Jun.-Prof. Dr. Thomas Schultz
Michael Ankele (ankele@cs.uni-bonn.de)
Amin Abbasloo (abbasloo@informatik.uni-bonn.de)

University of Bonn
Institute of Computer Science II
November 9, 2015

Winter term 2015/16

Bioinformatics II

Assignment Sheet 3

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

We strongly encourage you to continuously work on the assignments and contact us with questions. However, you will only have to hand in your results (for all sheets of the first project) on December 1.

Exercise 1 (Producing a Scatterplot Matrix, *25 Points*)

In the previous assignment, you wrote a reduced dataset to disk that is limited to four classes and expression levels of the five proteins that distinguish most strongly between normal (c) and Ts65Dn (t) mice. This week, you will create and interpret a basic visualization of that data.

Your final visualization should be a 5×5 matrix whose rows and columns are the five proteins you selected last week. Diagonal cells visualize how expression levels of the corresponding protein are distributed; off-diagonal cells visualize the relationship between expression levels of pairs of proteins.

Please proceed in the following steps and submit your final script, the final image, and answers to the questions:

a) Each diagonal cell should contain four overlaid density plots, one each for the four classes c-CS-m, t-CS-m, c-CS-s, and t-CS-s. In the density plot, protein expression levels should be on the x axis, the frequency of observing that expression level in each class should be on the y axis. Use different colors and markers to distinguish between the classes, and add a legend. Your visual design should make it easy to answer the following questions (5P for implementation, 1P for justifying choice of colors, 3P for answering questions):
   - Comparing expression levels with the drug memantine (m) to ones with saline solution (s), for which proteins did the drug have a clear effect of making the expression levels in Ts65Dn mice more similar to those found in normal mice?
   - For which proteins did the drug have a strong effect on the normal mice?
   - For which protein(s) did the drug leave expression levels almost unchanged, independent of the genotype?

b) In each non-diagonal cell, display a scatter plot that visualizes the expression levels of the corresponding pair of proteins. Use colors and markers so that it is simple to relate these scatter plots to the density plots on the diagonal. (5P for implementation, 2P for answering questions):
   - Point out a pair of proteins whose expression levels have a clear positive correlation overall.
   - Can you identify a pair of proteins for which expression levels are highly correlated in one group of mice (e.g., Ts65Dn), but less so in the other group?

c) Compute the distance consistency of all scatter plots. Which pair of proteins leads to the highest distance consistency? (6P)

d) Imagine that, given only the expression levels of two proteins, you will be asked to decide whether they are from a normal control who received saline solution, or a Ts65Dn mouse who received memantine. Which pair of proteins would you choose to make that decision? Modify the visualization to best answer the question, and justify both your answer and your modification. (3P)

Hint: You can use the Python toolkit matplotlib to create plots. More information on it is available from http://matplotlib.org/.

Good Luck!