

Bioconductor Project Working Papers

Bioconductor Project

Year 2004

Paper 6

Error models for microarray intensities

Wolfgang Huber* Anja von Heydebreck†
Martin Vingron‡

*Department of Molecular Genome Analysis, German Cancer Research Center

[†]Global Technologies, Merck KGaA

[‡]Department of Computational Molecular Biology, Max-Planck-Institute for Molecular Genetics

This working paper site is hosted by The Berkeley Electronic Press (bepress).

<http://www.bepress.com/bioconductor/paper6>

Copyright ©2004 by the authors.

Error models for microarray intensities

Abstract

We derive the additive-multiplicative error model for microarray intensities, and describe two applications. For the detection of differentially expressed genes, we obtain a statistic whose variance is approximately independent of the mean intensity. For the post hoc calibration (normalization) of data with respect to experimental factors, we describe a method for parameter estimation.

Error models for microarray intensities

Wolfgang Huber^{1*}, Anja von Heydebreck² and Martin Vingron²

¹ German Cancer Research Center, Department of
Molecular Genome Analysis, 69120 Heidelberg, Germany
Tel. +49 6221 42 4709, Fax +49 6221 4252 4709, w.huber@dkfz.de

² Max-Planck-Institute for Molecular Genetics, Department of
Computational Molecular Biology, 14195 Berlin, Germany
Tel. +49 30 8413 1168, Fax +49 30 8413 1152, [heydebre,vingron]@molgen.mpg.de

Abstract We derive the additive-multiplicative error model for microarray intensities, and describe two applications. For the detection of differentially expressed genes, we obtain a statistic whose variance is approximately independent of the mean intensity. For the post hoc calibration (“normalization”) of data with respect to experimental factors, we describe a method for parameter estimation.

Keywords error model; microarrays; differential expression; normalization; calibration; variance stabilization; parameter estimation.

*to whom correspondence should be addressed

1 Motivation

An error model is a description of the possible outcomes of a measurement. It depends on the true value of the underlying quantity that is measured and on the measurement apparatus. For microarrays, the quantities that one wants to measure are the abundances of specific molecules in a biological sample. The measurement apparatus consists of a cascade of biochemical reactions and an optical detection system with a laser scanner or a CCD camera. Biochemical reactions and detection are performed in parallel, allowing up to a million measurements on one array.

What is the purpose of constructing error models for microarrays? There are three aspects:

1. An appreciation of the distribution of all possible outcomes of a measurement is necessary for basing **inference** on one or a limited number of measurements. Consider an experiment in which we want to compare gene expression in the colons of mice that were treated with a certain substance and mice that were not. If we have many measurements, we can simply compare their empirical distributions. For example, if the values from ten replicate measurements for the DMBT1 gene in the treated condition are all larger than ten measurements from the untreated condition, the Wilcoxon test tells us that with a p -value of 10^{-5} the level of the transcript is really elevated in the treated mice. But often it is not possible, too expensive, or unethical, to obtain so many replicate measurements for all genes and for all conditions of interest. Often, it is also not necessary. If we are sufficiently confident in an error model, we are able to draw significant conclusions from fewer replicates.

2. An error model is an efficient tool for the summarization and **reporting** of experimental results. If we have reason to be confident that the measured outcomes follow a certain distribution, then they are sufficiently described by that distribution's parameters, e. g. mean and standard deviation; it may then not be necessary to report all of the individual measurements.

3. An error model is a summary of past experience and of our understanding of the measurement apparatus. It can be used for **quality control**: if the distribution of a new set of data greatly deviates from the model, this may direct our attention to quality issues with these data.

2 The additive-multiplicative error model

Consider the following generic observation equation

$$z = f(x, y), \quad (1)$$

where z is the outcome of the measurement, x is the true underlying quantity that we want to measure, the function f represents the measurement apparatus, and $y = (y_1, \dots, y_n)$ is a vector that contains all other parameters on which the functioning of the apparatus may depend. The functional dependence of f on some of the y_i may be known, on others it may not. Some of the y_i are controlled by the experimenter, some are not. If the measurement apparatus is well-constructed, then f is a well-behaved, smooth function, and we can write Eqn. (1) as

$$z = f(0, y) + f'(0, y) x + O(x^2), \quad (2)$$

where $f(0, y)$ is the baseline value that is measured if x is zero, f' is the derivative of f with respect to x , $f'(0, y)$ is a gain factor, and $O(x^2)$ represents non-linear effects. By proper design of the experiment, the non-linear terms can be made negligibly small within the relevant range of x . Examples for the parameters y in the case of microarrays are the efficiencies of mRNA extraction, reverse transcription, labeling and hybridization reactions, amount and quality of probe DNA on the array, unspecific hybridization, dye quantum yield, scanner gain, and background fluorescence of the array. All of these have an influence either on the baseline $f(0, y)$ or the gain $f'(0, y)$. Ideally, the parameters y could be fixed exactly at some value $\bar{y} = (\bar{y}_1, \dots, \bar{y}_n)$. In practice, they will fluctuate around \bar{y} between repeated experiments. If the fluctuations are not too large, we can expand

$$f(0, y) \approx f(0, \bar{y}) + \sum_{i=1}^n \frac{\partial f(0, \bar{y})}{\partial y_i} (y_i - \bar{y}_i) \quad (3)$$

$$f'(0, y) \approx f'(0, \bar{y}) + \sum_{i=1}^n \frac{\partial f'(0, \bar{y})}{\partial y_i} (y_i - \bar{y}_i). \quad (4)$$

