

Prediction of the Gene Expression Measure by Means of a Generalized Linear Mixed Model

Previsione della misura dell'espressione genica mediante modello lineare generalizzato ad effetti misti

Angelo M. Mineo, Luigi Augugliaro, Calogero Fedè,
Mariantonietta Ruggieri

Dipartimento di Scienze Statistiche e Matematiche *Silvio Vianelli*,
Università di Palermo, e-mail: elio.mineo, augugliaro, fede, ruggieri@dssm.unipa.it

Riassunto: Negli ultimi anni la tecnologia microarray ha rivoluzionato la ricerca medica ed ha fornito la possibilità di identificare un sottoinsieme di geni, la cui variazione nel livello d'espressione può essere utilizzata per prevedere, ad esempio, il trattamento medico più idoneo da somministrare a un gruppo di pazienti. Uno dei problemi legati all'analisi dei dati microarray, ottenuti mediante piattaforma oligonucleotidica ad elevata densità, è la definizione di un'opportuna misura del livello d'espressione dei geni in esame. Scopo del presente lavoro è quello di definire una nuova misura del livello d'espressione dei geni derivata dal modello per la correzione del background proposto da Mineo *et al.* (2006).

Keywords: Microarray, background correction, gene expression measure, GLMM.

1. Introduction

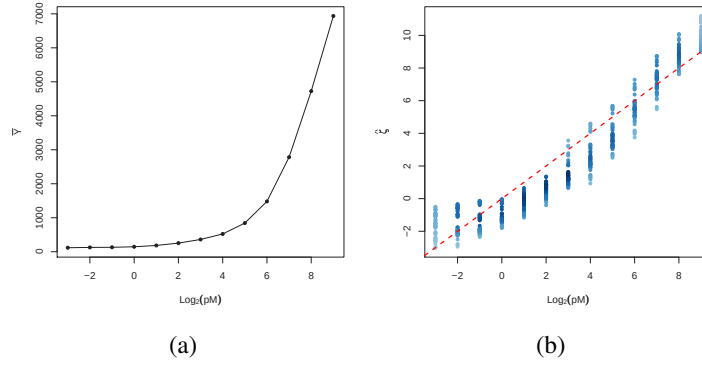
Microarrays permit to scientists the screening of thousands of genes simultaneously to determine, for example, whether those genes are active, hyperactive or silent in normal or cancerous tissues. A primary task in microarray analysis is to obtain a *good* measure of the gene expression that can be used for a so called *higher level analysis*. Different methods have been proposed for high density oligonucleotide arrays (see Cope *et al.* (2004) for a review). Aim of this paper is to obtain a new gene expression measure based on the background correction model proposed by Mineo *et al.* (2006). The proposed method is validated by means of a free available data-set called *Spike-In133* experiment, where 42 genes are spiked in 42 arrays at known concentration from 0 to 512 pico-Molar (pM).

2. Gene expression measure based on a GLMM

The background correction model, proposed by Mineo *et al.* (2006), is based on a generalized linear mixed model in order to take into account the data structure and the skewness of the distribution of the probe intensities. In order to obtain the gene expression matrix is necessary to define a summarization step, where the intensities read from a probe-set are summarized in a single measure. The proposed method is based on three steps; first of all, by means of the *Spike-In133* experiment, we estimate the calibration curve for the

⁰Work partially supported by grants of the University of Palermo.

Figure 1: The panel (a) shows the relationship between the mean intensity ($\bar{Y}$) and the amount of concentration in pico-Molar (pM) on $\log_2$ scale. The panel (b) shows the relationship between calibrated gene expression measure ($\hat{\zeta}$) and $\log_2 pM$; the dashed line represents the ideal situation.



logarithm of the mean intensities (figure 1(a)); then, we use a generalized linear mixed model to obtain a non-calibrated gene expression measure and finally we use the calibration curve to calibrate the gene expression measures. Formally, let Y_{ijk} be the intensity read from the i -th probe of the j -th gene of the k -th array and ζ_{jk} the calibrated expression measure; we assume that Y_{ijk} is Gamma distributed and for the j -th gene of the k -th array we use the model:

$$\log[E(Y_{ijk}|\alpha_i)] = \alpha_i + \theta_{jk}, \quad \alpha_i \sim N(0, \sigma^2), \quad (1)$$

where θ_{jk} is the non-calibrated expression measure. By using a linear calibration curve for the logarithm of the mean intensity, a calibrated expression measure for the j -th gene on the k -th array is defined in the following way

$$\hat{\zeta}_{jk} = \frac{\hat{\beta}_0 - \hat{\theta}_{jk}}{\hat{\beta}_1}. \quad (2)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the calibration line parameters. Figure 1(b) shows the calibrated gene expression measures obtained with the proposed method. The results are better than those obtained with other popular methods. Moreover, this procedure can be used to link the preprocessing step with the usual *higher order analysis*; for example, to identify a subset of differentially expressed genes, we can define a new linear predictor (1) in order to use all the information contained in the experimental design.

References

- Cope L., Irizarry R., Jaffee H.A., Wu Z. and Speed T.P. (2004) A benchmark for Affymetrix GeneChip expression measures, *Bioinformatics*, 20, 3, 323–331.
- Mineo A.M., Fede C., Augugliaro L. and Ruggieri M. (2006) Modelling the background correction in microarray data analysis, in: *COMPSTAT 2006 - Proceedings in Computational Statistics*, Rizzi A. and Vichi M., eds., IASC, Physica-Verlag, 1593–1600.