Expression arrays, normalization, and error models

There are a number of different array technologies available for measuring mRNA transcript levels in cell populations, from spotted cDNA arrays to in-situ synthesized oligoarrays, and other variants. Our goal here is to illustrate basic computational methods and ideas involved in teasing out the relevant signal from array measurements. For simplicity we will focus exclusively on spotted cDNA arrays.

Spotted cDNA arrays are geared towards measuring relative changes in the mRNA levels across two populations of cells, e.g., cells under normal conditions and those undergoing a specific treatment (e.g., nutrient starvation, chemical exposure, temperature, gene deletion, and so on). The mRNA extracted from the cells in each population is reverse transcribed into cDNA and labeled with a fluorescent dye (Cye3 or Cye5) specific to the population. The resulting populations of differently labeled cDNAs are subsequently jointly hybridized to the matrix of immobilized probes, complements of the cDNA targets we expect to measure. Each array location or spot contains a number of probes specific to the corresponding target to ensure efficient hybridization. We won't consider here the question of how the probes are/should be chosen, for example, to minimize potential cross-hybridization (target hybridizing to a probe other than the intended one). By exciting the fluorescent dyes of the hybridized targets on the array, we can read off the amount of each cDNA target (hybridized to a specific location on the array) corresponding to each population of interest. By jointly hybridizing the two populations we can more directly gauge any changes in the mRNA levels across the two populations without necessarily being able to capture the actual transcript levels in each. This type of internal control helps determine whether a gene is up or down regulated relative to the control.

Array measurements are limited by the fact that we have to use a large number of cells (10,000 or more) to get a reasonable signal. When the cell population of interest is relatively uniform this typically doesn't matter. However, when there are two or more distinct cell types in the population, we might draw false inferences from the aggregate measurements. Suppose, for example, that gene A is active and gene B is inactive in cell type 1 and that the converse holds for cell type 2. We would see both genes active in the array measurement but this conclusion matches neither of the two underlying cell types. We will return to this issue later on in the course.

Normalization

To use the arrays we have to first normalize the signal so as to make two different array measurements or the two channels (specified by the fluorescent dyes) within a single array mutually comparable. By normalizing the signal we aim to remove any systematic experimental variation. For simplicity we will consider here only normalization to remove biases arising from the differences between the two dyes. The effect can be easily seen by swapping the dyes between the control and treatment populations or by using two identical populations with different dyes. The biases we observe may arise during sample preparation due to differences in how effectively the dyes are incorporated, or later due to different heat and light characteristics of the dyes.

The simplest way to normalize the signal from the two channels would be to equalize the total measured intensity across the spots. The assumption here is, of course, that the overall amount of mRNA transcript is the same in two the cell populations (control and treated). A slightly better approach would be to normalize based on the total intensity across genes unlikely to change due to the treatment (e.g., housekeeping genes); defining this set may be difficult, however.

We will consider here a slightly different approach. Suppose first that for genes that remain unchanged due to the treatment, the measured intensities, in the absence of noise, are proportional to each other: $R = kG$, where R and G are the signals from the red and green channels, respectively, and k is an unknown constant we have to estimate. We do not assume that we know which genes remain unchanged and which ones (or how many) have widely different expression levels due to the treatment. For this reason we cannot equalize the total signal from the two channels. A robust alternative is to set k so that $\log R/kG$ (log-ratio of the corrected signals) has median zero. While this value can be computed directly, we formulate the problem in terms of robust estimation for later utility: we'd like to find $c = \log(k)$ that minimizes

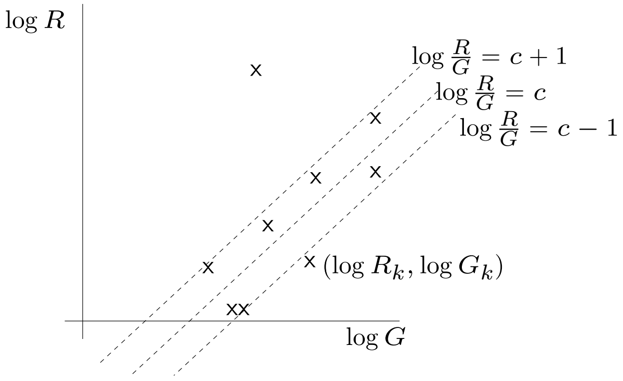
$$\sum_{j=1}^{n} |\log R_j - c - \log G_j|$$

where the summation is over n probes/spots on the array and $|\cdot|$ denotes the absolute value. To see that we get the same answer in this case, we can take the derivative of this objective with respect to c (recalling that the derivative of the absolute value is ± 1

depending on the sign of the argument[1]) and set it to zero

$$\frac{d}{dz}\sum_{j=1}^{n} |\log R_j - c - \log G_j| = \sum_{j:\log R_j/G_j > c} (+1)\cdot(-1) + \sum_{j:\log R_j/G_j < c} (-1)\cdot(-1) = 0$$

This condition ensures that, at the optimum, we balance the number of probes for which $\log R_j/G_j$ is greater than c and those for which it is less than c. The optimal c is therefore the median value of $\log R_j/G_j$. Geometrically, setting c corresponds to centering a 45-degree line in the $\log R$ versus $\log G$ scatter plot from origin (no bias) to the center of the cloud of points (see figure below).

