Published in final edited form as:
Stat Methods Med Res. 2009 December ; 18(6): 533–541. doi:10.1177/0962280209351924.

A Review of Statistical Methods for Preprocessing Oligonucleotide Microarrays

Zhijin Wu
Center for Statistical Sciences and Department of Community Health, Brown University. zhijin_wu@brown.edu, Phone: (401)8631230; 121 South Main street, Box G-S121, Brown University, Providence RI 02912

Abstract

Microarrays have become an indispensable tool in biomedical research. This powerful technology makes it possible to quantify a large number of nucleic acid molecules simultaneously, but also produces data with many sources of noise. A number of preprocessing steps are therefore necessary to convert the raw data, usually in the form of hybridization images, to measures of biological meaning that can be used in further statistical analysis. Preprocessing of oligonucleotide arrays includes image processing, background adjustment, data normalization/transformation and sometimes summarization when multiple probes are used to target one genomic unit. In this paper we review the issues encountered in each preprocessing step and introduce the statistical models and methods in preprocessing.

1 Introduction

In oligonucleotide microarrays, short DNA molecules (25-60 bases long, depending on technology) are used as *probes* to query complex biological samples, such as transcriptomes, that contain a large number of RNA or DNA species (the *targets*). The probes are usually fixed on a solid surface, the *microarrays*. The target molecules are labeled with fluorescent dyes and bind to probes with complementary sequences during the hybridization step. A scanner is used to read the fluorescent intensity off each probe and the intensities represent the relative abundance of targets for each probe. The raw data from a microarray experiment is therefore an image. The goal of a microarray experiment is, however, to obtain measurements on some genomic units, such as levels of gene expression, genotypes of single nucleotide polymorphism (SNP), or DNA copy number. The process that converts the raw data into the form of biological meaning is referred to as *preprocessing*.

Preprocessing is often not straightforward because microarray data contain noise from many sources. In addition to the signal determined by actual abundance of the target, a number of other factors can affect the measurements taken from microarrays. The experimental procedures, such as extraction, amplification and labeling of the biological samples, can affect the quality of the target sample. The probes have different efficiency and specificity in hybridization, thus the intensities on each probe are a result of both specific and nonspecific binding. Optical and electrical noise are also unavoidable in the scanning step. The rest of the paper is organized as follows. In Section 2 we introduce how image processing is done to obtain numeric values for probe intensities. In Section 3 we present some common statistical models for probe level data and discuss the motivation for the other preprocessing steps. which are discussed in detail in the following sections.

2 Image Processing

The original hybridization data obtained from the scanner is an image. To turn this image into probe level intensity values involves several steps of image processing. The first step attributes pixels to probes and is referred to as segmentation. However, not all pixels in a probe cell are used in reporting the intensity. Pixels around the perimeter of a probe cell may be discarded to minimize possible segmentation errors, as in Affymetrix arrays (1). A summary, for example, the 75*th* percentile, of intensities from the remaining pixels is reported as probe intensity. Segmentation for high density oligonucleotide arrays are easier than for spotted cDNA arrays since the probe cells are more regular than the cDNA spots. In most cases, grid alignment is used and accurate segmentation can be achieved. When there is no strong image blur or large defects, the summarized probe intensities adequately summarize the hybridization information from the image file (1). When there are image defects, they often affect more than one probe cell. The pseudo-image from the probe level data can still reveal spatially coherent problems. Therefore, the probe level intensities are often treated as raw data.

Several methods have been developed to detect and remove or reduce impact of image defects. Harshlight (2) detects three types of image defects (extended, compact and diffuse defects) using pattern recognition methods, and replace the defective probe intensities by missing values. Potentially these missing values can also be filled in by imputation. (3) scan the pseudo-images with a sliding window to detect “blob-like” regions that exhibit high intensities, possibly resulting from artifacts such as bubbles. An overview of the methods that detect and correct for these image level abnormalities is given by (1).

