Background adjustment of cDNA microarray images by Maximum Entropy distributions

Christos Argyropoulos a,b, Antonis Daskalakis a, George C. Nikiforidis a, George C. Sakellaropoulos a,*

a Department of Medical Physics, School of Medicine, University of Patras, GR-26504 Rion, Greece
b Renal and Electrolyte Division, University of Pittsburgh Medical Center, A919 3550 Terrace Street Pittsburgh, PA 15261, USA

A R T I C L E   I N F O

Article history:
Received 13 January 2009
Available online 31 March 2010

Keywords:
Maximum Entropy
cDNA
Microarray
Image segmentation
Image restoration
Gene expression

A B S T R A C T

Many empirical studies have demonstrated the exquisite sensitivity of both traditional and novel statistical and machine intelligence algorithms to the method of background adjustment used to analyze microarray datasets. In this paper we develop a statistical framework that approaches background adjustment as a classic stochastic inverse problem, whose noise characteristics are given in terms of Maximum Entropy distributions. We derive analytic closed form approximations to the combined problem of estimating the magnitude of the background in microarray images and adjusting for its presence.

The proposed method reduces standardized measures of log expression variability across replicates in situations of known differential and non-differential gene expression without increasing the bias. Additionally, it results in computationally efficient procedures for estimation and learning based on sufficient statistics and can filter out spot measures with intensities that are numerically close to the background level resulting in a noise reduction of about 7%.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

The advent of complementary DNA (cDNA) microarray technologies enabled the simultaneous and specific [1] assessment of the expression levels of thousands of genes [2,3]. The basic microarray procedure involves hybridization of complementary nucleic acid molecules, one of which (target) has been immobilized in a solid substrate (e.g. glass) using a robotically controlled device (arrayer).

Such targets form spots at the vertices of a rectangular lattice on the solid substrate surface; each spot then serves as a highly specific and sensitive detector of the corresponding gene. It is widely appreciated that analysis of such datasets is confounded by a number of technical factors which operate at the different stages, of the microarray pipeline [4]. These factors not only result in as irregularities of spot features [5] but also impart the final microarray image with a non-negligible background.

The most common procedure for correcting for the presence of the background takes place after segmentation of the microarray image into spot (foreground) and background (regions), and involves the subtraction of an estimate (usually the mean or median) of the latter’s magnitude from the former. Despite the popularity of this approach, which is implemented as default in many software packages i.e. Scanalyze [6], UCSF Spot [7], and TIGR’s Spotfinder [8], it is not entirely without problems. Empirical studies have shown that this form of subtractive background correction may hinder the ability to detect differential gene expression [9–11] and impact the performance of signal calibration algorithms [12,13]. Even more concerning is the fact that different methods used to extract a sufficient statistic for the magnitude of the background show very little agreement among them [14]. To overcome these objections, alternative approaches for background correction have been put forward, e.g. subtractive correction using an estimate of the global rather than the local image background [9], morphological opening filters [14] and post-segmentation probabilistic subtractive methods [10].

A common feature of the aforementioned methods is that they fail to explicitly acknowledge extraneous constraints stemming from the discrete bounded nature of the microarray image scale, the positivity as well as the different magnitude of hybridization phenomena across cDNA microarrays. The most common manifestation of this error of omission is the generation of negative values in gene expression profiles, in direct contradiction to the physics of the experiment (additivity of fluorescent sources) and the acquisition apparatus (digital positive scale). To counteract such effects, a number of intuitive ad hoc solutions have been proposed. These range from arbitrarily setting negative corrected foreground values to one [11], to a small percentile of the expression of all other spots [15] and finally running the analysis without background correction.

In the present work we take a different approach to the problem of background adjustment and propose a statistical correction...
algorithm that is constrained by the discrete bounded nature of the microarray image scale and the additivity of fluorescent sources. Under a set of weak assumptions about the global background noise of a microarray image we derive an approximation to the unknown distribution of this error source using the method of Maximum Entropy [16,17]. This approximation is used to (a) estimate the magnitude of the background noise by segmenting the image histogram and (b) correcting individual pixels for the presence of noise using the maximum likelihood estimator. To illustrate the effects of our approach we apply it on publicly available microarray image sets and compare its effects to the default subtractive approach. It is shown that the technique presented may have significant benefits in terms of decreasing the bias and variance of replicate measurements under conditions of non-differential and differential gene expression.

