Available online at www.sciencedirect.com

SCIENCE DIRECT®

C. R. Biologies 327 (2004) 175–180

COMPTES RENDUS BIOLOGIES

Biological modelling / Biomodélisation

Statistical analysis of oligonucleotide microarray data

Ziad Taib

Department of Biostatistics, AstraZeneca R&D, Mölndal, 431 83 Mölndal, Sweden

Received 15 April 2003; accepted 23 May 2003

Presented by Pierre Auger

Abstract

Microchip arrays have become one of the most rapidly growing techniques for monitoring gene expression at the genomic level and thereby gaining valuable insight about various important biological mechanisms. Examples of such mechanisms are: identifying disease-causing genes, genes involved in the regulation of some aspect of the cell cycle, etc. In this article, we discuss the problem of estimating gene expression based on a proper statistical model. More precisely, we show how the model introduced by Li and Wong can be used in its full bivariate generality to provide a new measure of gene expression from high-density oligonucleotide arrays. We also present a second gene expression index based on a new way of reducing the model into a simpler univariate model. In both cases, the gene expression indices are shown to be unbiased and to have lower variance than the established ones. Moreover, we present a bootstrap method aiming at providing non-parametric confidence intervals for the expression index. ***To cite this article: Z. Taib, C. R. Biologies 327 (2004).***

Résumé

Analyse statistique de données de puces d'ADN à base d'oligonucléotides. Les puces d'ADN sont devenues l'une des techniques les plus utilisées pour observer l'expression génétique à l'échelon génomique et ainsi mieux comprendre certains mécanismes biologiques. Comme exemples de ces mécanismes, citons l'identification de gènes causant certaines maladies héréditaires, ou l'identification de gènes impliqués dans la régulation de certains aspects du cycle cellulaires etc. Dans cet article, nous discutons le problème d'estimation de l'expression génétique à l'aide d'un modèle statistique adéquat. En effet, nous montrons comment le modèle introduit par Li et Wong peut être utilisé dans toute sa généralité multivariée pour obtenir une nouvelle mesure de l'expression génétique dans le cas de puces d'ADN d'oligonucléotide à haute densité. D'autre part, nous présentons une autre mesure de l'expression génétique basée sur un nouveau concept réduisant le modèle en un autre, univarié et plus simple. Nous montrons que les nouvelles mesures sont sans biais et que leurs variances sont plus petites que les variances des mesures existantes. Nous présentons aussi une méthode de *bootstrap* qui vise à créer des intervalles de confiance non-paramétriques pour les mesures d'expression génétique. ***Pour citer cet article : Z. Taib, C. R. Biologies 327 (2004).***

Keywords: gene expression; microchip array; model-based expression index; bootstrapping

Mots-clés : expression génétique ; puces d'ADN ; modèle statistique de l'expression génétique ; bootstrapping

E-mail address: ziad.taib@astrazeneca.com (Z. Taib).

1631-0691/$ – see front matter © 2004 Académie des sciences. Published by Elsevier SAS. All rights reserved.
doi:10.1016/j.crvi.2003.05.003

1. Introduction

In recent years, it has become possible to make simultaneous measurements of the expression levels of tens of thousands of genes using the so-called microarray technology. Through various microarray experiments, scientists try to answer various basic biological questions ranging from understanding some aspect of the cell cycle mechanism of a given organism to the identification of individual genes involved in some biochemical pathway related to some biological function or some disease of interest.

In this article, we discuss statistical modelling of data from high-density oligonucleotide arrays of the type manufactured by Affymetrix. The point of view we adopt is that statistical methods can and should be used in order to achieve a better understanding of experimental data. Although statistical methods can be useful at all the stages of a typical microarray experiment (experimental design, image analysis and artefact detection, normalization, calculation of gene expression, making comparisons between arrays, dimension reduction, clustering, etc.), we will only discuss the fundamental problem of estimating the expression index (cf. [1,2]).

The article is organized as follows. In Section 2, we describe how the data arise in some detail. The bivariate Li–Wong model (cf. [3,4]) is presented in Section 3. This model is used in Section 4 to provide maximum likelihood estimates of various parameters. A new way of reducing the model into a univariate one and the resulting estimates are discussed in Section 5. Comparisons between these two estimates as well as others are discussed in the discussion section.

The main conclusion of the article is that the new estimates seem to be better than previously established ones.