3 Models for probe level intensity

After image defects are removed, the probe intensities are usually taken as the raw data for oligonucleotide arrays. The desired reading on each probe is the amount of fluorescence from molecules that are complementary to the probe, the intended target. However, the hybridization sample consists of a complex mixture of nucleotide molecules, and non-complementary sequences also bind to the probes. This phenomenon is referred to as nonspecific binding and is a major reason for background in microarray data. Optical noise is another source of background, but is usually smaller than non-specific binding background and appears not to be probe specific. Calibration data suggest that background noise is additive (4). As a result, the observed probe intensity is a sum of all the above components. The relative quantity of a target across samples can be seriously biased if background is not accounted for (4). Another phenomenon observed by many investigators is that the stochastic noise on specific binding appears to be proportional to its mean. Since many traditional statistical methodologies are based on the assumption of variance not depending on the mean level, data transformation is often a useful, if not necessary, procedure in preprocessing. There are also factors that affect the observed intensities on an entire array, such as sample preparation, hybridization condition and scanner differences. These “obscure variations” (5) are of no biological interest but exist in almost all microarray data. Normalization, a procedure that aims to remove such obscure variations and makes data across arrays comparable, is thus another key step in preprocessing.

Putting all the above mentioned components together gives the common additive-background-multiplicative-measurement-error model (6; 7; 8; 9) for intensity Y_{ij} of probe j on array i:

$$Y_{ij}=B_{ij}+S_{ij}=\left(B_j+\delta_{ij}\right)+\left(\exp\left\{a_j+\gamma_i+\mu_{ij}+\varepsilon_{ij}\right\}\right). \tag{1}$$

Here $B_{ij} = B_j + \delta_{ij}$ represents background noise for probe *j* on array *i*, with *Bj* being the mean background on probe *j*. S_{ij} represents the specific binding intensity, a_j represents the efficiency of probe *j* in measuring the intended target, γ_i represents an array effect for all probes, μ_{ij} represents the log concentration of the target, and ε_{ij} represents the multiplicative measurement error while δ_{ij} represents an additive measurement error and is sometimes absorbed into the background term.

In this model, the need for normalization is demonstrated by the inclusion of an array effect γ_i in the exponential term, equivalent to a scaling factor on linear scale. In general, if the nature of the array effect is left unspecified, we can use f_i to describe the array effect:

$$Y_{ij}=B_j+f_i\left(\exp\left\{a_j+\mu_{ij}+\varepsilon_{ij}\right\}\right)+\delta_{ij} \tag{2}$$

When multiple probes are used to measure one genomic target, they form a *probe set*. All probes within a probe set *g* measure the same target quantity μ_{gi}, and Model 1 becomes

$$Y_{gij}=B_{gj}+f_i\left(\exp\left\{a_{gj}+\mu_{gi}+\varepsilon_{gij}\right\}\right)+\delta_{gij}.$$

The above models are motivated mostly by Affymetrix array data. Data generated from other platforms appear to share very similar stochastic properties. For example, the additive background and multiplicative measurement error model seem to be applicable to Illumina data as well (10). In the following sections we discuss how the quantity of interest, μ_{gi}, is estimated by our preprocessing procedures. We will continue to use Affymetrix arrays as our main example.

4 Background estimation and adjustment

Since the introduction of the oligonucleotide arrays, various attempts to correct for background noise have been proposed. For example, the manufacturer Affymetrix designed *Mismatch (MM) probes*, probes that differ from the *PerfectMatch (PM)* probes by only the middle nucleotide, to directly measure probe and array specific background. Measurements on MM probes were expected to contain background B_{ij} similar to their PM counterparts, but little specific binding intensity. These direct measurements are simply subtracted from the PM intensities to adjust for the additive background in several generations of preprocessing methods provided by Affymetrix, including MAS 5.0. (11) also use PM-MM to correct for background in their model-based expression index. However, a number of investigators have noted that the assumption of MM measuring the expected background on PM probes, and background only, is often violated in reality (12; 13). (14) report that a considerable proportion of MM probes had consistently higher intensity than the paring PM probes. (15) show that the direct adjustment of background using MM values results in inflated variance of the resulting gene expression estimates. A number of preprocessing methods have avoided the problem by using PM-only measurements. Without a direct measurement, estimation of the background is often done by borrowing information from other probes using an empirical Bayes approach. For example, the *Robust Multiarray Analysis (RMA)* assumes a global background distribution common for all probes that is normally distributed. From each array *i*, probe intensities that are smaller than the empirical mode is used to estimate the mean and variance of the background. The *Variance Stabilization Normalization (VSN)* (16) also estimates a common mean background (baseline) value for all probes on an array using an iterative procedure. Assuming a global background for all probes often results in a conservative estimate of background noise and thus sacrifices the accuracy of some probes. However, compared to PM-MM methods, the gain in precision in PM-only procedures often outweighs the loss in accuracy in identifying biological variations (17).

