Technische Universität München
Lehrstuhl XII: Bioinformatik

Prof. Burkhard Rost
Tatyana Goldberg
Maximilian Hecht

Exercise 'Protein Prediction II'
Winter Term 2012/13

Sheet 4

General information

- Contact:
  `goldberg@rostlab.org, hecht@rostlab.org`
- Send an email (one per group) to **the two of us** including the paths only (no files as attachments!) to your results and answers in **PDF** format and your **program code**. Everything has to be readable by us, so please check the permissions of your directories/files.
- This is the **final** assignment sheet
- Sheet 4 due is on Jan. 24 2013 at 11AM.
- Your prediction method should get as input a protein sequence in FASTA format and output the predicted localization class. It should be executable for us.
- We will have two more meetings: on **Jan. 24** we will discuss your methods and results and on **Jan. 31** you will give a 5 minutes presentation (present algorithm and results) in front of the group and Prof. Rost

Assignment (250 Points)

1. Select and familiarize yourselves with one of the following **homology search tools**:
   - PSI-BLAST `(man blastpgp)`
   - HHblits `(man hhblits)`
   - Jackhmmer `(man jackhmmer)`

2. In the last exercise we created three splits for each dataset (at HVALs=0, 30 and 70). For the development of your prediction model we will use the HVAL=70 data set only. Now, we will use the three splits of the HVAL=70 dataset to create **search databases** for your tool of choice:
   - for PSI-BLAST use `formatdb`
   - HHblits (please contact us)
   - for Jackhmmer simply use a multi sequences Fasta file

3. All three methods heavily rely on the quality of sequence profiles. This means that for each sequence you should create a search profile upon a large number of protein sequences. This can be done by running a few iterations against the 80% non-redundant database combining Swiss-Prot, Trembl and PDB (big_80 database) and saving out the profile. (The option for PSI-BLAST is `'-Q'` and for Jackhmmer `'-chkhmm'`). This profile can then be used as input to the corresponding method instead of the amino acid sequence. Thus, for each sequence in your entire HVAL=70 set make a run against the `big_80` database and store their profiles for further use:

- use 3 iterations on `/opt/rost_db/data/big/big_80` for PSI-BLAST
- use 3 iterations on `/opt/rost_db/data/big_80.fasta` for Jackhammer
- no need to do this for HHblits (if you plan on using HHblits, please contact us)

4. Think of an accurate prediction algorithm for finding hits with correct localization annotations. Your prediction could be based on the localization annotation of:

- a top hit
- max. of top n hits
- weighted max. of top n hits (e.g. scoring by e-value, HVAL, seq identity, ...)
- weighted max. of top n hits and properties (e.g. organism, sequence length, amino acid composition, …)
- be creative! :)

5. Now you are ready to train your homology inference based localization prediction method. We will use the three-fold cross validation set-up, which consists of three splits that we in the following call: **train**, **cross-train** and **test** sets. The training procedure is as follows:

- Use **cross-train** as your query sequences (or profiles for PSI-BLAST and Jackhmmer searches [step3])
- Search against **train** (i.e. the database consisting of the **train** sequences [step 2])
- Score and rank the hits you find such that the prediction of query sequences becomes optimal
- Now, use **test** as query against **train** applying **the same scoring scheme as before**. Evaluate the performance (see Hints: Evaluation)
- Rotate your sets and repeat this procedure such, that each of the three splits is used for training, cross-training and testing exactly once
- The method performance is the average over the test results of the entire rotation

6. Repeat steps 3-5 until you find your best performing algorithm

7. Evaluate your best performing algorithm on other HVAL sets (HVAL=0 and 30) using the three-fold cross validation set-up again (Hint: you do not have to do the training again, only use the corresponding **test** set of each rotation as query against **cross-train**). Report the average performance over the three **test** sets for each HVAL in a table (See Hints: Evaluation)

8. Evaluate the performance of your best performing algorithm on **independent test sets** (i.e. sets that were not involved into any step of the training procedure). These test sets are the 'After May 2011' sets. The database is the entire HVAL=70 'Before May 2011 set'. Does the performance change with the changing HVAL of the 'After May 2011' sets? How well does your method perform when using other databases (Before May 2011: HVAL=0/HVAL=30)?

9. Compare the performance of your method to the following sequence-based localization predictors on the three independent data sets:

- LocTree2 (https://rostlab.org/~loctree2/)
- WoLF PSORT (http://wolfpsort.org/)
- CELLO v.2.5 (http://cello.life.nctu.edu.tw/)

Hint: don't forget to split your data into plant, fungi and animal if necessary for comparison
Summarize your results for questions 8 and 9 in a table and briefly describe the algorithm of the three predictors.

10. Prepare a PDF report describing the algorithm of your prediction method, its usage (what are the input, the output and the parameters) and the evaluation results (no more than 3 pages).

Hints:
Hhblits database: `/opt/rost_db/data/pp2exercise/uniprot20_02Sep11`

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

- the overall performance Q10:

$$Q_{10} = 100 * \frac{TP}{TP + FN} = 100 * \frac{correct \quad predictions}{total \quad predictions}$$

- Accuracy for every class - Acc(L):

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

- Coverage for every class - Cov(L):

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

- Geometric average of accuracy and coverage for every class – gAv(L):

$$gAv(L) = \frac{\sqrt{Acc(L) + Cov(L)}}{100}$$

Note:
TP (True positives): Correctly predicted proteins of localization L.
FP (False positives): Proteins predicted to be localized in L but in fact are **not in L**.
FN (False negatives): Proteins that are localized in L but **predicted to be not in L**.

A complete model evaluation consists of 31 values: Acc(L), Cov(L) and gAv(L) for each of the ten localization classes and one combined Q10 value.

Method presentation:
Make it short! Max. 5 slides / 5 min. Describe the algorithm (scoring scheme, idea behind it, …) and the results (cross-validation and independent test set).

Final method:
Send us a path to the fully functional method. In your report, describe exactly how to call the program. Make sure it takes a FASTA file as input and returns the localization class as output. Handle everything else inside your code. Test your program, if it does not work you will not get credits!

Need help?
Feel free to contact us if you have any questions or problems. You can also come to our offices during the usual lecture/exercise time.