

## Data set normalization, part II

The normalization methods that we explored in the first problem set were some of the heuristics used to solve the problem of normalization in the early days of GeneChip experiments. Since then, more sophisticated statistical analyses have been performed. In this problem set we will compare a couple of those methods, maximum likelihood estimation and MAP estimation, with the normalization methods we used in problem set 1. The reference for our approach is the paper by Hartemink et al. which is posted on the course website.

The data for this problem set are the same as for problem set 1. That is, they are taken from experiments in which a total of 48 yeast chip “sets” were run. 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.

Data and partial matlab code for this problem can be downloaded from the website. The only additional data file is a “test” matrix to use as a “positive control” when debugging the code. The columns of the test matrix are approximately a constant multiple of each other.

Partial code that is available on the web:

- PS2\_solution.m – contains the code for the main program, with some missing lines you will need to fill in.
- WC\_normalize.m – code for whole-chip normalization. No modifications needed.
- AM\_normalize.m – code for arithmetic-mean normalization. No modifications needed.
- ML\_norm.m – code for maximum likelihood normalization. You need to fill in missing lines.
- MAP\_norm.m – code for MAP normalization. Again, you need to fill in missing lines.

Note that WC and AM normalize are slightly different from what you wrote last week, so please use the downloaded code. And at the very least, reflect on why the code is altered from its most basic possible implementation.

1. Read over the different M-files and look for the comments that ask you to insert code. We will be flooring the data and then taking the log of the floored data. We will use a different flooring function this time, for which you will need to write code:

<u>OLD VALUE</u>	<u>RESCALED VALUE</u>
$x > 40$	$x$
$-35 < x \leq 40$	$x^2 / 160 + x/2 + 10$
$x \leq -35$	.15625

You will be adding code to normalize the data using maximum likelihood estimation and MAP estimation of the control probes, as described in the Hartemink et al. paper. Finally, the code for normalization using the WC and AM methods from PS1 is already written; for these two methods the raw data is floored, normalized, and then logged.

2. Pick 2 different control probes. Plot scatter plots of one control probe against the other for a

fixed normalization method. Do this for each normalization method, as well as for unnormalized (log) data. Based on these plots, which normalization method looks the “best”?

3. Now assess the quality of normalization using hypothesis testing. One way to measure the success of a normalization method is to determine whether it reduces the variance of a given control across experiments. For example, is the variance of the MAP data for control probe 1 smaller than the variance of the unnormalized data for that same control? And is this reduction statistically significant? One classical statistical test for this question is termed the F-test. Here the test statistic is the ratio of sample variances. If the “true” variances are identical, then the sampling distribution of this statistic is an F-distribution with  $n-1$  and  $m-1$  degrees of freedom, where  $n$  is the sample size of the numerator, and  $m$  is the sample size of the denominator.

Determine whether the variance of the MAP normalized controls are statistically significantly different from the variance of the unnormalized controls, on a control by control basis – i.e. MAP control 1 vs. unnorm control 1 etc. Additionally, perform this comparison for MAP vs. MLE as well as AM vs. unnormalized. (Use the logged data for all of these calculations).

Comment on the pros and cons of the two methods above for comparing normalization methods.

4. What was the motivation for employing MAP estimation? Is there any direct evidence from the computed normalized data to support its original(in the paper) justification? How about from the test data?

Please turn in your legibly written / typed answers as well as your source code. The due date for this problem set is Thursday March 14<sup>th</sup>, by 4PM.