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, 14:10-15:40

Project: Prediction of binding site residues using SNAP-based features (with Machine Learning)

Content of the exercise will be part of the exam

We will offer a bonus of 0.3 for the final exam, requirements to achieve the bonus:

- active participation in the exercise until the end of the semester
- final presentation of the exercise at the end of the semester
- exam grade without bonus is 4.0 or better
General overview - subject to change

<table>
<thead>
<tr>
<th>Date</th>
<th>Exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td>02.05.19</td>
<td>Introduction, Distribution of groups and topics</td>
</tr>
<tr>
<td>09.05.19</td>
<td>Presentation of dataset and features</td>
</tr>
<tr>
<td>16.05.19</td>
<td>Introduction to ML (with Python and Scikit-learn)</td>
</tr>
<tr>
<td>23.05.19</td>
<td>Data preparation, decide on ML model</td>
</tr>
<tr>
<td>30.05.19</td>
<td>No exercise <em>(Christi Himmelfahrt)</em></td>
</tr>
<tr>
<td>06.06.19</td>
<td>ML presentation, performance measurements</td>
</tr>
<tr>
<td>13.06.19</td>
<td>Performance evaluation</td>
</tr>
<tr>
<td>20.06.19</td>
<td>No exercise <em>(Fronleichnam)</em></td>
</tr>
<tr>
<td>27.06.19</td>
<td>Present results, refinement</td>
</tr>
<tr>
<td>04.07.19</td>
<td>11.07. 19 Final talks (with Burkhard Rost) [Tentative]</td>
</tr>
</tbody>
</table>
Dataset and features

- Dataset obtained from PDB, BioLip [1], and information from [2]
- Task: Prediction of binding site residues and catalytic residues
- Prediction based on effect prediction (SNAP2) compared to BLOSUM
- Possible extension: include additional features like solvent accessibility or conservation

Material (relevant for everyone):

Dataset - Download

- FASTA files: http://rostlab.org/~bindpredict/fasta_seqs.tar.gz
- SNAP2 predictions: http://rostlab.org/~bindpredict/snap_predictions.tar.gz (17 proteins still running, will be updated)
- Binding annotations: http://rostlab.org/~bindpredict(binding_residues_2.5_0.txt)
Features (1)

Single SNAP scores:

○ For each position, use the 19 SNAP scores for mutations to a non-native amino acid
○ How to deal with mutation to itself?

Single SNAP scores compared to BLOSUM:

○ For each position, compare SNAP score and BLOSUM value
○ Are they predicting the same or something different?
○ How to put the comparison into a number?
Features (2)

Average/cumulative SNAP score:

○ For each position, calculate average or cumulative SNAP score
○ For cumulative SNAP score: Only use positive scores? (or similar)

Fraction of high SNAP scores:

○ For each position, how many of the 19 possible mutations are predicted with a strong effect?
○ How to define strong effect?
### Groups

<table>
<thead>
<tr>
<th>Feature</th>
<th>Group members</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single SNAP scores</td>
<td></td>
</tr>
<tr>
<td>Single SNAP scores compared to BLOSUM scores</td>
<td>Paul, Thomas E, Thomas H</td>
</tr>
<tr>
<td>Average/Cumulative SNAP score per position</td>
<td>Kyra</td>
</tr>
<tr>
<td>Fraction of high SNAP scores</td>
<td>Felix, Amrei, Kaya</td>
</tr>
</tbody>
</table>
Tasks until next week

- familiarize yourself with the data sources, SNAP2, and the dataset
- prepare your feature
- perform a basic dataset analysis
  - size of the dataset
  - distribution of the states (binding/non-binding)
  - distribution of your feature
- present your results next week
  - discussion about data sources and SNAP2
  - present dataset analysis in max. 10 minutes per group, max. 5 slides