Karl-Stephan Baczkowski
Martin Davy
Alice Rogier

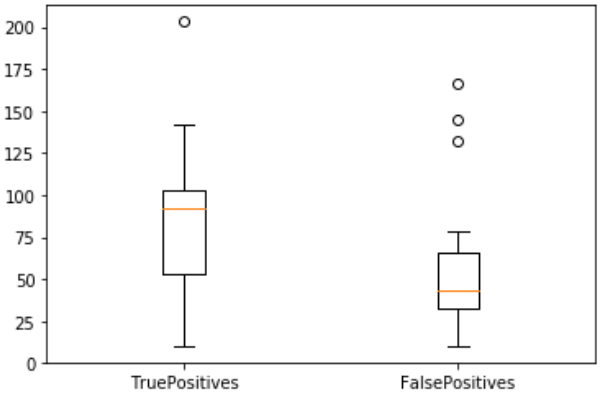
Notre équipe a pour objectif d'élaborer une pipeline permettant de répondre à la partie amont du projet Meet-U . Le but de la partie amont est d'identifier, à partir d'une séquence query donnée dont la structure est inconnue, ses familles de protéines les plus proches parmi les familles de la banque Homstrad, par homologie décroissante. À cette pipeline est incluse la partie avale du projet, programme qui lui a pour objectif de prédire la structure tertiaire de la séquence query à partir des templates trouvés par le programme de la partie amont, par des méthodes de threading. Nous avons choisi d'ajouter le programme de l'équipe 11, car son approche «biophysique» est originale et donne des résultats prometteurs.

Un psi-blast contre la banque Uniclust30 est réalisé. Un alignement multiple des meilleurs hits trouvés est effectué avec Muscle. Pour sélectionner les meilleurs hits, la e-value peut être modifiée. Si spécifié, les séquences sont ensuite pondérées. Puis, une PSSM (7 ou 21 lettres) est calculée. Trois à quatre lignes prenant en compte les probabilités de structures secondaires et d'accessibilité au solvant peuvent lui être ajoutées. On peut également choisir de retirer les gaps de la séquence query. Un alignement du profil de la séquence query avec les profils des familles Homstrad est réalisé. Trois méthodes de scoring ont été implémentées pour classer les familles par ordre d'homologie décroissante (corrélation de Pearson, dot product et dot product normalisé). Les résultats sont donnés sous la forme d'un fichier .foldrec et d'un fichier .rank contenant le classement des familles.

Pour évaluer la qualité des résultats, des courbes d'enrichissement ont été réalisées. Le programme a été évalué en plusieurs étapes. Les méthodes de scoring ont d'abord été testées. Aucune ne s'est avérée significativement meilleure, mais la corrélation de Pearson a été écartée car elle demandait plus de temps. L'apport des structures secondaires et de l'accessibilité au solvant à la PSSM a ensuite été testé avec JPred. Ils augmentent significativement la qualité de nos résultats.

La méthode de scoring dot product permet de discriminer efficacement les séquences HOMSTRAD correctement classés (cf. **boxplot**).

Boxplot : Comparaison des scores des séquences HOMSTRAD correctement classées des autres séquences pour DotPFreq, sans gaps, prédiction Jpred + AS, pondération, gap=0,5.



Puis la PSSM à 7 lignes a été testée seulement avec les meilleurs paramètres trouvés pour la PSSM 21 lettres. Elle a montré une qualité de résultats comparable à ceux obtenus avec la PSSM 21 lettres, et demande moins de temps de calcul.

Enfin, le programme de l'équipe 11 a été choisi pour la partie avale du projet. Les sorties foldrec obtenues avec le meilleur jeu de paramètres trouvé pour notre pipeline ont été mis en entrée du programme de l'équipe 11. Cependant, les résultats obtenus sont moins bons que lorsqu'on utilise notre pipeline seule.

Equipe 10 (Amont)

Karl-Stephan Baczkowski
Martin Davy
Alice Rogier