

Exercises & Answers: DNA Microarray Data Analysis

You will use the TIGR MultiExperiment Viewer (TMEV) Java software package for the analysis of microarray data. You will look at the distribution of measured microarray fluorescence intensities and see the influence of normalization. Furthermore, you will apply some unsupervised learning techniques such as hierarchical clustering, *k*-means clustering, and principal component analysis on simulated microarray data. If there is still time left, you will also apply hierarchical clustering on a yeast cell-cycle microarray data set as published by Spellman *et al.*

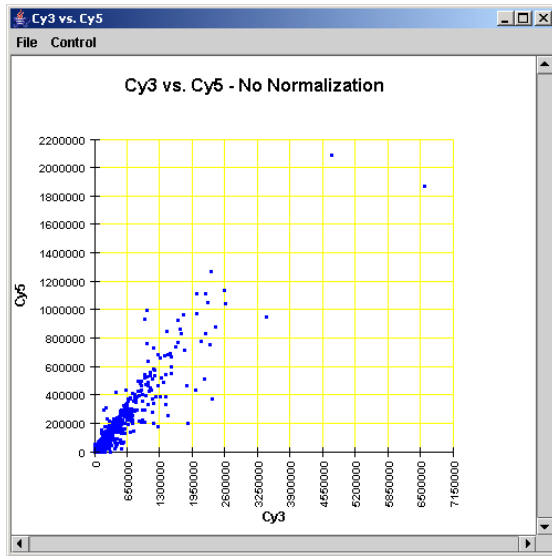
Data for the exercises and some extra information are on <http://www.bioinformaticslaboratory.nl> under Education.

A: Normalization

Normalization refers to adjusting the relative fluorescence intensities between the two scanned channels (red: Cy₅; sample tissue and green: Cy₃; reference tissue). A correction for various types of systematic bias, *e.g.*, differences in labeling and detection efficiencies for the two fluorescent dyes, is often necessary. If in the TIGR software total intensity normalization is selected, an intensity correction is performed under the assumption that the total measured Cy₅ intensity should be equal to the total measured Cy₃ intensity. A multiplication factor is applied on all Cy₃ intensities to do exactly this.

1. Start the TIGR software package (Start – Geneeskunde – Bio Informatica – TMEV – TMEV); the main program window labeled “TIGR MultiExperiment Viewer” appears together with a TIGR Multiple Array Viewer Window, which you can close for now.
1. In the main window, select the top menu option (File - New Single Array Viewer).
2. In the “Single Array Viewer” window, do “File-Open Experiment From File”. Select TMEV Preferences and then open the file “Single_array.tav”.
3. On the right in the “Single Array Viewer”, an image of a microarray is depicted in which each rectangle represents an array spot and different colors code for the value of the log-ratios. The color overlay viewing option computes a color for Cy₃ and Cy₅ separately and then overlays the two colors to get a resulting color. The further the log-ratio from zero, the brighter the element is. Gray elements have invalid values (Cy₃ and Cy₅ values are both zero); they are not used in any analysis.
4. Select a “spot” in the image by single clicking with the left mouse button. A new window with experimental spot information will appear. Which properties are displayed?
5. If you are interested in only those spots with absolute $^2\log(\text{ratio})$ larger than two, you can select the “Expression Ratio” button. You can use the slider to change the “critical” expression ratio value. Try this out for yourself.

6. Now, we will look at some Cy_3/Cy_5 plots. Select “Views - View Graph - Intensity Scatterplot”. Can you deduce from this plot whether normalization is necessary? Why or why not?



It can be concluded that normalization still should be applied because the slope is not equal to one (remember the definition of the total intensity correction in the introduction of these exercises).

7. An alternative plot is “Views - View Graph - Ratio Histogram”. What is depicted in this histogram? Keep this window open for the next question.

Depicted is the ratio frequency.

8. Now make the log-transform histogram: “Views - View Graph - Ratio Histogram (log)”. What are the differences before and after log-transformation? Why are there negative values after log-transformation? Keep both windows open for the last question.

Differences: negative values do exist and histogram is more bell-shaped.

Negative values on the x-axis do exist because the $\log(\text{ratio})$ of ratios smaller than 1 is negative, e.g., $^{10}\log(0.1) = -1$ and $^{10}\log(0.01) = -2$.

9. Make a “Log Ratio \times Log Product” plot. How can you see that normalization is necessary?

The average log-intensity $\log(Cy_3 \times Cy_5)$ is plotted on the x-axis. The log-ratio is plotted on the y-axis. This is a so-called MA plot. The mean of the cloud is not centered around zero (horizontal axis), so normalization should still be applied.

