Protein Prediction 2

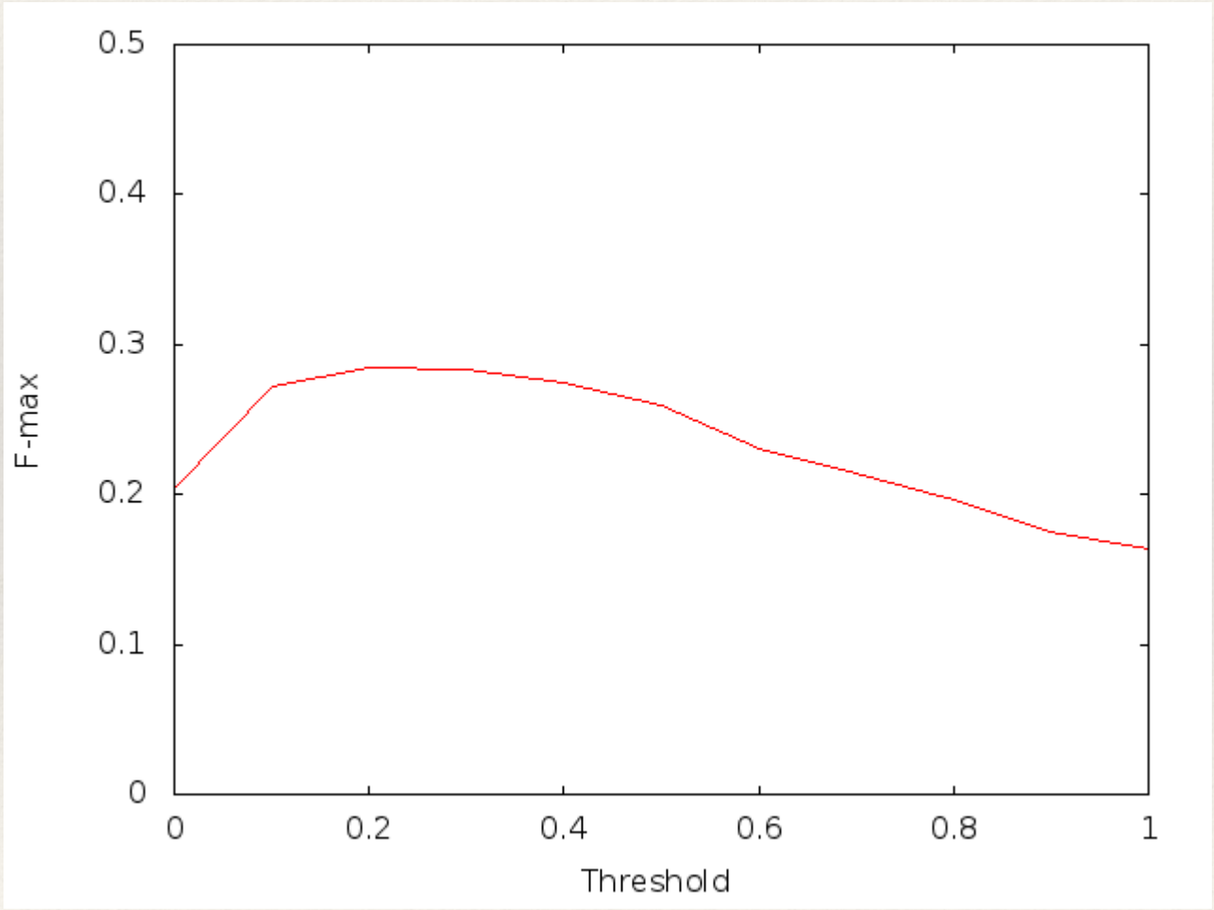
Exercise
Team 6
19.12.2013

Merging Method

- Using blast “score” value
- Score is added for a term seen more than once
- Score normalized [0, 1]
- Testing over whole training set
- Testing variables:
  - Hits N (4 → 10)
  - Threshold T (0.0 → 1.0)

Results

- Best result at N=7, T=0.2, **F-max=0.2842**



“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 sequences
- F-max 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	...	

Thank you