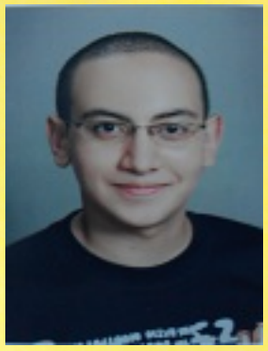
Protein Prediction 2

23.01.2013

Team



Omar

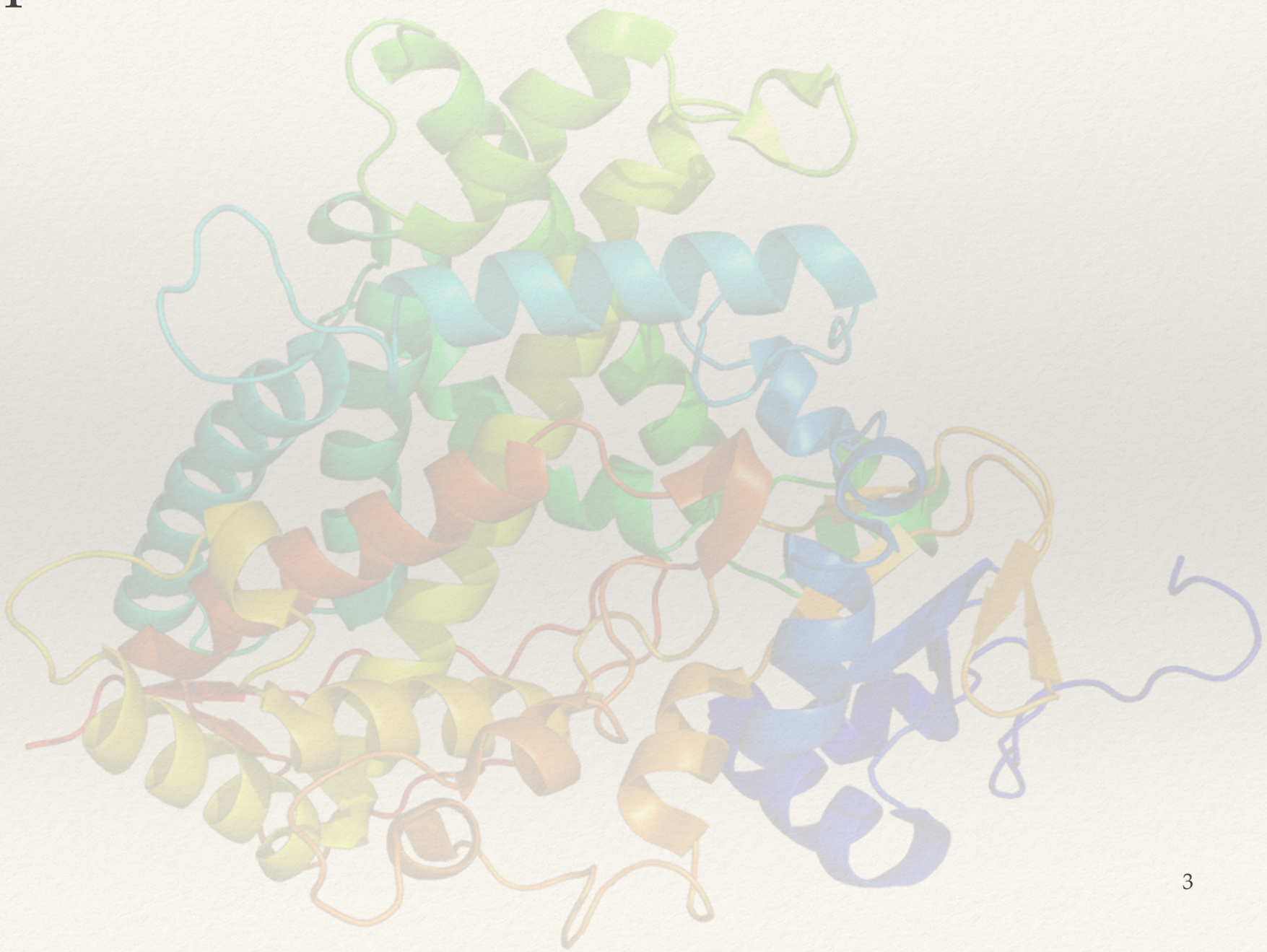
Meghana

Kshitija

Sebastian

Agenda

- Problem Description
- Approach
- Evaluation
- Conclusion



Problem Description

- Assuming direct relation:

  Protein sequence → Protein function

- UniProtKB ≈ 49M proteins (as of Dec 2013)

- Swiss-Prot ≈ 541K proteins (as of Dec 2013)

  - ✔ Manually annotated and reviewed

- Need for *automated function prediction*

Approach

Input: protein sequence

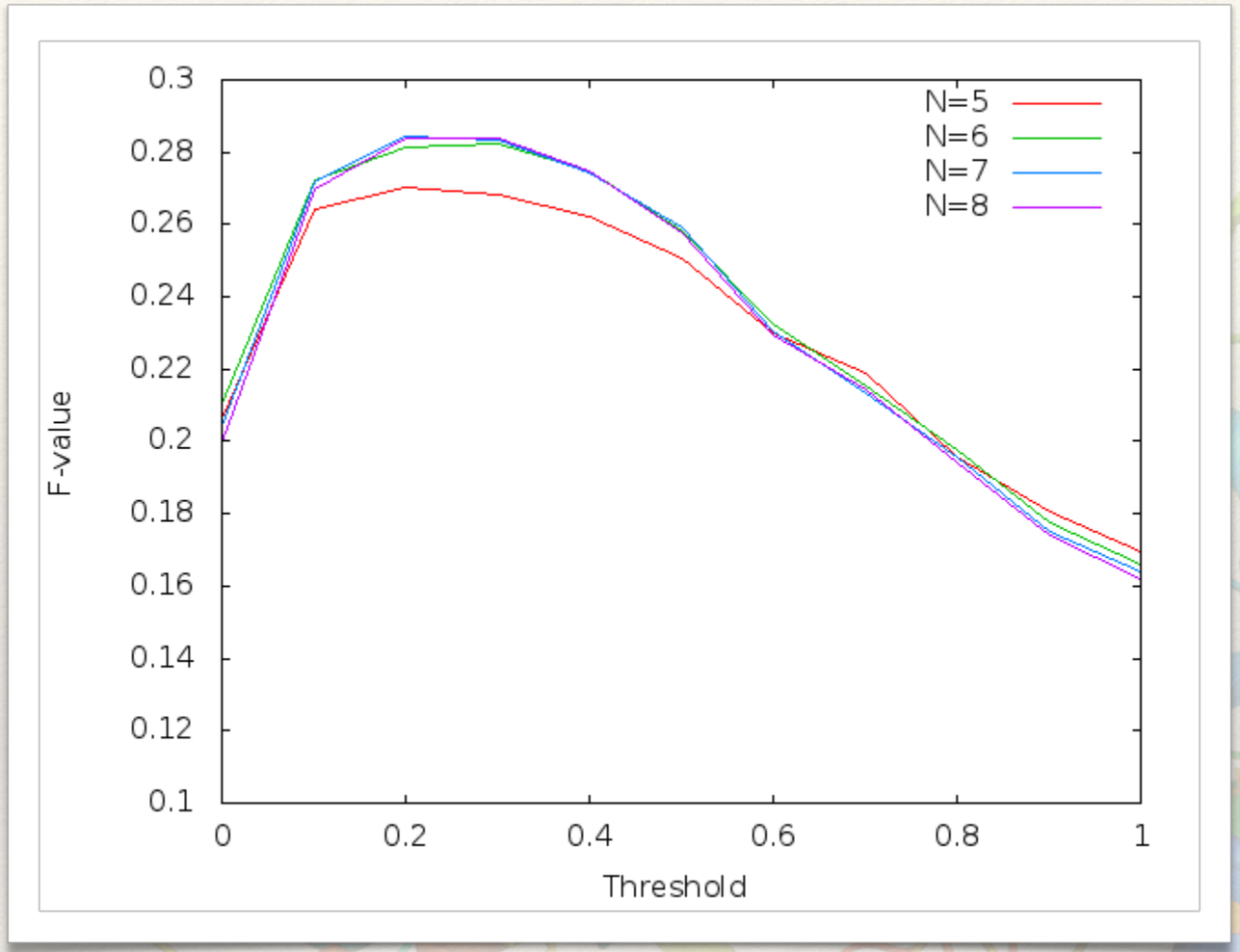
1. Get Blast top N hits
2. Create HPO tree for each hit
3. Merge into one tree – add Blast **score**
4. Normalize score values [0, 1]
5. Apply threshold T

