Bioinformatics III

Prof. Dr. Volkhard Helms
Andreas Denger
Winter Semester 2019/2020

Saarland University
Chair for Computational Biology

Exercise Sheet 6

Due: Nov 28, 2019 14:15

Submit your solutions on paper, hand-written or printed at the *beginning* of the lecture. Alternatively, you can send an email with a single PDF attachment to andreas.denger@bioinformatik.uni-saarland.de. Your submission should include code listings for programming exercises. Additionally, hand in a .zip file with your source code via email.

Differential Gene Expression Analysis and Boolean Networks

Exercise 6.1: Differential Gene Expression Analysis (50 points)

In this exercise, you are going to perform a differential gene expression analysis in order to find genes which are significantly up- or downregulated between two conditions. Specifically, you are going to investigate the transcriptomes of bronchial epithelial cells in two types of human donors: those who are currently smoking cigarettes (labeled S), and a control group of people who have never smoked in their lifetime (labeled N).
The file provided in the supplement contains a table of expression values from single-channel microarrays. The columns correspond to donors, the rows to the probes on the microarray platform. The first two columns contain the identifier for each probe, as well as the associated gene symbol. Note that gene symbols can occur more than once.
So far, the dataset has been downloaded from Gene expression omnibus (GEO), the number of samples has been lowered in order to improve the running times of your programs, and a *quantile normalization* has been applied. Finally, the probes were annotated with gene symbols and those without an associated gene name, as well as reference probes used for normalization, were removed.

(a) Briefly explain when and why performing the normalization step is necessary to get meaningful results from your analysis.

(b) Write a Python program that reads the table into a data structure. Create methods that calculate the following values for every probe $i \in \{1, \ldots, n\}$ in the dataset:

(1) Let S_i and N_i be the lists of expression values for probe i with condition S and N, respectively. Calculate the *arithmetic means* μ_{S_i}, μ_{N_i} and the *variances* $\sigma^2_{S_i}, \sigma^2_{N_i}$.

(2) Calculate the *log fold change* ($logFC$) between the two conditions:

$$logFC_i = log_2\left(\frac{\mu_{S_i}}{\mu_{N_i}}\right)$$

What is the benefit of taking the logarithm of the fold change?

(3) Since $logFC_i$ alone is not a statistical test, we need some way of calculating a confidence level for the fold changes. For this reason, we are going to perform a gene-specific t-test to calculate a p-value p_i from S_i and N_i. You are can use the *stats.ttest_ind* function from the *scipy* package (with parameter $equal\_var = True$).

Shortly explain the meaning of the *p-value* returned by this function. Which problems could this specific function for calculating p-values face on this dataset? What would be a possible solution? *(Hint: Take a look at the previously calculated variances)*

(4) The next step is multiple testing correction. For simplicity, we are going to implement the *Bonferroni*-method. The adjusted p-value p_i^{adj} is calculated from for p-value p_i as:

$$p_i^{adj} = min(1.0, n * p_i)$$

where n is the number of tests (i.e. the number of p-values). Why is p-value adjustment necessary? Which benefits and disadvantages would the *Benjamini-Hochberg*-method bring compared to *Bonferroni*?

(c) Next, write a function that creates a *volcano plot* from the previously calculated values. To get the typical volcano plot shape, you first need to transform the *unadjusted* p-values to the negative tenth logarithm:

$$p_i^* = -log_{10}(p_i), p_i \in \{p_1, \ldots, p_n\}$$

Then, create a *scatter plot* with the package *matplotlib*, with the $\{logFC_1, \ldots, logFC_n\}$ values on the horizontal axis and the corresponding transformed p-values $\{p_1^*, \ldots, p_n^*\}$ on the vertical axis. Include the plot in your submission and briefly describe what you see.

(d) Finally, filter your table for those probes with a statistically significant *adjusted p-value* (i.e. probes i with $p_i^{adj} < 0.05$). Write a tab-separated file that, for each of these probes, contains the following columns:

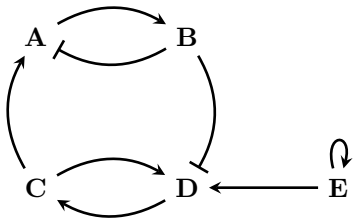
- Probe identifier
- Gene symbol
- Mean and variance for conditions N_i and S_i
- Log fold change $logFC_i$
- Adjusted and unadjusted p-value (p_i^{adj}, p_i)

Include the file in your submission.

(e) Take a look at the probes i in your filtered table from Ex6.1(d) that are up-regulated in smokers by at least a factor of four (i.e. where $logFC_i \geq 2$). Can you find any genes where this up-regulation makes intuitive sense in the context of exposure to cigarette smoke? You can find short descriptions of gene functions by searching UniProt or NCBI for the gene symbols.

Exercise 6.2: Boolean Network (50 points)

Consider the following network, which describes the mutual regulation of the hypothetical genes **A** to **E**. A line with an arrowhead denotes an activation while a flat end denotes an inhibition, i.e., if **A** is high, **B** is activated, whereas high levels of **B** inhibit the expression of **A**.



To investigate the behavior of this network use a dynamic simulation as introduced in the lecture with a synchronous update scheme.

Assume that an activation has a weight of 1, while an inhibition is always 3 times stronger than an activation. Set all thresholds to 0.

(a) **Weighted Interactions**
Set up the propagation matrix that relates the states of the genes **A** to **E** in the next iteration to the current state.

(b) **Implementation**
Write a program to simulate the Boolean Network.

To enumerate the initial states, convert the binary levels of the genes into an integer where **A** determines the least significant bit and **E** the most significant one. An initial state where, e.g., only **A**, **C**, and **D** are on high levels would translate into $1 + 4 + 8 = 13$.

(1) When does it make sense to stop the propagation and why?

(2) Which sequences do you get when you start from states 7, 13, 17, and 23?

(c) **Periodic Orbits**
To determine the attractors and the corresponding basins of attraction, go through all possible initial states and save at which state the Boolean network closes its orbit.

(1) List these orbits with their respective lengths and basins of attraction.

(2) Give the relative coverages of the state space by the basins of attraction.

(d) **Interpretation**

(1) Give the attractors in terms of active genes and characterize them with a few words.

(2) Which are the special genes and what are their respective effects on the behavior of the network? For this, explain what is determining the period of the orbits.