Predicting signal peptides

This project is for students with experience of machine learning. If you start from scratch, i.e., without training in machine learning, you will have a hard time.

Background

Many proteins need to be localised properly in the cell to do their job and they are guided there by a molecular address tag, a **signal peptide** (http://en.wikipedia.org/wiki/Signal_peptide). The signal peptide, if it exists, is found in a limited region in the **N-terminal** (http://en.wikipedia.org/wiki/N-terminus) part of the protein and has a relatively simple structure. First there are a variable number of amino acids with a tendency to be positively charged, known as the n-region. Then comes a hydrophobic region with **α-helical** (http://en.wikipedia.org/wiki/Alpha_helix) structure, known as the h-region, providing the strongest statistical signal. Next comes the c-region of 3-8 amino acids, often polar but not charged, and the signal peptide is typically cleaved off after the c-region. A few positions further from the cleavage site, away from the signal peptide, there is a tendency of small and uncharged residues.

When trying to understand an organisms **proteome** (http://en.wikipedia.org/wiki/Proteome), part of the annotation task in a a typical gene project, there is often an interest in determining where proteins are located. If there is no or limited experimental data, a bioinformatic approach is necessary. Many prediction tasks are regarding proteins and their functions are extremely difficult, but the simple structure of signal peptides has enable quite accurate sub cellular classification. The main obstacle in signal peptide prediction has been that the h-region can look like a transmembrane region, which has prompted joint prediction of signal peptide and transmembrane structure.

Problem

Your task is to build, train and test a signal peptide classifier based on the data below, and apply it to the proteome of two organisms. The main questions to adress are:

1. How well can your classifier distinguish signal peptides from non-signal peptides?
2. Does your classifier behave differently on TM and non-TM proteins?
3. How many signal peptides do you identify in the two proteomes you have chosen?

Data

Training and test data

Here is labelled **training data for signal peptide prediction** (https://kth.instructure.com/courses/3845/files/582075/download?wrap=1)

(https://kth.instructure.com/courses/3845/files/582075/download?wrap=1) containing both positive and negative examples, also divided into transmembrane (tm) and non-transmembrane (non_tm) examples.

The files are in a Fasta-like format, in which each protein sequence is followed by a annotation line (prefixed with #) in which the n-, h-, and c-region positions are marked with n, h, and c. The actual cleavage site is marked with a upper case C. The markup after that is not important for this assignment, but it details whether a position is found in a TM region, or prefers the cytoplasmic side or non-cytoplasmic side ("inside or outside") of a TM protein, or whether it is "just" globular.

A note based on experience: you will not pass if you base your prediction on the labels instead of the amino acids.

Proteome data

You get to choose organisms yourselves, but it is suggested you download the proteome data from Ensembl's BioMart service.

Requirements

- Perform proper training and testing of your classifier, separating data for the two tasks.
- Detail your results from the testing. How well do you think you can classify signal peptides?
- Try to visualise what a typical signal peptide looks like using a **sequence logo** (http://en.wikipedia.org/wiki/Sequence_logo) . There are **online services** (http://weblogo.berkeley.edu/logo.cgi) that can help you.
- Show statistics from the application on proteomes. For example, what is the distribution of prediction scores, if you have such information.
- You must apply you predictor to two full proteomes and summarise your findings.