Technische Universität München
Lehrstuhl XII: Bioinformatik

Prof. Burkhard Rost
Tatyana Goldberg
Maximilian Hecht

Exercise 'Protein Prediction II'
Winter Term 2012/13

Sheet 2

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 (answers, program code, figures, …).
  Scripts should be executable and readable for us so that we can reproduce your results. Please check the permissions of your directories/files.
- Sheet 2 is due on November 21, 2012 at 11AM.

Exercise 1: Uniqueprot (10 Points)

According to the paper http://nar.oxfordjournals.org/content/31/13/3789.full.pdf please answer the following questions:

- Is it possible to infer functional similarity from sequence similarity? Discuss.
- The standard value to measure sequence similarity is the HSSP-value. Describe this measure with your own words. How are it's parameters derived?
- The default setting for clustering the protein sequences in the Uniqueprot algorithm is 'largest first'. What does it mean? What is it's advantage?

Exercise 2: Redundancy Reduction (12 Points)

In the last exercise, we parsed the Swiss-Prot database for protein sequences with localization annotations. Usually, protein sequence databases contain a lot of redundancy in terms of identical or similar sequences. 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).

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

Additionally, answer the following questions:

- How many sequences do you find now in your redundancy-reduced sets? What is the ratio compared to the unreduced set?

Exercise 3: HSSP curve (20 Points)

The HSSP curve is a curve that is used to discriminate between proteins of similar and non-similar structure by relating alignment length and pairwise sequence identity. The HSSP curve has also been proved to accurately reflect similarity in sub-cellular localization.

According to the HSSP-value formula in the Uniqueprot paper, please draw using a programming language of your choice the HSSP curves at following thresholds:

- HSSP-value = 0
- HSSP-value = 30
- HSSP-value = 70

Now, take the set of proteins of a localization class with the least number of members extracted from the Swiss-Prot database. Using this set, perform an all-against-all BLAST search and visualize each pairwise alignment as a point in the HSSP graph.
Hint: a set of 300 protein sequences will result in 90.000 points in the graph.

Mark the alignments of proteins left after the redundancy reduction at PID=90% green, the alignments of proteins left after the redundancy reduction at PID=50% blue, and leave the remaining alignments black.

Please visualize all your results in a single graph.