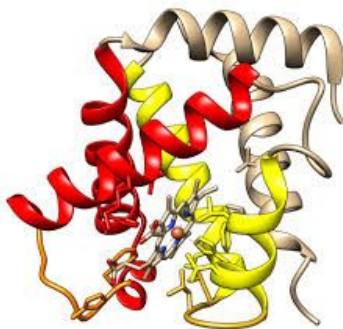




SORBONNE
UNIVERSITÉ

Final Presentation

January 2019



MEET-U

Upstream - Team 6

Yasser **Mohseni Behbahani**

Bénédicte **Colnet**

Gabriela **Lobinska**

Irène Mauricette **Mendy**

Amandine **Sandri**



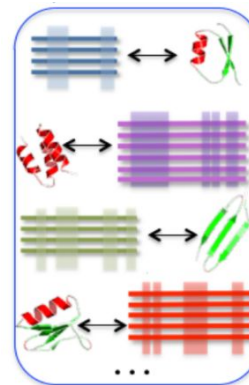
Goal

Find best templates for a target protein.

Expected output

Profile-Profile Alignment file and Scoring

Data supplied



HOMSTRAD Database (reduced to 405)

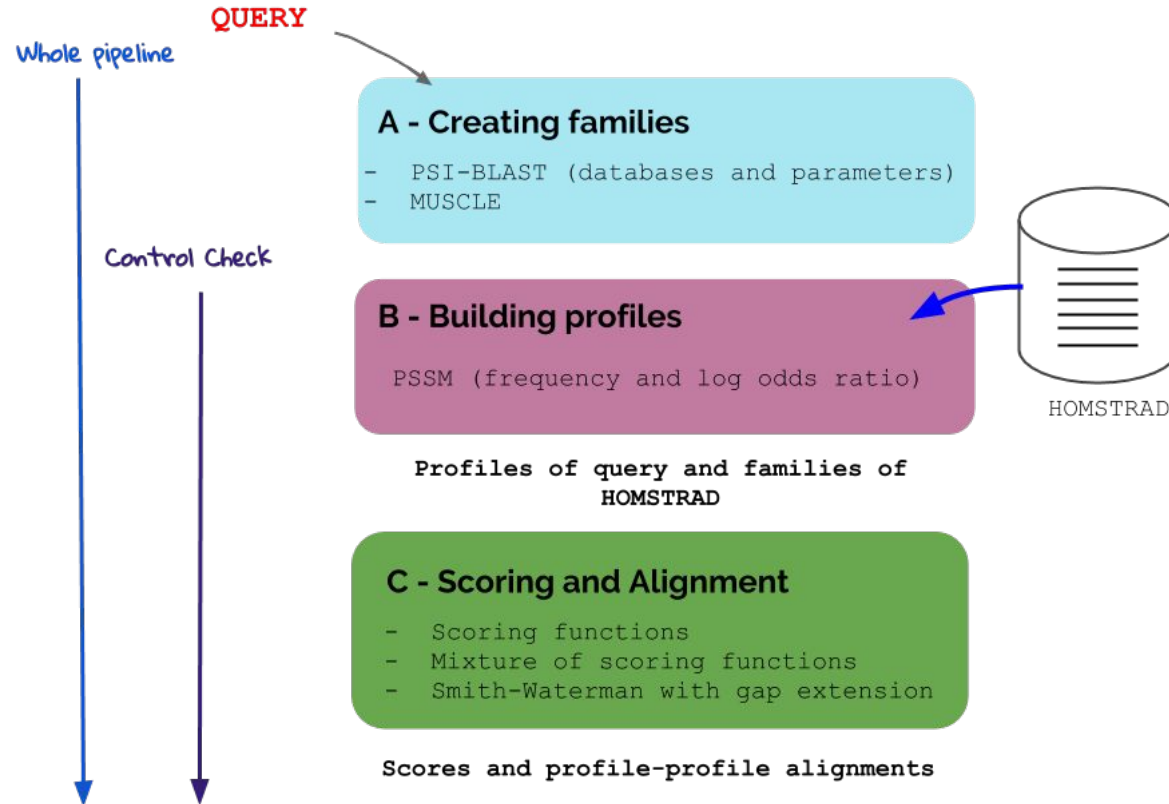
+



Benchmark

Hemery → SRP54 (Fold)
Agglutinin → intb (Fold)
Lipoprotein_4 → mofe (Superfamily)
...

