Integrative Transcriptomics
Dr. K. Nieselt, Faculty of Science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lecture: Microarrays Bioinformatics — WS 2017/18

Assignment No. 8 (5 points)

Hand out:	Thursday, January 11
Hand in due:	Thursday, January 18, 10:00
Tutorial date:	Tuesday, January 16, 10:15-11:45
Direct inquiries to:	fabian.moertter@student.uni-tuebingen.de

Theoretical Assignments

1. **Mutual information of expression time series data** (5p)

   During a microarray experiment a time series measuring the expression of bacterium during growth was measured. Here are the expression profiles of two genes x, y:

	0h	1h	2h	3h	4h	5h	6h	7h	8h
x	8.4395	8.2963	7.9803	7.7198	8.8918	10.6271	10.8426	11.6703	10.7754
y	5.3752	5.6375	5.5026	5.9107	5.5489	5.6988	5.782	5.5134	5.4937

   Here, we ask you to compute the mutual information distance of these two gene expression profiles. In the lecture we started discussing several alternatives to discretize continuous ($\log_2$) expression data. However, here in addition we have time series data, thus a dependence of the order of each experiment. Use this feature to discretize the expression data to for example binary vectors. Then (manually) compute the MI distance of the two discretized expression profiles. Hand in the computations, the profile plot of the two genes, the MI distance value and a short conclusion of the computed value.

Practical Assignments

The material is available from `ILIAS` (`A08.zip`).

In the material you will find the so-called Spellman cell cycle data. It contains data from a genome-wide cell cycle microarray experiment in *Saccharomyces cerevisiae* (Spellman *et al.*, 1998) as well as a list of genes that have been found to be significantly associated with the cell cycle. These genes show a (mostly periodic) expression behavior along the cell cycle. The experiments encompassed two consecutive cell cycles of synchronized yeast cells.

2. **Distance functions**

   Select the set of all cell cycle genes (S_{cycle}) and select a random set of genes (S_{random}) from the complete data such that the size of both sets (cycle and random) is identical.

   (a) Implement a function to compute a distance matrix of a set of gene expression profiles. Allow the user to choose between the Euclidean, the Pearson correlation and the Spearman correlation distance via a function parameter. You may use existing R functions.

   (b) Apply your function to compute the distribution of Euclidean distances on S_{random} and S_{cycle}. Visually compare the two distributions using a QQ-Plot.

   (c) do the same for the Pearson correlation distances.

   (d) do the same for the Spearman correlation distances.

   (e) Describe similarities and differences between (b), (c), and (d).

3. Hierarchical Clustering

In this task you will produce hierarchical clusterings of the cell-cycle specific genes only.

(a) Write a function that provides a simple interface to R's `hclust` method. Let the user choose which method (single, average, complete linkage) and which distance method should be used. Use your distance matrix function from the previous task.

(b) Select one of the linkage methods and hierarchically cluster the cell cycle specific genes. Produce one clustering using the Euclidean distance and one using the Pearson Correlation distance. Create a heatmap (`heatmap()`) for each clustering. (2 plots)

(c) Can you (visually) confirm the periodicity of (some of) the genes? Which distance metric is 'better' suited for this hypothesis in your eyes?

Please read the questions carefully. If there are any questions, you may ask them during the tutorial session or via e-mail to your tutor. You will usually get an answer in time, but late e-mails (e.g. on Thursday morning before class) might not be answered in time. Please upload your solutions in the Ilias system. Please pack your source code, the plots, as well as the theoretical part into **one single archive file (zip)**. Source code should compile correctly.