After the probe sequences became public, the thermodynamics between RNA targets and DNA probes on the arrays have been more intensively studied. The reason that MM probes do not offer the expected background measurements was better understood when (18) discovered that the middle base being purine or pyrimidine affects the hybridization stability. They also fit a linear model that uses the position and type of nucleotides in the probes to predict the amount of hybridization, and showed strong relationship between probe sequence and intensity. (19) adapted their sequence model to explain the nonspecific binding background on different probes. The non-specific binding affinity of a probe is described by the sum of position-dependent base contributions

$$\sum_{k=1}^{25} \sum_{j\in\{A,C,G,T\}} \mu_{j,k} 1_{(b_k=j)}, \tag{3}$$

where *k* indicates the position and *j* represents the base type. The base contributions $\mu_{j,k}$ for each base *j* is modeled as a smooth function of position $k{:}\mu_{j,k}=\sum_{l=0}^{3}\beta_{j,l}k^l$. The parameters $\beta_{j,l}$ are estimated by fitting the model to experimental data containing nonspecific binding alone. Affinity values for any probe sequence can then be calculated from the probe sequence and used to predict and adjust for the probe specific background. This background adjustment is used in preprocessing procedure GCRMA and shown to decrease the bias in expression measurements without much increase of variance (19). Physical models that use the stacking energy of DNA/RNA hybridizations have also been used to predict the hybridization stability of specific and nonspecific bindings. (12) use a stacking energy model and computed both gene-specific and non-specific binding energies of a probe as weighted means of nearest neighbor stacking energies. However, the physical model using thermodynamic parameters (20) does not appear to predict non-specific binding as well as stochastic models using empirically estimated parameters (4).

5 Normalization and transformation

In addition to background noise, there are often other sources of variation affecting the observed intensities that are not of biological interest. For example, arrays scanned on one scanner can in general give higher readings than these from another scanner. The variation due to scanner is of no biological interest and should be removed whenever possible. Such “obscuring variations” (5) can also be introduced by the preparation of the samples or the processing of arrays and often affect the observations on entire arrays. To make the observations across arrays comparable, normalization is a usually a necessary step in preprocessing. Data transformation is also common practice for several reasons. As shown in Model 1, the variance of the specific binding intensity *S* is proportional to the mean level because of the error *ε* in the exponential term. After background adjustment a log transformation is often found to remove or reduce the dependency between the mean and the variance. A log transformation also gives the difference in transformed data the easy interpretation of log ratio. Other variance stabilization transformations are the generalized logs, that are similar to logs for high values and linear for low values (7; 16; 21).

Most normalization methods equalize some summary statistics of the distribution of measurements across arrays. The simplest ones, such as MAS 5.0, scale the arrays so that each array has the same mean or median intensity. The scaling normalization implicitly assumes that biological variations of interest may affect a number of measurements but should not change the mean or mode of the distribution of intensities on each array. Since non-linear relationships between arrays are common, normalization methods that use a non-linear smooth curve have also been introduced (22). Using a baseline array, a smooth normalization curve can be estimated from the scatter plot of two arrays (23). Without a baseline array, one can also use all arrays available in a dataset, and iterates over distinct pairwise combinations of arrays so that all arrays are normalized to an “average” array. The *cyclic loess normalization* (24) and *contrast normalization* (25) are two examples. Another approach is *quantile normalization* (26; 5), which makes all arrays have the same empirical distribution of intensities after normalization. One can use a baseline array for the reference distribution, or use all arrays to generate an “average” distribution for reference. Quantile normalization is used in RMA, GCRMA and recommended for Probe Logarithmic Intensity Error (PLIER) (27).

All normalization methods make assumptions on what is expected to be constant across arrays, although not all make the assumptions explicit. Many normalization procedures assume that the biological effects do not change the basic location and/or shape of the distribution of data. This is often the case if only a small subset of targets are expected to be affected, or, the effects on increasing or decreasing the target amounts are symmetric. When these conditions are at least approximately met, one can use data from the entire array to robustly estimate the location and shape of distribution and perform normalization. However, when the up- and down-regulation is not symmetric across arrays, most methods that normalize to a common distribution would fail. An alternative is to use only the probes that are expected to show no difference across arrays in normalization (28; 29). Since there is often limited number of control genes to cover the range of observed data and some control genes (such as housekeeping genes) actually show biological variations, an “*invariant set*” of probes is often selected to be the normalizing elements. (23) selected their invariant set as probes that show little changes of ranks across arrays. Recently, (30) use a robust linear model to estimate array-to-array variation and define a *least-variant set* of genes.