21 queries → Best hit HOMSTRAD



Goal

Assess the quality of the scoring

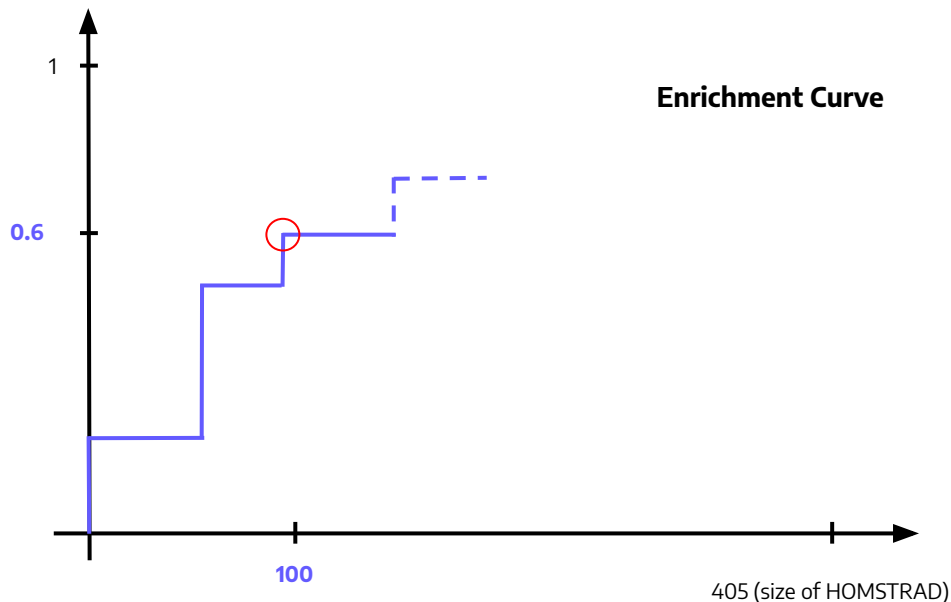


Benchmark

21 queries → Best hit HOMSTRAD

For each query
Launch the pipeline
Register the ranking

Plot the enrichment curve



Meaning : In the top 100 (25%) of the HOMSTRAD database, 60% of the hits are found

Goal

Assess the quality of the scoring

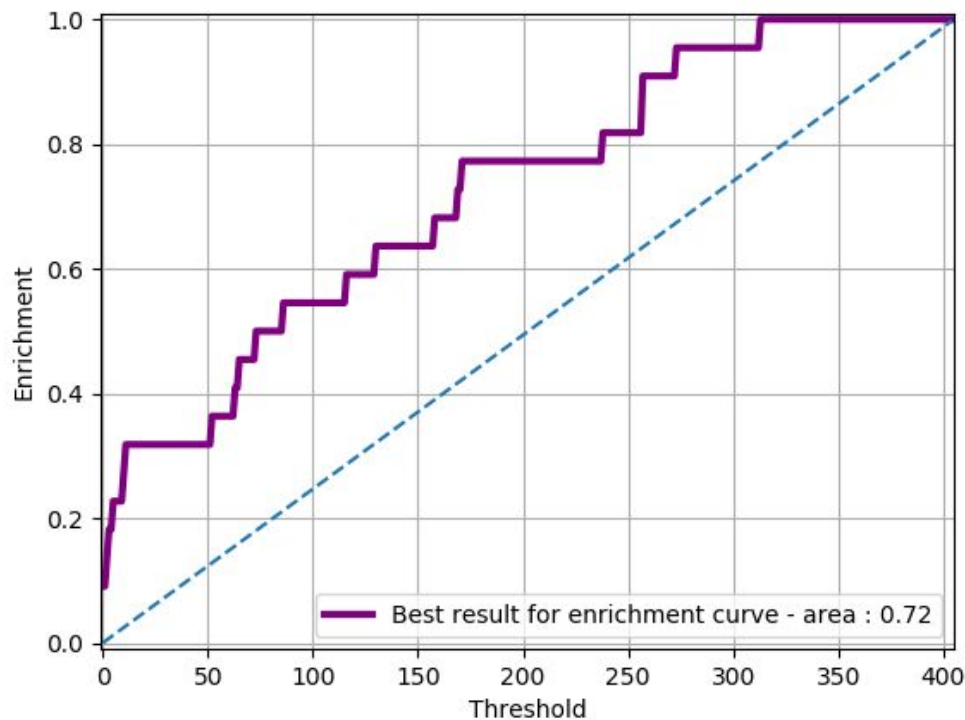


Benchmark

21 queries → Best hit HOMSTRAD

For each query
 Launch the pipeline
 Register the ranking

Plot the enrichment curve



Classic Smith & Waterman

- Classic version: one gap penalty
- Affine Gap extension: one penalty for opening a gap sequence, and one penalty for extending an existing gap

