Team 1: Rebecca Clodion - Inka Leroy - Saida Hajjou - Samantha Samson
Upstream team

SALUT : Prediction of protein structure

Context. For this Meet-U edition, the goal for bioinformatician teams was to be able to predict the fold of a protein starting from its sequence. To do this, the work was divided in 2 parts. We chose the first part that consisted on building profile-profile alignments with scores based on the HOMSTRAD database.

Strategy. With the aim of completing successfully that project, we wanted to obtain the most similar foldrec file that we can for each query to compare to the one hold up as an example. We did that this way in order to facilitate the reading by the downstream teams which have used the foldrec file example to build their program.

To obtain the foldrec file we proceeded by several steps. First, we have run a PSI-BLAST with the provided fasta sequences against the uniref50 database. The results of the PSI-BLAST for the round chosen were used to do a multiple alignment with the software MUSCLE. Then we computed a program that create a position-specific scoring matrix (PSSM). To do that, the strategy was to use background frequencies derived from BLOSUM62 matrix to calculate the frequencies of amino acids at each position with the HENIKOFF & HENIKOFF method which allow to assign frequencies depending on the weight of the sequences (their information degree). We used the same method for the HOMSTRAD database from alignments that were provided and then we compared each PSSM from the database to the query PSSM. Finally, we used the affine gap penalty method with the algorithm of Needleman & Wunsch to generate the profile-profile alignment and information about it. As some downstream teams needed the secondary structure, we did the assignation with DSSP for the templates and the prediction with PSIPRED for the queries. Once we handled this part, we choose the team Fold U to predict the three-dimensional structure.

Results. Most of the hits (correspondence between the structural classification of a query and a template) for the fold were concentrated in the first ranks. We tested different parameters (database, number of rounds, background frequencies) to see which one return the best accuracy in terms of class and fold. To check our results, we plot a graph representing the hits accumulated sum depending on the rank of the templates. We noticed that neither the type of background frequencies (from BLOSUM62 or equi-probable) nor the database for the PSI-BLAST (uniref50 or uniref90), show a significant difference about the number of hits by rank. Also, adding a round does not change the results, however, we noticed that using 2 rounds instead of 3 reduces considerably the time for the program to run for a big sequence.