The sums on the right hand sides of Eqns. (3) and (4) are linear combinations of a large number n of random variables with mean zero. Thus, it is a reasonable approximation to model $f(0, y)$ and $f'(0, y)$ as normally distributed random variables with means $a = f(0, \bar{y})$ and $b = f'(0, \bar{y})$ and variances σ_a^2 and σ_b^2 , respectively. Thus, omitting the non-linear term, Eqn. (2) leads to

$$z = a + \varepsilon + b x (1 + \eta), \quad (5)$$

with $\varepsilon \sim N(0, \sigma_a^2)$ and $\eta \sim N(0, \sigma_b^2/b^2)$. This is the **additive-multiplicative error model** for microarray data, which was proposed by [Ideker et al., 2001]. [Rocke and Durbin, 2001] proposed it in the form

$$z = a + \varepsilon + b x \exp(\eta), \quad (6)$$

which is equivalent to Eqn. (5) up to first order terms in η . Models (5) and (6) differ significantly only if the coefficient of variation σ_b/b is large. For microarray data, it is typically smaller than 0.2, thus the difference is of little practical relevance.

One of the main predictions of the error model (5) is the form of the dependence of the variance $\text{Var}(z)$ of z on its mean $E(z)$:

$$\text{Var}(z) = v_0^2 + \frac{\sigma_b^2}{b^2} (E(z) - z_0)^2, \quad (7)$$

that is, a strictly positive quadratic function. In the following we will assume that the correlation between ε and η is negligible. Then the parameters of Eqn. (7) are related to those of Eqn. (5) via $v_0^2 = \sigma_a^2$ and $z_0 = a$. If the correlation is not negligible, the relationship is slightly more complicated, but the form of Eqn. (7) remains the same.

3 Nesting

Consider a situation in which the quantity x from Eqn. (6) is itself the result of a process whose outcome is approximately described by an additive-multiplicative error model,

$$x = a' + \varepsilon' + b' x' \exp(\eta'). \quad (8)$$

The resulting distribution of z can again be described by an additive-multiplicative model with new parameters. This means that the class of models of the form (6) is closed under such hierarchical nesting, and the range of its applicability can be quite large.

For example, Eqn. (6) could be used as a model for microarray measurements in a study of diseased tissues, with x the true abundance of a certain gene transcript in the tissue from an individual patient, while (8) could model the population distribution of this gene's transcript levels in a set of similar tissues from different patients.

4 Detection of differentially expressed genes

Suppose we want to compare two measurements z_1, z_2 that are distributed according to Eqn. (6) with the same parameters a, b, σ_a , and σ_b , but possibly with different values of x_1, x_2 . We want to find a function $h(z_1, z_2)$ that fulfills the two conditions:

$$\begin{aligned} \text{(i) antisymmetry: } \quad h(z_1, z_2) &= -h(z_2, z_1) & \forall z_1, z_2 \\ \text{(ii) homoskedasticity: } \quad \text{Var}(h(z_1, z_2)) &= \text{const.} & \text{independent of } z_1, z_2 \end{aligned} \quad (9)$$

An approximate solution is given by [Huber et al., 2003]

$$h(z_1, z_2) = \text{arsinh} \left(\frac{z_1 - a}{\beta} \right) - \text{arsinh} \left(\frac{z_2 - a}{\beta} \right) \quad (10)$$

with $\beta = \sigma_a b / \sigma_b$. If both z_1 and z_2 are large, this expression coincides with the log ratio

$$q(z_1, z_2) = \log(z_1 - a) - \log(z_2 - a). \quad (11)$$

However, $q(z_1, z_2)$ has a large, diverging variance for $z_i \rightarrow a$, a singularity at $z_i = a$, and is not defined in the range of real numbers for $z_i < a$. These unpleasant properties are important for applications: many genes are not expressed or only weakly expressed in some, but not all conditions of interest. That means, we need to compare conditions in which, for example, x_1 is large and x_2 is small. The log ratio (11) is not a useful quantity for this purpose, since the second term will wildly fluctuate and be sensitive to small errors in the estimation of the parameter a . In contrast, the statistic (10), which is called the *generalized log-ratio* [Rocke et al., 2004], is well-defined everywhere and robust against small errors in a . It is always smaller in magnitude than the log ratio (see also Fig. 1),

