Problem Based Learning

SoSe18
Course outline

1. **Intermediate talk** (mid June): 20 minutes talk + 10 minutes discussion. 2 students working on the same feature have a joint presentation of this feature. Presentations will take place in 2 slots, each roughly 2 hours (=8 students per slot).

2. **Final talk** (beginning of February 2019): 20 minutes talk + 10 minutes discussion. 4 slots, each roughly 2 hours (=4 students per slot). This talk should summarize the complete project including the dataset, the algorithm (incl. tuning of meta-parameters) and the performance assessment of the final model.

3. **Final report:** is due 1 week after final talk. Around 5 pages in Bioinformatics Oxford style. Should summarize results in paper-like style.

Final dates and link to template will be published on www.rostlab.org
Presence

● if not stated otherwise, no regular presence is required for students during the lecture slot on Monday
● however: supervisors are available during this time for questions regarding the project
● you can work in our student room
Topics

● implement a simple machine learning prediction method to predict a feature relating to protein structure and function
● we will use Python/sci-kit learn
● each student will either apply an SVM (Support Vector Machine) or an ANN (Artifical Neural Network)
● each project will use ProtVec as a representation of protein sequences (except sarcoma group)
● 2 students will predict the same feature
Radiomics: soft tissue sarcoma

- CT image data of sarcoma patients
- 1358 image features
- 7 clinical features

Literature:
Transmembrane Protein Prediction

● Data set: 1607 proteins
  ○ transmembrane and soluble proteins
  ○ with and without signal peptides

● Task is to...
  ○ ...predict transmembrane proteins
  ○ ...predict transmembrane helices
  ○ ...predict inside/outside topology

Literature:
Prediction of transmembrane protein expression success in a structural genomics pipeline

Extension: Predict other stages such as crystallization, or predict for non-transmembrane proteins
Prediction of protein-protein interactions

Extension: Predict DNA- or RNA-binding sequences
Secondary Structure assignment with DSSP

- Standardization of secondary structure assignment (no prediction!)
- Original publication: [1]
- Newest version: [2]
- Currently maintained by: [3]

4. Figure taken from: [https://www.rcsb.org/structure/3on1](https://www.rcsb.org/structure/3on1), accessed on 16.04.2018.
### 8-states vs. 3-states secondary structure DSSP

<table>
<thead>
<tr>
<th>8-states</th>
<th>3-states</th>
</tr>
</thead>
<tbody>
<tr>
<td>3\textsubscript{10} Helix (G)</td>
<td>Helix (H)</td>
</tr>
<tr>
<td>Alpha Helix (H)</td>
<td></td>
</tr>
<tr>
<td>PI Helix (I)</td>
<td></td>
</tr>
<tr>
<td>Beta Bulge (E)</td>
<td>Strand (E)</td>
</tr>
<tr>
<td>Beta Bridge (B)</td>
<td></td>
</tr>
<tr>
<td>Region of high curvature (S)</td>
<td>Loop (C)</td>
</tr>
<tr>
<td>Turns (T)</td>
<td></td>
</tr>
<tr>
<td>Others/Loops (C or blank or '-')</td>
<td></td>
</tr>
</tbody>
</table>

Prediction of subcellular localization in bacteria

**Dataset:** bacterial proteins with localization annotation from LocTree2/3

**Task:** Prediction of 6 classes for bacterial proteins
   (cytosol, plasma membrane, periplasmic space, outer membrane, fimbrium, extra-cellular)

**Possible extension:** Prediction of
   - eukaryotic proteins (18 classes)
   - archaeal proteins (3 classes)

**Literature:**

https://www.thoughtco.com/prokaryotes-meaning-373369
Prediction of nuclear traveler proteins

**Dataset:** proteins known to localize to
- only the nucleus (non-traveler)
- at least 1 other compartment (traveler)

**Task:** Distinction between nuclear traveler and non-traveler proteins

**Possible extension:**
- Prediction of traveler proteins for important model organisms
- Distinction of traveler proteins for more/all compartments

Literature:

http://biology4isc.weebly.com/a-cell--unit-of-function.html
Important material for every group

1. ProtVec and word2vec (not for Sarcoma-group!):

2. General introduction to Python:
   https://docs.python.org/3/tutorial/

3. General introduction to ML:
   https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3