Bioinformatics III

Prof. Dr. Volkhard Helms — Saarland University
Daria Gaidar, Markus Hollander, Duy Nguyen, Thorsten Will — Chair of Computational Biology
Summer Semester 2018

Exercise Sheet 7

Due: 12.06.2018 10:15

Submit your solutions on paper, hand-written or printed at the *beginning* of the lecture or in building E21, Room 3.03. Alternatively you may send an email with a single PDF attachment. If possible, please include source code listings. Additionally hand in all source code via mail to duy.nguyen@bioinformatik.uni-saarland.de.

Missing Data Imputation and GRNs Reconstruction

Exercise 7.1: Missing Data Imputation (50 points)

Missing observations are frequently encountered in biological data. There are several reasons which may cause missing data: the protein/gene is actually present, but it's not detected or falsely identified; or the protein/gene abundance level is below the detection limit of the instrument; or the protein/gene is not present at all. Many imputation methods have been developed to tackle this problem, but due to the complicated nature of the missing data, different imputation methods should be used depending on which mechanism that led to missing data.

In this exercise, we will perform imputation based on a given data distribution. The main idea behind this method is to impute missing proteomics data which have expression below the detection limit (http://www.nature.com/nmeth/journal/v13/n9/fig_tab/nmeth.3901_SF3.html). Basically, the imputation can be broken down into the following steps:

1. Calculate the mean and standard deviation of the current data.
2. Derive the new mean and standard deviation for the missing data based on the current distribution. The new mean should be in the lower quantile of the distribution since we want to simulate the low expression data. The new standard deviation could be derived by taking a fraction of the current standard deviation.
3. Generate the new data based on the new mean and standard deviation from the previous step.

(a) The *ms_toy.txt* file contains an example of proteomics data with 6 samples (columns). Use any of the 6 samples and write a script to impute the missing data (which are indicated by "NA") for the sample of your choice by following the steps mentioned above. (25 points)

Hint: in R, use function `qnorm` *to derive the new mean and function* `rnorm` *to generate the data.*

(b) Play around with different new means and standard deviations. Plot the distribution of the sample with the imputed data in a similar fashion as Fig. 1. What is the effect of different means and standard deviations? (25 points)

Hint: in R, use function `hist` *and function* `plot`

Exercise 7.2: DREAM challenge (50 points)

Apply one of the three models introduced in the lecture (Noise, Linear and Sigmoidal) to predict the directed unsigned GRN topology of E.coli from steady state and time series gene expression data. The target network is of size 10 genes without self-regulatory interactions.
Download the gene expression dataset and use the following:

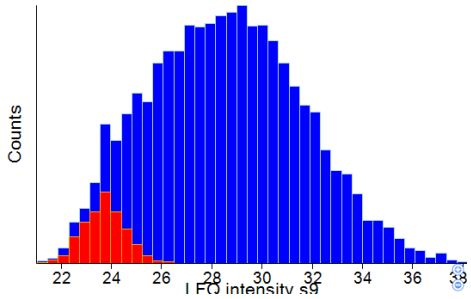

Figure 1: Sample Plot (blue is the overall distribution, red is the imputed data)

- *heterozygous.tsv* contains the steady state levels for the wild-type and the heterozygous knock-down strains for each gene. Thus, for a network of size 10 there are 11 experiments (wild-type plus knock-down of every gene).
- *null-mutants.tsv* contains the steady state levels for the wild-type and the null-mutant strains for each gene. Thus, for a network of size 10 there are 11 experiments (wild-type plus knock-down of every gene).
- *trajectories.tsv* contains time courses of the network recovering from several external perturbations. For the network of size 10, you have 4 perturbations (each one with 21 time points).

Finally, the expected output to be submitted should be a ranked list of regulatory link predictions ordered according to the significance of each prediction (sample output file is in the supplementary).

For example: G1 G2 score

Where G1 and G2 are two different genes (no self-interactions). Links are directed: the gene in the first column regulates the gene in the second column. (If both G1 regulates G2 and G2 regulates G1, then both lines should be included). Score is between 0 and 1 and indicates the confidence level you set to this link prediction.

Have fun!