6 Summarization

In some platforms multiple probes are used to quantify the same genomic target. For example, Affymetrix GeneChips use a set of 11-20 probes to measure expression levels of a gene and on average 4 probes for an exon. Illumina arrays use one probe for each gene but include technical replicates (approximately 30) of the same probe. After preprocessing, a summary of these multiple measurements is given as the measure of gene expression.

Since it is known that a number of factors can contribute to the noise in microarray data, most methods choose to do a robust summary that is resistent to outliers. Single array summaries, such as MAS 5.0 for Affymetrix arrays, use Tukey's biweight to calculate a robust mean of background corrected, log transformed intensities as expression measure. One drawback for this approach is that probes with high efficiency (large a_j in Model 1) in detecting specific binding may have consistently higher intensities. These are the most informative probes but may be down weighted as outliers. Multiarray analyses, such as

RMA (15), GCRMA (19) and model-based expression index (MBEI) (23), identify outliers using information across arrays. For each probe set, consider a linear model (31) on the background adjusted, normalized and log transformed data of probe *j* on array *i*:

$$s_{ij}=\mu+a_j+\theta_i+\varepsilon_{ij}, \tag{4}$$

where a_j is the probe efficiency for j^{-th} probe and $\mu + \theta_i$ represent the gene expression on array *i*. The estimate $\hat{\mu} + \hat{\theta}_i$ is the is the quantity of interest in a gene expression experiment, and the final output from all previous preprocessing steps. RMA and GCRMA use median polish, a fast procedure to fit the above linear model robustly by removing row and column medians iteratively. In the robust multi-array model, observations on probes with very large or small $a_j s$ are not mistakenly treated as outliers, as long as their log expression levels are parallel to those of the other probes. The MBEI considers a model in linear (un-logged) scale:

$$S_{ij}=c_i\phi_j+\delta_{ij,}\text{with}\sum_j a_j^2=J,$$

where c_i is equivalent to $e^{\mu+\theta i}$ and ϕ_j is equivalent to e^{aj} in Model (4). Least squares estimates of the parameters are obtained and the probes with large standard errors are identified as outliers and removed to add robustness. PLIER achieves robustness by using a Geman and McClure function to down weight extreme residuals in the M-estimation procedure.

In Illumina arrays, since all replicate probes share the same sequence, there is no probe effect a_j. The *BeadStudio* output reports a trimmed mean by removing observations 3 median absolute deviations (MAD) away from the median.

7 Discussion

In the previous sections we reviewed the major steps of preprocessing for microarray data. Many preprocessing methods are developed originally for gene expression arrays since monitoring gene expression has been the primary application of microarrays. However, the principle of all microarrays, regardless of their biomedical applications, is the same: nucleic acid molecules are labeled and separated by hybridizing to probes with known sequences and quantification is done by measuring intensities carried on the labels. Therefore, the main sources of systematic and random noise are similar in microarrays generated in applications other than gene expression. Many preprocessing methods have proven to be useful in new platforms with small changes in implementation.

For example, non-specific binding continues to be a main source of background. Using probe sequence information to predict affinity to non-specific binding is shown to be useful in tiling arrays for ChIP-chip experiments. (32) use a model similar to Model 3 used in GCRMA, with additional quadratic terms of the number of each base in the probe. (33) show that the same model is useful in predicting and adjusting for non-specific binding for exon arrays.

Normalization is again a necessary component in preprocessing. The assumptions for normalization in some applications are met more easily than in gene expression. For examples, when microarrays are used to measure copy number variation, the assumption that the majority of the genes have two copies is probably acceptable in most situations.

The connections between various applications of microarrays allows preprocessing methods developed for gene expression to be a useful start, but specific development for different platforms are still necessary. For example, robust summaries similar to those used in gene expression arrays may be used for exon-specific expression levels. But with much smaller probe sets (average 4 probes per exon instead of over 10 per gene), it is much harder to define outliers and reach a balance between robustness and efficiency. The probe sequence, especially the GC content of probes, has been shown to affect the nonspecific binding. Background adjustment methods that adjust for the sequence bias largely remove the GC content effect However, for applications in epigenetic studies, since GC content are also related to methylation sites, simply removing the sequence effect in the same fashion as in expression or Chip-chip arrays may bleach out the desired signal. For all these reasons, preprocessing remains an active research field as microarray technology evolves.

