Technische Universität München
Lehrstuhl XII: Bioinformatik

Prof. Burkhard Rost
Tatyana Goldberg
Maximilian Hecht

Exercise 'Protein Prediction II'
Winter Term 2012/13

Sheet 3

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 (text answers in **PDF**, program code, fasta files).
  Scripts should be executable for us so that we can reproduce your results.
  Everything has to be readable by us, so please check the permissions of your directories/files.
- Sheet 3 due is on **Monday, December 10, 2012 at 11AM**.

Exercise 1: Redundancy Reduction

Since our performance results should not be biased by those over-represented sequences, the next step would be to reduce this redundancy. The common tool for this purpose is Uniqueprot.

- Familiarize yourself with the command line options of Uniqueprot
- Now, take the Swiss-Prot annotated data set and run the redundancy reduction on a
  1. 20%,
  2. 50% and
  3. 90% sequence identity threshold (PID value in the Uniqueprot paper).

  Hint: use one round of redundancy reduction for PID=90% at e-value=1e-3; two rounds for PID=50% at e-values of 1e-3 and 1; three rounds for PID=20% at default e-values

  Since the maximal length of known localization signals can reach up to 35 residues, please set the parameter regulating the minimal number of aligned residues to this threshold
- Store the reduced sequence unique sets in three separate files
- How many sequences do you find now in your redundancy-reduced sets?
- What is the ratio compared to the unreduced set?
- How many instances are there for each localization class?

For the successful runs of Uniqueprot:

- use one biolab machine per group
- run one set at a time
- configure the Uniqueprot configuration file to use all four CPUs in parallel and a temp folder that you need to create in your home directory as a working directory. Hand in the configuration file together with your fasta files.

Exercise 2: Data preparation

We are one step away from the development of our homology-based prediction method. Now, we need to split our protein sets into training and test data.

Data set for Testing

Since current state-of-the-art methods for localization prediction, the methods we are going to compete against (e.g. LocTree2), were developed before Mai 2011 we will split each of our redundancy reduced data into two subsets: one containing proteins added to Swiss-Prot before May 2011 and the other one with all others. We will use the former set for the training of our prediction method and the latter for the performance comparison against other methods.

- For each redundancy level split your data into the two sub-sets: 'before May11' and 'after May11'
- For each redundancy level: how many proteins remain in each sub-set?
- What is the class distribution in the sub-sets?

Data set for the Development

For the development of our prediction method we will put the 'after May11' set aside and will only use the 'before May11' set. We will develop our method on the PID 90% reduced set and then apply the same model to the 20% and 50% reduced sets.

- For this purpose, take the 'before May11' set of the PID 90% reduced data and split it into three equally sized sub-sets such that each sub-set has the same class distribution as the original set (i.e. **stratified** splitting).
- How many instances of each localization class do you have in the three sub-sets? Display the class distribution in a matrix.

In the next exercise, you will use one sub-set for training, one for cross-training and the last one for testing. You will rotate your sets such each of them will be used for testing exactly once. This procedure is called **cross-validation**.

Exercise 3: Sequence homology search tools

Please name three sequence homology search tools of your choice.

- Briefly describe their main idea and how they work.
- How are their results being displayed?
- What are the parameters defining sequence similarity?
- Do you have exeperience working with these methods?
- Don't forget to provide literature references for each of them.