Technische Universität München
Lehrstuhl XII: Bioinformatik

Prof. Burkhard Rost
Tatyana Goldberg
Maximilian Hecht
Lothar Richter
Christian Schäfer

Exercise 'Protein Prediction I'
Summer Term 2013
Computational Biology Block

General information

- Our course homepage, containing lecture slides, announcements, group lists, etc.: http://rostlab.org/cms/teaching/teaching-overview/pp-ss-13/

- Grading („Schein"):
  60% Exercise (programming),
  40% Final exam (lecture/exercise content)

- Please work on your project from the student computers in room 01.08.021. There are ten of them: i12k-biolab01 through i12k-biolab10

- You can access them also remotely using `ssh`. For example:
  `ssh ppgroup1@i12k-biolab01.informatik.tu-muenchen.de`

- At every meeting, each group has to give a maximum 5 minutes presentation on its current progress

- Each milestone is discussed in detail in the exercise – refer to the corresponding slides for more information.

Project:

During this semester you will implement a novel method for the prediction of alpha-helical transmembrane proteins. This exercise sheet shall provide a rough guidance for a successful completion of your project. The attendance of the exercises (at least one member per group) and understanding of the lecture content are indispensable.

Data Set:
`/mnt/opt/data/pp1_13_exercise/dataset/`

Arff file containing amino acid – based features:
`/mnt/opt/data/pp1_13_exercise/tmps.arff`

Weka Workbench. For implementation of your method please **only** use this version.
`/mnt/opt/data/pp1_13_exercise/weka-3-6-9/`

Meetings and Milestones:

Milestone	Group A presentations	Group B presentations
1	16.05.	23.05
2	**06.06.**	**13.06.**
3	20.06.	27.06.
4	**04.07**	**04.07**

Milestone 1: Prepare for Machine Learning

- Understand the data – the tmps.arff holds all the information
- Partition the data in three folds for training, cross-training and evaluation
- Select the minimal set of Features that provides good performance
- You must use at least all 'pssm' and 'chemprop_hyd' features
- Select a window length (range 17-25 residues) that leads to a (near) optimal solution
- Optimize the parameters of your model
- Cross-train a good model (and evaluate on test sets – see **Hints**)

Milestone 2: Prepare an executable.

- Provide a standalone command line executable
- Make sure its executable on the student machines
- For any protein sequence, your method should take the corresponding '.arff ' file as input and for each residue provide the (hopefully correct) class assignment
- Provide a RESTful interface for your method

Milestone 3: Evaluate your methods

- Each group will evaluate the method of another group
- We will provide an independent data set
- Provide figures that compare your own method to that of the other group
- This Milestone will be different for Comput. Biol. Track

Milestone 4: Final Presentation

- Every group presents its method to the audience
- Presentations should be no longer than 5 minutes (max 3 slides)
- It should cover the feature set, the algorithm and the results of your evaluation
- Until July 30th: Prepare and submit your final report as Bioinformatics Application Note. It is obligatory to stick to the guidelines and use the Word or Latex templates given in: Oxford Journals - Bioinformatics

Happy coding!

Hints:

Evaluation:
For all evaluation purposes, you should use the following measures:

- the overall performance Q2:

$$Q_2 = 100\frac{TP+TN}{TP+FP+TN+FN} = \frac{correct \quad predictions}{total \quad predictions}$$

- Accuracy positive - Acc(TM):

$$Acc(TM) = 100\frac{TP}{TP+FP}$$

- Coverage positive - Cov(TM):

$$Cov(TM) = 100\frac{TP}{TP+FN}$$

Note:
TP (True positives): Residues correctly predicted to be **in TM regions**.
FP (False positives): Residues predicted **but not observed in TM regions.**
TN (True negatives): Residues correctly predicted **not to be in TM regions.**
FN (False negatives): Residues observed but **not predicted in TM regions.**

- Accuracy negative – Acc(non-TM) and Coverage negative – Cov(non-TM) are calculated by exchanging TP by TN and FP by FN.

A complete model evaluation consists of 5 values: Q2, Acc(TM),Cov(TM), Acc(non-TM) and Cov(non-TM);
Use all five values when visualizing your method performance (plots).