Meet-U 2018-2019: Downstream

Team n°8 with students of Master BIM at Sorbonne University :

- Antoine Gagelin (M2 - BIM)
- André Lanrezac (M2 - BIM)
- Gaspar Roy (M2 - BIM)
- Florian Specque (M2 - BIM)
- Yvan Sraka (M2 - Info)

Abstract

As a downstream team, our goal was to predict the most probable templates from a *Multiple Sequence Alignment* (*MSA*) matrix, built around a query sequence. This file was output by the upstream teams. To fulfill our objective, we threaded the query sequence on all the templates in order to compute a series of scores on the various threads (*Hydrophobic score*, *DOPE score*). They were used to estimate the similarity between the query sequence and the obtained 3D structure. Finally, the different scores were combined into a single one, using a model of linear regression, learnt on a big dataset. This last score was used to rank the templates, allowing us to assess the efficiency of our method.

Flowchart

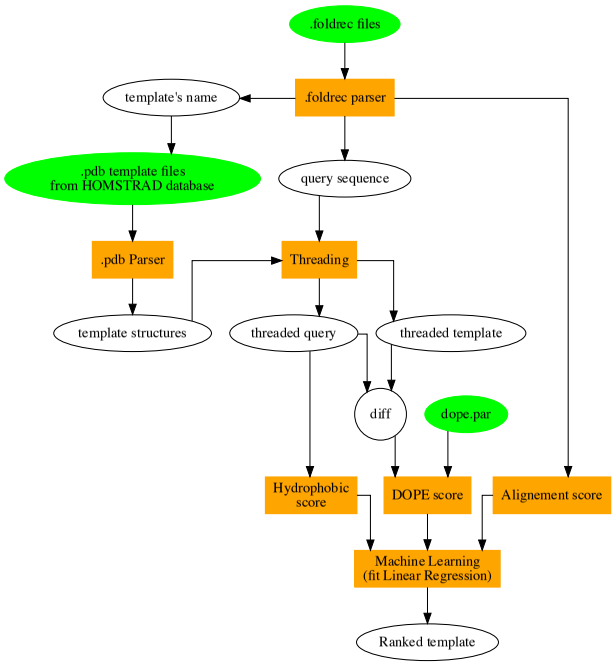

Figure 1: FlowChart

DOPE Score

- Mean of statistical potentials between pairs of residues (*Alpha Carbons*).
- To make DOPE scores of different threading's comparable: subtraction to get the increase of the score of the template and the thread. The lower is the difference, the better.

HP Score

- *HP* stands for *Hydrophobic-Polar score*
- Simple score that tends to capture the number of hydrophobic contacts. All pairs of residues that are separated by a distance bellow a threshold increase the score with 1 if they are both hydrophobic. So, it's a very simple score that we chose because we know that the hydrophobic contacts are the driving force the protein folding.

Alignment Score

- Recovered from the upstream team (from the `.foldrec` file), score of similarity between the query sequence and the template.

Linear regression model fitting with Machine Learning

- Used a dataset of 1000 `.foldrec` files.
- Computed the 3 scores for all the templates in all the `.foldrec`, and the *TM score* between the native template structure and the threaded one.
- Trained a model of Linear Regression using the *TM score* as the target variable.

Results

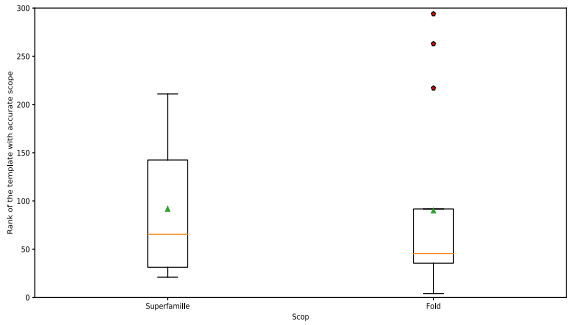

Figure 2: Boxplot des rangs des folds et superfamilles correctes pour les foldrec du benchmark prédits avec la régression linéaire.