

A model-based normalization for cDNA microarray experiments¹

Una normalizzazione basata su modello per esperimenti di tipo cDNA microarray

Federico M. Stefanini

Dipartimento di Statistica 'G.Parenti', viale Morgagni 59, Firenze

Università degli Studi di Firenze

stefanin@ds.unifi.it

Riassunto: In un esperimento microarray migliaia di ibridazioni tra sequenze di DNA sono esaminate misurando l'intensità della fluorescenza che dipende dal numero di molecole ibridate e dalla variabilità del processo di misura. L'integrazione dei dati inerenti più esperimenti e il controllo statistico della qualità richiedono un'adeguata normalizzazione dei risultati. In questo lavoro viene proposto un modello grafico Bayesiano che sfrutta l'informazione contenuta nelle *blank spot* e nel *background* per ottenere la distribuzione dei parametri di normalizzazione. I dati sono preventivamente trasformati con una funzione motivata in questa sede. I risultati di un caso di studio inerente campioni di tessuto vegetale mostrano che potenziali artefatti sono identificabili impiegando il modello e che una larga parte della variabilità è attribuibile alle sorgenti incluse nel modello.

Keywords: Bayesian inference, DAG, DNA microarrays, MCMC.

1. Introduction

The possibility of monitoring RNA cell expression levels represents a huge advance in the genetical research (Brown and Botstein 1999) because RNA levels are widely determined by gene regulation. In a typical cDNA microarray experiment, templates for genes of interest (probes) are obtained from a sequence database or by amplification of biological material. Then a robot prints aliquots of purified DNA on a coated glass slide, thus a sequence of spots is placed onto the array. Print heads carry several print tips, e.g. 2 by 4 pins, thus spots printed by the same pin belongs to the same subarray. Total RNA from reference and treatment samples is labelled using fluorescent dyes, like Cy3 and Cy5 dUTP, by reverse transcription, then the two labelled samples, called targets, are pooled and allowed to hybridize with the DNA probes covalently linked to the array. A laser source excites dyes that emit light characterized by a dye-dependent spectrum. A scanning confocal laser microscope collects light emission from one dye and stores results as a monochrome 16 bit image. The two monochrome images are merged into one pseudocolored TIF 16 bit image, then an imaging software produces the raw fluorescence data.

Whether the interest focuses on the comparison between control and treatment samples or on the dynamical events and DNA motifs that determine the genetic regulation

¹ This paper is partially supported by COFIN 2001: "L'impiego di modelli grafici per avanzamenti nella biologia molecolare, nello studio dei processi di punto, e nell'analisi causale"

of a metabolic pathway, raw data must be normalized to account for several sources of experimental error: varying efficiency in transcription and amplification, fluctuations in labelling, print tip geometry, hybridization parameters, image analysis parameters, background noise and outshining from neighboring spots, localization and identification of spots (and many others). The normalization is even more important if the results obtained in different experiments on a given biological system are introduced into gene expression databases (Bassett, Eisen and Boguski 1999, Quackenbush 2001).

In this work, a graphical model is proposed for the fluorescence measured in blank spots and in the background. Blank spots and the background do not contain spotted DNA sequences, therefore fluorescence is due to experimental noise, e.g. SDS crystals, dust, etc.. The final distribution of model parameters is obtained by MCMC simulation. The proposed approach is tested on a numerical case dealing with vegetable tissues.

The start of this paper brings up the transformation applied to raw data, then the graphical model is described. In the subsequent section an actual case study is introduced, and the last section contains some open problems in view of a full analysis including thousands of different arrayed probes.

2. Methods

Let d, a, s, k be indices respectively for dye, array, subarray and replicated spots. Let $X_{R,d,a,s,k}$ be the random variable for the raw fluorescence of dye d in a spot indexed by a, s, k and $X_{B,d,a,s,k}$ be the random variable for the background. Transformed background values are denoted by $X_{\tilde{B},d,a,s,k}$, while centered transformed values by $X_{\underline{\tilde{B}},d,a,s,k}$.

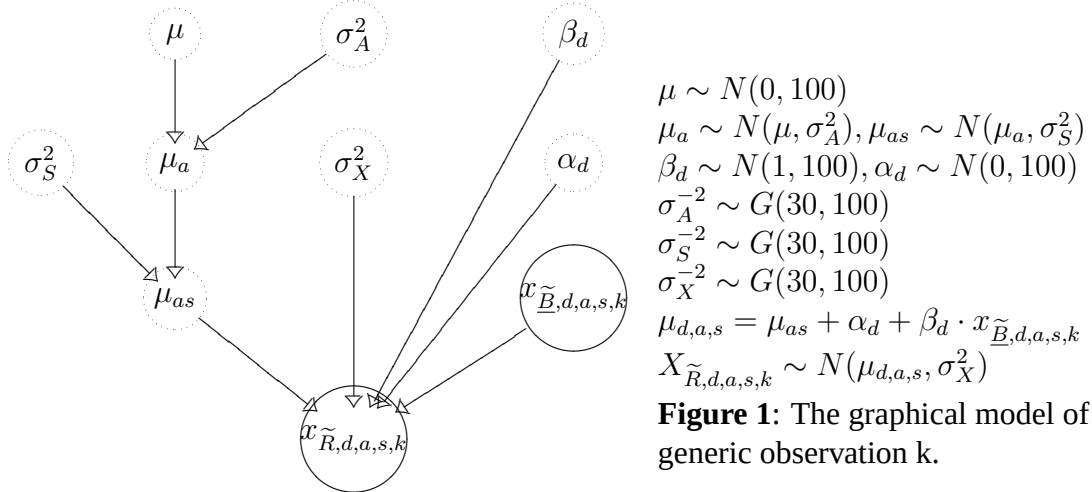


Figure 1: The graphical model of a generic observation k .

The transformation proposed in this work is motivated by the finite range of measurements. A 16 bits coded image carries fluorescence values in the set $\{0, 1, \dots, 2^{16} - 1\}$. The logarithmic transformation has been empirically found to be effective in making the distribution of fluorescence values approximately symmetric, and the base 2 logarithm is widely used at this purpose. Finally, although the calibration of fluorescence intensity is often unfeasible, the expected value of mRNA counts should be a monotonically increasing function of fluorescence intensity. Therefore a location parameter is meaningful before transforming raw values and it should be also informative after transformation.

The proposed transformation is $X_{\tilde{R},d,a,s,k} = \log_2 \left\{ \frac{X_{R,a,s,k}}{65535 - X_{R,a,s,k}} \right\}$.

The fluorescence of a regular spot that contains a probe is partially due to local protocol fluctuations (e.g. salt precipitation) that may be also present in the close neighborhood of the spot. The background fluorescence might be used to predict the amount of fluorescence due to local noise. A simple regression may be interpreted as an empirical approximation to the true but eventually complex model, but regression parameters may depend on the specific features of dyes. Moreover, array manufacturing and print tip wear may introduce local bias in fluorescence values. In Figure 1, a directed acyclic graph represents the conditional independence structure of the model and details on conditional distributions are given.

3. A case study

The experiment deals with cereal tissues collected from plants of two genotypes. Original data are unpublished, but it suffices to remark that two commercial arrays were used thus researchers did not set the array design and printing. The array layout is made by 2 by 4 subarrays, and each subarray is a square of 45 by 45 spots. The top two and the bottom two rows of each subarray contain control spots, in particular the blank spots analyzed in this work. It follows that $d = 1, 2$, $a = 1, 2$, $s = 1, 2, \dots, 8$ and $k = 1, \dots, n_s$.

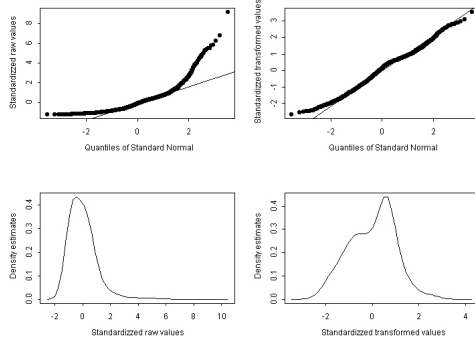


Figure 2: Graphical summaries of raw fluorescence.