$$|h(z_1, z_2)| < |q(z_1, z_2)| \quad \forall z_1, z_2. \quad (12)$$

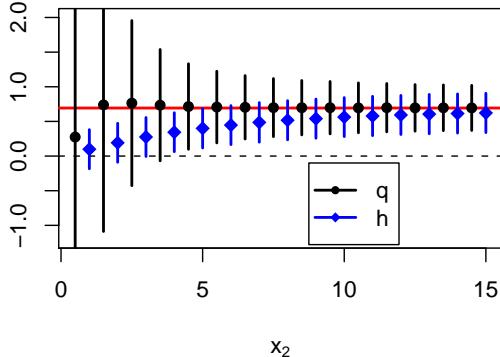


Figure 1: The shrinkage property of the generalized log ratio h . Blue diamonds and error bars correspond to mean and standard deviation of $h(z_1, z_2)$, cf. Eqn. (10), black dots and error bars to $q(z_1, z_2)$, cf. Eqn. (11). Data were generated according to Eqn. (6) with $x_2 = 0.5, \dots, 15$, $x_1 = 2x_2$, $a = 0$, $\sigma_a = 1$, $b = 1$, $\sigma_b = 0.1$. The horizontal line corresponds to the true log ratio $\log(2) \approx 0.693$. For intensities x_2 that are larger than about ten times the additive noise level σ_a , h and q coincide. For smaller intensities, we can see a *variance-bias trade-off*: q has no bias but a huge variance, thus an estimate of the fold change based on a limited set of data can be arbitrarily off. In contrast, h keeps a constant variance – for the price of systematically underestimating the true fold change.

The exponentiated value

$$\widehat{FC} = \exp(h(z_1, z_2)) \quad (13)$$

can be interpreted as a shrinkage estimator for the *fold-change* x_1/x_2 . It is more specific, i. e. leads to fewer false positives in the detection of differentially expressed genes, than the naive estimator $(z_1 - a)/(z_2 - a)$ [Huber et al., 2002, Durbin et al., 2002].

5 Normalization and parameter estimation

The explanatory power of the model (6) can be greatly increased if we take into account the systematic dependence of its parameters on known experimental factors. This is often called *normalization*. A parametrization that captures the main factors that play a role in current experiments is

$$z_{ip} = a_{i,s(p)} + \varepsilon + b_{i,s(p)} B_p x_{j(i),k(p)} \exp(\eta). \quad (14)$$

Here, p indices the probes on the arrays and $k = k(p)$ the genes. Each probe is intended to represent exactly one gene, but one gene may be represented by several probes. B_p is the probe affinity

of the p -th probe [Li and Wong, 2001, Irizarry et al., 2003]. i counts over the arrays and, if applicable, over the different dyes. $j = j(i)$ labels the biological conditions (e. g. normal/diseased). $a_{i,s(p)}$ and $b_{i,s(p)}$ are normalization offsets and scale factors that may be different for each i and possibly for different groups of probes $s = s(p)$. Probes may be grouped according to their physico-chemical properties [Wu and Irizarry, 2004] or array manufacturing parameters such as print-tip [Dudoit et al., 2002] or spatial location. In the simplest case, $a_{i,s(p)} = a_i$ and $b_{i,s(p)} = b_i$ are the same for all probes on an array [Beißbarth et al., 2000]. The noise terms ε and η are as above.

A method for the estimation of these parameters that uses the variance stabilizing transformation (10) was described by [Huber et al., 2002, Huber et al., 2003]; software is available as an R package [Huber et al., 2004].

References

- [Ideker et al., 2001] T. Ideker, V. Thorsson, A.F. Siegel, and L.E. Hood. Testing for differentially expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, 7:805–818, 2000. [3](#)
- [Rocke and Durbin, 2001] D.M. Rocke and B. Durbin. A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 8:557–569, 2001. [3](#)
- [Rocke et al., 2004] D. Rocke, W. Huber, B. Durbin, A. von Heydebreck, and M. Vingron. Interpretability and data transformations for gene expression microarray data. Preprint, 2004. [4](#)
- [Huber et al., 2002] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl. 1:S96–S104, 2002. [5](#), [6](#)
- [Durbin et al., 2002] B.P. Durbin, J.S. Hardin, D.M. Hawkins, and David M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18 Suppl. 1:S105–S110, 2002. [5](#)
- [Li and Wong, 2001] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS*, 98:31–36, 2001. [6](#)
- [Irizarry et al., 2003] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonelis, U. Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–64, 2003. [6](#)
- [Wu and Irizarry, 2004] Z. Wu and R.A. Irizarry. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. Proceedings of RECOMB 2004, 2004. [6](#)

[Dudoit et al., 2002] S. Dudoit, Y.H. Yang, T.P. Speed, and M.J. Callow. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 97:77–87, 2002. 6

[Beißbarth et al., 2000] T. Beißbarth, K. Fellenberg, B. Brors, R. Arribas-Prat, J.M. Boer, N.C. Hauser, M. Scheideler, J.D. Hoheisel, G. Schütz, A. Poustka, and M. Vingron. Processing and quality control of DNA array hybridization data. *Bioinformatics*, 16:1014–1022, 2000. 6

[Huber et al., 2003] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical Applications in Genetics and Molecular Biology*, 2(1), 2003. 4, 6

[Huber et al., 2004] W. Huber. Vignette: Robust calibration and variance stabilization with vsn, 2004. The bioconductor project, <http://www.bioconductor.org>. 6