Affine gap is more biologically plausible

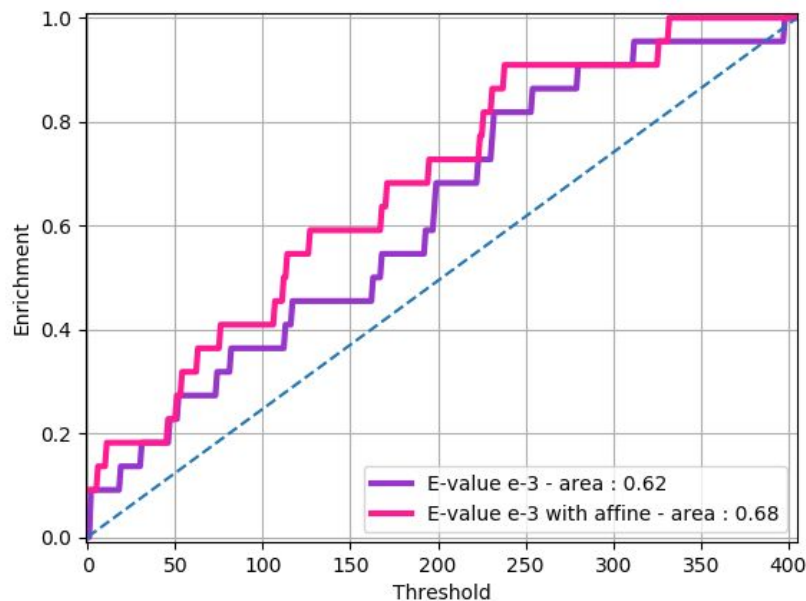
| | F | A | T | C | A | T | Y |
|---|---|----|---|----|----|----|----|
| T | 0 | 0 | 5 | 0 | 0 | 5 | 0 |
| C | 0 | 0 | 0 | 14 | 8 | 2 | 3 |
| A | 0 | 4 | 0 | 8 | 18 | 12 | 6 |
| G | 0 | 0 | 2 | 2 | 12 | 16 | 10 |
| S | 0 | 0 | 5 | 1 | 6 | 17 | 14 |
| F | 6 | 0 | 0 | 3 | 0 | 11 | 20 |
| A | 0 | 10 | 4 | 0 | 7 | 5 | 14 |

