Protein Prediction I Exercise

Develop a Transmembrane Predictor

- Each group:
  - Train a machine learning device
  - Predict which residues are in transmembrane region
  - Optimize and evaluate your tool
- Possibly: combine all of your tools into one big predictor (e.g. Random Forest)

Milestones

- Only one exercise sheet, but:
- Divided into milestones with deadlines
- Milestones are presented by at least one group member
- Milestone presentation consists of:
  - Steps taken
  - Problems encountered (and their solution if available)
  - Assessment of quality (if possible)
  - Next steps planned
  - No more than 3 min

Typical steps for ML development

- Data extraction
- Redundancy reduction (and fold splitting)
- Feature extraction
- Feature selection (feature optimization)
- Cross-training (parameter optimization)
- Evaluation (holdout set)
- Evaluation (competing methods)

WEKA

- Machine Learning (ML) workbench
- Includes implementations of various ML algorithms
  - Neural networks
  - Support vector machines
  - Random forests
  - Linear regression
  - Bayesian networks
- We suggest to use WEKA unless you are already familiar with specific ML implementations

WEKA

Some suggestions:

- Familiarize yourself with WEKA
- Read the tutorial on their webpage
- Download WEKA to your local machine
- Use their data examples and play around in the GUI

- http://www.cs.waikato.ac.nz/ml/index.html

The data

- Several features with different biological background (evolutionary, biochemical, annotation..)
- Features come in a window around the central position
- Each line corresponds to one position in the protein
- The first feature corresponds to the protein name and residue number
- The last feature is the target class (+) means transmembrane and (-) not in membrane

The data

```
@ATTRIBUTE swiss_annotation_available   {0,1}
@ATTRIBUTE pssm_diff                    NUMERIC
@ATTRIBUTE pssm_sign                    {0,1}
@ATTRIBUTE perc_diff                    NUMERIC
@ATTRIBUTE perc_sign                    {0,1}
@ATTRIBUTE sift_tolerated               {0,1}
@ATTRIBUTE sift_deleterious             {0,1}
@ATTRIBUTE sift_score                   NUMERIC
@ATTRIBUTE psic_diff                    NUMERIC
@ATTRIBUTE psic_sign                    {0,1}
@ATTRIBUTE pfam_diff                    NUMERIC
@ATTRIBUTE pfam_sign                    {0,1}
@ATTRIBUTE pfam_indomain                {0,1}
@ATTRIBUTE pfam_annotated               {0,1}
@ATTRIBUTE class                        {+,-}

@DATA
KP4T_UMV4_63,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
KP4T_UMV4_89,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0
Q9P2H3_swissvar_104,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,
VG6_BPPH2_10,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0
VG6_BPPH2_17,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1
VG6_BPPH2_2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
VG6_BPPH2_14,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
VG6_BPPH2_14,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
VG6_BPPH2_5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
VG6_BPPH2_5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
VG6_BPPH2_6,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
VG6_BPPH2_9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
VG6_BPPH2_9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

Your task

- The data is (largely) prepared
- You will be assigned a specific ML method
- Select the minimal set of Features that provides good performance
- Cross-train a good model (and evaluate on test sets)
- Provide a standalone, command-line executable
- Evaluate the method of one other group against the holdout set (Cooperate with the others)
- Give a final presentation – show the performance of your tool

Milestone 1 – method development

- Split the big data file into 3 subsets
- Use PSSM and hydrophobicity as base features – feel free to add others to increase performance (don't use all/too many – use the smallest possible set)
- Window size between 17-25 residues (typical size of TM helix;) Select a good window length
- Cross-train and optimize parameters of your model
- Visualize the performance of your method

Next meeting will be

- 16.05 – Group A
- 23.05 – Group B

(Presentation of Milestone 1)