Output: tree leaves + score

Parameter Optimization

- Goal: find optimum
  - Hits N (5 → 8)
  - Threshold T (0.0 → 1.0)
- Testing over 2800 sequences (training set)
- Average F-value for each **parameter combination**

Parameter Optimization – Results



At N=7, T=0.2 → F-value=0.2842

Interfaces - CL

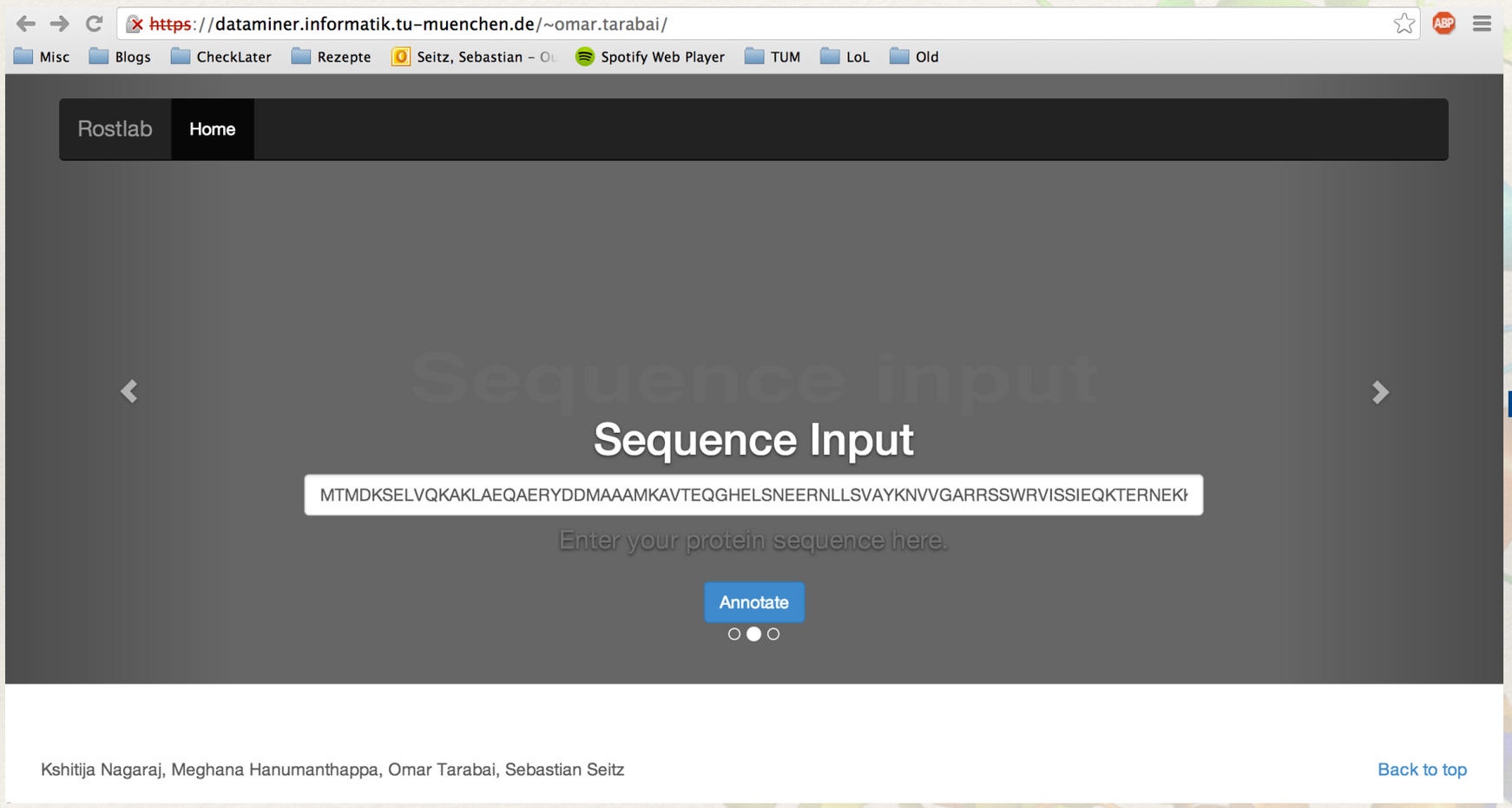
1. sseitz@dataminer: /mnt/home/student/omar.tarabai/bin (ssh)