Source: Wikipedia

FATCA-TY
||| ::
TCAGSFA

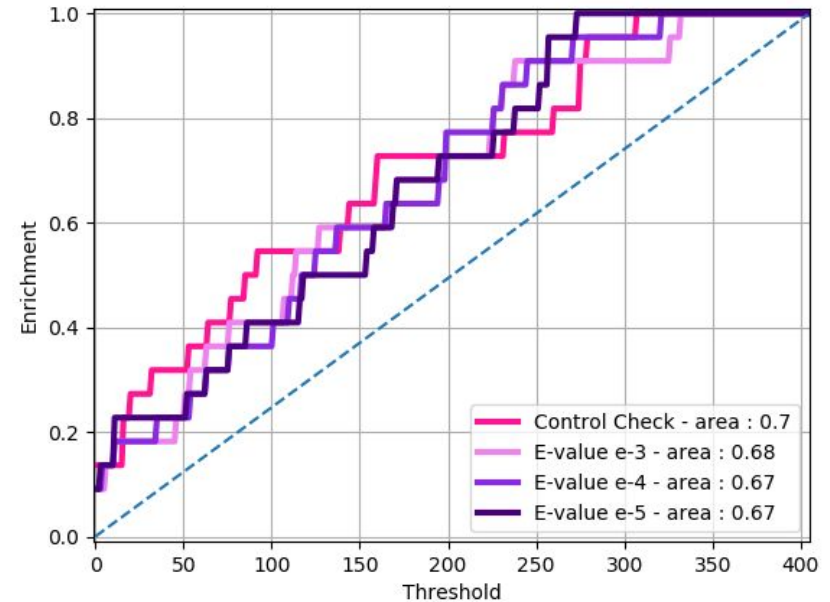
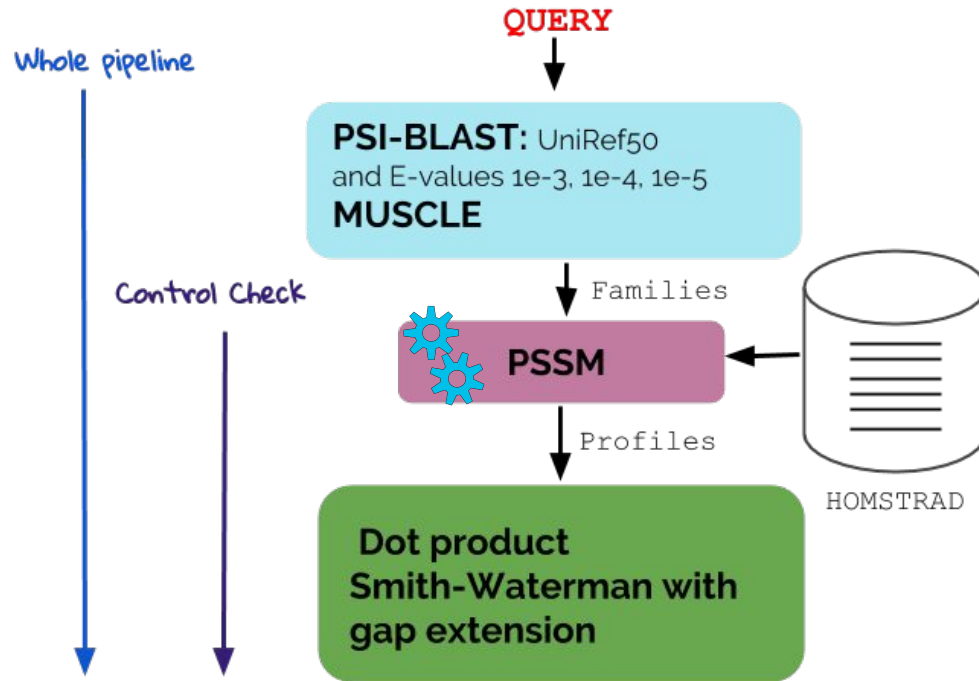
In our version, **profile-profile alignment**

