Bioinformatics Workshop on S. aureus-related genomics

Prof. Dr. Volkhard Helms
Ruslan Akulenko
Summer Semester 2014

Saarland University
Chair for Computational Biology

Exercise Sheet 5

August 29, 2014

Clustering *Staphylococcus aureus* data

During this assignment you will learn how to apply clustering to *S. aureus* data using R, visualize clusters and annotate them using DAVID.
In the first part you will study clustering methods available in R. Among them: hierarchical agglomerative clustering, k-means clustering and Affinity Propagation. Different methods for assessing results and choosing number of clusters will be covered as well.
In the second part of the assignment, you will draw plots to see how well clustering has performed.
In the final part, several clusters will be chosen to be annotated using DAVID.

Exercise 1.
Run RStudio and open existing script using File → Open File and select **Exercise5.R**. Load the data from the previous assignment (mssa_completed.RData). If this file is not available, execute lines 47-85 from **Exercise4.R**.
As in the previous assignment, current script contains complete code used in this assignment.
Perform Hierarchical Agglomerative clustering of samples and plot the dendrogram. Visually determine clusters and draw red borders around those clusters.
In order to perform k-means clustering of genes you need to determine the number of clusters first. Use all following approaches:
(a) Rule of thumb.
(b) Choose number of clusters using Silhouette. Create Silhouette plot (one axis contains silhouette width, another genes). Determine average silhouette and the number of genes in each cluster, compute total average silhouette for all clusters.
(c) Sum of squared errors scree plot for a number of cluster solutions. Create the plot and define the elbow – the number of clusters.
Pick up two cluster solutions and validate them with cluster.stats function from fpc package.
Run Affinity Propagation to cluster genes.

Exercise 2.
One way to visualize k-means results is to plot genes using first two principle

components (you can obtain them using prcomp function from stats package) and color them according to the clusters.
Plot heatmap of Affinity Propagation results.

Exrcise 3.
DAVID is the Database for Annotation, Visualization and Integated Discovery, which provides a comprehensive set of functional annotation tools. http://david.abcc.ncifcrf.gov/
Take 2-nd cluster of genes obtained by Affinity Propagation and run pathway enrichment analysis using DAVID.