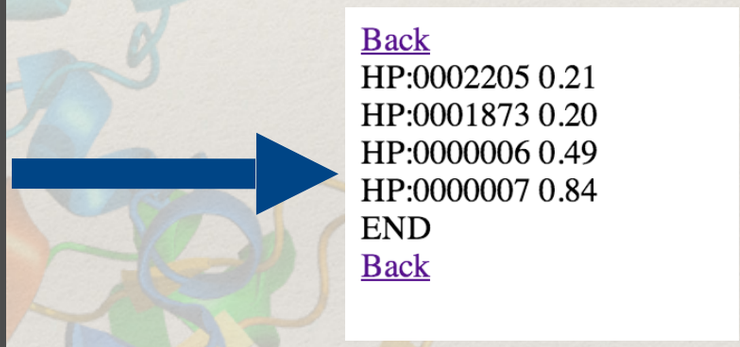
```
sseitz@dataminer:/mnt/home/student/omar.tarabai/bin$ ./predicthpo -h
\usage: predicthpo [-h] [-s SEQUENCE-STRING] [-n] [-p]

Rostlab PP2 2013 Team6 protein predictor

optional arguments:
  -h, --help            show this help message and exit
  -s SEQUENCE-STRING, --sequence SEQUENCE-STRING
                        Predict protein sequence function
  -n, --hits            Return number of hits used
  -p, --others          Return sensitivity, precision and F-measure
sseitz@dataminer:/mnt/home/student/omar.tarabai/bin$ ./predicthpo -s MTMDKSELVQKAKLAEQAERYDDMAAAMKAV
TEQGHELSNEERNLLSVAYKNVVGARRSSWRVISSIEQKTERNEKKQQMGKEYREKIEAELQDICNDVLELLDKYLIPNATQPESKVFYLKMKGDYFRYL
SEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSPEKACSLAKTAFDEAIAELDTLNEESYKDSTLIMQLLRDNLTLWT
SENQGDEGD
HP:0002205      0.21
HP:0001873      0.20
HP:0000006      0.49
HP:0000007      0.84
END
sseitz@dataminer:/mnt/home/student/omar.tarabai/bin$
```

