Modelling the background correction in microarray data analysis

Angelo M. Mineo, Calogero Fede, Luigi Augugliaro, and Mariantonietta Ruggieri

Dipartimento di Scienze Statistiche e Matematiche “S. Vianelli” di Palermo
elio.mineo; fede; augugliaro; ruggieri@dssm.unipa.it

Summary. Microarray technology has been adopted in many areas of biomedical research for quantitative and highly parallel measurements of gene expressions. In this field, the high density oligonucleotide microarray technology is the most used platform; in this platform oligonucleotides of 25 base pairs are used as probe genes. Two types of probes are considered: perfect match (PM) and mismatch (MM) probes. In theory, MM probes are used to quantify and remove two types of error: optical noise and non specific binding. The correction of these two types of error is known as background correction. Preprocessing is an essential step of the analysis in which the intensity, read from each probe, is manipulated in order to obtain an expression measure for each gene. In this paper, we introduce a new method for the background correction by using a calibration approach based on a generalized linear mixed model (GLMM).

Key words: microarray, background correction, Affymetrix, GLMM

1 Introduction

In the last decade, microarray technology has been adopted in many areas of biomedical research for quantitative and highly parallel measurements of gene expressions. In this field, the high density oligonucleotide microarray technology is the most used platform; in this platform oligonucleotides of 25 base pairs are used as probe genes. Two types of probes are considered: perfect match (PM) are probes that match exactly to a specific target sequence; mismatch (MM) probes, which differ from the corresponding PM only by a single base in the 13th position of the base sequence. Typically, an Affymetrix GeneChip uses 16-20 probe pairs (probe set), each of them investigating a different part of the gene sequence. In theory, MM probes are used to quantify and remove two types of error: optical noise, that is the read intensity value when biological sample is not labeled, and non specific binding, that occurs when a single stranded DNA sequence binds to a probe sequence which is not completely complementary. The correction of these two types of error is known as background correction. From this point of view, preprocessing is an essential step of the analysis in which the intensity, read from each probe, is manipulated in order to obtain an
expression measure for each gene. In this paper, we introduce a new method for the background correction by using a calibration approach based on a generalized linear mixed model (GLMM). In the next section, we show the main approaches used in preprocessing; then we show our method, giving some results.

2 Review on preprocessing

Data from a microarray experiment are manipulated in order to obtain a single measure for each gene. This data manipulation is known as **preprocessing**; preprocessing acts as a link between probe level data and higher level measurements, used, for example, to identify differentially expressed genes, class discovery and class prediction. Preprocessing can be viewed as an operation involving three different steps:

1. **background correction**, in which probe intensities are corrected for optical noise \((O)\) and non-specific binding \((N)\);
2. **normalization**, in which probe intensities are corrected in order to assure that measurements from different arrays are comparable;
3. **summarization**, in which intensities obtained for a particular probe set are summarized in a single measure.

Usually, researchers use a combination of models, each one proposed originally to solve just one single step. A common model used for background correction [Aff01] is based on the following deterministic relations

\[
P_{Mj} = O_j + N_j + S_j, \quad M_{Mj} = O_j + N_j, \quad \text{with} \quad j = 1, \ldots, J
\] (1)

where \(P_{Mj}\) and \(M_{Mj}\) are the intensity values for the \(j\)-th probe pair, \(O_j\) is the optical noise, \(N_j\) is the non-specific binding and \(S_j\) is the specific signal. The measure of expression, proposed by Affymetrix in 1999 and based on the transformation \(P_{Mj} - M_{Mj}\), is called Average Difference.

In [Ial03] the authors have shown that the Average Difference produces an expression estimate with large variance; moreover, for about 1/3 of probes the intensity read by \(MM\) probes is larger than the \(PM\) intensity. For this reason, Affymetrix has introduced the Ideal Mismatch (IM). The adjusted \(PM\) intensity, proposed by Affymetrix [Aff01] and called MAS 5.0, is based on \(PM - IM\).