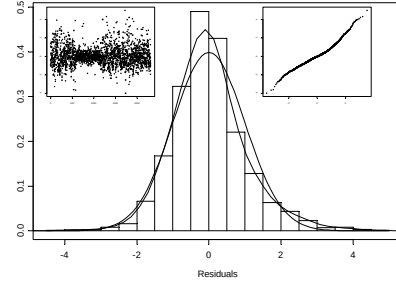


Figure 3: Graphical summaries of residuals.

In Figure 2, a quantile-quantile plot (top left) and a kernel density estimate (bottom left) are shown for standardized raw fluorescence values. The skewness is evident and no structure is apparent in the graph. In Figure 2 quantile-quantile plot (top right) and a kernel density estimate (bottom right) of transformed (and standardized) fluorescence values are also shown. The long right tail disappeared and evidence for the presence of major components of the variability, e.g. dye dependence, was found.

The Bayesian specification of the model was obtained by assigning a weakly informative normal prior distribution to all but the variance parameters. The initial distribution of the inverse of each variance was gamma with parameters 30 and 100. Markov Chain Monte Carlo simulation was used to approximate the posterior distribution of model parameters. The simulation software BUGS (Spiegelhalter, Thomas, Best and Gilks 1996) run for 51000 updates (9630 seconds, burn-in 1000 updates) with initial values selected

quite far from maximum likelihood estimates. A thin factor equal to 5 gave 10000 values to approximate the final joint distribution of parameters. Standard methods of output analysis to assess the convergence of chain (Spiegelhalter et al. 1996) were applied on all model parameters and no relevant violation was found. Among the relevant results, the posterior marginal means of β_1 and β_2 are, respectively, 0.648 and 0.758, suggesting that background subtraction is a suboptimal normalization procedure. Moreover, the posterior means of variances $\sigma_X^2, \sigma_A^2, \sigma_S^2$ are, respectively, 0.339, 3.400, 2.780, thus array and print tip variances are absolutely relevant in comparison with the pure error term.

The analysis of residuals was performed to assess the agreement of data with model assumptions. The histogram of standardized residuals (Figure 3) is overlapped by the standard normal probability density function and by a non-parametric density estimate. The estimated density shows a slightly right heavy tail and leptokurtosis. These features are also confirmed by the quantile-quantile plot in Figure 3, top right. The leptokurtosis is explained by the pattern of reduced variability shown by residuals from array 2, dye 1 in all the eight print tips (index values from 550 to 1100 in Figure 3 top left). This pattern might be due to wrong image processing or chemical anomalies for dye 1.

4. Discussion

The proposed model has meaningful parameters and it allows to assess the uncertainty of estimates. The analysis of residuals may be performed to investigate for the presence of experimental artifacts or model failures. The normalization of raw fluorescence data from spotted sequences may be based on the expected values of model parameters, or on the mode of the final distribution. Otherwise, the normalization of a spot corresponds to obtain the final marginal distribution of its sequence effect after extending the proposed graphical model accordingly.

Some model testing on a larger number of arrays and on spotted housekeeping probes is in progress to validate and to expand the proposed model. Further studies might focus on Bayesian model comparison, especially as regards the selection of distributions with heavier tails and of a larger class of graphical models in which the dependence among some model parameters is considered.

References

- Bassett, D. E., Eisen, M. B. and Boguski, M. S. (1999). Gene expression informatics - it's all in your mine, *Nature Genetics (suppl.)* **21**, 51–55.
- Brown, P. O. and Botstein, D. (1999). Exploring the new worlds of the genome with DNA microarrays, *Nature Genetics suppl.* **21**, 33–37.
- Quackenbush, J. (2001). Computational analysis of microarray data, *Nature reviews Genetics* **2**, 418–427.
- Spiegelhalter, D. J., Thomas, A., Best, N. G. and Gilks, W. R. (1996). *BUGS: Bayesian inference using Gibbs Sampling, Version 0.5*, MRC Biostatistics Unit, Cambridge.