Computers in Biology and Medicine 42 (2012) 19–29

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Evaluating the effect of various background correction methods regarding noise reduction, in two-channel microarray data

E.G. Sifakis [a], A. Prentza [b], D. Koutsouris [a], A.A. Chatziioannou [c,*]

[a] Biomedical Engineering Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, Greece
[b] Department of Digital Systems, University of Piraeus, Piraeus, Greece
[c] Institute of Biological Research & Biotechnology, National Hellenic Research Foundation, 48 Vassileos Constantinou Ave., Athens 11635, Greece

ARTICLE INFO

Article history:
Received 26 June 2010
Accepted 13 October 2011

Keywords:
Background estimation
DNA microarrays
Normalization techniques
Self-versus-self hybridizations
Transcriptomic analysis

ABSTRACT

In this work, two novel background correction (BC) methods, along with several commonly used ones, are evaluated regarding noise reduction in eleven two-channel self-versus-self (SVS) hybridizations. The evaluation of each BC method is investigated under the use of four statistical criteria combined into a single measure, the polygon area measure. Overall, our proposed BC approaches perform very well in terms of the proposed measure for most of the cases and provide an improved effect regarding technical noise reduction.

1. Introduction

DNA microarrays represent a fairly recent, yet powerful high-throughput technology, allowing the simultaneous measurement of the expression levels of thousands of genes in a single experiment. Maturation of the printing technologies enables nowadays full coverage of an organism's genome within a slide, thus alleviating costs and reducing batch related noise. There are numerous variations regarding DNA microarrays experimental implementation, yielding numerous technical protocols, which can generally be categorized in single channel and double channel experiments, regarding popularity of use, though there are experimental implementations entailing a larger number of dyes. Regarding two-channel microarrays, their analysis encompasses various steps, which are summarized in the following scheme. The collection of biological material represents the starting step, a stage encompassing RNA isolation and labeling with fluorescent dyes (i.e. Cy3, Cy5) excited at different wavelengths. The next step is that of hybridization with the reporters fixed on the surface of the microarray slide, followed by that of image acquisition, where two independent images, each one corresponding to a specific dye, are generated. These images are then segmented in order to identify the arrayed features, and to measure the relative hybridization intensities in a pixel–by-pixel basis for each channel from which subsequently comparative gene expression is inferred through comparison of channel values.

Most DNA microarrays software, commercial or freely distributed, provides a variety of summary statistics outputs per channel and feature (usually Green (Cy3) for the reference sample, and Red (Cy5) for the treated sample), encompassing estimates of total intensity, the mean and median of the pixel intensity distribution (foreground intensity), as well as an estimate of these for the local background (background intensity) [1]. Among these, the median intensity values tend to be more popular, since they represent a robust measure of central tendency of the data (especially when compared to the mean which is sensitive to the presence of outliers) [2]. In each channel, the foreground intensities f^R and f^G represent estimates of the specific binding of the labeled mRNA to the spotted reporters, whereas the local background intensities, b^R and b^G, comprise measures of the infiltrating noise in each channel, which encumbers the interpretation procedure.

The sources of this noise are multiple and varying in nature, like (i) fluorescence radiation due to non-specific hybridization, (ii) over-shining (fluorescence from neighboring features), or (iii) technical imperfections throughout various stages of the experimentation (like incomplete washing after hybridization, undesired noise due to laser over-excitation of the fluorescent dyes, scanner operation close to its saturation phase, or imprecision in feature segmentation during image analysis) [3–5]. It is thus important for the definite interpretation of the experiment that an accurate estimate of the background noise is derived for the scope of the correction of the foreground intensities and the

* Corresponding author. Tel.: +30 210 7273751; fax: +30 210 7273758.
E-mail address: achatzi@eie.gr (A.A. Chatziioannou).

0010-4825/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2011.10.003

most accurate possible signal estimation of the expression of the transcripts. Noteworthily, a systematic evaluation of the impact of the background noise-correction estimating methodologies remains elusive despite its potentially critical role as regards the biological interpretation and validation of the experimental results.

The most usual approach for background correction (BC) is based on an additive background noise model. Specifically, this model assumes that any foreground intensity is the sum of two components [6], one representing latent true gene expression, and the other representing background noise. The true unknown gene expression is assumed to increase proportionally with the true gene concentration (in the corresponding sample), while the background noise is assumed to be mathematically independent of it. Moreover, it is assumed that true gene expression and background noise are additive and probabilistically independent.

Extrapolating from the additive background noise model, the local background subtraction (LBS) method utilizes the feature background estimates in each channel, b^R and b^G, and directly subtracts them from the foreground intensities, f^R and f^G, respectively. However, this conventional method produces very often, feature intensity distributions with undesirable statistical properties. For example, in cases where the background intensities are measured larger than the corresponding foreground ones (i.e. in presence of local background artifacts), then non-positive background-corrected intensities are produced (the production of negative background-corrected intensities is nonsensical and suggestive of a flaw in using local background to estimate nonspecific hybridization, as characterized by Brown et al. [4]), leading to missing log-ratios for a rather nontrivial number of features. Besides, even in the case where local foreground intensities are larger than the local background estimates, background subtraction tends to pile together feature measurements with extremely different qualitative characteristics, i.e. considering two-channel signals, where the local foreground intensities have been measured 100 and 1000, whereas the local background ones 50 and 950, respectively. Then by applying the additive approach, we would have for both cases that their expression is equal, namely 50. What seems to elude in this approach is an assessment of the quality of the measured signal, where in the former case the signal is two times the background, whereas in the latter it is just 5% over the background level (1.0526). By taking into consideration the variation of noise, which in microarray experiments is pretty high (far more than 20%), it can be easily conjectured that the latter estimate represents an artifact. Moreover, as the subsequent stages of microarray analysis are built upon these results, this leads to a heavy distortion of the total signal population. Henceforth, the statistical techniques used are incapable of dealing with the problem of the potentially massive introduction of artifact values.

Therefore, the LBS approach generates highly variable, low intensity, feature values (appearing as so-called 'fishtail' or 'fan' patterns in the scatter plots of log-ratios versus the average of log intensities), as pointed out by many researchers [7–10]. All these facts ground a pinpointed skepticism regarding the widespread use of the LBS correction method, and stress the necessity of tackling the problem of background correction in alternative fashions.

An alternative to the LBS method is the constant background subtraction (CBS) approach. Instead of subtracting local background estimates for each feature, a global background is estimated and subtracted for all features. This global background may be estimated either by a set of negative control features or by a percentage of all feature foreground intensities [7,11]. However, not all microarray platforms support negative control features, whereas the selection of the percentage threshold is a rather empirical one, based on case-specific assay of each slide and channel. As it was already explained for the LBS method, a critical limiting factor for the CBS approach is the fact that background signal distributions in microarray experiments can be pretty inhomogeneous, due to intensity dependencies, originating directly from the experimental technology applied.

Another approach, recommended by many researchers [7,8,12,13] is rather to use uncorrected (no background correction altogether—NBC) compared to local background subtracted foreground feature intensities. The approach is suggested as a better option, since NBC does not depend on potentially problematic background estimates, while allegedly it is resilient with respect to low intensities [13]. On the other hand, NBC reduces the ability to identify differentially expressed genes [7,11], while it is severely underperforming in cases where significant spatial artifacts are encountered, induced only in one channel [13].

More sophisticated image approaches incorporate the utilization of morphological features, like for instance the morphological opening (MO), which appears in Spot software (CSIRO, North Ryde, Australia, publicly available at http://www.hca-vision.com/product_spot.html), and is a non-linear filter. According to [7,11], MO provides a better balance in the bias-variance trade-off when compared with LBS, CBS, and NBC, while at the same time, enhances the identification of differentially expressed genes by increasing the magnitude of t-statistics. Another approach of this category is the $TV+L^1$ model developed by Yin et al. [14], which is also a non-linear filter. The proposed model estimates closer background-corrected intensity to the true foreground signal, when compared with MO. However, despite the enhanced performance of these non-linear filters, there are few individual limitations and disadvantages of each algorithm [14]. Moreover, from the technical point of view, such non-linear filter approaches are extremely demanding computationally when compared to other methods. But more importantly, they fail providing a functional interface with usual (commercial or freely distributed) microarray analysis software, and consequently, in most cases, are unavailable for the processing of DNA microarray data, which averts their widespread incorporation to DNA microarray analysis pipelines.

In order to avoid the defects induced by straightforward subtraction of raw background intensities, other BC methods are proposed. Like log-linear interpolation [15], empirical Bayesian modeling based on the additive nature of background noise [16], or model-based methods of stabilizing variance, incorporating additive components [17,18]. Smyth [9] (limma software) adapts the BC approach originally introduced by Irizarry et al. [19] for Affymetrix data for use with two-channel microarrays. Silver et al. [20] further develop the normal-exponential convolution model (*normexp*) by improving the estimation of its parameters. Schutzenmeister and Piepho [21] propose local background smoothing with either 2D locally weighted regression or ordinary kriging using an isotropic model of spatial correlation prior to applying a BC algorithm. Argyropoulos et al. [22] derive an approximation to the unknown distribution of the infiltrating fluorescent using the method of maximum entropy, while utilizing it to estimate the magnitude of background noise by segmenting the image histogram and to correct individual pixels for the presence of noise using the maximum likelihood estimator.

The multiplicative background correction (MBC), as proposed by Zhang et al. [10], appears to be one of the simplest and most promising strategies for BC. In contrast to the additive background noise model, MBC assumes that the background noise affects the feature intensities in a multiplicative manner, a notion fully complying with the perception of the experimentalist about signal quality, emphasizing in the strength of the signal compared to that of the noise. MBC represents a background correction method, based upon the concept of the signal-to-noise ratio (SNR)