

Matlab Problems.

Data set characterization and normalization

Any analysis of gene chip data across different experiments requires that the data be normalized somehow. In this problem set, you will explore different methods of normalization, and see what effects the different techniques have on the final interpretation of the data. We will be examining genes in the galactose metabolism pathway, the 12 GAL genes.

The data for this problem set are taken from experiments in which a total of 48 yeast chip *sets* were run (under 48 different experimental conditions). Each set consists of 4 chips (YA, YB, YC, YD), which make up the yeast genome. Each chip has the same 15 control probes, and each experiment had the five *B. subtilis* genes complementary to these probes spiked in.

The data for this problem can be downloaded from the course website. The data are in three text files: PS1_data, labels and WC_data. PS1_data is a 27x192 matrix. Each column is a different chip: the first 48 columns are the YA chips, the 2nd 48 are the YB chips, etc. The first 15 rows contain data for each of the fifteen controls run on each chip. The remaining twelve rows contain the data for the twelve GAL genes – the remaining genes on the chips are not relevant for this problem set. Since probes for the gal genes are spread out among the chip set, for the chips on which a gal gene is not present, the expression level is indicated by -1000; each gal gene is present on only one chip – YA,B,C,or D. The labels for each row in this data set are contained in the array “labels”. Finally, WC_data is a 1x192 matrix, containing the total signal(for ALL genes, not just controls + gal genes) on each chip, with the data arranged in the same columnwise format as PS1_data.

1. Floor the data in the problem set, by making negative values (except for the -1000 which indicate missing data) a small positive number, such as 25
2. Take the log of the data. We will compare using log-transformed to non-transformed data.
3. Normalize both the log transformed and non-transformed data sets using two methods: whole chip signal, and arithmetic mean of the controls. You should now have four different data sets: log-transformed WC normalized, untransformed WC normalized, log-transformed AM normalized, and untransformed AM normalized.
4. For each data set, calculate the correlation coefficients of the 12 GAL genes. Which gene pairs have a positive relationship (as one gene level increases, so does the other)? Which have a negative relationship (as one gene increases, the other decreases)? Do these differ between the different normalization methods? Is there anything that stands out to you about the relationships?
5. Plot histograms of the correlation coefficients of the four data sets.
6. Choose 2 controls, and show scatter plots of one normalization method vs. the other for each. (Four graphs, log transformed WC normalized vs. log transformed AM normalized, non-transformed WC normalized vs. non-transformed AM normalized etc.)

7. Attempt to describe the differences between the normalization methods. Does one seem “better” than the others? What criteria would you use for making a determination about which method to use? What are some possible improvements that could be made to these techniques?

Please turn in your written answer to questions 4 and 7, and your graphs for questions 5 and 6, as well as your source code. These are due by 4pm on Tuesday March 5th.