Prediction of signal peptides

Supervisor: Konstantin Weißenow
Students: Tba
Signal peptides

- Signal peptides are small sequences of amino-acids (residues) in the N-terminal region of a protein
- Used to identify proteins for translocation, e.g. into membranes or for extracellular transport
- Usually cleaved off by enzymes after translocation has finished
Task overview

- You are supposed to develop simple machine learning devices, which, given a short amino-acid sequence, predict for each residue
  - 1) whether it is part of any signal peptide or not (binary classification)
  - 2) whether it is part of a specific signal peptide or not (4-state classification)
  - 3) whether it is part of a specific signal peptide or where the residue in the mature protein is located (6-state classification; optional, we will focus on the first two tasks)
- The possible signal peptide types we are using are Sec/SPI (S), Sec/SPII (L) and Tat/SPI (T).
- The location types are extracellular (O), membrane (M) and cytoplasm (I).
Example

Input sequence:
MNKGLCNWRLFSLFGLGMMALLLAGCGKPLSTLQPAGEVADMQYSLMLIALLSSTSIMVLIIVVVAIIFVYVVR

Prediction:
LLLLLLLLLLLLLLLLLLLLLLLL0000000000000000MMMMMMMMMMMMMMMMMMIII

Interpretation: The first few residues form a signal peptide of type Sec/SPII, which is cleaved off after the protein reaches its destination. The following residues of the mature protein indicate that we’re dealing with a transmembrane protein, since respective parts are located in extracellular space, membrane and cytoplasm. This example would be similar to the case illustrated on slide 2 (extracellular space instead of ER, though).

For the first two tasks, you would just map O, M, I to one class, which indicates ‘not a signal peptide’, while loading the data.
Literature

● Overview

● Dataset
  ○ [http://www.cbs.dtu.dk/services/SignalP-5.0/data.php](http://www.cbs.dtu.dk/services/SignalP-5.0/data.php)

● Methods: