

A new proposal for microarray background correction by means of a GLMM ^(*)

Una nuova proposta per la correzione di background in microarray attraverso un modello GLMM

Angelo M. Mineo, Calogero Fedè, Luigi Augugliaro, Marian Antonietta Ruggieri

Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università di Palermo
e-mail: fede@dssm.unipa.it

Riassunto: La tecnologia microarray ha il grosso pregio di misurare simultaneamente il livello di espressione di migliaia di geni. All'elevata quantità d'informazione fornita da un singolo chip si contrappone la necessità di un adeguato pretrattamento dei dati grezzi al fine di ottenere una misura "affidabile" del livello di espressione genetico. Scopo del lavoro è analizzare, attraverso un modello lineare generalizzato misto, il legame esistente fra il livello d'intensità osservato ed il livello di concentrazione, attraverso l'utilizzo degli esperimenti Spike-In forniti dall'Affymetrix. Si propone, quindi, un nuovo metodo per la correzione del background.

Keywords: microarray; preprocessing; background correction; error model, 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, it is considered the high density oligonucleotide microarray technology, where oligonucleotides of 25 base pairs are used as probe genes. Two types of probes are used: *perfect match* (*PM*) are probes that match exactly to a specific target sequence; *mismatch* (*MM*) probes, which differ from the corresponding *PMs* only for the 13th position. Typically, an Affymetrix GeneChip[®] uses 16-20 probe pairs (*probe set*) and each of them investigates different parts of the gene sequence. In theory, *MM* probes are used to quantify the non-specific binding; we have non-specific binding when a single stranded DNA sequence binds to a probe sequence which is not completely complementary. From this point of view, preprocessing is an essential step of the analysis: in this phase the intensity, read from each probe, is manipulated in order to obtain the "real" expression measure for each gene. Preprocessing can be viewed as an operation involving three different steps called *background correction*, *normalization*, *summarization*. Usually, researchers use a combination of models, proposed originally to solve just one single step. The most common model used for background correction (Affymetrix, 2001) is based on the following relations

$$PM_j = O_j + N_j + S_j; \quad MM_j = O_j + N_j; \quad \text{with } j = 1, \dots, J, \quad (1)$$

^(*) Work partially supported by grants of University of Palermo.

where PM_j and MM_j 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. Irizarry et al. (2003) have shown that the transformation $PM_j - MM_j$ produces an expression estimate with large variance; moreover, for about 1/3 of probes the intensity obtained for MM probes is larger than PM intensity. For this reason, Affymetrix (2001) has introduced the *Ideal Mismatch (IM)*. The adjusted PM intensity, proposed by Affymetrix and known as MAS 5.0, is based on the difference $PM - IM$. A similar approach is used by Li and Wong (2001): this method is known as Model-Based Expression Index (MBEI). 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 Irizarry et al. (2003) it is proposed a model for only PM probes, called Robust Multi-array Average convolution (RMA). This model considers the PM intensity as sum of two components: a signal component (S) with an exponential distribution and a background component (bg) with a normal distribution. Anyway, the assumptions of this model seem to us too much simplistic and maybe unrealistic; moreover, in this approach the background correction step is unlinked with the normalization step, so we do not know how normalization affects the model parameters. An unified approach is proposed by Huber et al. (2002), in which the authors use a simple linear calibration model with variance stabilization, to obtain a new transformation of raw data, called generalized logarithm (glog). A similar result is obtained by Durbin et al. (2002). In this paper, we introduce a new method for background correction by using a generalized linear mixed model (GLMM).

2. A new statistical model for background correction

The background correction purpose is to correct data both for optical noise and non specific binding. To make this correction, we propose a model for the measurement errors in PM probe intensity as a function of the gene expression level. We use the data set from the *SpikeIn95* experiment, where 16 genes 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 Irizarry et al. (2003).

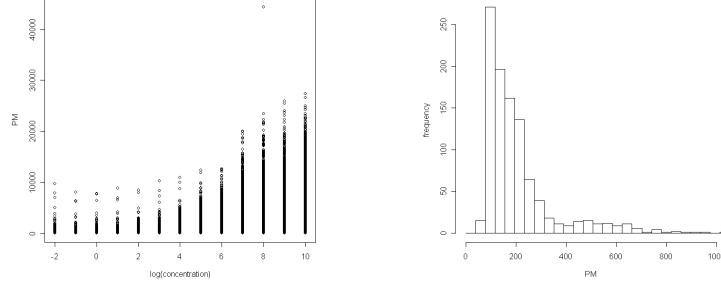
A model, widely used to describe the relationship between observed intensity and gene expression level, is the *additive-multiplicative error model* (Rocke and Durbin, 2001). Let y be the intensity observed and x the real gene expression level; then, we can write the model as

$$y = \alpha + \beta x e^{\eta} + \epsilon, \quad \text{where } \eta \sim N(0, \sigma_{\eta}^2) \text{ and } \epsilon \sim N(0, \sigma_{\epsilon}^2). \quad (2)$$

Figure 1 shows that this model explains the increase of the standard deviation with the mean, but is not able to explain the asymmetry of the data at null concentration.

For this reason, we propose a model that considers the asymmetric distribution of PM observed for each value of x and the hierarchical structure of data, where more than one probe is related to the same gene. Let y_{ijk} be the intensity of the i -th probe of the j -th gene on the k -th array, x_j the logarithm of the concentration of the j -th gene. Then, the model is:

Figure 1: The first graph shows the logarithm of gene level and PM. The second graph is the histogram of PMs from 0 to 1000, related to null concentration.



$$y_{ijk} = e^{\alpha_i + \beta x_j} + \epsilon_{ijk}, \quad (3)$$

where $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$ and

$$E_\epsilon[y_{ijk}] = y_{ij} \sim \text{Gamma}\left(r, \frac{r}{e^{\alpha_i + \beta x_j}}\right). \quad (4)$$

By considering that ϵ_{ijk} has a normal distribution, we can summarize the data for each probe at every concentration level by means of the arithmetic mean of replicates, that is, by means of the ML estimator. To estimate the model parameters, we have first used a GLM with Gamma distribution and logarithmic link function. This model has the drawback that does not consider the data dependence structure. So, we propose a GLMM model with random intercept (Diggle et al., 2000). Then, we use the estimated model parameters to derive a transformation of the *PM* for background correction.

By using a fixed effect model (GLM), the transformation proposed is

$$\hat{x}'_{ijk} = \frac{\log(PM_{ijk}) - \hat{\alpha}}{\hat{\beta}}, \quad (5)$$

whereas using, more correctly, a mixed effect model (GLMM) we have

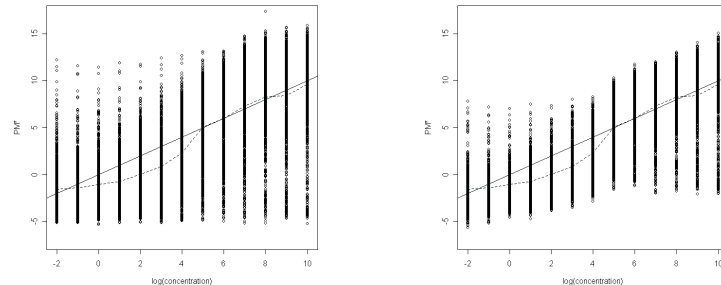
$$\hat{x}''_{ijk} = \frac{\log(PM_{ijk}) - \hat{\alpha}_i}{\hat{\beta}}, \quad (6)$$

where we use $\hat{\alpha}_i$ instead of $\hat{\alpha}$. To obtain the transformation linked to the GLMM, we use information about the position of the probe in the probe set, ordered by *PM* intensity.

3. Results and Conclusions

We have verified the background correction on the same data derived from the *SpikeIn95* experiment by using a cross validation technique. The figure 2 shows how the correction

Figure 2: The graphs show the relationship between gene expressions and correct PM intensities for both transformations, the first one related to the fixed effect model, the second one to the mixed effect model. The dashed lines show the means of the estimates.



works well for concentrations from 32 to 1024 pM , whereas underestimates for lower concentrations. At last, we can see how the second transformation improves the accuracy without increasing the bias.

The proposed correction is based on the real concentrations that are known in the *SpikeIn95* data set: this could seem a limitation of this method. In fact, we think that the estimated parameters and the relative background correction relations could be used for the same gene type as that used in this paper. 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.

References

- Affymetrix (2001) *Statistical algorithms reference guide*, Technical report, Affymetrix, Santa Clara, CA.
- Diggle P., Heagerty P. Liang K.Y., and Zeger S. (2000) *Analysis of Longitudinal Data*, University Press, Oxford.
- Durbin B.P., Hardin J.S., Hawkins D.M. and Rocke D.M. (2002). A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics*, 18 Suppl. 1, 105–110.
- Huber W., Von Heydebreck A., Sultmann H., Poustka A. and Vigneron M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, 18 Suppl. 1, 96–104.
- Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U. and Speed T.P. (2003) Exploration, Normalization and Summaries of High Density Oligonucleotide Array Probe Level Data, *Biostatistics*, 4(2), 249–264.
- Li C. and Wong W.H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proceedings of the National Academy of Sciences USA*, 98(1), 31–36.
- Rocke D.M. and Durbin B. (2001) A Model for Measurement Error for gene expression arrays, *Journal of Computational Biology*, 8(6), 557–569.