A similar approach, used by [LW01], is known as Model-Based Expression Index (MBEI). This model is useful because the variation in \(PM - MM\) differences, due to probe effect, is larger than the variation due to arrays. For any given gene, the model for \(PM - MM\) is given by:

\[
y_{ij} = P_{Mij} - M_{Mij} = \phi_j \theta_i + \epsilon_{ij}, \quad \sum_j \phi_j = J, \quad \text{with} \quad \epsilon_{ij} \sim N(0, \sigma^2)
\] (2)

where \(\theta_i\) denotes an expression index for the \(i\)-th sample and \(\phi_j\) denotes the effect of the \(j\)-th probe pair. A problem with this model is that some \(MM\) probes may poorly respond to changes in the expression level of the target gene. This phenomenon raises doubts on the correctness of using \(MM\) probes. In [Ial03] is also proposed a model
for only PM probes, called Robust Multi-array Average convolution (RMA). This model is motivated by looking at the probe intensity distribution; in particular, the PM intensity is modelled as sum of two components: a signal component (S) with an exponential distribution and a background component (bg) with a normal distribution, so

\[ PM = bg + S. \]  

By considering this model, the background correct probe intensities is given by

\[
E(S|PM = pm) = a + b \frac{\phi(pm/a) - \phi(pm/b)}{\phi(a) - \phi(pm/b)} - 1.
\]

Anyway, the assumptions of this model seem too simplistic; moreover, in this approach the background correction step is unlinked with the normalization step, so we do not know how normalization affects the estimation of the model parameters.

An unified approach is proposed in [Hal02], in which the authors use the simplest linear calibration model with a variance stabilization to obtain a new transformation of raw data called “generalized logarithm” (glog). A similar result is obtained in [DHHR02].

In [ZR05], the authors propose a new expression index based on the generalized logarithm and justified on the observation that the mean is a summary function that performs well when observations have the same variance.

### 3 A new statistical model for background correction

The background correction purpose is to correct data for both optical noise and non specific binding. To make this correction, by considering the relation (1), we propose a model for measurement errors in PM probe intensity as a function of the gene expression level. We suppose that both specific binding (B) and non specific binding (N) depend on gene level, whereas the optical noise (O) does not depend on it.

To estimate the model parameters, we use the SpikeIn95 data set from the SpikeIn experiment, where 16 genes are spiked in 59 arrays at known concentration from 0 to 1024 picoMolar (pM) in a Latin Square design. More information on this experiment can be found in [Ial03].

#### 3.1 The additive-multiplicative error model

A model widely used to describe the relationship between observed intensity and gene expression level is the additive-multiplicative error model that has been proposed for the first time by [RD01].

Let \( y \) be the observed intensity and \( x \) the real gene expression level, then the model is

\[
y = \alpha + x e^n + \epsilon
\]
where $\eta \sim N(0, \sigma_\eta^2)$ and $\epsilon \sim N(0, \sigma_\epsilon^2)$. See how $\eta$ and $\epsilon$ have different variance, of course.

The figure 1 shows how this model explains the increase of the standard deviation with the mean, whereas the figure 2 shows that the assumptions of this model are not suitable to explain the skewness of the intensities at null concentration.

Fig. 1. The graph shows the relation between observed $PM$ and gene level logarithm

Fig. 2. The histogram of $PM$s from 0 to 1000, related to null concentration
3.2 A hierarchical model

By choosing to measure the gene level on logarithmic scale $(x)$ and by observing the data structure of *SpikeIn95* dataset, we note that expected value of $PM$ intensity increases like an exponential function of $x$ (see figure 1). Furthermore, we observe that the variance can be considered proportional to the square of expected value

$$\text{Var}[PM] \propto E[PM]^2$$

and that the distribution of $(PM|x)$ is always skewed. The following model takes into account these observations.

Let $y_{ijk}$ be the intensity of the $i$-th probe of the $j$-th gene of the $k$-th array, $x_j$ the logarithm of the concentration of the $j$-th gene; then, we have

$$y_{ijk} = e^{\alpha + \beta x_j} + \epsilon_{ijk},$$

where $\epsilon_{ijk} \sim N(0, \sigma)$ and

$$E[y_{ijk}] = y_{ij} \sim \text{Gamma} \left( r, \frac{r}{e^{\alpha + \beta x_j}} \right)$$

3.3 Parameter estimation

By considering that $\epsilon$ has a normal distribution, we can summarize the data $y_{ijk}$ for each probe at every concentration level by means of the arithmetic mean of replicates, i.e. by means of the ML estimator of the location parameter: these are the $y_{ij}$ values of our model.

