Technische Universität München
Lehrstuhl XII: Bioinformatik

Prof. Burkhard Rost
Tobias Hamp
Esmeralda Vicedo

Exercise 'Protein Prediction II'
Winter Term 2011-2012

Sheet 2

General information

- Our course homepage, containing lecture slides and exercise sheets: http://rostlab.org/cms/teaching/teaching-overview/hs-ws-11-12/
- Time and place: **Friday, 12:00 – 13:30**, room **MI 01.09.034**
- Grading („Schein"):
  - **60% Exercise** (theory/programming),
    40% Final exam (lecture/exercise content)
  - T.b.d.: minimum of exercise points required for exam attendance
- Contact:
  hampt@rostlab.org, vicedo@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 ,...) until **February 1, 10:00 am**. 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.

Exercise 1: Development of a Protein Function Predictor

1. Introduction

During the course of this Exercise, we will establish methods for predicting Gene Ontology (GO) annotations for given protein sequences. These programs are supposed to use various sources of information which individually capture different aspects of protein function. As such, they are going to complement or even replace existing predictors which work by exploiting sequence homology as found by BLAST. The great goal is to create an integrated state-of-the art function predictor ranking among the best of its kind in the next CAFA challenge.

Due to the extent of the task, we plan this to be the last Exercise sheet. 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 Friday. Please understand that questions sent in

after Thursday morning might not be answered on the Friday of the same week, but instead on the Friday 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 1 page can be enough for a simple, yet correct solution. You can use pseudocode to describe more sophisticated issues.

2. Rules
CAFA wants the methods and results to follow certain formats. You can find the official rules here:
http://biofunctionprediction.org/node/20

Please make sure your solutions follow these rules. In case you are not sure about certain details, again do not hesitate to ask us.
Beside the CAFA rules, please also make your programs comply with the following conditions:

- No inference of GO terms by BLASTing them against a database with already annotated protein (this is already implemented)
- Only 3 input parameters (Simple wrappers of programs with more parameters are allowed):
    1. a path to the required external data (e.g. a SwissProt database)
    2. a path to a fasta file with target protein sequences
    3. a path to the output folder
- Output in CAFA format, excluding the accuracy line.
- No hard coded sequences (or similar) together with their GO annotations. The only source for annotations and information must come from data provided by parameter 1.
- [to be extended]

Excellent solutions should optimize free parameters and implement so-called *individual predictions* (see 4.).

We will use independent data sets after the hand-in to validate your results.

3. Hints
Resources helpful for the prediction of protein function include (but are not limited to):
- HHblits (instead of BLAST)
- EC Number Predictors
- Protein Interaction Data
- Expression Data

You can find various helpful tools and data in the folder /mnt/opt/data/pp2_11_12. See the corresponding readme file for more information.

All of the above is going to be introduced in more detail throughout the course of the lecture.

4. Restriction to the Scoring Scheme

In order to make predictions more consistent with the underlying biology, we introduce the concept of individual predictions. An individual prediction is defined as a set of GO nodes which contains exactly one leaf node and all of its parents. A leaf node is a predicted(!) node which is not a parent of any other predicted(!) node.
Make your scoring scheme adhere to individual predictions by implementing the following steps:

1. Calculate a ranking of all individual predictions (e.g. based on the scores of the terms it contains).
2. In the second best individual prediction: exclude all nodes which are already contained in the best individual prediction.
   In the third best individual prediction: exclude all nodes which are already contained in the second best individual prediction etc.
3. Make sure that the score of any term in the best individual prediction is higher than every score in the second best prediction etc.

The distribution of scores within one individual prediction is still up to you.

Happy coding!