2. Methods

2.1. Attributes of signal and background in microarray images

In this section we discuss the general attributes of the signal and the background in a microarray image that are encoded in the background adjustment algorithm. These attributes are:

1. The intensity values for all pixels in a microarray image are expressed in a discrete, positive, integer scale with $M + 1$ elements (e.g. $M = 65,535$ for a 16 bit image).
2. The observed intensity level ($D$) of each pixel in the microarray image obeys the additive noise model:

$$D = S + B | 0 \leq D \leq M \land 0 \leq S \leq M \land 0 \leq B \leq M$$

(1)

In this equation, $S$ denotes the signal generation process, whereas $B$ stands for the background process.

1. The latent (unobserved) components of Eq. (1) are generated by distinct physical processes in a microarray image. The signal generating process is due to the specific hybridization between labeled mRNA species and spotted target molecules. The background process results from a combination of non-specific hybridization phenomena, presence of non-cDNA fluorescent molecules, and diffusion of labeled particles.
2. Each microarray image can be segmented into two classes of pixels i.e. foreground or spot ($I_s$) and non-spot (background, $I_b$).

These four statements may be viewed as cogent prior information that should be taken into account in describing the processes that operate on microarray images. Such information is also relevant in specifying a mathematical error model for the intensities of microarray pixels, and a set of constraints regarding the variables that appear in this model. For example the first statement summarizes the features of the measurement apparatus used to generate microarray images. The optical instruments that are used are digital devices that generate positive images with finite dynamic range ($M$) that is specific to each system. Fluorescent intensities that fall below the lower detection limit are set to zero, while pixel intensities that exceed the upper detection limit are set to $M$.

The second statement specifies a constraint for the additive noise model i.e. that the magnitudes of the latent processes should be positive entities. Due to the finite range of microarray images this particular additive model implies the existence of saturation phenomena. As the data ($D$) and the generating processes ($S, B$) are constrained to lie in the set of positive integers $\{0, 1, 2, \ldots, M\}$, an increase in the intensity of the noise will shift the recorded pixel intensity upwards. If that value is larger than the upper cutoff of the measurement apparatus, the intensity recorded will saturate to $M$.

The fourth statement postulates the existence of a unique class assignment or labeling ($L$) of all distinct pixel intensities in the histogram of a microarray image ($I_b$) into two distinct classes $I_s$ and $I_b$.

Such a classification can be justified as follows: if a particular intensity belongs to the background class, then the magnitude of the $S$ component of that pixel is zero. More formally we have that $L = I_s \Rightarrow S = 0$ for each pixel in the image and this logical equivalence relation determines the following conditional class assignments: $P(S|L = I_b) = 0$ for $S \neq 0$ or $P(S|L = I_b) = 1$ for $S = 0$.

2.2. Maximum Entropy distributions for cDNA microarray images

The general attributes discussed in the previous sections constrain the distributions for the spot and background processes in a finite sub-set of the non-negative integers. Since the support of the distributions is bounded from below (non-negativity of fluorescent signals) and above (finite dynamic range of the measurement apparatus), the domain of support of the unknown distributions is bounded. As a consequence, these distributions could be uniquely characterized by their power moments i.e. the infinite sequence of real numbers: $\mu_k = \sum_{x=0}^{M} x^k \cdot p(x)$ for background and spot, respectively [18]. On the other hand recovering of a probability density of which a finite number of moments is known, is an in-determined problem as there are, in general, an infinite variety of functions with the same first $n$ moments. To stabilize the problem it is necessary to impose some further condition which will act as a constraint within the infinite dimensional function space and lead to a unique solution. The Maximum Entropy (MaxEnt) approach offers a definite procedure for the reconstruction of this approximant by selecting the distribution that maximizes the entropy $H[p]$ functional:

$$H[p] = -\sum_{x=0}^{M} p(x) \log p(x)$$

(2)

under the condition that the first $v$ moments be equal to the true moments. Such constraints are expressed in the form of numerical values regarding the expectations of functions of a discrete random (uncertain) variable $x$. By employing standard variational analysis arguments involving Lagrange multipliers it can be proved [19,20] that the MaxEnt distribution assumes the form:

$$p_i \equiv p(x_i) = \frac{1}{Z} \exp(-\sum_{j=1}^{v} \lambda_j x_i^j),$$

$$Z = \sum_{x=0}^{M} \exp \left( -\sum_{j=1}^{v} \lambda_j x^j \right), \quad \mu_k = -\frac{\partial \log Z}{\partial \lambda_k}$$

(3)

In applications the values of the Lagrange multipliers are computed by equating the theoretical moments $\mu_k$ to the finite sample moments of the same order. From an implementation standpoint, entropy maximization corresponds to a convex optimization problem [21] whose unique solution i.e. the values of the Lagrange multipliers $\lambda_k$ may be obtained by efficient numerical algorithms even for spaces of extremely high dimensionality [22]. Even though the quality of the approximation to the unknown distribution increases with the number of moment constraints imposed [23], in practice a very small number of moments (e.g. one or two) is sufficient for the purpose of the background adjustment of microarray images as we show in Section 4.

2.3. Specification of the background adjustment algorithm

Having established the constraints on the domains of definition of the latent components $S, B$ and the corresponding entropic
approximations we turn our attention to the problem of background correction of a single microarray image. Since we are concerned with correcting for the presence of a global background, the development we will pursue from this point only considers the intensities of the image pixels and not their position within the image. Equivalently we will work with the histogram of the microarray image \((I_k)\) defined as a vector of length \(M + 1\) whose ith element is equal to the number of pixels in the image with intensity equal to \(i\). \((n_i)\) i.e. \(I_k = (n_0, n_1, \ldots, n_M)\). The components of a background adjustment algorithm are the following:

- The functional form of the MaxEnt approximations \(p_B(i|\lambda_B)\) and \(p_S(i|\lambda_S)\) for spot and background, respectively, parameterized by the Langrangian vectors \(\lambda_B = \{\lambda_{B1}, \lambda_{B2}, \ldots, \lambda_{B8}\}\) and \(\lambda_S = \{\lambda_{S1}, \lambda_{S2}, \ldots, \lambda_{S8}\}\) with \(v\) the number of moment constraints.
- A procedure for the estimation of the parameters \(\lambda_B\) and \(\lambda_S\) from the data of the image histogram.
- A procedure for the classification of the intensity levels into background \((I_B)\) and spot classes \((I_S)\). Such a classification may be written as \(I_k = [0, \ldots, t_i, \ldots, b_k, \ldots, M]\) where \(t_i = 1\) if the ith histogram level has been classified as signal and \(t_i = 0\) if it has been classified as background.
- An estimator for the intensity of the latent component \(\hat{S}\) for all intensity levels in the histogram. This estimator is conditional on the classification of each intensity level that was determined in the previous step. For intensity levels classified into the \(I_B\) class \(S\) is zero by definition, so we set \(\hat{S} = 0\). For intensity levels \((D)\) classified into the \(I_S\) class, \(\hat{S}\) can be calculated by the “plug-in” estimator: \(\hat{S} = D - \mu_{B1}\), i.e. the difference between \(D\) and the estimate for the mean of the background distribution.

The parameters \(\lambda_B\), \(\lambda_S\) and \(\lambda_0\) can be estimated simultaneously by adopting a “missing data” perspective which considers the histogram class labels \(I_k\) as missing descriptors of the process (spot or background) that most likely generated the various intensity levels. By introducing the “missing data” descriptors, the second and the third components of the background adjustment algorithm can be implemented by iterating the steps of the Expectation–Maximization algorithm (EM) [24]. At convergence the EM will also provide an estimate for the first moment of the background distribution. Hence a technique that introduces a smaller bias should lead to more negative values for these particular spots. In the presence of simultaneous measurements of gene expression ratios, both negative control spots printed in the surface of microarrays (single channel assessments) and all spots in self–self hybridizations (two channel assessments) under experimental conditions associated with known differential expression (e.g. deletion mutants/antisense mRNA experiments). In the first case the “true” intensity of the spots is zero, while in the second case the log-expression ratio should be zero for all gene probes spotted in the surface of the microarray. In the last case the expected log-expression ratio will deviate from zero for these spots approaching the limit of detection for a particular experimental setup. Consequently depending on the specificity of the targets used one should expect to get a very small value for the corresponding gene expression ratio. Hence a technique that introduces a smaller bias should lead to more negative values for these particular spots.
expression changes by a non-microarray technique (e.g., RT-PCR), the bias introduced by background adjustment strategies could be quantified more accurately comparing the expression ratios to the values obtained by the gold standard technique.

In order to quantify the bias and variance introduced by the various background adjustment algorithms, we calculated the deviations of gene expression measures from their true value using replicate measurements. For negative control spots in any experiment and all spots in self–self hybridization experiments, the normal theory empirical estimates of the bias and variance are given by the average and root mean squared from the true value ($x_{\text{true}}$) of zero:

$$\text{Bias}_{\text{Normal}} = \frac{1}{N_j} \sum_{j} \left( \sum_{i} (x_{ij} - x_{\text{true}}) \right) = \frac{1}{N_j} \sum_{j} x_{ij} = \bar{x} \quad (8)$$

$$\text{Variance}_{\text{Normal}} = \frac{1}{N_j} \sum_{j} \sum_{i} (x_{ij} - x_{\text{true}})^2 = \frac{1}{N_j} \sum_{j} \sum_{i} x_{ij} = \text{RMS}(x) \quad (9)$$

where $i$ and $j$ index the spots in an array and the arrays in an experimental dataset.

These measures coincide with the average and the root mean square error summary statistics from the sample of replicate values. Since these statistics are not robust with respect to deviations from normality, we also report their robust theory alternatives [28]:

$$\text{Bias}_{\text{Robust}} = \text{median}(\{|x_{ij} - x_{\text{true}}|\}) = \text{median}(\{x_{ij}\}) \quad (10)$$

$$\text{Variance}_{\text{Robust}} = \frac{1}{N_j} \sum_{j} \sum_{i} |x_{ij} - x_{\text{true}}| = \frac{1}{N_j} \sum_{j} \sum_{i} |x_{ij}| = \text{MAE}(x) \quad (11)$$

which coincide with the sample median and the mean absolute error, respectively. To apply these formulas for negative control spots, we let $x_{ij}$ stand for the un-normalized, raw intensity obtained after image analysis. In the case of self–self hybridization experiments $x_{ij}$ is the log-transformed channel expression ratio after a suitable global normalization transformation.

In other categories of control spots (e.g., salmon sperm DNA, hybridization buffer or carryover negative control spots) found in microarray images, the true value ($x_{\text{true}}$) of the hybridization signal is likely to be different from zero. For such control spots, the sample mean and median do not coincide with measures of bias (since $x_{\text{true}}$ is unknown). However, the sample mean and median can still be used to rank background adjustment strategies, since these statistics are location measures for the residual error after noise reduction by background adjustment. Furthermore, measures of dispersion around the mean and the median do provide an empiric assessment of the performance variability of each background adjustment strategy and thus of variance. Assuming normally distributed measurements, variability around the mean is quantified by the sample standard distribution:

$$S = \sqrt{\frac{1}{N_j} \sum_{j} \sum_{i} (x_{ij} - \bar{x})^2} \quad (12)$$

The mean absolute deviation from the sample median (MAD-MD) and the median absolute deviation (MAD) are measures of scale that are robust with respect to deviations from normality and can be used to quantify variability around the median:

$$\text{MADMD}(x) = \frac{1}{N_j} \sum_{j} \sum_{i} |x_{ij} - \text{median}(x_{ij})| \quad (13)$$

$$\text{MAD}(x) = \text{median}(|x_{ij} - \text{median}(x_{ij})|) \quad (14)$$

A more formal assessment of the noise reduction properties of the different background adjustment algorithms in the case of known differential expression is afforded by variance component (VarComp) or mixed-effects models. In their most general use, mixed-effects models are utilized to describe relationships between a response variable (log-expression ratio in our case) and certain explanatory covariates and/or grouping (classification) factors in the data. VarComp models have been previously utilized to normalize microarray experiments (in the form of analysis of variance (ANOVA) [27,28], and linear mixed models (LMM) [29]). They have also been used to detect differential expressions while accounting for global (experimental conditions, print-tip or dye effects) as well as gene specific sources of variability [28,30–33]. The statistical analysis of different background adjustment strategies may be undertaken by linear models that incorporate (a) fixed effects (type of algorithm used), which are associated with an entire population or with certain repeatable levels of experimental factors and (b) random effects, which are associated with individual experimental units (array) drawn at random from a population and heteroscedastic error arising from different probes and algorithms. Assuming a gene that is quantified by $N_j$ different probes in the surface of a single microarray, $N_j$ repetitions of an experiment that quantifies the relative expression of that gene in two channels (e.g. red and green) and $N_j$ different background algorithms (3 in this case) to process the same images, we can invoke the following model for the log-expression ratio of gene $y_{ij,k}$:

$$Y_{ijk} = \beta_i + b_k + \epsilon_{ijk} \quad b_k \sim N(0, \sigma_k^2) \quad \epsilon_{ijk} \sim N(0, \sigma^2_1) \quad (15)$$

In the aforementioned equation, $\beta_i$ is the gene expression ratio estimated by each of the $N_j$ algorithms (fixed effect), while $b_k$ is the random effect representing the deviation of the $k$th replicate from the population mean $\beta_i$. Finally, $\epsilon_{ijk}$ is a random variable modeling the error in the replicate measurements from each microarray arising from using multiple arrays and background adjustment algorithms. Stated in other terms, the variance terms in Eq. (15) $\sigma_k^2$ and $\sigma^2_1$ thus characterize the “between-array” and “within-array” variability. As a first approximation, one can postulate that the two sources variability implicit in $\sigma_k^2$ are independent and thus this term factorizes as a product of two variances (p. 209, 213 in [34]): $\sigma_k^2 = \sigma_3^2 \sigma_2^2$. By fitting these models to replicated datasets analyzed by different background methods one can account for three major sources of variability arising from differences in individual arrays ($\sigma_2^2$), replicate spots within each array ($\sigma_3^2$) and possibly heteroscedastic algorithm error ($\sigma_1^2$).