2. The nature of the data

First we take a look at some of the features of oligonucleotide-based microarrays. A single array (1.28 cm × 1.28 cm) can contain probe sets for tens of thousands of genes and ESTs. Every probe set consists of 10–20 probe pairs. Every probe pair contains a perfect match (PM) probe and a mismatch probe (MM). The perfect match probe is a small DNA subsequence (25 bases long), which is assumed to be specific for a particular gene, the expression of which we want to measure. The mismatch probe is identical to the perfect match probe, except for the base in the middle (13th) position.

The oligonucleotides in the probes are typically chosen from the so-called 3 prime end of the gene, but have been until recently unknown to the user. The expression level of the gene has to be inferred from the amount of hybridisation of the PM and MM probes. The latter are measured using a scanner, which is sensible to the fluorescence intensities of the probes. Up to now, expression indices have been based on PM–MM differences at the probe level. Contrary to what one expects, such differences are often negative. A trial study reveals that although very few probe sets will only have negative differences, a vast majority will have some. In fact, the median number of negative probes per probe set is around 8 (out of a total of 20). In many cases, the overall average of these differences for a probe set will be negative. Since this means that the corresponding gene has negative expression, it is obvious that such an average difference cannot be used. Various conditions are often imposed to prevent this from happening. These include using PM values only (cf. [1,2,5]), excluding probe pairs with negative differences, using PM $-$ c MM with some suitably chosen constant c (cf. [6]), treating negative differences as missing data and using imputation, etc. All these methods suffer from drawbacks such as being ad hoc, being inconsistent in that not all probe pairs are used in the same way, etc. This obscures the analysis.

At the most basic level, the data is in the form of pixel intensities (36 pixels/cell for the Mu11k mouse chip) for the individual probe cells. It goes without saying that one can (and should) try to model the data already at this level using ideas from statistical image analysis. In these notes, however, we will only consider models on the next level, namely that of PM and MM intensities. Another issue that we will not discuss here is that of background calculation and subtraction.

3. The model

The basic idea behind the Li–Wong model (cf. [3, 4]) is that the PM intensities are expected to be higher than the MM intensities. One can thus assume the following model:

$$\begin{aligned} M_{ij} &= \nu_{ij} + \theta_i \alpha_j + \varepsilon_{ij}^M \\ P_{ij} &= \nu_{ij} + \theta_i \alpha_j + \theta_i \phi_j + \varepsilon_{ij}^P \\ & i = 1, 2, \ldots, I, \text{ and } j = 1, \ldots, J \end{aligned} \tag{1}$$

where I is the number of arrays and J is the number probe pairs (usually 20), ν stands for the base line intensity of a probe pair due to non-specific hybridisation, α is the rate of increase in MM intensity of a probe pair and ϕ is the additional rate of increase in the corresponding PM response. The error terms are assumed to have zero mean and some common variance. The most important parameter of this model is the expression index θ.

All these quantities are assumed to be non-negative. Moreover in order to avoid 'unidentifiability', we impose some additional conditions like $\sum_j \phi_j^2 = J$. Although this model was introduced in [3], the authors have only treated the reduced case

$$Y_{ij} = P_{ij} - M_{ij} = \theta_i \phi_j + \varepsilon_{ij}^P - \varepsilon_{ij}^M \tag{2}$$

The authors of [7] use (1) explicitly, but assume that the PM and MM values are independent; so their model describes the marginal distributions. We propose to augment the model so as to take into account the empirically observed correlation between PM and MM, which is usually rather high. For example, the average correlation coefficient for 30 probe sets was 0.77 with a standard deviation of 0.14. The maximum value was 0.94 and the minimum 0.36 (cf. Fig. 1).

The rationale for this is that the probe pairs corresponding to the same gene are scattered all over the array, while the two components of the same probe are always adjacent to each other. More precisely, we assume that the error terms in (1) follow a bivariate normal distribution according to:

$$\begin{pmatrix} \varepsilon_{ij}^M \\ \varepsilon_{ij}^P \end{pmatrix} \approx N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c \\ c & \sigma^2 \end{pmatrix} \right) \tag{3}$$

where c is the covariance term and σ is the corresponding correlation coefficient.

How do we know that such a model fits the data? It is possible to examine the adequacy of the model by using the residuals to measure the lack of fit. The reader will find more information about this at the end of the next section.

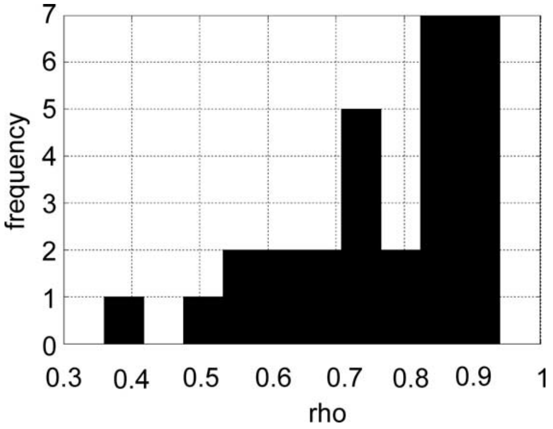

Fig. 1. Shows the distribution of the PM/MM correlation for 30 probe sets (ρ on the x-axis and the frequency on the y-axis).

4. The estimates

Given data (P_{ij}, M_{ij}), we can find parameter estimates using the maximum-likelihood method. The likelihood function has the form:

$$\begin{aligned} & L\big(P_{ij}, M_{ij}, \theta_i, \alpha_j, \rho, \nu_{ij}, \varphi_j, \sigma^2\big) = \prod_{i,j} K \\ & \quad \times \exp\left[-\frac{1}{2} \left(\begin{pmatrix} P_{ij} \\ M_{ij} \end{pmatrix} - \begin{pmatrix} \nu_{ij} + \theta_i \alpha_j + \theta_i \varphi \\ \nu_{ij} + \theta_i \alpha_j \end{pmatrix} \right)^T \right. \\ & \quad \left. \times \Sigma^{-1} \left(\begin{pmatrix} P_{ij} \\ M_{ij} \end{pmatrix} - \begin{pmatrix} \nu_{ij} + \theta_i \alpha_j + \theta_i \varphi \\ \nu_{ij} + \theta_i \alpha_j \end{pmatrix} \right) \right] \end{aligned}$$

The corresponding log likelihood function is

$$\begin{aligned} l = {} & \sum_i \sum_j \log K - \sum_i \sum_j \frac{1}{2\sigma^2(1-\rho^2)} \\ & \times \big(\big(P_{ij} - (\nu_j + \theta_i \alpha_j + \theta_i \varphi_j)\big)^2 \\ & + \big(M_{ij} - (\nu_j + \theta_i \alpha_j)\big)^2 - 2\rho\big(M_{ij} - (\nu_j + \theta_i \alpha_j)\big) \\ & \times \big(P_{ij} - (\nu_j + \theta_i \alpha_j + \theta_i \varphi_j)\big)\big) \end{aligned}$$

Taking the partial derivatives with respect to the parameters and setting the resulting expression equal to zero leads to maximum-likelihood estimates of the parameters.

$$
\begin{aligned}
\hat{\varphi}_j &= \frac{\sum_i \theta_i (P_{ij} - \rho M_{ij} - (1-\rho)(\nu_{ij} + \theta_i \alpha_j))}{\sum_i \theta_i^2} \\
\hat{\theta}_i &= \frac{\sum_j (P_{ij} - \nu_j - \rho(M_{ij} - \nu_{ij}))(\alpha_j + \varphi_j)}{\sum_j \varphi_j^2 + 2(1-\rho)\sum_j \alpha_j(\alpha_j + \varphi_j)} \\
&\quad + \frac{\sum_j (M_{ij} - \nu_{ij} - \rho(P_{ij} - \nu_{ij}))\alpha_j}{\sum_j \varphi_j^2 + 2(1-\rho)\sum_j \alpha_j(\alpha_j + \varphi_j)} \\
\hat{\alpha}_j &= \frac{\sum_i (P_{ij} - \nu_{ij} - \theta_i \varphi_j)\theta_i (1-\rho)}{2\alpha_j \sum_i \theta_i^2 (1-\rho)} \\
&\quad + \frac{\sum_i (M_{ij} - \nu_{ij})\theta_i (1-\rho)}{2\alpha_j \sum_i \theta_i^2 (1-\rho)} \\
\hat{\nu}_{ij} &= \frac{(2-\rho)\sum_i (PM_{ij} - \theta_i(\alpha_j + \varphi_j))}{2I} \\
&\quad + \frac{\sum_i (M_{ij} - \theta_i \alpha_j)}{2I}
\end{aligned}
\tag{4}
$$

