

## Bioinformatics III

Prof. Dr. Volkhard Helms  
Andreas Denger  
Summer Semester 2021

Saarland University  
Chair for Computational Biology

### Exercise Sheet 7

Due: June 10, 2021 12:00

Submit your solutions to [andreas.denger@bioinformatik.uni-saarland.de](mailto:andreas.denger@bioinformatik.uni-saarland.de) with two attachments: (1) A ZIP file containing all your source code files, potential result files, figures and whatever else is needed to generate your solution, (2) a PDF file containing your answers. Subject of the email should be in the following format: BI3 A7 *LastName1 LastName2*.

## Co-expression Correlation and Master Regulatory Genes

### Exercise 7.1: Co-expression based on Correlation and Mutual Information (60 points)

*Mutual information* ( $I$ ) and *Pearson correlation coefficient* ( $Corr$ ) between two random variables are defined as:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \times \log \left( \frac{p(x, y)}{p(x) \times p(y)} \right)$$
$$Corr(X, Y) = \frac{\sum_{i=1}^n (x_i - \mu_X) \times (y_i - \mu_Y)}{\sqrt{\sum_{i=1}^n (x_i - \mu_X)^2} \times \sqrt{\sum_{i=1}^n (y_i - \mu_Y)^2}}$$

, where  $p(x, y)$  is the *joint probability distribution* of expression levels  $x$  and  $y$ ,  $p(x)$  is the *marginal probability* of expression value  $x$ , and  $\mu_X$  is the (arithmetic) mean expression for gene  $X$ .

- (a) Calculate the Pearson correlation coefficient and mutual information for the data given below. Here, the data is comprised of two genes whose expression were measured over 6 time series. An expressed gene is denoted by value 1. Solve this task by hand.

Gene	t1	t2	t3	t4	t5	t6
gene1	0	0	1	1	1	0
gene2	0	1	0	0	1	1

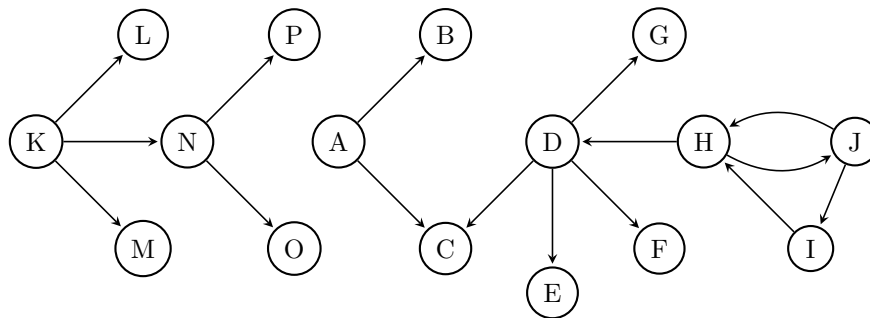
- (b) Explain the main difference between mutual information and Pearson correlation.
- (c) What is the advantage of using rank-based correlation coefficients?
- (d) Write a program that reads the time-series gene expression data given in the supplement and calculates the Pearson correlation coefficients for all pairs of genes.
- (e) Plot the distribution of correlation coefficients between pairs of genes, for example by using the `displot` function from the Python package *seaborn*. Ignore pairs that contain the same gene twice.
- Interpret the shape of the plot and include it in your submission.
- (f) Take a look at the correlation scores between the gene *ITGB2* and the other genes. Write a function that finds the gene with the:
- Highest correlation to *ITGB2*
  - Lowest correlation to *ITGB2*

- Correlation to *ITGB2* that is closest to zero

Next, for each of these three genes, create a scatter plot with a linear regression model fit between its expression values and those of *ITGB2*, for example with the [regplot](#) function from the Python package *seaborn*.

Include the three plots in your submission and describe what you see.

### Exercise 7.2: Identification of master-regulatory genes (40 points)



- Which *dominating sets* exist in the network shown above?
- What is the *minimum dominating set* (MDS) of this network?
- List the following sets of nodes and their sizes:
  - Largest connected component in the directed graph
  - Largest strongly connected component in the directed graph
  - Largest connected component in the underlying undirected graph

Find the *minimum connected dominating set* (MCDS) for each of the three sets.

- Compare the MDS and MCDS in terms of size.