Technische Universität München
Lehrstuhl XII: Bioinformatik

Prof. Burkhard Rost
Tatyana Goldberg
Maximilian Hecht

Exercise 'Protein Prediction I'
Winter Term 2012/13

Sheet 1

General information

- Our course homepage, containing lecture slides and exercise sheets:
  `http://rostlab.org/cms/teaching/teaching-overview/pp2-ws-12-13/`
- Time and place: **Thursday, 12:30 – 13:15 (we may start earlier, depending on the lecture!)**, room **MI 01.09.034**
- Contact:
  `goldberg@rostlab.org, hecht@rostlab.org`
- The email distributor for all exercise participants: `pp2.2012@rostlab.org`
- Grading („Schein“):
  - **40% exercise** (theory/programming), **60% final exam** (lecture/exercise content)
  - minimum of exercise points required for exam attendance
- Please build groups of 2-3 people
- 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 goldberg@i12k-biolab01.informatik.tu-muenchen.de`
- 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 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 1 due is on November 08, 2012 at 11AM.

Exercise 1: Protein databases (15 Points)

The UniProt Knowledgebase is one of the largest online resources for protein sequence data.
Referring to http://www.uniprot.org please answer the following questions:

a) What is the UniProt Consortium? Who are its members? Where do they come from?
b) The UniProtKB is the database we shall be using for the development of our projects. It consists of two sections.
   - What are their names and what is the difference?
   - For each section: How many sequences does it contain?
   - For each section: Which are the 3 most represented species?
   - For each section: What is the most frequent type of evidence for protein existence?
   - Explain each type of evidence with one short sentence. Which one do you trust most? Why?

c) The UniProt offers more than only sequences.
  - ○ What is UniRef? What does UniRef90 or UniRef50 mean?
  - ○ What is a sequence cluster? What do sequences in the same cluster have in common?
  - ○ Take a look into the UniRef50: How many sequences are alone in their cluster? What does that mean?

Exercise 2: Text search (12 Points)

We will find some UniProt entries for human proteins using simple text mining. Please use the database homepage http://www.uniprot.org to answer the following questions.

a) How many reviewed entries do you find?

b) How many of them are annotated to be localized in the nucleus? Describe your search criteria.

c) Pick one entry:

- When was it added to UniProt? When was it last modified?
- What is the location of the described molecule in the cell? How was the location determined? How reliable do you consider this type of annotation?
- What is the length of the molecule?
- How many accession numbers does the entry have? How many can it have? Why?
- How many entry names does the entry have? How many can it have? Why?
- How many related protein sequences produced by alternative splicing of the same gene are there? Do they have localization annotations?

Exercise 3: Data set retrieval (23 Points)

Now we will build the data set from the current release of the Swiss-Prot database 2012_09 for the development and evaluation of our prediction methods. The Swiss-Prot database is located in `/opt/rost_db/data/trembl/taxonomic_divisions`.

Please extract eukaryotic protein sequences that:

- have exactly one subcellular localization annotation (How do you check this?)
- do not have non-experimental localization annotations (Which keywords tell you that?)
- are longer than 10 and shorter than 2045 amino acid residues
- have their origin in plants, fungi or animals
- belong to one of the following ten localization classes: exra-cellular (also called secreted), Golgi body, endoplasmic reticulum, chloroplast, mitochondria, peroxisome, vacuole, cytoplasm, nucleus, plasma membrane (please find definitions of the locations in `http://www.uniprot.org/docs/subcell`)

Your data set should be written in a FASTA formated file and contain for each protein sequence the following information:
>entry_id`(ID line)`#[plant|animal|fungi]`(OC line)`#entry integration date`(DT line)`#subcellular location annotation `(CC -!- SUBCELLULAR LOCATION line)`
amino acid sequence `(SQ line)`

How many protein sequences for each localization class did you find? Please summearize your result in a table.

Hints

- During your work, you will certainly come across questions, which are not answered on this sheet and you cannot answer yourselves. In such a case, feel free to email them to us, we will collect and try to solve them in the Exercises on Thursday. Please understand that questions sent in after Wednesday morning might not be answered on the Thursday of the same week, but instead on the Thursday of the week after. The same holds true for issues regarding the computational resources of the student machines

- To enable us to follow your work, please also write a short report containing *rough* descriptions of the steps you have taken during the implementation and evaluation and hand it in together with the usual paths to your solutions. It should not exceed 5 pages and 2 pages can be enough for a simple, yet correct solution. You can use pseudocode to describe more sophisticated issues.

Happy codding!