10. Select “Normalization - Total Intensity” and select again the 'G/R bar display'. Make some new scatterplots and histograms. Do you see what you would expect in the new graphs?

Red and green intensities in the bars are equally distributed.

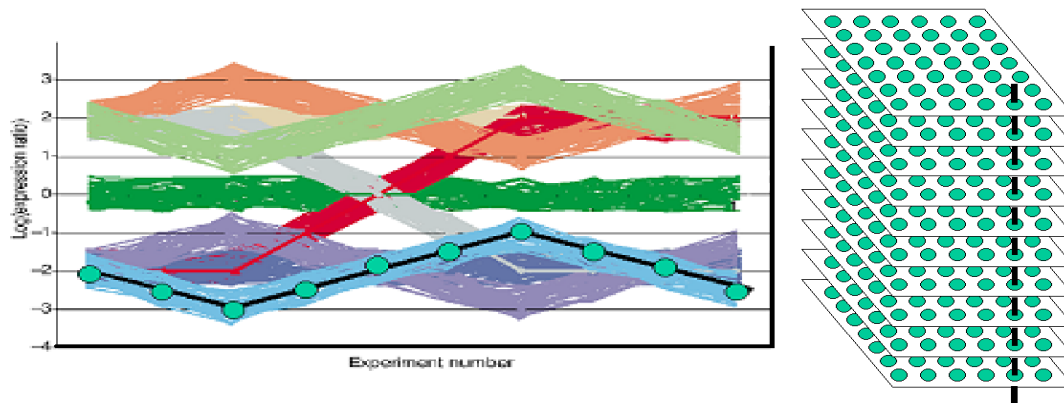
In the “Log Ratio \times Log(Product)” plot, the cloud is now centered around zero. Other plots confirm that data is now properly normalized.

B: Unsupervised learning on a simulated dataset

Cluster analysis includes multiple methods to find genes that have similar expression profiles. It is assumed that genes with similar profiles are somehow in the investigated sample cells, e.g., they share a biological pathway.

The left-hand figure shows the simulated dataset introduced previously. The black line depicts one single gene expression profile and the vertical black line in the right-hand image figure illustrates how

the black gene expression profile is constructed: the gene expression levels of the same spot (=gene) over multiple microarray experiments are taken. Only ten arrays are considered in this simulated dataset. Therefore, the left-hand figure has ten experiments on the x-axis. The data set consists of nine different profiles; round each profile 50 noisy profiles were generated. Therefore, each simulated array contains $9 \times (50+1) = 459$ different profiles. The log-ratio of these gene profiles is indicated on the y-axis.



1. Close all windows except the main “TIGR Multiple Experiment Viewer” window; select “New Multiple Array Viewer”. Do “File - Load Data” and select “Load expression files of type: Tab-Delimited Multi-Sample (TDMS)”. Click on “File Browser” in the “Expression File Loader” window and load the file “JQ_stanford.txt”.¹ This is the simulated dataset shown in the figure above.
2. After loading the JQ data, the “Expression File Loader” window contains the log-ratios of all ten arrays and some columns with extra information. Click on the upper-leftmost log-ratio (in the column labeled “Ex1” on the second row) and load the data. The right-hand pane of the “TIGR Multiple Array Viewer” now shows the expression matrix of the experiment. Click on a spot and look at the information. Make a “Gene graph”, what do you see?

In a gene graph, a single gene expression profile is depicted: the log-ratio versus the hybridizations.

3. Click on the “HCL” pictogram. Perform an “Average linkage clustering” on both genes and experiments. Can you interpret the resulting dendrogram (accessible from the folder structure in the left-hand pane of the “TIGR Multiple Array Viewer”)? How many clusters does the dendrogram reveal?² To which group of gene expression profiles does the first cluster correspond? Reminder: the green, black, and red colors represent negative, about zero, and positive ²log(ratios) respectively.

One can easily distinguish nine different clusters corresponding to the nine different base profiles.

In the first cluster, the gene expression profile goes up, then goes down, and finally goes up again. This is equal to the orange group of gene expression profiles.

4. Now perform a single-linkage hierarchical clustering. An undesired effect that you should see is *chaining*. What does that mean? Can you explain why this effect might be troublesome during an analysis? Is it a problem for this data set?

¹ Available on the course website. Save the file somewhere where you can find it and then load in TMEV.

