Meet-U
a meeting story
This is a collective adventure...
edition 2019
Meet-US

Meet-U - Team 6 - AMONT

Protein structure prediction

Bénédicte Colnet, Gabriela Lobinska, Irène Mauricette Mendy, Yasser Mohseni Behbahani, and Amandine Sandri

Abstract

In order to infer a 3D-structure of a protein from homology modeling, also known as comparative modeling, we provide the following methodology. A profile-profile comparison, instead of sequence-to-profile comparison, is used to mine databases and find an appropriate structural template [1]. With an incremental approach we succeeded to distinguish two comparison methods that classify superfamily and fold. Therefore our original result is to propose a classifier mixture to best fit the benchmark provided.

Methods

On its core, the intrinsic problematic faced in this upstream study is:

- **Part A** To gather homologous sequences
- **Part B** To build profiles from several aligned sequences
- **Part C** To compare two profiles and to score their similarity

The pipeline build is presented figure 1.

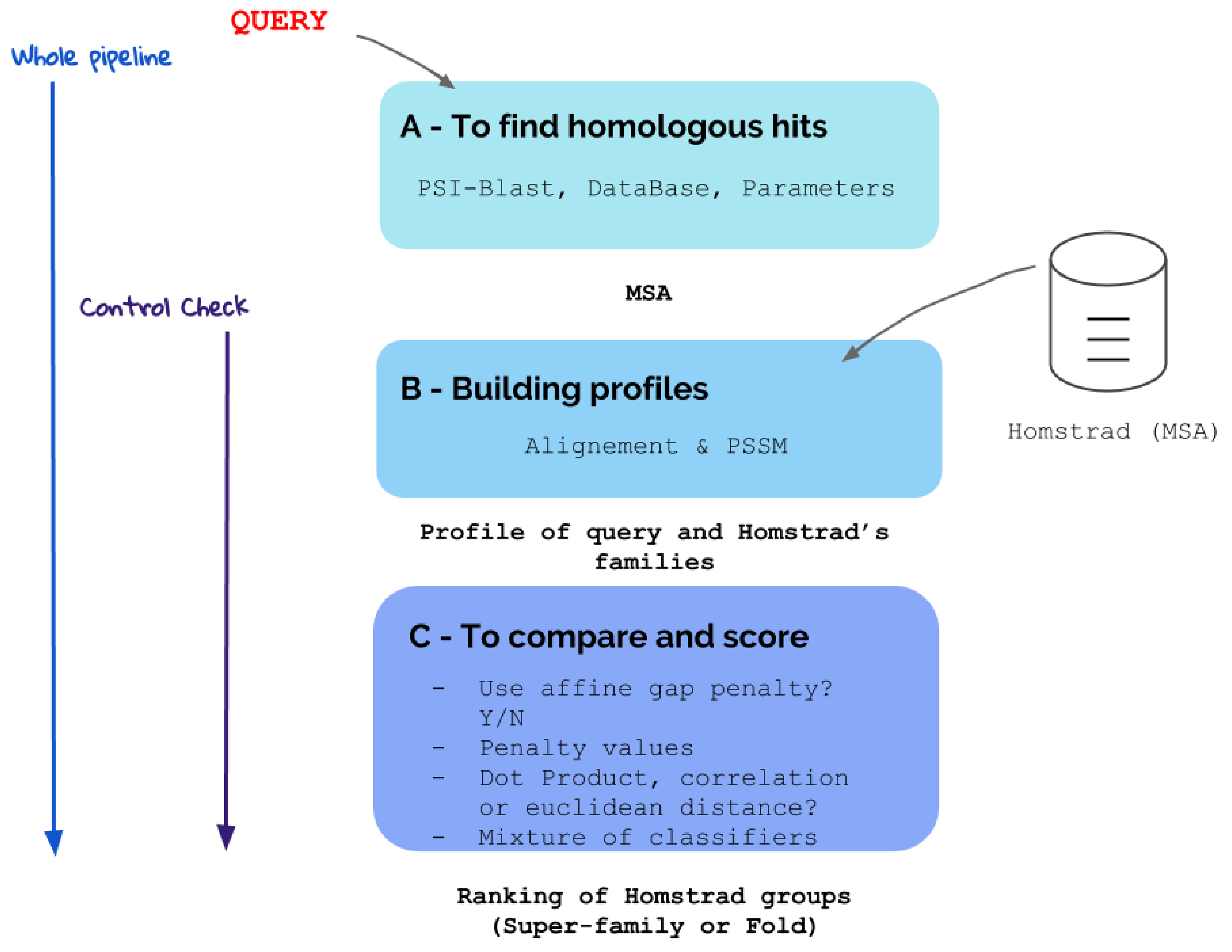
QUERY
Whole pipeline
Control Check
A - To find homologous hits
PSI-Blast, DataBase, Parameters
MSA
B - Building profiles
Alignement & PSSM
Homstrad (MSA)
Profile of query and Homstrad's families
C - To compare and score
- Use affine gap penalty? Y/N
- Penalty values
- Dot Product, correlation or euclidean distance?
- Mixture of classifiers
Ranking of Homstrad groups (Super-family or Fold)

Figure 1: Schematic of the whole pipeline

All the results come from the 21 sequences proposed in the benchmarklist. The results are presented either as an enrichment curve with one single good result - considering or not the specificity of the fold or superfamily - or as a dispersion plot showing the rank for each query. This last analysis allows to distinguish whether a method has an effect on the spreading of scores (therefore on the confidence in results), even if the rank does not change a lot.