Interfaces - Web





Additional Optimization

- Caching:
  - Saving annotation map
  - Saving HPO tree

Fast access to combined full HPO tree

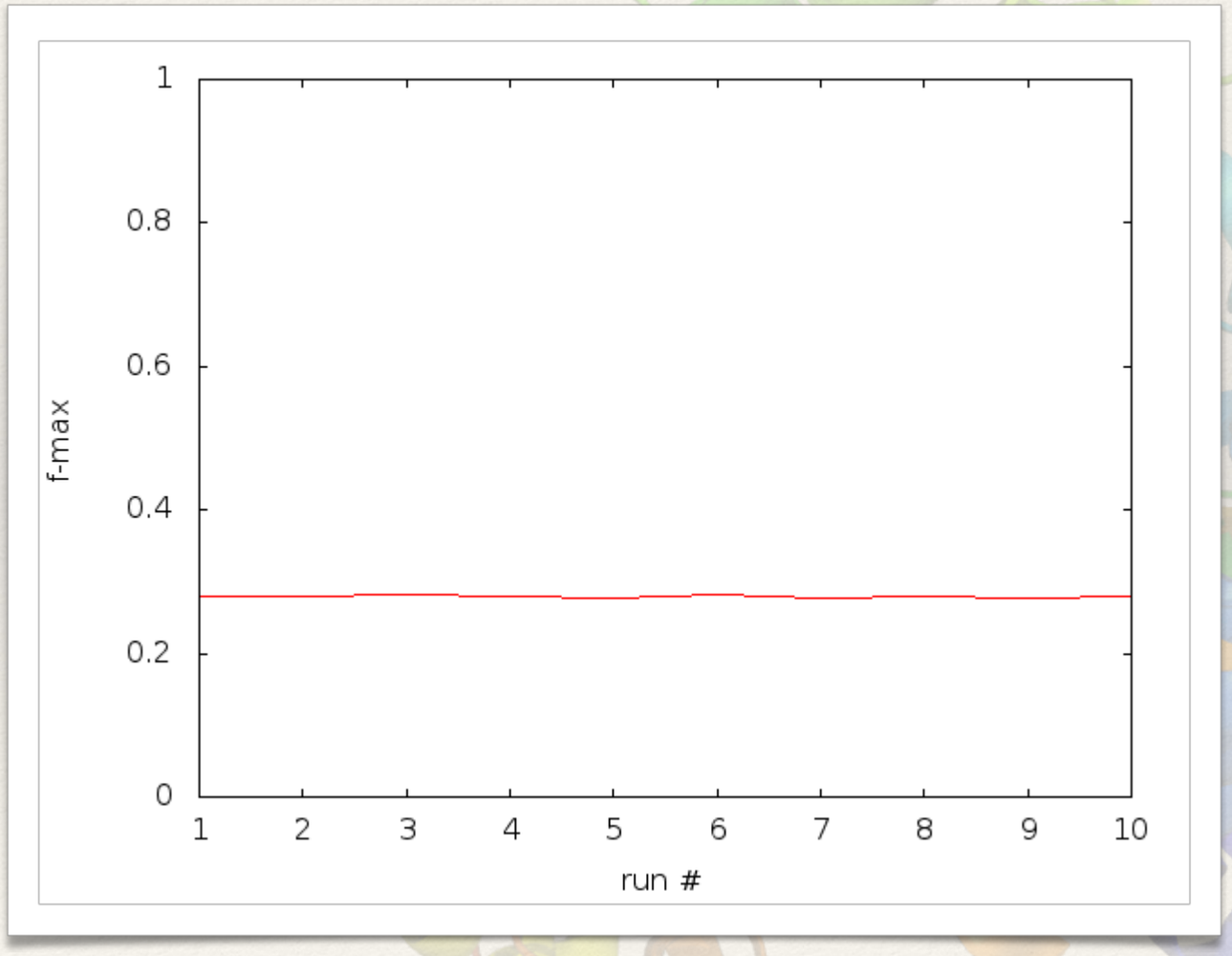
“No hits” problem

- Training set: 89 / 2805
- Targets set: 1008 / 20257
- Statistics from training set:
  - Average terms per sequence: 73
  - Top occurring terms
- “Default” tree with top 73 terms
- F-value 0.2736 → 0.2842

1	HP:0000001	2805
2	HP:0000118	2777
3	HP:0000005	2601
4	HP:0000707	1712
5	HP:0002011	1600
6	HP:0000007	1595
7	HP:0000152	1373
8	HP:0011446	1365
9	HP:0000234	1362
10	HP:0000478	1288
11	HP:0100543	1234
12	HP:0000271	1201
13	HP:0001939	1190
14	HP:0001438	1159
15	HP:0000924	1130
16	HP:0011842	1099
17	HP:0000006	1096
18	HP:0001574	1085
19	HP:0001507	1069
20	HP:0003011	1028
21	HP:0000119	1020
22	HP:0000598	975
23	HP:0011804	940
24	HP:0001626	924
25	HP:0000951	920
26	...	
27	...	
28	...	