² There are several options to investigate the dendrogram in more detail. You can click on a branch to see the subtree under this branch highlighted. You can also right click anywhere in the right window pane and select “Gene Tree Properties”. Then you can select where to cut the dendrogram with the “distance range” slider.

Chaining is the sequential addition of single objects to an existing cluster. This effect is clearly visible on this data set. However, in the current example it is not a problem because still the correct clusters are detected (despite the chaining effect). It is possible though that, due to sequential addition of single objects, no valid clusters are obtained.

- Click on the KMC pictogram. Perform two runs of k -means with nine clusters: one with the default (Euclidean) distance and one with the cosine correlation distance (Distance-Cosine Correlation). Inspect the clusters found by k -means, accessible from the folder structure in the left-hand pane of the “TIGR Multiple Array Viewer” under Expression Graphs-All Clusters. Which of the two distance measures gives better results? Can you explain why one of the distance measures fails?

Cosine correlation distance discovers all nine clusters although the profiles around the horizontal axis are partly attributed to other clusters. The Euclidean distance only discovers five clusters. Cosine correlation takes both shape and absolute value into account and, therefore, is more suitable in this case.

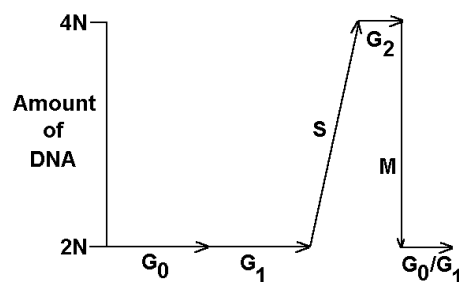
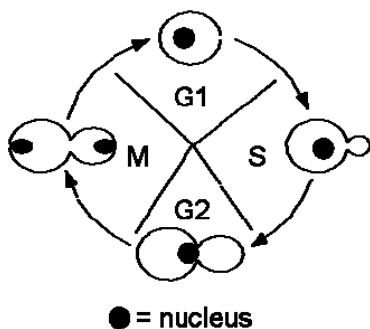
C: Unsupervised learning on the Spellman yeast cycle data set

In 1998, the following paper appeared:

Spellman *et al.*, (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9**, 3273-3297.

A website is dedicated to this publication: <http://genome-www.stanford.edu/cellcycle/>

This was one of the first publications that demonstrated the power of microarrays. On the microarray, genes were included that play an important role in the yeast mitotic cell cycle regulation, see the figures below:



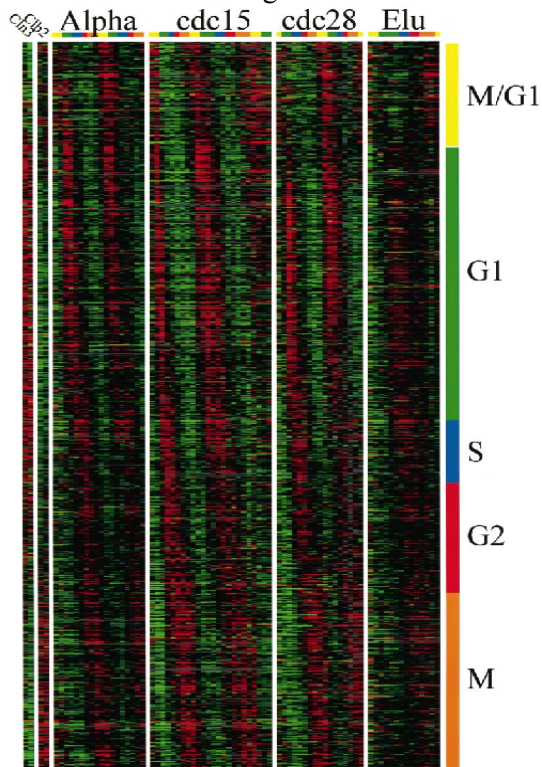
In short, a yeast mitotic cell cycle is usually divided into 4 phases: G_1 , S, G_2 , and M. During G_1 the yeast cell is growing until it is large enough to enter the cell cycle. The conditions during the G_1 stage (*e.g.*, food and temperature) determine whether the yeast cell will go into a stationary phase (also called G_0) or will enter a new cell cycle. During the S phase, the DNA of the yeast cell is replicated. G_2 denotes the phase where the DNA is replicated but the mitosis has not yet started. Finally, in the M-phase the cell undergoes mitosis: partitioning of DNA between the mother and daughter cells followed by cytokinesis: separation of the mother cell from the daughter cell.

