Jun.-Prof. Dr. Thomas Schultz
Michael Ankele (ankele@cs.uni-bonn.de)
Shekoufeh Gorgi Zadeh (gorgi@cs.uni-bonn.de)

University of Bonn
Institute of Computer Science II
November 28, 2016

Winter term 2016/17

Bioinformatics II

Assignment Sheet 5

If you have questions concerning the exercises, please write to our mailing list:
vl-bioinf@lists.iai.uni-bonn.de.

Please submit your results for this and all previous assignment sheets by December 6th, before class (9.30 a.m.). *Submit a single .zip or .tar.gz archive that should contain all your code, as well as screenshots and answers to questions, clearly sorted by assignment. Please avoid leaving answers as comments inside your code. Either let your code print out the answer or write it down in a separate text file. Send that archive by e-mail to gorgi@cs.uni-bonn.de.*

Exercise 1 (Graph Visualization, *25 Points*)

In the final part of the first project, you will learn how to use the graph visualization package Graphviz via its Python bindings.

a) Compute the correlation between all the variables. Using your data set, write down a code to fill in the missing values and then compute the pairwise Pearson correlation between variables in a matrix. (5P)

b) Install the Graphviz library and its Python interface. Generate and visualize some simple graph. You can find the software and its documentation at `https://pypi.python.org/pypi/graphviz`. (5P)

c) Create a graph from the correlation matrix and visualize it. Represent each variable as a node in the graph. Insert an edge between two variables whenever the Pearson correlation between them exceeds the threshold $\rho > 0.5$. (4P)

d) Modify the visual attributes of edges to reflect the magnitude of the correlation. (3P)

e) Produce an alternative visualization in which there are three subgraphs, one containing all the nodes that have at least one correlation more than 0.65 to other nodes, another subgraph for nodes with $0.5 < \rho_{\max} <= 0.65$ and the last subgraph for the remaining nodes. The "class" node should be outside of all subgraphs. (5P)

f) Answer the following questions:

- At the selected threshold, which nodes are disconnected from the rest of the graph and what do they indicate? (1P)
- If two nodes A and B are strongly correlated, and node C is strongly correlated with node B, can we conclude that node C will be also strongly correlated with node A? (1P)
- Where have you already seen most of the nodes in the subgraph with $\rho > 0.65$? (1P)

Good Luck!