MOL3022 – Some possible projects

Some of these projects can be solved using traditional machine learning approaches. It is in these cases encouraged to use existing tools / libraries, rather than coding everything from scratch.

For projects with a component of machine learning, it is important that the performance is measured on a dataset that has not been used for training, e.g. a separate part of the original dataset (cross validation).

Maybe some inspiration:

https://se.mathworks.com/help/bioinfo/examples/predicting-protein-secondary-structure-using-a-neural-network.html

https://se.mathworks.com/help/bioinfo/examples.html

https://www.scipy.org/

http://hplgit.github.io/bioinf-py/doc/pub/html/index.html

https://medium.com/activewizards-machine-learning-company/top-15-python-libraries-for-data-science-in-in-2017-ab61b4f9b4a7

Signal sequences in proteins

Make a tool for predicting signals in protein sequences.

SignalP - Predicting the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms.

http://www.cbs.dtu.dk/services/SignalP/

http://www.cbs.dtu.dk/services/SignalP/data.php

ChloroP - Presence of chloroplast transit peptides (cTP) in protein sequences and the location of potential cTP cleavage sites.

http://www.cbs.dtu.dk/services/ChloroP/

http://www.cbs.dtu.dk/services/ChloroP/pages/datasets.php

Mapping known transcription factor binding sites

Make a tool that can use data on a transcription factors from Jaspar http://jaspar.genereg.net/, scan a DNA sequence, and identify the most likely transcription factor binding sites in the DNA sequence. For example draw this both as a graph showing the score at each position along the sequence, or as just symbols showing the most likely binding sites.

Predicting de novo transcription factor binding sites

Find motifs (de novo) for binding sites for a transcription factor, using for example the test set for MEME http://meme-suite.org/meme-software/example-datasets/crp0.fna.

Predicting secondary structure

Make a tool for predicting secondary structure of a protein. It is known that it is possible to use data from e.g. PSI-Blast (multiple alignment) to improve such predictions. It is not necessary to do that in this project, but this aspect should at least be discussed in the report.

http://www.compbio.dundee.ac.uk/jpred/about.shtml

https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Protein+Secondary+Structure)

Enrichment analysis of genomic tracks

Use genomic tracks from e.g. UCSC (BED files) and test if the overlap between tracks is greater than expected by random chance. It may be relevant to do this by a randomization test. It is then relevant to discuss and compare in the report different strategies for randomization, e.g. what is actually randomized, and how. It may be possible to implement different strategies for randomization, and compare results. It may be relevant to use BEDtools http://bedtools.readthedocs.io/en/latest/ for some steps in this project. Suitable data sets can be made available, for example based on overlap between tracks for histone modifications.

Enrichment analysis of genes based on genomic tracks.

Use genomic tracks from e.g. UCSC (BED files) and gene lists (e.g. affected vs un-affected genes), and test whether a given track overlap with features of the track more often than expected by random chance. It may be relevant to use BEDtools http://bedtools.readthedocs.io/en/latest/ for some steps in this project. Suitable data sets can be made available, for example based on genes associated with specific diseases or cellular processes.