match = dot product of the profile's columns

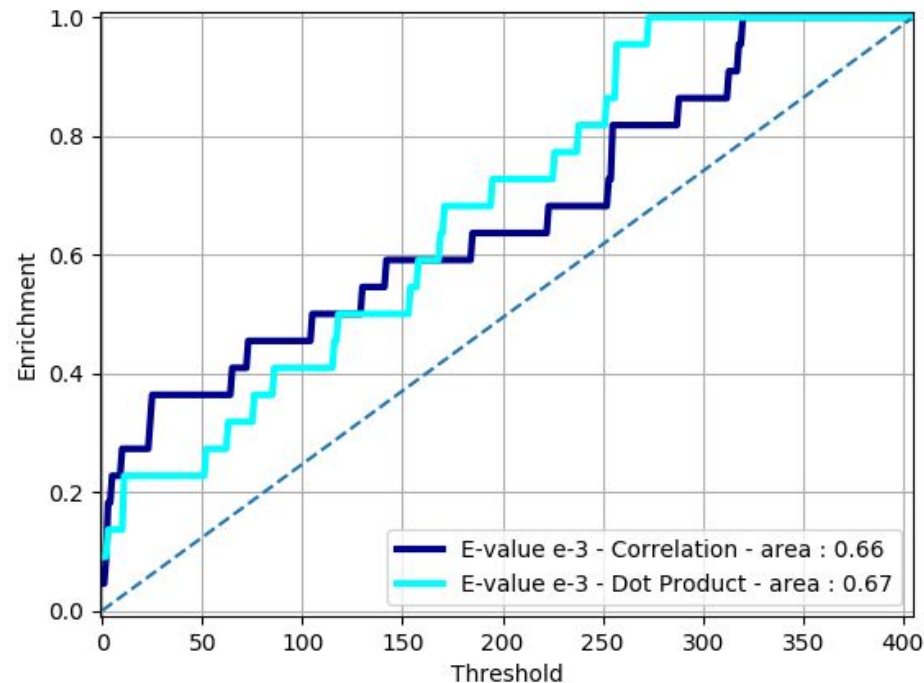
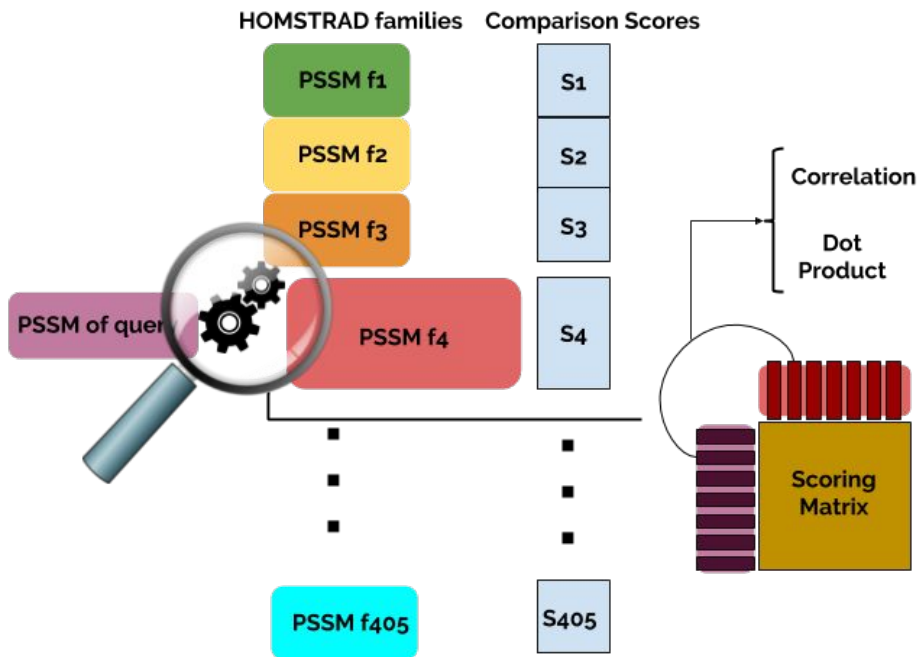


| | | | |
|----------|---|--|-----|
| Query | 3 | AGLPVIMCLKSNHOKYLRYSQSDNIQQYGLLQFSADKILDPLAQFEVEPS-----KTYDGLVHIKSRYNKYLVWRSPNHYWITASANEPEDENKSNWA----CTLFKPLYVEEGNMK-KVRLLVQLGHYQNYTVGGSFVSYLFAESSQIDTGSKDVFHVID- | 151 |
| Template | 1 | ---ATWTCINQLEDKRLLYSQAQAE-----SNSHHAPLSDGKTGSSYPHWFTNGYDG-----NGKLIKGRTPIKFGKADCDRPPKHSQNGMGKDDHYLLEFPDGDHDKFDSKKPKENPGPARVIYT---YPNKVFCEGIVAHQRGNQGDRLRLCSH | 149 |

