Protein Prediction 1 for Computational Biologists - Exercise

Exercise 1 - Introduction
Hi :) 

Maria Schelling: maria.schelling@tum.de

PhD student @Rostlab

Contact me via email if you have any questions/problems
About the exercise

Time slot: Thursday, 14:00-15:30

Project: Prediction of per-residue features using high-resolution PDB structures and ProtVec

All results will be saved on our wiki, please also upload your code there

Content of the exercise will be part of the exam
# General overview - subject to change

<table>
<thead>
<tr>
<th>Date</th>
<th>Exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.04.18</td>
<td>Introduction to ProtVec, dataset and features</td>
</tr>
<tr>
<td>26.04.18</td>
<td>Introduction to ProtVec in Python, data preparation</td>
</tr>
<tr>
<td>03.05.18</td>
<td>Introduction to ML in Python, decide on ML model, develop model</td>
</tr>
<tr>
<td>10.05.18</td>
<td><em>No exercise (Christi Himmelfahrt)</em></td>
</tr>
<tr>
<td>17.05.18</td>
<td>ML presentation and refinement, performance measurements</td>
</tr>
<tr>
<td>24.05.18</td>
<td>Performance evaluation</td>
</tr>
<tr>
<td>31.05.18</td>
<td><em>No exercise (Fronleichnam)</em></td>
</tr>
<tr>
<td>07.06.18</td>
<td>Present results, refinement</td>
</tr>
<tr>
<td>14.06.18</td>
<td>Final talks (with Burkhard Rost) [Tentative]</td>
</tr>
<tr>
<td>21.06.18</td>
<td>Final talks (with Burkhard Rost) [Tentative]</td>
</tr>
</tbody>
</table>
Features (1)

Secondary structure:

- experimental annotation taken from DSSP in 3 or 8 states
- X: unresolved structure, Y: conflicting annotations for different chains

Solvent accessibility:

- experimental annotation taken from DSSP (absolute solvent accessibility)
- normalized to relative solvent accessibility
Features (2)

B-Values:

- experimental annotation taken from BDB
- normalized to have a mean of 0 and a standard deviation of 1

Unresolved structure/disorder:

- X-characters in the secondary structure: residues without a measured structure in PDB → indicator for disorder
<table>
<thead>
<tr>
<th>Topic</th>
<th>Group members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction of secondary structure</td>
<td>Martin, Luna, Mariana, Binh</td>
</tr>
<tr>
<td>Prediction of solvent accessibility</td>
<td>Nathalie, Tatjana, Pia</td>
</tr>
<tr>
<td>Prediction of B-values</td>
<td>Sebastian, Joel, Nabil, Lukas</td>
</tr>
<tr>
<td>Prediction of disorder</td>
<td>Klaudia, Daniel, Florian, Luise</td>
</tr>
</tbody>
</table>
Tasks until next week

● familiarize yourself with ProtVec (and Word2Vec in general)
● familiarize yourself with your feature to be predicted
  ○ biological meaning of the feature?
  ○ how many states are possible?
  ○ how many states do you want to predict?
  ○ where did the data came from and how was it preprocessed?
  ○ what are problems/decisions you are facing when predicting this feature?
● familiarize yourself with the dataset
  ○ record simple statistics like the number of sequences and similar
  ○ distribution of your feature states in the dataset
Next exercise

● Present your results - except ProtVec
  ○ short presentation for each group
  ○ 10 minutes, max. 5 slides

● Q&A about ProtVec

● You will get access to the wiki page
  ○ after that: put all information to this page
Material

● ProtVec:

● Dataset (sequences + feature):
  ○ Secondary structure: [www.rostlab.org/~evcfunc/dssp8.tar.gz](www.rostlab.org/~evcfunc/dssp8.tar.gz)
  ○ Solvent accessibility: [www.rostlab.org/~evcfunc/rel_asa.tar.gz](www.rostlab.org/~evcfunc/rel_asa.tar.gz)
  ○ B-value: [www.rostlab.org/~evcfunc/bdb_bval.tar.gz](www.rostlab.org/~evcfunc/bdb_bval.tar.gz)
  ○ Disorder: [www.rostlab.org/~evcfunc/disorder.tar.gz](www.rostlab.org/~evcfunc/disorder.tar.gz)
  ○ some of the targets saved as “memmap” - can be read with python using the following code:

```python
import numpy as np
np.memmap(path, dtype=np.float32, mode='r', shape=shape)  # path: path to memmap file
# shape: length of corresponding input sequence
```