Protein Prediction II
Winter 18/19

for Computer Scientists
for Bioinformaticians

13.12.2018

Timeline

25.10.	Introduction
1.11.	No exercise
8.11.	Build a machine learning pipeline
15.11	Optimize Word2Vec parameters
22.11.	Discuss W2V optimizations
29.11.	Present ‘final’ W2V optimizations
6.12.	No exercise, Dies academicus
13.12.	Present current progress
20.12.	No exercise session

10.1.	Feedback about your progress (no presentation necessary)
17.1.	
24.1.	Final presentations w/ Prof. Rost Deadline for method submission
31.1.	Q&A Session Method evaluation
7.2.	**Exam**

Exercise groups

1	2	3	4	5	6	7	8
Nabil	Sofie	Kayalvizi	Marla	Jinlong	Felix	Muhammed	Francesco
Vanessa	Ghalia	Amrei	Vagram	Reza	Elisabeth	Mustafa	Muhammad
Lukas	Tobias	Corinna	Michaela	Chris	Issar	Abdulrahman	Silvia
Nathalie	Marco	Daniel	Rinita	Julian	Omar	Nail	Aynesh
						Martin	

Tasks until January I

- Create CV-splits by protein not by sample, if protein A is in training, all samples of protein A must be in training and none in test
- Try out the other ideas you already mentioned in your talks

Tasks until January II

- Prepare your methods to use the following input and output formats

Input:

- A (multi-)FASTA file containing one or more sequences
- Each sequence is in a single line, preceded by a header in the previous line

Output:

- A single file with predictions in a “tab separated values”-format
- Each sequence is preceded by the same header it had in the input file
- For each position list the amino acid, binary prediction, and prediction score

Example Input: (multi-)FASTA file

```
>P0A8Q6
MGKTNDWLDFDQLAEEKVRDALKPPSMYKVILVNDDY...
>P15927
MWNSGFESYGSSSYGGAGGYTQSPGGFGSPAPSQAEK...
>Q8VPC3
MKRGFTLLEVMLALAIFALSATAVLQIASGALSNQHV...
...
```

Example Output: TSV (tab-separated) file

```
>P0A8Q6
M   -   0.13
G   -   0.32
K   +   0.79
...
>P15927
M   -   0.03
W   +   0.54
N   +   0.92
...
```

← same header(s) as in input file

← amino acid, binary prediction (“-” or “+”), score* ∈ [0, 1]

*high score = binding, low score = non-binding