Evaluation – Cross Validation

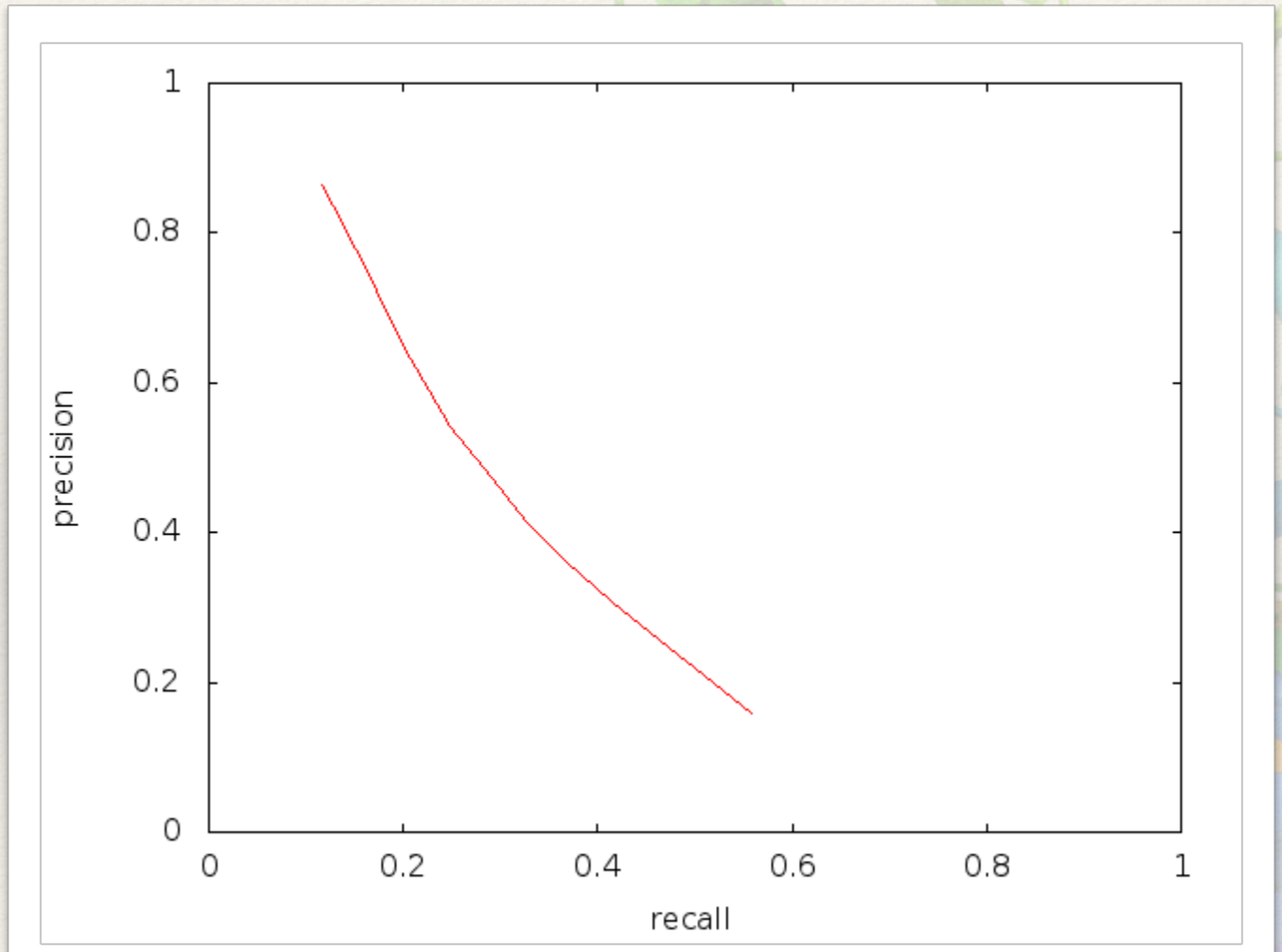
- 10 folds: 9 for training, 1 for evaluation
- Randomized and tested 10 times
- Results:



- Average F-value: 0.278975

Evaluation – Precision / Recall

- Precision / recall curve for 1 run:



