Daniel DE MURAT - Madeleine DE SOUSA-VIOLANTE - Mei-Shiue KUO - Sonia TIEO
Equipe 2 - Partie amont - Université Paris Diderot (M2BI)

Introduction:

Le thème du concours inter-université Meet-U de 2019 est l'analyse d'une séquence protéique et de la prédiction de sa structure 3D la plus stable/probable. Cette édition se présente en deux étapes: (i) une annotation de domaine basée sur une comparaison profil-profil (partie amont) et (ii) une identification du "fold" 3D le plus stable par "threading" (partie aval). Notre équipe travaille sur la partie amont en collaboration avec l'équipe aval n°4.

Stratégie:

Tout d'abord nous avons recherché les séquences homologues à une séquence requête avec PSI-BLAST 2.2.31 contre la base de données UniRef90, construit en regroupant des séquences de UniRef100 aux niveaux d'identité de séquence de 90%. Un alignement multiple des séquences homologues et de la requête est réalisé avec MAFFT, utilisé pour sa rapidité à aligner de grandes séquences en conservant une bonne qualité. Ensuite, un profil a été construit pour les séquences homologues de la requête et parallèlement pour les alignements multiples de séquences de chacune des 405 familles de HOMSTRAD. Ce profil comprend une matrice PSSM (position-specific scoring matrix) auquel est concaténé la fréquence des brèches et la structure secondaire produite par PSIPRED. Nous avons affiné en ajoutant un pseudo-count pour chaque acide aminé (1/20) et une pondération pour favoriser les séquences rares. Le profil de la requête est aligné par la suite avec chacun des 405 profils selon un alignement semi-global. Un score d'alignement (fonction Dot Product ou coefficient de corrélation de Pearson) est ainsi attribué à chaque alignement.

Résultats:

Tout d'abord, notre programme a été lancé sur les séquences requêtes tests (21) , afin de pouvoir comparer les méthodes de scoring (Dot Product et corrélation de Pearson) et les bases de données(UniRef50 et UniRef90) et en déduire la méthode la plus appropriée. Globalement, les familles des séquences requêtes sont retrouvés à des meilleurs rangs en utilisant le Dot Product, il est donc choisi comme fonction de scoring. Concernant les bases de données, bien que les résultats se sont révélées similaires pour les deux bases de données, UniRef90 a été choisie car permettait d'obtenir plus de séquences alignées avec la séquence requête pour la partie aval.

En analysant plus en détail nos résultats obtenus avec le Dot Product et UniRef90, 21 familles des séquences tests étudiées ont été retrouvées aux classements variés. Nous avons remarqué une bonne prédiction pour des requêtes ayant de longs alignements qui partage un pourcentage élevé d'identité relatif avec le représentant de sa famille HOMSTRAD. La partie aval a été ensuite exécutée à partir des alignements obtenus pour comparer et choisir la meilleure méthode de scoring. C'est le sum-scores qui a prédit les familles avec un meilleur classement, elle a donc été choisie.

Enfin, le programme entier (notre partie et celle du groupe 4) a été lancé sur les séquences de nos collègues de biologie santé. Parmi les familles prédites, COX1, COX3 et cytochrome b (tous des composants de la chaîne respiratoire mitochondriale) sont souvent retrouvés. Cependant, il est difficile de conclure la pertinence de nos résultats sans avoir plus de contexte biologique de leur expérience.