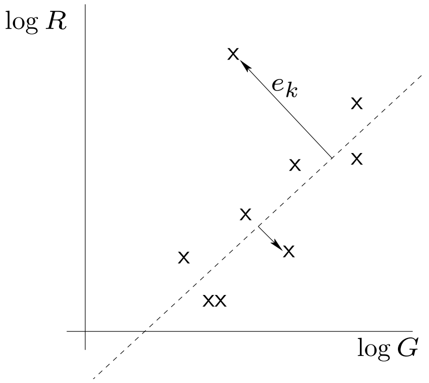


The dye biases may depend on the intensity as well. For example, we might suspect that the intensities for identical transcript levels would relate as $R = kR^p$ for some k and p. As before, we could estimate the parameters p and $c = \log(k)$ by minimizing

$$\sum_{j=1}^{n} |\log R_j - c - p\log G_j|$$

This is, however, not quite correct. We should measure the deviations orthogonally to the line $\log R = c + p\log G$, not vertically, since both measurements $\log R$ and $\log G$ involve errors (see figure below).

[1] We omit here the special cases when the argument is exactly zero.



For any given slope p the orthogonal distance is proportional to the vertical distance and so we can easily correct the above objective:

$$\sum_{j=1}^{n} \overbrace{|\log R_j - c - p \log G_j| \cdot \frac{1}{\sqrt{p^2+1}}}^{e_j}$$

By defining $c' = c/\sqrt{p^2+1}$ and $p' = p/\sqrt{p^2+1}$ the optimization problem we have to solve still has the same form:

$$\sum_{j=1}^{n} |\log R_j - c' - p' \log G_j|$$

with the constraint that $p' \in [-1, 1]$. After solving for c' and p' we can reconstruct c and p to perform the normalization.

Differential expression

Suppose now that we have removed all the systematic variations, e.g., due to the dye biases. The first type of inference we'd like to make on the basis of the array measurement is to determine which genes are differentially expressed (up or down regulated) due to the treatment. Here we consider making these decisions on the basis of a single array; the decisions can obviously be strengthened by carrying out the experiment in duplicate or triplicate as is typically done in practice. The methodology we discuss here can be directly extended to include multiple replicates.

A simple approach would be to assume that for genes that are not differentially expressed $\log R/G \sim N(0, \sigma^2)$. In other words, the null hypothesis for each gene is that the log-ratios follow a normal distribution with mean zero and variance σ^2, where the variance does not depend on the spot. This normal distribution summarizes the experimental variation we expect to see on spots that should return identical intensity values from the two channels. We could estimate the variance from the measured log-ratios corresponding to housekeeping genes that are assumed not to change their expression due to the treatment. For other genes we could use the normal distribution to evaluate a p-value for differential expression: the probability mass of the tail of the normal distribution at the observed value of the log-ratio.

The problem with this approach is that it only pays attention to the log-ratio $\log R/G$. Thus decisions are made as confidently when the intensity measurements R and G are very low (at noise level) as when they are high (clear signal).

We follow here a bit more sophisticated hierarchical Bayesian approach. The basic idea is very simple. We wish to gauge, for each gene, whether we can explain the observed intensities from the two channels by assuming a common biological signal, or whether we have to accept that the intensity differences are too large to be explained by experimental noise. But we have to define "experimental noise" and what type of biological signal to expect so they can be compared.

Let's start by specifying the experimental variation for a single channel (say red). We expect the intensity measurements for a given biological signal to be distributed according to a Gamma distribution: $R_j \sim \text{Gamma}(a, \theta_j)$, where a is called the shape parameter and θ_j is the (inverse) scale parameter. The shape parameter is common to all genes, while the scale parameter characterizes the underlying biological signal and varies from gene to gene. The mode (peak) of this distribution occurs at intensity $(a-1)/\theta_j$; the mean is a/θ_j and the variance is given by a/θ_j^2. The distribution has "heavy tails" meaning that larger errors are also permitted. We don't expect to know either of the parameter values; a will be estimated on the basis of the array measurements.