Method - Evaluation of Profile Construction



Method - Correlation vs Dot Product



Method - Correlation vs Dot Product



Benchmark

21 queries → Best hit HOMSTRAD
Either Superfamily & Family (8) or Fold (13) hit

Fold

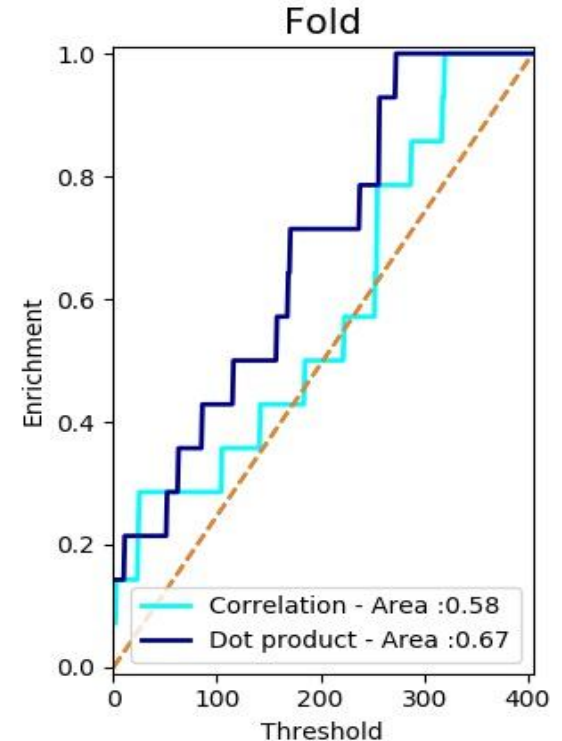
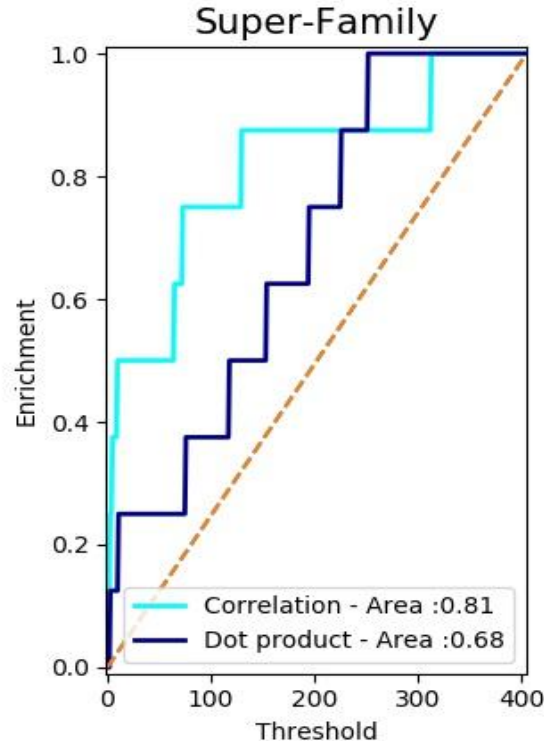
- Major structural similarity
- SSE's in similar arrangement

Superfamily

- Probably common ancestry
- Homology
- Sequence similarity

Family

- Pairwise sequence similarity > 25%



Method - Mixture of Scoring Functions

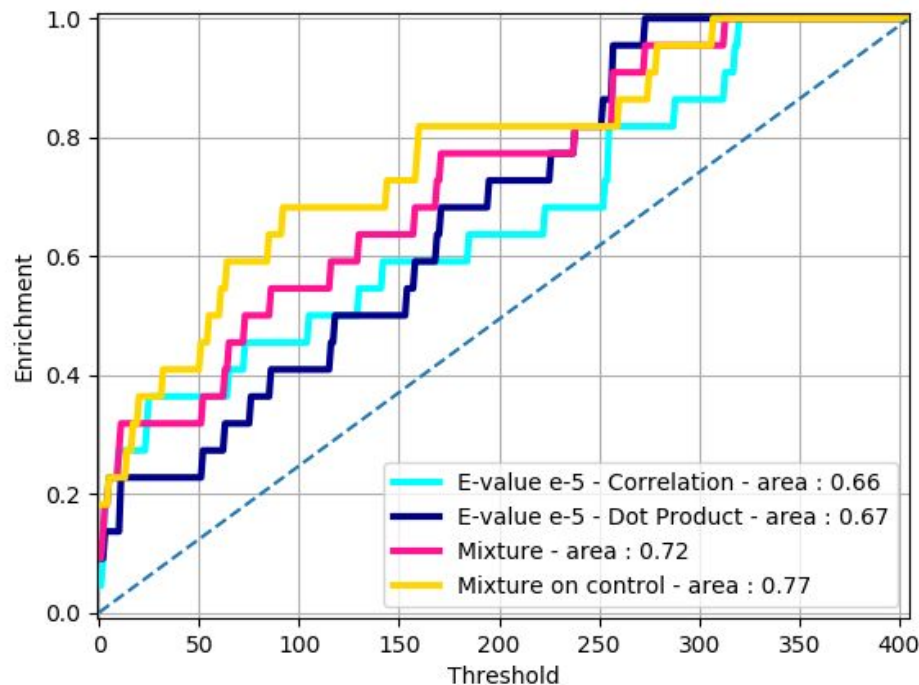
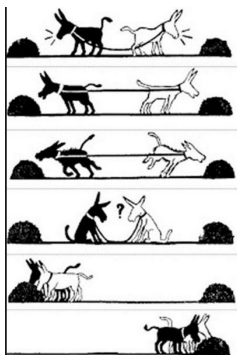
Dot product

