BIOS 7659 Journal Club:

A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. (Bolstad et al., 2003)

Tim Vigers

9/22/2020

Introduction

- The goal of normalization is to separate the interesting biological variation from the variation that is a result of sample preparation, array production and processing, etc.
- Affymetrix proposes scaling the arrays so that each one has the same mean expression summary measure.
    - This does not work well when there are non-linear relationships between arrays.

Alternatives

- Other approaches, such as non-linear smooth curves or transforming data to standardize the distribution of intensities across arrays, rely on picking a “baseline” array.
- Bolstad et al. compare three different approaches, all of which combine data from every single array rather than relying on a baseline.

Cyclic loess

- Basically an extension of the M vs. A plots discussed in class, but applied to pairwise combinations of Affymetrix arrays.
- M is the difference in log expression values and A is the average (a Bland-Altman plot).

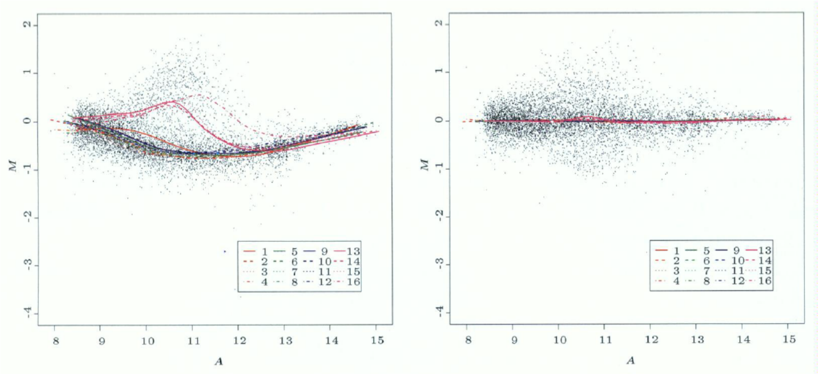

Figure 1: Dudoit et al., 2002

Cyclic loess

1. Take two arrays i and j, each with probes $k = 1, ..., p$.
2. Create an MA plot for these two arrays, and fit a loess curve through these data:

$$M_k = log_2(\frac{x_{ki}}{x_{kj}}), A_k = \frac{log_2(x_{ki}x_{kj})}{2}$$

3. Subtract the normalization curve fits $M'_k = M_k - \hat{M}_k$ and obtain adjusted probe intensities:

$$x'_{ki} = 2^{A_k + \frac{M'_k}{2}}, x'_{kj} = 2^{A_k - \frac{M'_k}{2}}$$

4. Take each of these adjustments (one for each pairwise comparison between arrays) and weight them equally across the set of arrays.

Contrast method

- Very similar to the cyclic loess method, because it's another way of normalizing based on M vs. A:

1. Data is converted to the log scale and the basis is transformed (this is just a fancy linear algebra step).
2. $n - 1$ normalizing curves are fit to the transformed basis as in cyclic loess.
3. Data is transformed again so that the normalizing curves lie on the horizontal, this time using a smooth function.
4. This normalized data is tranformed back to the original basis and exponentiated.

- This is slightly faster than cyclic loess but fitting the curves can still be slow.

Contrast method

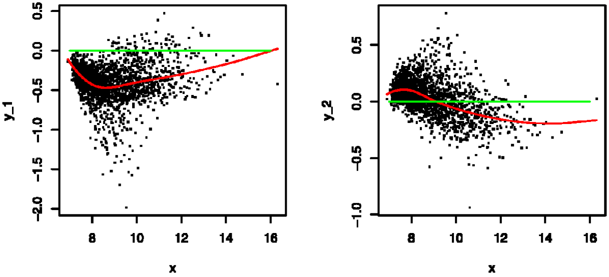

FIG. 2. Contrast plots. Scatter plots of the 2 contrasts against the mean for 3 arrays, A, B, and C, prior to normalizing. The red curve is the fitted normalizing curve, and the green line is the reference line.

Figure 2: Åstrand, 2003

The quantile method

- The goal of this method is to standardize the distribution of probe intensities across all arrays.
- The approach is an *n*-dimensional extension of the fact that, given a quantile-quantile plot where all of the points are on a straight diagonal line, you can be fairly sure that the two data vectors have the same distribution.
- We want to project our data onto the unit vector $(\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$.
- Let $\mathbf{q}_k = (q_{k1}, ..., q_{kn})$ be the vector of k^{th} quantiles for $k = 1, ..., p$. Then:

$$proj_{\mathbf{d}}\mathbf{q}_k = (\frac{1}{n}\sum_{j=1}^{n} q_{kj}, ..., \frac{1}{n}\sum_{j=1}^{n} q_{kj})$$

Quantile normalization algorithm

1. Given n arrays with p probe intensity measurements, make the $p \times n$ matrix X, where each column has all the data from a single array.
2. Sort each column of X to produce X_{sort}. So, each row in X_{sort} is a quantile.
3. Take the mean of each row, and replace every value in the row with the mean to produce X'_{sort}
4. Put each column of X'_{sort} back in the original ordering from X to produce $X_{\text{normalized}}$

- This approach could theoretically be a problem for probes that have the same value across all arrays, but in practice this isn't an issue.

Scaling method

- Based on the approach suggested by Affymetrix, but this paper uses a probe-level version.

1. Choose a baseline array x_{base}: Usually this is the median array, but doesn't necessarily have to be.
2. For each other array, calculate the mean trimmed intensity $\tilde{x}_i$ and find
$$\beta_i = \frac{\tilde{x}_{\text{base}}}{\tilde{x}_i}$$
3. Normalized intensities are $x_i' = \beta_i x_i$

Non-linear method

- The scaling method is the same as fitting a straight line with intercept 0 between x_{base} and x_i.
- This can be extended to non-linear methods, usually a loess curve such that:

$$x_i' = \hat{f}_i(x_i)$$

where $\hat{f}_i(\cdot)$ is the curve mapping from array i to baseline.

References

1. Dudoit, S., Yang, Y. H., Callow, M. J., & Speed, T. P. (2002). STATISTICAL METHODS FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENES IN REPLICATED cDNA MICROARRAY EXPERIMENTS. Statistica Sinica, 12(1), 111–139. JSTOR.
2. Åstrand, M. (2003). Contrast Normalization of Oligonucleotide Arrays. Journal of Computational Biology, 10(1), 95–102. https://doi.org/10.1089/106652703763255697