Parameters

The parameters used for the gap penalty opening and extension are presented Table 1, these parameters were derived from bibliography relative to the subject [2] and had to be changed to find better results on the benchmark.

Parameters	Value	Bibliography value [2]
Gap opening penalty Dot Product	12	0.07
Gap extension penalty Dot Product	1	0.005
Zero Shift Dot Product	-0.03	-0.05
Gap opening penalty Correlation	5	1.39
Gap extension penalty Correlation	0.5	0.07
Zero Shift Correlation	-0.2	-0.21

Table 1: Optimized parameters for scoring functions

Results

Affine Gap Penalty Experiment with and without affine gap penalties on the benchmark were launched using the dot product comparison[a], and the enrichment curve was recorded (figure 2 (a)). As expected, the enrichment curve is better with the affine gap penalty. Moreover, the affine gap penalty increases the spread in scores which is promising for scoring protein and giving an estimation of the certainty and confidence we have on the score. The figure 2 (b) highlights the spread.

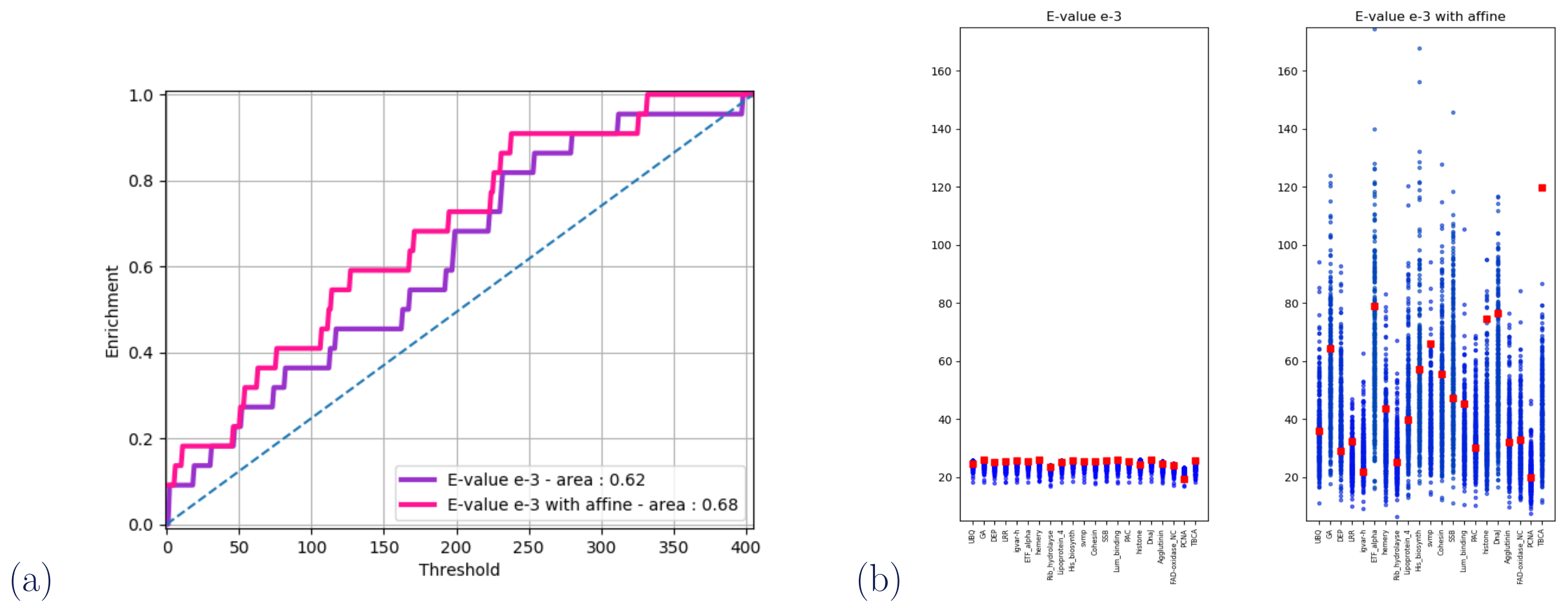

Figure 2: (a) The enrichment curves for two tests: with and without affine gap penalties, all other parameters remaining equal. (b) Visualization of the distribution of scores over HOMSTRAD for the algorithm with and without affine gap penalties.