We expect the biological signals (scale parameters) θ_j to be independent for each gene (spot) and also follow a Gamma distribution $\theta_j \sim \text{Gamma}(a_0, v)$, with a common shape parameter a_0 and scale v. So, according to this model, we could generate biological signals for each spot on the array by drawing independent samples from this Gamma distribution. The actual intensity values that we would expect to see on the spots would be subsequently sampled from $R_j \sim \text{Gamma}(a, \theta_j)$, separately for each spot, as discussed above. The fact that we have chosen to use Gamma distributions is largely for mathematical convenience albeit they have some appropriate qualitative features (e.g., normal looking peak, heavy tails). As a first approximation we have also chosen to consider each spot independently of others; in other words, we decide whether a gene is differentially expressed largely by looking only on the intensity values from the two channels for that gene (save the parameters of the Gamma distributions that are estimated on the basis of all the spots).

We are now ready to define what expect in terms of intensity measurements from the two channels when they are/are not assumed to be differentially expressed:

H_0: same biological signal at j

$$\begin{aligned} &\theta_j \sim \text{Gamma}(a_0, v) \\ &R_j \sim \text{Gamma}(a, \theta_j), \quad G_j \sim \text{Gamma}(a, \theta_j) \end{aligned}$$

H_1: differential expression at j

$$\begin{aligned} \theta_j^R &\sim \text{Gamma}(a_0, v), R_j \sim \text{Gamma}(a, \theta_j^R) \\ \theta_j^G &\sim \text{Gamma}(a_0, v), G_j \sim \text{Gamma}(a, \theta_j^G) \end{aligned}$$

In other words, we sample different biological signals for the two channels if they are differentially expressed; otherwise we sample a common signal. These sampling schemes give rise to two different (continuous) mixture distributions over the observed intensities:

$$P(R_j, G_j | a_0, v, a, H_0) = \int_{\theta_j} \text{Gamma}(\theta_j; a_0, v) \Big[\text{Gamma}(R_j; a, \theta_j) \text{Gamma}(R_j; a, \theta_j) \Big] d\theta_j$$

where, for example, $\text{Gamma}(R_j; a, \theta_j)$ is the pdf for a Gamma distributions with parameters a and θ_j. The integration over θ_j represents the fact that we need to account for different levels of underlying biological signal, in proportion to our expectations. The measurements R_j and G_j are not independent since they share a common biological signal. Similarly,

$$P(R_j, G_j | a_0, v, a, H_1) = P(R_j | a_0, v, a, H_1) P(G_j | a_0, v, a, H_1)$$

where, for example,

$$P(R_j | a_0, v, a, H_1) = \int_{\theta_j^R} \text{Gamma}(\theta_j^R; a_0, v) \text{Gamma}(R_j; a, \theta_j^R) d\theta_j^R$$

Note that in this case R_j and G_j are independent since they have no common co-variate.

We have now two competing explanations for measurements from the two channels, one that assumes a common biological signal, and the other that doesn't (differential expression). For any particular gene we don't know a priori whether it is differentially expressed. We will therefore have to entertain both possibilities and infer which explanation is more likely. Assuming a prior probability p for the fact that any specific gene is differentially expressed, our model over the intensity measurements (R_j, G_j) is given by

$$P(R_j, G_j|a_0, v, a, p) = p\, P(R_j, G_j|a_0, v, a, H_1) + (1-p)\, P(R_j, G_j|a_0, v, a, H_0)$$

This is again a mixture model, a mixture of two distributions that are themselves mixtures. What's left to do for us is to estimate the four parameters in this model a_0,v, a and p. We find the setting of these parameters that maximize the probability of reproducing the observations across the spots on the array (maximum likelihood fitting):

$$\prod_{j=1}^{n} P(R_j, G_j|a_0, v, a, p)$$

This task may look a little daunting but can be done analogously to the EM-algorithm we have used previously for estimating mixture models (motifs). Here, in the E-step we evaluate for each gene the posterior probability that it is differentially expressed and what the biological signal might be. Once we have these posteriors, we can, in the M-step, essentially just estimate the Gamma distributions from weighted observations. We omit the details of the algorithm.

We are now ready to make decisions concerning differential expression. Given the estimated parameters $\hat{a}_0$,$\hat{v}$, $\hat{a}$ and $\hat{p}$ (where the hat denotes the fact that they were estimated), we find genes that are differentially expressed on the basis of the posterior probabilities:

$$P(j \text{ is differentially expressed}|R_j, G_j, \hat{a}_0, \hat{v}, \hat{a}, \hat{p}) = \frac{\hat{p}\, P(R_j, G_j|\hat{a}_0, \hat{v}, \hat{a}, H_1)}{P(R_j, G_j|\hat{a}_0, \hat{v}, \hat{a}, \hat{p})}$$

These posterior probabilities approapriately discount high log-ratios corresponding to low intensity measurements as such measurements are easily explained by experimental noise.

References

Yee Hwa Yang et al., “Normalization for cDNA microarray data: a robust composite method for addressing single and multiple slide systematic variation”, Nucleic Acid Research, Vol 30, No 4, 2002.

Newton et al., “On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data”, Journal of Computational Biology, Vol 8, No 1, 2001.