A model based analysis of different background adjustment algorithms would then directly compare the fixed effects coefficients $\beta_i$ to the log-expression ratio estimated by a gold standard technique to ascertain the bias. Additionally, the variances due to the use of different algorithms may be quantified by examining the variances of each background adjustment algorithm $\sigma^2, j = 1, 2, 3$. Statistical tests of significance (or equivalently confidence intervals) may be applied on both bias and variance terms to formally compare the performance of the various algorithms.

### 3. Materials

3.1. Microarray datasets, image segmentation and normalization of gene expression profiles

We applied the proposed background correction methods in the following microarray datasets from the Stanford Microarray Database (Table 1):

- a. The images from the DeRisi et al. experiment concerning the Diauxic shift of Saccharomyces cerevisiae [35].
- b. Fourty self–self hybridizations obtained from the quality control program of the Arabidopsis Functional Genomics...
Table 1
Microarray datasets utilized.

<table>
<thead>
<tr>
<th>Microarray experiment alias</th>
<th>SMD experiment set ID (experiment slide ID)</th>
<th>Number of microarray slides</th>
<th>Type and number of control spots per array</th>
<th>Design and replication structure</th>
<th>Experiment year</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Diauxic” [35]</td>
<td>8 (NA)</td>
<td>7</td>
<td>“Empty” (82)</td>
<td>Common Reference Sample (Cy3)</td>
<td>1997</td>
</tr>
<tr>
<td>“AFGC” [36,37]</td>
<td>NA (10,750–10,789)</td>
<td>40</td>
<td>“SSDNA” (576)</td>
<td>Self–self Hybridization Experiments. 8 plant clones in addition to the SSDNA Hybridizations of HeLa cells expressing the tetracycline transactivator against HeLa cells transfected with sense (13/19) and antisense (6/19) clones for the human BAT3 (HLA-B-associated transcript 3) gene.</td>
<td>1997</td>
</tr>
<tr>
<td>“BAT3” [38]</td>
<td>3898 (45,331, 47,648, 45,415, 46,205, 46,206, 46,271, 47,647, 45,414, 47,644, 48,824, 48,825), 45,330, 47,645, 48,826, 49,842, 48,844</td>
<td>19</td>
<td>“Empty” (192)</td>
<td>Common Reference Sample (wild type V. Cholerae at stationary growth phase)</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“SSDNA” (48)</td>
<td>• Deletion Mutant for rpoS (stationary phase alternative sigma factor) (9/13)</td>
<td>2004</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“Carryover” (32)</td>
<td>• Mid Log phase (4/13)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“rRNA” (32)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“Chromosomal DNA” (80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Heat shock”</td>
<td>3207 (32,129–32,132)</td>
<td>4</td>
<td>“Empty” (64)</td>
<td>Common Reference Sample (Cy3) against non heat stressed (2/4) and heat stressed E. histolytica (2/4)</td>
<td>2005</td>
</tr>
</tbody>
</table>

Characteristics of microarray datasets, control spots and replication numbers of microarray datasets utilized in the present study. The first column lists the alias used in the text to refer to each microarray experiment.

Microarrays in all five experiments utilized at least one negative control spot (blank spots, SSDNA spots, 3XSSC spots) allowing the assessment of the biases introduced by the background adjustment algorithms in microarray signal detection process. One experiment (“Cholera”) also used positive (rRNA spots, chromosomal DNA) and non-specific hybridization controls (carryover spots) allowing a more detailed assessment of the noise suppression properties of the proposed algorithm. The degree of replication and the identity of the eight non-control targets used in the experiment from the AFGC facilitated the analysis of bias and variance of gene expression ratios across a wide range of mRNA concentrations (from transcription factors to metabolic genes).

For two of the datasets (BAT3 and Cholera), additional information permitted an exploration of the bias/variance properties of the three background adjustment algorithms in situations of biological signal change. In the case of the BAT3 dataset, the authors had quantified and reported the changes in BAT3 mRNA resulting from stable transfection of sense and antisense clones under the same conditions used in the microarray experiments. In the Cholera dataset, wild type strains in the mid log phase and deletion mutants for the rpoS gene at the stationary phase were compared against wild type strains at the stationary growth phase. Since rpoS is induced at conditions of metabolic stress found in the stationary growth phase [40,41], the two comparisons involve biological states where hybridization signals for the rpoS targets are very low or absent (wild type in mid log phase/deletion mutants in stationary phase) versus those which are present (wild type strains growing in stationary phase).

For these datasets the raw microarray images were downloaded from SMD with the exception of the Diauxic set that was downloaded from the MicroArray Genome Imaging and Clustering Tool (MAGIC) website [42] (http://www.bio.davidson.edu/projects/magic/magic.html).

Wherever available, gridding information was extracted from the GenePix Results (GPR) files stored for each microarray image in the SMD. Microarray images were then segmented with the seeded region growing algorithm implemented in the software SPOT [14]. To implement the EM/MAXENT strategy, background adjusted images were processed by SPOT, while the original images were used for the SBC and NONE algorithms. For the SBC algorithm, the mean of the background region was subtracted from the mean foreground value for each spot after the images had been segmented. Unless stated otherwise, the channel median was used as a measure of gene expression. Comparisons with empty spots (“zero”) were carried out in the raw microarray scale, while logarithms (base 2) were taken prior to analyzing the results of all other spots in order to symmetrize the scale and improve the normality of measurements. Non-positive measurements generated by SBC and EM/MAXENT were excluded from analyses in the log transformed scale. cDNA targets mapping to the same gene were modeled as independent and identically distributed variates in linear mixed model regressions.

In order to normalize the gene expression profiles we used the global median normalization method applied to a single channel or both channels separately after log-transformation. More advanced normalization algorithms (e.g. lowess [43]) that propose corrections for spatial, intensity and dye dependent biases were not utilized for the comparison of control spots in this report, since such algorithms would hinder our ability to detect subtle biases introduced by the proposed method. Normalization implicitly assumes that the majority of the log-expression ratios is zero and effects a local correction to the raw data to shrink them back to the null. Lowess is a very effective way to effect this correction in a manner that exploits regional features in the M–A plots of any given microarray hybridization. Hence, if the proposed method introduces any biases, they would likely be masked by the application of advanced normalization algorithms. On the other hand, model based comparisons were carried out with both normalization methods to assess the behavior of the proposed method under conditions similar to typical microarray analysis practice. All statistical analyses were carried out in R version 2.9.1 running on Linux Ubuntu.
version 8.04. For linear mixed-effects regression models we used the R package \textit{nlme}.

3.2. Software implementation, availability and requirements

The methods presented in this study have been implemented as a dynamically linked library in the C programming language accessible via a PERL interface. The software tool accepts as input TIFF microarray images and generates background adjusted images in the same format suitable for processing in existing microarray software tools. Processing of a 16-bit $1024 \times 1024$ image is complete within 15 s on a system with a Pentium 4 2.4 GHz Processor, 512 MB RAM under Windows XP Home SP2. The software has been tested under Windows Vista Premium on a system with a Pentium Duo T9300 at 2.5 GHz. Executables, source code, PERL scripts and example batch files are available at http://stat.med.upatras.gr/cDNA under the GNU GPL license.

4. Results

4.1. Visual exploration of the EM/MAXENT algorithm

The results of the restoration algorithms for a typical section of a microarray image in the “Diauxic” dataset are shown in Fig. 1 (this corresponds to Green Channel OD 1.8 in [35]). At its convergence, the binary classification of the image pixels by the EM algorithm has more or less correctly labeled the high intensity pixels as signal (Fig. 1B). After labeling, the background correction (step 3 of the proposed algorithm) effects a contrast enhancement in the original image emphasizing the distinction between background and spot regions. Although this contrast enhancement is imperceptible in the two dimensional images it can readily be appreciated in three dimensional intensity plots. In Fig. 1C and D we plot the log (base 2) intensity of the $3 \times 3$ sub-grid demarcated by the white rectangle in Fig. 1A. Application of the EM/MAXENT algorithm (Fig. 1D), leads to an increase in the contrast of the restored images across the spot boundaries of about three orders of magnitude (~10 bits) when compared to the original image (Fig. 1C).

Increasing the number of moment constraints led to labelings that most closely resembled the expected appearance of a microarray image for many slides (Fig. 2A–F). However, this increase was not noted in all images (Fig. 3A–C) or even in the images of the two channels of a given microarray experiment e.g. contrast Fig. 3D–F and Fig. 2D–F. Since two moment constraints led to labelings that were subjectively evaluated as adequate we used the latter images as the source of all subsequent results.

