Microarray Normalization Protocols

James C Costello,[1,2] Mehmet M Dalkilic,[2] Justen R Andrews[1,2,3]

[1] Department of Biology, [2] School of Informatics, [3] Center for Genomics and Bioinformatics
Indiana University, Bloomington, IN

Data acquisition

The Arbeitman, et al.[1] raw data were downloaded from the Stanford Microarray Database as compressed .xls files. Data were read into the statistical programming environment, R, and subsequently analyzed following the two channel protocol described below. The probes used in this experiment are spotted cDNA. Further details can be found at: http://www.sciencemag.org/cgi/data/297/5590/2270/DC1/1

The FlyAtlas (Chintapalli, et al.[2]) raw data were downloaded in .CEL format from the FlyAtlas website (http://www.flyatlas.org/). Data were read into the statistical programming environment, R, and subsequently analyzed following the Affy protocol described below.

Two channel normalization protocol (Arbeitman, et al.)

The data were normalized using the Optimized Local Intensity-dependent Normalization (**OLIN**) algorithm developed by Futschik and Crompton [3]. This R package is maintained in the Bioconductor repository for bioinformatics software (http://www.bioconductor.org/). The data were normalized in two ways. First, *within array* normalization was performed to correct for location-dependent, intensity-dependent, and dye-dependent biases. Second, *between array* normalization, or *scaling*, was done to balance the distribution of intensity values across all the slides within the experiment. This was performed as follows:

1. A target file was created and read into R using the readTargets() function in the package **marray**, which is a base component of the Bioconductor software suite. The target file is a tab-delimited text file that contains a list of filenames of the arrays to be normalized, a description of each array, and the two channel sample descriptions.
2. Each file listed in the target file was read into the **marray** class, **marrayRaw**. The **marrayRaw** class is a data structure used to logically store and manage raw two channel microarray data.
3. Normalization of the raw data was performed with **OLIN** using the default parameters and with scaling turned on. This results from each array, stored in a **marrayNorm** object, are normalized in one batch to accommodate between array scaling. Any flagged spots were not considered in the normalization calculations.
4. The final output of **OLIN** is a log-transformed ratio value (base 2), referred to as the M-value, and a log-transformed average intensity value (base 2), referred to as the A-value. Values for technical and biological replicate spots representing the same target (spotted DNA) under the same condition are averaged. A target under any condition may be flagged for several reasons. It is important to note that this MAY be due to

corresponding transcript not being expressed under that condition, the amplification of the printed cDNA was reported as “failed” in the original data, OR the data is missing for technical reasons.

Spotted DNA array Presence/Absence calls for the Arbeitman, et al. data

In order to provide an indication of whether a gene, which hybridizes to its corresponding target, is expressed under a given condition, we performed a statistical test to determine if the expression of a labeled sample is significantly above the distribution of background values. This was done by creating a distribution of background intensities over the entire slide and testing each individual spot’s foreground intensity against the background distribution in the form of a t-test. This test was done on a channel-by-channel basis. The calculated p-values were then adjusted for multiple hypotheses testing using a Bonferroni correction, giving a corrected p-value. The null hypothesis is that the spot is within the distribution of background noise. To reject the null hypothesis a value of 0.001 was used, so spots with corrected p-values less than or equal to 0.001 were considered present. Spots with corrected p-values in the interval (0.01 to 0.001) were considered marginal. Spots with a corrected p-value greater than 0.01 were considered absent (or within the distribution of background noise). It is important to note that this is a statistical test and should be interpreted as such.

The Arbeitman, et al. expression data were hybridized to two channel spotted arrays. One channel is pooled sample (control) and the other is the sample from a given time point (experimental), therefore the above test was done for the channel containing the experimental sample only. A presence/absence call can then been interpreted as a statistical test of the foreground intensity of a spot in the experimental sample in relation to the background intensity distribution.

Affymetrix normalization protocol

The Affymetrix platform is a series of 25-mer oligonucleotide probes. A set of roughly 14 probes is intended to map with a one-to-one nucleotide correspondence to one transcript, or one exon in one transcript. Each probe also has a corresponding mismatch probe, where the middle nucleotide in the 25-mer is altered. This mismatch probe is used to measure the level of non-specific hybridization.

GCRMA was used to normalize and scale the data. For details on **GCRMA**, see Wu and Irizarry[4]. Briefly, **GCRMA** corrects for background noise and non-specific hybridization, scales all the arrays in a batch through quantile normalization, and reports log-transformed, background adjusted intensity values. **GCRMA** also considers the nucleotide composition as a measure of binding affinity when normalizing for non-specific hybridization.

The “fullmodel” **GCRMA** normalization was run using the **Affy** and **GCRMA** packages in R, as follows:

1. Using the **readAffy()** function in the **Affy** package, read in all the .CEL files.
2. The **GCRMA** normalization method requires that an affinity score be calculated for each

probe on the array. The affinity scores are calculated with respect to the melting temperature of the oligonucleotide sequence. This is done with the **compute.affinities()** function.

3. Perform the **GCRMA** normalization with type = "fullmodel", optical correction is turned on, and the previously calculated probe affinities are included.
4. The final output of **GCRMA** is the log-transformed intensity values for the background corrected probe sets.

Affymetrix Presence/Absence calls

The presence/absence calls provide a statistical measure of the presence of a transcript within the tested biological sample. Details of the calculations can be found in the Affymetrix Microarray Suite User Guide[5]. Briefly, p-values for each probe set with respect to both the perfect match and mismatch probes are calculated using the Wilcoxon signed rank test. If the p-value is less that 0.04 the probe is considered to be present. If the p-value is in the interval [0.06, 0.04], then the probe is considered marginal, and if the p-value is greater than 0.06, the probe is called absent. The Affymetrix Expression Console software was used to make presence/absence calls. The raw data are read in this software and the **MAS5.0** normalization method was run. Of the resulting output only the presense/absence calls were extracted. These calls were then mapped to the expression values output from the **GCRMA**.

References

1. Arbeitman, M.N., et al., *Gene expression during the life cycle of Drosophila melanogaster.* Science, 2002. 297(5590): p. 2270-5.
2. Chintapalli, V.R., J. Wang, and J.A. Dow, *Using FlyAtlas to identify better Drosophila melanogaster models of human disease.* Nat Genet, 2007. 39(6): p. 715-20.
3. Futschik, M.E. and T. Crompton, *OLIN: optimized normalization, visualization and quality testing of two-channel microarray data.* Bioinformatics, 2005. 21(8): p. 1724-6.
4. Wu, Z. and R.A. Irizarry, *Preprocessing of oligonucleotide array data.* Nat Biotechnol, 2004. 22(6): p. 656-8; author reply 658.
5. Affymetrix, *Affymetrix Microarray Suite User Guide*. 2001, Affymetrix: Santa Clara, CA.