Correlation, Dot product, & Classifier mixture The correlation and dot product comparison methods are compared on the benchmark, and the enrichment curve is presented in figure 3 (b). We observe that the correlation is a most powerful tool to distinguish conserved positions on profiles, therefore this method is more performing on superfamily recognition as it is linked to sequence homology. Then we proposed, from two methods to build a mixture[b] As expected the mixture has the best performances as showed on figure 3 (b). We also notice that, we can estimate the maximum performance using control check. It reveals the method to be sensitive to the MSA used.

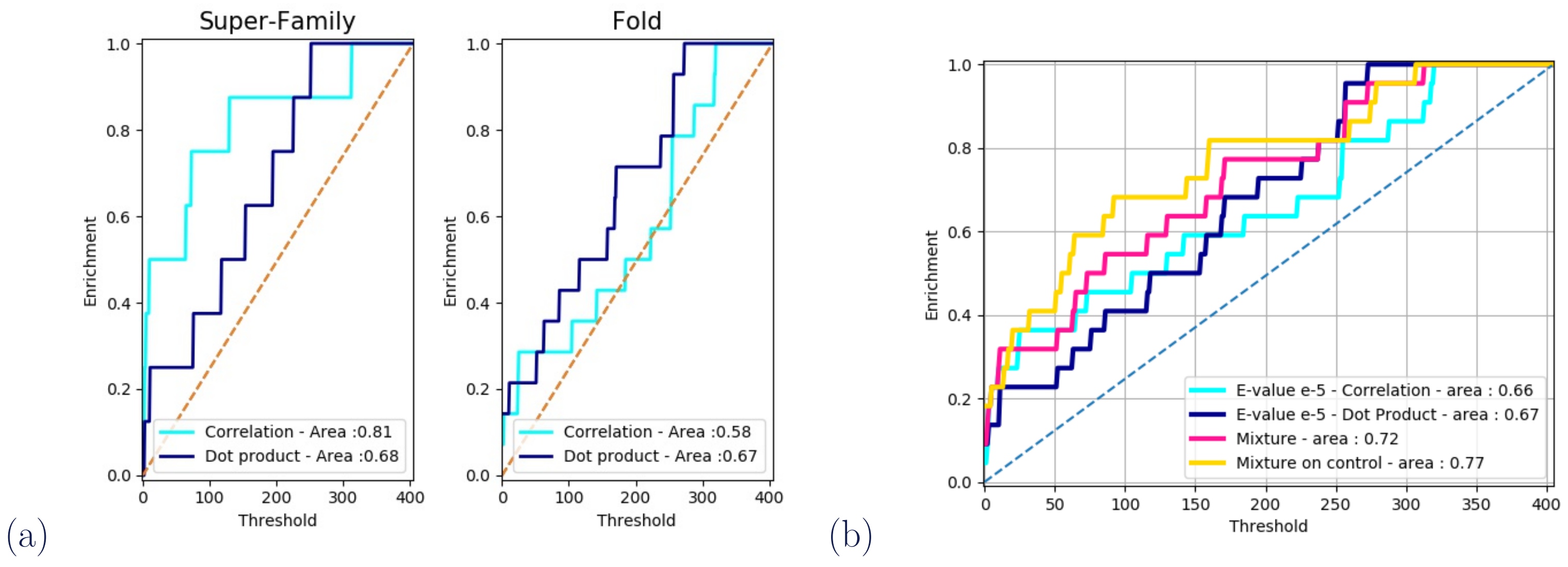

Figure 3: (a) Enrichment curves with correlation and dot product comparison methods, all other parameters remaining equal (e-value equals $1e-5$). (b) A mixture of the two classifiers.

Conclusion

The profile construction efficiency is assessed using a *Control Check* procedure. It reveals the profiles to be of good quality, but improvements are still required. Beyond the profile construction, our pipeline underscores the intrinsic difference when ranking a superfamily or a fold, which leads us to the idea of a mixture to be the most efficient classifier.

[a] This procedure was repeated for several e-values with no significant effect on the results, therefore only the e-value $e-3$ is presented in this figure.

[b] Using either dot product comparison and correlation, for e.g. when the best result in benchmark is supposed to be a superfamily then the output of correlation is taken.

Mentoring

LCQB

This project was realized with the help of Élodie Laine, Juliana Bernardes, and Hugues Richard.

References

[1] Y. Ghouzam, G. Postic, P. Guerin, A. de Brevern, and J. Gelly. Orion: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. *Sci Rep.*, 2016.

[2] G. Wang and R. Dunbrack. Scoring profile-to-profile sequence alignments. *Protein science : a publication of the Protein Society*, 13:1612–26, 07 2004. doi: 10.1110/ps.03601504.