4.2. Empirical analysis of bias and variance using control spots

In order to empirically assess the performance of the EM/MAXENT algorithm relative to SBC and the original data, we computed empirical estimates of the bias and the variance for the datasets of

Fig. 1. Typical Background Adjustments by the EM/MAXENT algorithm. Original Image (A), Output of EM algorithm at convergence (B). The intensities in bits of the pixels the area demarcated by the white rectangle in (A) has been expanded as 3-D intensity plot for the original (C) and the EM/MAXENT adjusted image (D). Note the different scales of the z-axis among the first two images (10–15) and the conditionally restored image (0–15).
Table 1 that employ the genes and probes presented in Table 2. Measures of bias are summarized in Table 3 for the five datasets assessed in this report. From this table several observations can be made: (a) the distribution of the raw (uncorrected) spot intensities is positively skewed (mean > median), (b) background correction by the EM/MAXENT leads to comparable or smaller bias for the majority of the datasets by virtue of smaller mean and median values control spots compared to the SBC strategy. Furthermore, this performance is consistent in both channels in spite the presence of dye specific bias. Using the robust (non normal) measure of bias, it can be seen that the proposed strategy eliminates at least 50% of all spots whose intensities are not expected to differ from the background in 4/5 datasets. Finally, the skewness of the distribution of the negative control spots is still evident after background adjustment since the median for both EM/MAXENT is substantially less than the mean value. In line with the last observation we would not expect to find a consistent pattern in the normal theory measures of variance (Table 4). Even though background adjustment was in general associated with decreased measures of variability, there were cases where normal measures of variance e.g.

Table 1 that employ the genes and probes presented in Table 2. Measures of bias are summarized in Table 3 for the five datasets assessed in this report. From this table several observations can be made: (a) the distribution of the raw (uncorrected) spot intensities is positively skewed (mean > median), (b) background correction by the EM/MAXENT leads to comparable or smaller bias for the majority of the datasets by virtue of smaller mean and median values control spots compared to the SBC strategy. Furthermore, this performance is consistent in both channels in spite the presence of dye specific bias. Using the robust (non normal) measure of bias, it can be seen that the proposed strategy eliminates at least 50% of all spots whose intensities are not expected to differ from the background in 4/5 datasets. Finally, the skewness of the distribution of the negative control spots is still evident after background adjustment since the median for both EM/MAXENT is substantially less than the mean value. In line with the last observation we would not expect to find a consistent pattern in the normal theory measures of variance (Table 4). Even though background adjustment was in general associated with decreased measures of variability, there were cases where normal measures of variance e.g.
the *Diauxic* and *Cholera* dataset did increase after background adjustment. Using robust measures of variance (e.g. MAE), there was a consistent decrease in the variability effected by any method of background adjustment (Table 5). Compared to the SBC, the EM/MAXENT method was characterized by smaller magnitudes of variance in almost all assessments; furthermore, the MAD statistic was zero for 5/6 comparisons of Table 4 since the method estimated an intensity of zero for a substantial estimation of negative control spots (Additional file 1 in http://stat.med.upatras.gr/cDNA). The relationship between the percentage of spots that are filtered out by the EM/MAXENT and SBC methods versus the raw spot intensities was evaluated graphically in the multiply controlled Cholera dataset (Fig. 4).

On a per array and channel basis there appears to be a non-linear (sigmoid) relationship between the percentage of replicate spots with positive intensities and the median raw intensity of the SBC method inflates the bias so that only 2/8 probes have medians less than 0.1 bits and three out of the eight probes have an absolute bias of >0.25 bits. In addition, the variance of the log-expression ratios increases after application of the SBC method. This is attested from the spread of the outliers in the plots and the magnitude of the MAD sample statistic (Additional file 2 in http://stat.med.upatras.gr/cDNA), which ranged from 1.01 to 1.82 times the magnitude of MAD in the original image. On the other hand, application of the EM/MAXENT method did not result in any appreciable increase in the bias, but was associated with decreased the variability of results compared to the SBC corrected and possibly the original images. To some extent the reduction in variability effected by the EM/MAXENT can be attributed to the filtering of low intensity spots out of the analysis. To see why, note that intensities from both channels have to be estimated as positive numbers in order to calculate the log-expression ratio. If zero (EM/MAXENT) or non-positive intensities (SBC) are generated during background adjustment the corresponding log ratios will be dropped from subsequent steps. Fig. 6 displays the scatterplots of the percentage of positive spot intensities after background adjustment versus the median intensity in the original image for the nine spots (eight genes and one negative) for each of the forty hybridizations.