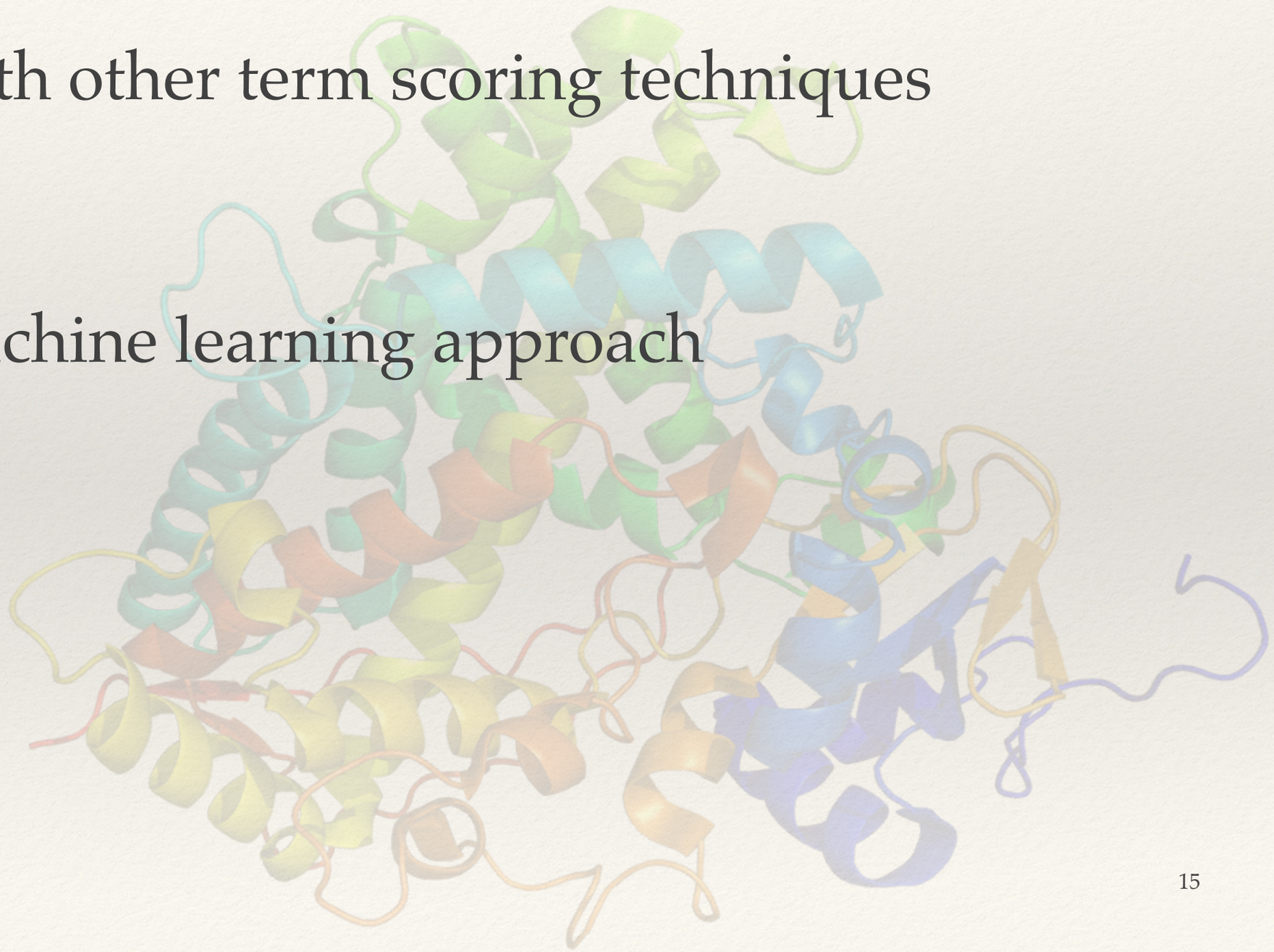
Conclusion

We used blast score as an indicative for protein sequence similarity, which reflects function similarity.

Achieved F-value ≈ 0.28

Future Work

- Experiment with other term scoring techniques
- Use *hhblits*
- Introduce a machine learning approach



Thank you!