... better on Fold

Correlation

... better on superfamily

Mixture of scoring



Prediction for the 11 sequences

| Mysterious query ID | Best hit template |
|--------------------------|-------------------|
| trQ6HKV8 (BT9727_1481) | BRCT |
| trQ6HLB4 (BT9727_1323) | END |
| feuA (BT9727_3483) | neur_chan_memb |
| G1G14-4353 | pot |
| recN (BT9727_3916) | COX1 |
| G1G14-3311 (BT9727_3234) | PCI1 |
| scdA (BT9727_1971) | PABP |
| hmp (BT9727_1331) | zf-CCHC |
| mfd (BT9727_0048) | COX1 |
| celB (BT9727_4888) | Ribosomal_L12 |
| arcD (BT9727_0754) | hormone5 |

- Despite the good quality of the profiles created, parameters improvement is still required (e-value & database)
- Further analysis on scoring method (euclidean distance, spearman correlation, ect.)
- Fold detection Improvement
 - Secondary structure prediction information & Hydrophobicity instead of sequence

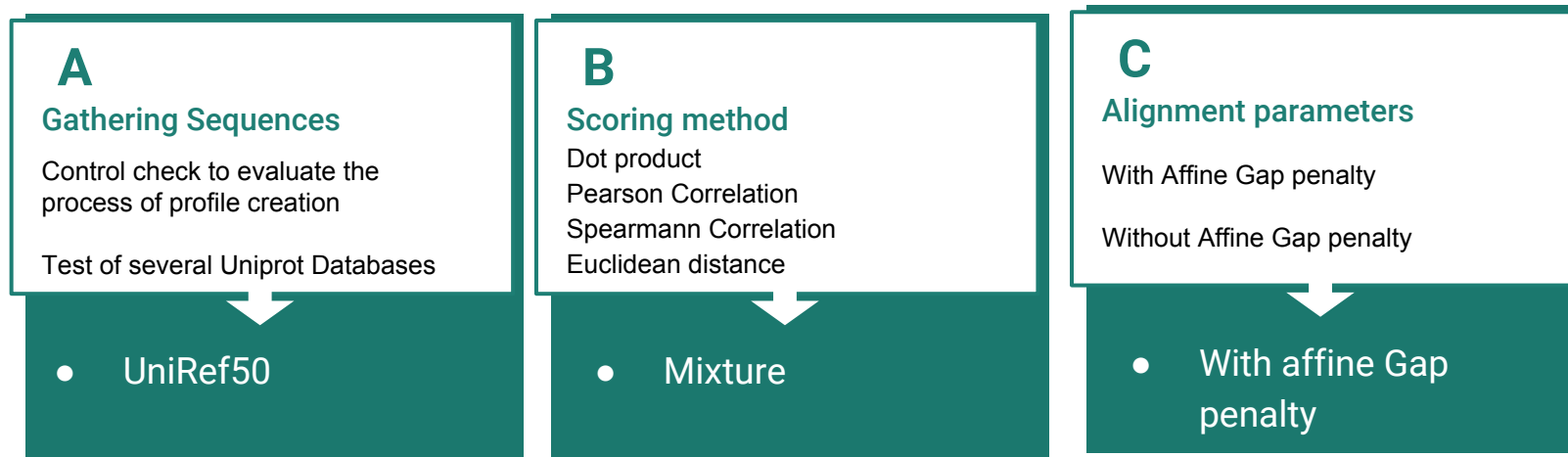
- Despite the good quality of the profiles created, parameters improvement is still required (e-value, database, gap penalties, scoring functions)
- Super-family and Fold types → influence scoring function performance
- Fold detection Improvement through:
 - Mixture classification approach (test on benchmark already positive)
 - Secondary structure prediction information

References

- [1] Moult J., Pedersen J. T., Judson R., and Fidelis K. A large-scale experiment to assess protein structure prediction methods. *Proteins*, 23, 1995.

- [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.