These formulas have to be understood as steps in an iterative procedure that will lead to the final estimates. Nonetheless they are quite useful when it comes to deriving various properties. It is thus easy to see that the estimate, $\hat{\theta}_i$ of the expression index is unbiased, i.e. that $E[\hat{\theta}] = \theta_i$. The usual expression index (the average difference or $\overline{Y_i} = \frac{1}{J}\sum_j (P_{ij} - M_{ij})$ advocated by Affymetrix) lacks this property, as is seen by the following argument. In terms of the Li and Wong model, we have $E[\overline{Y_i}] = \theta_i \frac{1}{J}\sum_{j=1}^{J} \phi_j$. From Steiner's theorem, we know that $\frac{1}{J}\sum_{j=1}^{J} \phi_j \leqslant 1$ so $E[\overline{Y_i}] = \theta_i \bar{\phi} \leqslant \theta_i$, i.e. $\overline{Y_i}$ is unbiased.

To obtain non-parametric confidence intervals for the expression level, we propose the following bootstrap method. First, estimate the parameters using the above maximum likelihood estimates; then, estimate the means:

$$
\begin{aligned}
\widehat{M}_{ij} &= \hat{\nu}_{ij} + \hat{\theta}_i \hat{\alpha}_j \\
\widehat{P}_{ij} &= \hat{\nu}_{ij} + \hat{\theta}_i \hat{\alpha}_j + \hat{\theta}_i \hat{\phi}_j
\end{aligned}
\tag{5}
$$

It is now possible to estimate the residuals $\hat{\varepsilon}_{ij} = \binom{\hat{\varepsilon}_{ij}^{P}}{\hat{\varepsilon}_{ij}^{M}}$. These can be used to generate new observations by adding (bivariate) bootstrap errors to the expectations in (5). For every new set of observations, the expression level can be estimated, so we end up with a large number of possible expression values, which can then be used to calculate a confidence interval.

Notice also that the common PM/MM variance can be (ML) estimated by:

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{\sum_i \sum_j (P_{ij} - \widehat{P}_{ij})^2 + \sum_i \sum_j (M_{ij} - \widehat{M}_{ij})^2}{2IJ(1-\hat{\rho}^2)} \\
&\quad - \frac{2\hat{\rho}\sum_i \sum_j (P_{ij} - \widehat{P}_{ij})(M_{ij} - \widehat{M}_{ij})}{2IJ(1-\hat{\rho}^2)}
\end{aligned}
$$

and that variants of this can be used for outlier detection, i.e. we can classify a probe pair as an outlier if its residual is much higher than expected. Lack of fit can be measured using the residuals and standard methods of model fitting. In essence, standardized versions of the residuals should behave as independent drawings from a bivariate normal distribution.

5. Reduced models

As mentioned earlier, the Li–Wong model has, until now, been only used in its reduced form (2). This has the advantage of being a much simpler model than the full one. This reduced model has been implemented in dChip (one of the most popular tools for analysing this type of data). Other ways to get reduced models have also been investigated: using PM only values or the average difference. It is possible to formulate a reduced version of the Li–Wong model along the lines proposed in [6], i.e. by considering $P_{ij} - cM_{ij}$ for some suitable value of c.

In what follows, we present a new way of reducing the full model. The rationale of this method is that the MM values can, in a sense, be considered as containing help information only. One way to use this help information is to condition on it, i.e. to consider the conditional distribution of the PM values given the MM values. Under the assumption that the bivariate values are normal, the conditional distribution is also normal according to (cf. [8]):

$$
\begin{aligned}
P_{ij} | M_{ij} \approx N\big(&\upsilon_{ij} + \alpha_j \theta_i + \phi_j \theta_i \\
&+ \rho(M_{ij} - \upsilon_{ij} - \alpha_j \theta_i), \sigma^2(1-\rho^2)\big)
\end{aligned}
$$

Again, maximum likelihood estimates of the parameters can be based on this model. The resulting estimate of the expression level is:

$$
\hat{\theta}_i = \frac{\sum_j (P_{ij} - \rho M_{ij} - \upsilon_{ij}(1-\rho))(\alpha_j(1-\rho) + \phi_j)}{\sum_j (\alpha_j(1-\rho) + \phi_j)^2}.
$$

It is rather easy to verify that this estimate is unbiased.