References

1. Arteaga-Salas JM, Zuzan H, Langdon WB, Upton GJ, Harrison AP. An overview of image-processing methods for Affymetrix GeneChips. Brief Bioinformatics. 2008 Jan.9:25–33. [PubMed: 18057073]
2. Suárez-Fariñas M, Pellegrino M, Wittkowski KM, Magnasco MO. Harshlight: a “corrective make-up” program for microarray chips. BMC Bioinformatics. 2005; 6:294. [PubMed: 16336691]
3. Song JS, Maghsoudi K, Li W, Fox E, Quackenbush J, Liu SX. Microarray blob-defect removal improves array analysis. Bioinformatics. 2007 Apr.23:966–971. [PubMed: 17332024]
4. Wu Z, Irizarry R. Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. Proceedings of RECOMB 2004. 2004
5. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinfromatics. 2003; 19(2):185–193.
6. Rocke DM, Durbin B. A Model for Measurement Error for Gene Expression Arrays. Journal of Computational Biology. 2001; 8(6):557–569. [PubMed: 11747612]
7. Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics. 2002; 18(Suppl. 1):S105–S110. [PubMed: 12169537]
8. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Parameter estimation for the calibration and variance stabilization of microarray data. Statistical Applications in Genetics and Molecular Biology. 2003; 2(1) Article 3.
9. Wu Z, Irizarry RA. A Statistical Framework for the Analysis of Microarray Probe-Level Data. The Annals of Applied Statistics. 2007; 1(2):333–357.
10. Lin SM, Du P, Huber W, Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res. 2008 Feb.36:e11. [PubMed: 18178591]
11. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Science U S A. 2001; 98:31–36.
12. Zhang L, Miles MF, Aldape KD. A model of molecular interactions on short oligonucleotide microarrays. Nature Biotechnology. 2003; 21(7):818–821.
13. Naef F, Hacker CR, Patil N, Magnasco M. Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. Genome Biol. 2002; 3:RESEARCH0018. [PubMed: 11983059]
14. Naef F, Lim DA, Patil N, Magnasco M. DNA hybridization to mismatched templates: A chip study. Physical Review E. 2002; 65(040902):65.
15. Irizarry RA, Hobbs B, C F, Beaxer-Barclay Y, Antonellis K, Scherf U, Speed TP. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics. 2003; 4:249–264. [PubMed: 12925520]

16. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002; 1:1–9.

17. Cope LM, Irizarry RA, Jaffee H, Wu Z, Speed TP. A Benchmark for Affymetrix GeneChip Expression Measures. Bioinformatics. 2004; 20:323–331. [PubMed: 14960458]

18. Naef F, Magnasco MO. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. Physical Review E. 2003; 68:011906.

19. Wu Z, Irizarry R, Gentlemen R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association. 2004; 99(468):909–917.

20. Sugimoto N, Nakano S, Katoh M, Matsumura A, Nakamuta H, Ohmichi T, et al. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry. 1995 Sep.34:11211–11216. [PubMed: 7545436]

21. Rocke DM, Durbin B. Approximate variance-stabilizing transformations for gene-expression microarray data. Bioinformatics. 2003 May.19:966–972. [PubMed: 12761059]

22. Schadt EE, Li C, Ellis B, Wong WH. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem Suppl. 2001 37:120–125. [PubMed: 11842437]

23. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol. 2001; 2:RESEARCH0032. [PubMed: 11532216]

24. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. Statistica Sinica. 2002; 12(1): 111–139.

25. Åstrand M. Contrast Normalization of Oligonucleotide Arrays. Journal of Computational Biology. 2003; 10(1):95–102. [PubMed: 12676053]

26. Amaratunga D, Cabrera J. Analysis of Data From viral DNA Microchips. Journal of the American Statistical Association. 2001; 96(456):1161–1170.

27. Affymetrix. Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. 2005. www.affymetrix.com/support/technical/technotes/plier_technote.pdf

28. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res. 2001 Jun.29:2549–2557. [PubMed: 11410663]

29. Reilly C, Wang C, Rutherford M. A Method for Normalizing Microarrays Using Genes That Are Not Differentially Expressed. Journal of the American Statistical Association. 2003; 98(464):868.

30. Calza S, Valentini D, Pawitan Y. Normalization of oligonucleotide arrays based on the least-variant set of genes. BMC Bioinformatics. 2008; 9:140. [PubMed: 18318917]

31. Tukey, JW. Exploratory Data Analysis. Addison Wesley; 1977.

32. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, et al. Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci USA. 2006 Aug.103:12457–12462. [PubMed: 16895995]

33. Kapur K, Xing Y, Ouyang Z, Wong WH. Exon arrays provide accurate assessments of gene expression. Genome Biol. 2007; 8:R82. [PubMed: 17504534]