Then, to estimate the parameters from *SpikeIn95* dataset of (8), we first used a GLM with Gamma distribution and logarithmic link function. This model, even though describes well the variability, does not consider the structure of dependence of the data.

For this reason, instead of a simple GLM, we have correctly used a GLMM with random intercept; we suppose that the random intercept is distributed as a normal distribution [DHLZ00]; then, we have to estimate in this model the following parameters: $E[\alpha]$, $\text{Var}[\alpha]$ and $\beta$.

In the next section, we shall use these models to derive a transformation of $PM$ for the background correction.

4 The proposed method for the background correction

The proposed method uses the estimation of the parameters of the model (8), relative to the *SpikeIn95* data set. By using a fixed effect model (GLM), we propose a calibration process that aims to estimate the real concentration, that is the value $x$. For this purpose, we invert the linear relation between the expected value of the $PM$ intensity and $x$, obtaining the following transformation:
\[
\hat{x}'_{ijk} = \frac{\log(PM_{ijk}) - \hat{\alpha}}{\beta}.
\] (9)

If we use, more correctly, a mixed effect model (GLMM), when we invert the estimated linear relation, problems arise with the use of \(\alpha\), because it is a random variable. To solve this problem, we follow these steps to compute the corresponding transformation (10):

- we order the probes in the probe set for PM intensities;
- we compute the corresponding \(\hat{\alpha}\) by using the hypothesis that the random intercept is normally distributed and then by using the 16 quantiles (in this case we have 16 probes in the probe set) of the normal distribution with mean and variance equal to the estimate of \(E[\alpha]\) and \(Var[\alpha]\), respectively.

Then, the transformation is

\[
\hat{x}''_{ijk} = \frac{\log(PM_{ijk}) - \hat{\alpha}_i}{\beta}.
\] (10)

### 4.1 Results

We have verified the background correction on the same data, derived from the SpikeIn experiment, by using a cross validation technique. Both transformations (9) and (10) give the same results for the expected value. The graphs in figure 3 show how the correction works well for concentrations from 32 to 1024 pM, whereas underestimates the intensity for lower concentrations. Furthermore, comparing the graphs in figure 3, we can see how the transformation (10) improves the accuracy without compromising the bias.

At last, we have used the SpikeInSubset data set [Gal05] to compare our results with the results obtained with the RMA method, that uses the relation (4). Figure 4 shows how our transformation performs better in terms of expected value, whereas it’s not so good in terms of variance.

### 5 Conclusions

In this paper, we have shown a background correction based on the real concentrations that are known in the SpikeIn95 data set, so to obtain a value that represents the logarithm of the gene concentration expressed in picoMolar. Then, we have compared this correction with that one based on the RMA model, obtaining promising results: indeed, the proposed correction performs better than the RMA method for the expected value, even though the results seem to present an higher variance.

The correction for background is the first step of preprocessing. We are working on synthesis measures to realize a correct summarization of probe set to reduce the variance. In the same way, we are working on parameter estimation methods that minimize directly the variance of the calibration estimates.

Using known concentrations to determine the background correction could seem a limitation of this method. Actually, we think that the estimated parameters and the relative background correction relations could be used for similar gene types of those ones used in the SpikeIn experiment. Of course, we hope to have some new data to verify the goodness of this proposal for other data set, but we think that the approach can be useful in every situation of the same type.
Fig. 3. Gene expressions and correct PM intensities, on logarithmic scale, for the transformation based on the fixed effect model (on the left) and on the mixed effect model (on the right). The dashed line shows the means of the estimates.

Fig. 4. Gene expressions and correct PM intensities, on logarithmic scale, for the RMA transformation. The dashed line shows the means of the estimates.
References


