallèle

Répartissez les 451 prédictions équitablement sur tous les processeurs.

Chaque processeur calcule la dRMSD entre chacune de ses prédictions et la structure cible (target.pdb) puis classe les prédictions et envoie les dix premières au maitre.

Dès que le maitre reçoit dix nouvelles prédictions, il réuni la liste de 10 meilleures prédictions qu'il a avec celle qu'il vient de recevoir et conserve les dix meilleures.

3) Ecrire le programme MPI qui affiche les identifiants des 10 meilleures prédictions et leurs dRMSD avec la cible.

4 Clustering

Il s'agit à présent de calculer la dRMSD entre tous les couples de prédictions, afin de pouvoir utiliser un algorithme de clustering sur ces données. Les dRMSDs peuvent être calculées en parallèle : chaque processeur calcule la dRMSD d'une fraction des structures contre tous les autres. Vous pouvez ensuite utiliser un algorithme de clustering pour regrouper les prédictions (par exemple k-means ou un algorithme hiérarchique).

4) Ecrire un programme MPI qui regroupe les prédictions en fonction de leur distance.

Regardez ensuite, par exemple, si les meilleures prédictions sont bien dans un même cluster et si les prédictions d'une même équipe sont regroupées ou dispersées.

5) Tracer/afficher les identifiants des équipes présentes dans chaque cluster ainsi que les clusters des 10 meilleures prédictions.