ITX Integrative Transcriptomics
Dr. K. Nieselt, Faculty of Science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lecture: Microarrays Bioinformatics — WS 2017/18

Assignment No. 9 (5 points + 3 Bonus)

Hand out:	Thursday, January 18
Hand in due:	Thursday, January 25, 10:00
Tutorial date:	Tuesday, January 23, 10:15-11:45
Direct inquiries to:	fabian.moertter@student.uni-tuebingen.de

Theoretical Assignments

1. **Key aspects of k-means** (3P)

   Clearly, the choice of the initial centroids is a key aspect of k-means. Consider k-means with random centroid initialization.

   (a) Construct a simple example of a two-dimensional data set (as a plot showing the data points) where random initialization can result in a clearly not optimal k-means clustering. (This proves that k-means sometimes only finds a local but not global optimum).

   (b) How could you use the result of hierarchical clustering in order to determine a possible k for the number of clusters when using k-means.

2. **Silhouette Plots** (2P + 1 Bonus)

   How can the Silhouette plot be used to assess whether the chosen number k of clusters using k-means is appropriate? Come up with a statistics.

   Bonus: Apply your findings to the results of your clusterings task no. 4(b). According to your method, which is the most appropriate k? Compare to the result of task no. 3(a).

Practical Assignments

3. **Implement k-means** (2 Bonus)

   Implement the original k-means algorithm, by writing a wrapper function around the R-function `kmeans`. The function should output the k clusters with its genes, as well as the intravariance sum. The number of clusters k should be a parameter that can be chosen by the user. Use the Spellman cell cycle data from the last assignment for the following tasks.

   (a) Perform a k-means clustering of the genes of the cell-cycle specific submatrix of the Spellman data set. Reconstruct clusterings for $k = 4, 6, 8, 10, 12, 14, 16$. According to the intravariance sum, which is the most appropriate k in your eyes?

   (b) Visualize the clustering of your most appropriate k using a heatmap.

   (c) **Bonus:** Visualize the clustering of your most appropriate k using profile plots. For the profile plot, also plot the profiles of the respective cluster centroids.

   (d) **Bonus:** Compare the results of the clustering to those of last week when you applied hierarchical clustering to the data. How could this be done „efficiently “?

4. **Silhouette plots**

(a) Write a function that creates a silhouette plot for a given k-means (partitioning) clustering (wrapper around the R function 'silhouette').

(b) Plot the silhouettes for every k you used in task no 3.

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.