

Exercises Algorithmic Systems Biology

Freie Universität Berlin, SoSe 2011

Roland Krause · Martin Vingron

Assignment 1 · to be handed in via E-Mail until 2011-04-28

Exercises can be performed in groups of two. Clearly label all sheets and provide them in a format suitable for publication. In particular, label axis of plots.

Exercise 1 (Normalization). Perform quantil-normalization on the following data set:

Gene	array1	array 2
g1	1	3
g2	19	11
g3	3	9
g4	5	17
g5	15	13
g6	11	21

Exercise 2 (Correlation coefficient). Calculate the Pearson correlation coefficient of the following data set: $(1, 2), (3, 4), (5, 3), (9, 5), (7, 5)$ using the product of the standard scores.

Exercise 3 (Differential genes and false discovery rate). 1. Simulate the gene expression of 1000 genes with 50 significantly changed between two samples a, b .

- Draw the non-differentially expressed genes from a distribution with $\mathcal{X} \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 10$ and $\sigma = 5$.
 - The differentially expressed genes are generated for sample a as before. For sample b , add values from $\mathcal{N}(5, 1)$.
 - Repeat the procedure to obtain 4 replicates. Use the following with 2, 3 and 4 replicates.
2. Calculate a t-test for each gene. Do you need to use a paired t-test?
 3. Plot the resulting test statistic in ascending order. Include a line from the origin to the highest point.
 4. Let's use an $\alpha = 0.05$. How many non-differentially expressed genes do we expect to see by chance and how many do we see?
 5. Apply the Bonferroni correction. How many genes are significantly expressed now?
 6. How many genes are significantly expressed if we apply a chosen FDR as introduced by Benjamini and Hochberg?
 7. Graphically interpret the procedure in the plot of the p-values.
 8. What is a reasonable FDR in this setting? What are the practical consequences?
 9. How would you have to select the FDR when using two replicates?

Exercise 4 (Protein-protein interaction data). Search the literature for a protein-protein interaction data set of your choice and perform a first pass analysis.

1. Search for a protein-protein interaction data set. Give the correct citation of the manuscript. Summarize the method.
2. Find the data in a suitable data base, e.g. IntAct¹.
3. Choose a graph representation of the data set and motivate your choice.
4. Determine clustering co-efficient, degree distribution and average path length in the graph.
5. Write a generative model to build a graph with the same ($\pm\epsilon$) number of edges and vertices. Describe your choice of the model.
6. Compute clustering co-efficient, degree distribution and average path length in the graph for the model graph.
7. Discuss.

¹<http://www.ebi.ac.uk/intact/>