6. Discussion

We have thus seen that using the Li–Wong model as a basis, one has the choice of using either the full bivariate model or some reduced univariate version. To make a rational decision as to the choice of a model, one can use both theoretical and empirical criteria as is done in [9]. One such criterion is unbiasedness. But perhaps the most natural of these criteria is the variance, i.e. preferring an estimate with a low variance over one with a large variance. Simple calculations show that according to this criterion the best model is the full bivariate model followed by the conditional and the reduced models in that order (cf. the Appendix). The comparison with the average difference is, however, not straightforward since that estimate has to be transformed to become unbiased. The transformed average difference has the largest variance of all reduced models.

A quite different way of comparing models is to use special experiments where known amounts of mRNA are used. Such data will become more available in the future.

It is interesting that the gene expression estimate based on the conditional model is of the type proposed by [6], i.e. of the form $P_{ij} - cM_{ij}$. In this estimate, the constant c is simply taken as ρ, the correlation coefficient. The conditional estimate has one additional nice feature, namely that it is a weighted average of $P_{ij} - \rho M_{ij}$ differences. Probe pairs with higher sensitivity (measured by $\alpha_j(1-\rho) + \phi_j$) are given a higher weight. The special cases $\rho = 0$ and $\rho = 1$ give the PM-only case and the reduced Li–Wong model respectively. To gain insight as to why ρ is a good choice of constant in the approach used in [6], one can argue as follows. The variance of each term $P_{ij} - cM_{ij}$ is simply $\sigma^2(1 + c^2 - 2\rho c)$ and is minimized when $c = \rho$.

To conclude, we have argued that, in the sense of unbiased estimates having low variance, the best estimate is simply the one based on the full bivariate model. Should one choose to use a univariate reduced model, then the estimate based on the conditional distribution is the next best choice. The latter has many desirable properties and for probe sets having high correlation coefficients, it is quite similar to the estimate based on the Li–Wong reduced model.

Appendix

In this appendix we give the variances of the different estimators mentioned in the article.

1. The estimate based on the full model:

$$\operatorname{var}[\hat{\theta}_i] = \frac{\sigma^2(1-\rho^2)}{\sum_j \phi_j^2 + 2(1-\rho)\sum_j \alpha_j(\alpha_j + \phi_j)} \tag{6}$$

2. The estimate based on the conditional model

$$\operatorname{var}[\hat{\theta}_i] = \frac{\sigma^2(1-\rho^2)}{\sum_j (\phi_j + \alpha_j(1-\rho))} \tag{7}$$

3. The estimate based on the reduced Li–Wong model

$$\operatorname{var}[\hat{\theta}_i] = \frac{2\sigma^2(1-\rho)}{\sum_j \phi_j^2} \tag{8}$$

4. The transformed average difference

$$\frac{2\sigma^2(1-\rho)}{J\bar{\phi}} \tag{9}$$

It is now not so difficult to show that (6) $\leqslant$ (7) $\leqslant$ (8) $\leqslant$ (9).

References

[1] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, T.P. Speed, Exploration, normalization, and summaries of high-density oligonucleotide array probe level data, Biostatistics 4 (2) (2002) 249–264.

[2] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, T.P. Speed, Summaries of Affymetrix GeneChip probe-level data, Nucleic Acids Res. 31 (4) (2003).

[3] C. Li, W.H. Wong, Model based analysis of oligonucleotide arrays: expression index computation and outlier detection, Proc. Natl Acad. Sci. USA 98 (2001) 31–36.

[4] C. Li, W.H. Wong, Model-based analysis of oligonucleotide arrays (II): model validation, design issues and standard error application, Genome Biol. 2 (8) (2001) 1–11.

[5] F. Naef, D.A. Lim, N. Patil, M.O. Magnasco, From features to expression: high-density oligonucleotide array analysis revisited, LANL e-print physics/0102010, 2001.

[6] B. Efron, R. Tibshirani, J.D. Storey, V. Tusher, Empirical Bayes analysis of a microarray experiment, Technical Report, Division of Biostatistics, Stanford University, 2001, JASA, in press.

[7] W.J. Lemon, J.J.T. Palatini, R. Krahe, F.A. Wright, Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays, Bioinformatics 18 (11) (2002) 1470–1476.

[8] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall, 1998.

[9] W.J. Lemon et al., 2001, in preparation.