In the experiment described in the paper of Spellman *et al.*, the yeast cell cycle has been arrested ('frozen') using 4 different techniques: α factor, CDC15-based blocking, CDC28-based blocking, and elutriation. How these methods work is not essential for the current application. It is important to know that these 4 techniques are responsible for obtaining a large amount of yeast cells that are in the same cell cycle stage. After releasing the arrest, microarray measurements are performed on samples collected on different time points.

The figure on the left depicts the $^2\log(\text{ratio})$ expression matrix. The different arrest methods are positioned in separate groups of columns and their names are indicated on top. The genes are divided into five somewhat arbitrary phase groups. This was done by ordering genes according to their time of peak expression and using published timing of gene expression for known genes in determining which genes belong in which phase group.

In the exercises below, hierarchical clustering techniques are used to find putatively co-regulated genes in one of the four yeast cell cycle stages.

1. Close all windows except the main "TIGR Multiple Experiment Viewer" window and load the file "Reduced_Spellman2.txt".³ Click on the upper-leftmost log-ratio (in the column labeled "alpha0" on the second row) and load the data. To visualize the gene names, click on Display - Element Size - 10x10. The gene names are now depicted on the right of the expression matrix. Click on some spots to get an impression of the dataset. Investigate some gene graphs. What time-dependent effect do you see?



The time-dependent effect that is visible is a periodic effect. The intensity goes up, goes down, goes up again, and so on. Note that four different arrest methods are depicted in a gene graph, so care must be taken when looking at a profile: it is better to focus on only one arrest.

2. Which distance measure is most suitable for this data set?

Since we are interested in grouping together genes with similar periodic behavior, we are looking for similarity in shape. This means that the Pearson correlation distance is most suitable.

3. Apply average linkage clustering and take a glance at the results. In the Appendix, a list of genes is presented that is retrieved from the web site on which the Spellman dataset is made

available:

<http://genome-www.stanford.edu/cellcycle/data/rawdata/>

This list gives names of genes known to be cell cycle regulated and the phase in which regulation occurs. One such group of cell cycle regulated genes is the histones. Are histones also grouped together in the dendrogram? If yes, all of them?

Yes. HHF1 is not in this cluster though.

³ Available on the course website. Save the file somewhere where you can find it and then load in TMEV.

4. The yeast data set that you have used so far is strongly reduced; it consists of only 92 genes. The data set “Reduced_Spellman.txt” contains 798 genes from the original yeast data set. Finding clusters in this larger data set is much more realistic (and difficult). Try to detect clusters in this larger data set (unfortunately, the TIGR software does not have a text search function). You can compare with Spellman’s paper that is accessible through the link given above.

For example, clusters similar to the methionine and MCM clusters of Spellman *et al.* can be found. Differences might be due to the Cln3 and Clb2 experiments that were left out in our data set.

Appendix

The genes listed below were determined to be cell cycle regulated by traditional methods.

M/G1 Boundary (SWI5 or ECB (MCM1) or STE12/MCM1 dependent):

AGA1, ASH1, CDC46, CDC47, CDC6, CHS1, CLN3, CTS1, EGT2, FUS1, MFA2, PCL2, PCL9, RME1, SIC1, SST2, STE2, SWI4, TEC1.

Late G1, SCB regulated:

CLN1, CLN2, FKS1/CWH53, GAS1, HO, KAR4, KRE6, MNN1, PCL1, PSA1, SWE1, TIP1, VAN2/GOG5.

Late G1, MCB regulated:

ASF1, ASF2, CDC21, CDC45, CDC9, CLB5, CLB6, DPB2, GIC2, MCD1, MSH2, MSH6, NIK1/HSL1, PDS1, PMS1, POL1, POL12, POL2, POL3/CDC2, POL30, PRI2, RAD27, RAD51, RAD54, RFA1, RFA2, RFA3, RNR1, RNR3, SPC110/NUF1, SPC42, SPK1, SRS2/HPR5, UNG1.

S-phase:

Histones: HHT1, HHT2, HHF1, HHF2, HTA1, HTA2, HTB1, HTB2.

S/G2-phase:

CIK1, CLB4, CWP1, CWP2, KAR3, NUM1.

G2/M-phase:

ACE2, ASE1, CDC20, CDC5, CLB1, CLB2, DBF2, FAR1, KIN3, MOB1, YRO2(MST1), YDR033w(MST2), SED1, SPO12, SWI5.