Annexe 1 - Method



Annexe 2 - Strategy Requirements

- Profile comparison
→ requires a **scoring method**

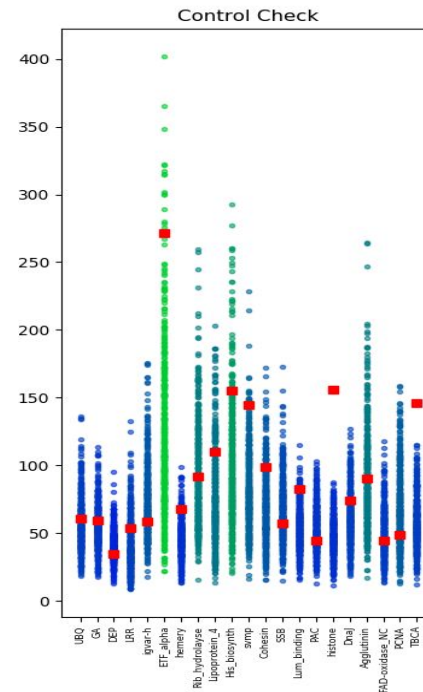
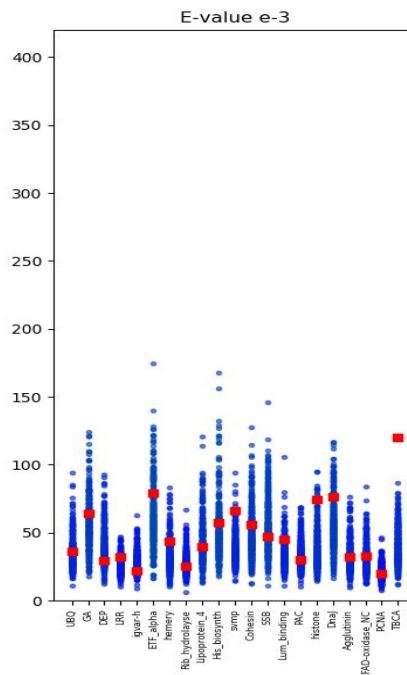


- Profile creation
→ requires **Protein Database** selection
→ Profile Creation method (PSSM, HMM...)
- Profile alignment
→ requires **strategic alignment algorithm**

Results (1) - Control Check of Profiles Construction

Evaluation of process of profile creation method (2)

- Running profile comparison on provided MSA and our MSA
- Scores are more spread for the provided MSA comparing to our MSA



Parameter Optimization

Gap penalties optimization

- Use of Smith-Waterman algorithm to align profiles
- Optimization for dot product and correlation comparison methods

| 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 |

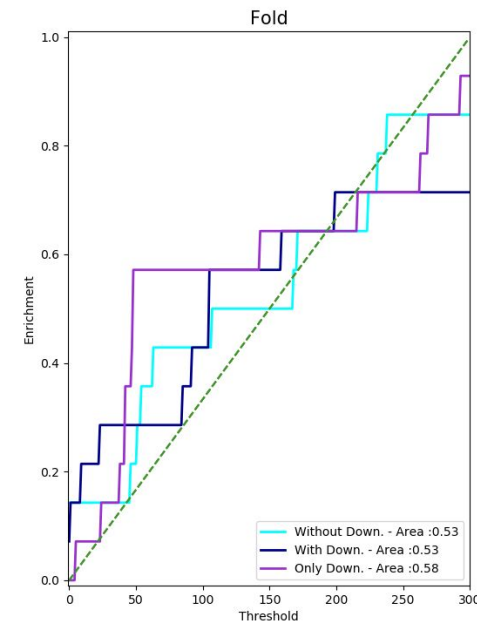
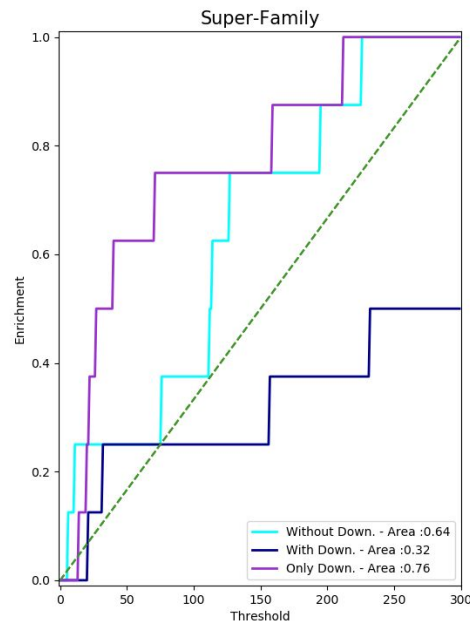
Table 1: Optimized parameters for scoring functions

→ Improvements of the scores with gap penalties

Results (4) Downstream evaluation

Final ranking scores with downstream program combination

- Downstream program uses machine learning on 3 scores (DOPE, H-P and Alignments scores)
- Superfamily: Poorer results with upstream & downstream combination
- hypothesis: Overfitting during learning on Orion dataset



Fold and Superfamily

