

Homework 4: Microarray data analysis (30 points)

Several steps are involved in the processing of microarray data, including filtering (to remove unreliable data points), log transformation, normalization, statistical and clustering analysis, and finally data interpretation. In this exercise, you will learn how to use several bioinformatics software programs for microarray data analysis.

1. Dataset: The dataset used in this exercise was a public yeast dataset, originally published in Eisen et al. (PNAS 1998), that contains 5,987 yeast genes (YORF, yeast open reading frame) from a number of different experiments. For this exercise, two simplified datasets were created that include only 15 samples (i.e., 15 columns representing 15 time points) from a time-course study. Dataset 1 contains all 5,987 YORF genes (in rows), whereas dataset 2 contains only a subset of 100 YORFs.

2. Programs to use:

Cluster (version 2.11) and TreeView (version 1.60), available in the Forestry student lab computer: via Start/programs/Cluster/Cluster, and Start/programs/EisenSoftware/TreeView. The manual is available on the course webpage. *Note: Sometimes the program freezes for unknown reasons. If so, kill all other jobs (processes) and try to restart the computer again.*

3. Analysis procedures: use the dataset 2 for the following analyses, and use dataset 1 for HCA analysis (or all of the steps for your own comparison).

- 1) Load data into the Cluster program.
- 2) Filter Data (“filter” → “accept”)
- 3) Adjust Data (log transformation and normalization)
- 4) HCA
- 5) K-means
- 6) PCA

4. View images by TreeView: save images as bmp files so that they can be viewed in other programs (e.g. Photoshop, PowerPoint, or Word).

5. Answer the following questions:

- 1) Are there any differences between the HCA array trees obtained from the two datasets? Describe the differences.
- 2) What does the gene tree by HCA from dataset 2 tell us? i.e. explain their expression patterns across different time points.
- 3) What tentative conclusion can you make based on the HCA analysis?
- 4) What did you find from other clustering methods?
- 5) Write 1-2 pages of the analysis procedures (use screen shots to show your parameter settings), and your observations from each clustering method (please refer to corresponding output files).
- 6) Attach two array trees and one gene tree by HCA. Zip all output files (using the PowerArchiver or WinZip programs) and submit them electronically.

Notes:

1. Data have already been log-transformed.
2. Save your normalized data (as output) and compare with the original one.
3. Use default settings, except for K-means
4. Use ‘Alt’ and ‘PrintScreen’ keys to obtain screen shots.