4.3. Empirical analysis of bias and variance using self–self hybridization experiments

To verify that the potential of the EM/MAXENT method to generate spot intensities with reduced bias and variance we analyzed the AFGC dataset consisting of a large number of replicate self–self hybridizations using a small number of targets. The distributions of the log-expression ratios in these self–self hybridization experiments are shown in the box and whisker plots in Fig. 5.

It can be appreciated that the raw data have a very small bias; the median absolute value is <0.1 bits with the exception of probe R89981 that had a median log-expression ratio of 0.15. Application of the SBC method inflates the bias so that only 2/8 probes have medians less than 0.1 bits and three out of the eight probes have an absolute bias of >0.25 bits. In addition, the variance of the log-expression ratios increases after application of the SBC method. This is attested from the spread of the outliers in the plots and the magnitude of the MAD sample statistic (Additional file 2 in http://stat.med.upatras.gr/cDNA), which ranged from 1.01 to 1.82 times the magnitude of MAD in the original image. On the other hand, application of the EM/MAXENT method did not result in any appreciable increase in the bias, but was associated with decreased the variability of results compared to the SBC corrected and possibly the original images. To some extent the reduction in variability effected by the EM/MAXENT can be attributed to the filtering of low intensity spots out of the analysis. To see why, note that intensities from both channels have to be estimated as positive numbers in order to calculate the log-expression ratio. If zero (EM/MAXENT) or non-positive intensities (SBC) are generated during background adjustment the corresponding log ratios will be dropped from subsequent steps.

Table 2

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Bias measure</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBC</td>
<td>EM/MAXENT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>EM/MAXENT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RED</td>
<td>Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPTY</td>
<td>Cholera</td>
<td>42.09</td>
<td>46.35</td>
</tr>
<tr>
<td></td>
<td>Heat shock</td>
<td>304.88</td>
<td>313.14</td>
</tr>
<tr>
<td></td>
<td>Diauxic</td>
<td>747.39</td>
<td>847.07</td>
</tr>
<tr>
<td></td>
<td>BAT3</td>
<td>25.01</td>
<td>50.90</td>
</tr>
<tr>
<td></td>
<td>SSDNA</td>
<td>234.3</td>
<td>269.02</td>
</tr>
<tr>
<td></td>
<td>AFGC</td>
<td>547.38</td>
<td>557.61</td>
</tr>
<tr>
<td>GREEN</td>
<td>Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPTY</td>
<td>Cholera</td>
<td>32.67</td>
<td>31.01</td>
</tr>
<tr>
<td></td>
<td>Heat shock</td>
<td>16.3</td>
<td>34.04</td>
</tr>
<tr>
<td></td>
<td>Diauxic</td>
<td>678.82</td>
<td>845.33</td>
</tr>
<tr>
<td></td>
<td>BAT3</td>
<td>38.1</td>
<td>61.61</td>
</tr>
<tr>
<td></td>
<td>SSDNA</td>
<td>244.43</td>
<td>233.93</td>
</tr>
<tr>
<td></td>
<td>AFGC</td>
<td>702.38</td>
<td>725.04</td>
</tr>
</tbody>
</table>

Empirical assessment of channel specific bias in quantification of negative control spots for two different types of such spots (blank and SSDNA). Both normal theory (mean) and robust (median) measures of bias are calculated for each of the three background assessment strategies assessed. The smallest (best performing) algorithm for each channel and dataset is indicated with bold face.

<table>
<thead>
<tr>
<th>Negative control spot</th>
<th>Variance measure</th>
<th>Algorithm</th>
<th>Red channel</th>
<th>Green channel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>EM/MAXENT</td>
<td>NONE</td>
</tr>
<tr>
<td>Red channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPTY</td>
<td>Dataset</td>
<td>Cholera</td>
<td>550.32</td>
<td>609.28</td>
</tr>
<tr>
<td></td>
<td>Heat shock</td>
<td>1684.07</td>
<td>1634.41</td>
<td>1583.02</td>
</tr>
<tr>
<td></td>
<td>Diauxic</td>
<td>4367.79</td>
<td>4647.39</td>
<td>4123.64</td>
</tr>
<tr>
<td></td>
<td>BAT3</td>
<td>632.44</td>
<td>680.80</td>
<td>659.03</td>
</tr>
<tr>
<td>SSDNA</td>
<td>Cholera</td>
<td>2775.24</td>
<td>1702.84</td>
<td>1615.74</td>
</tr>
<tr>
<td></td>
<td>AFGC</td>
<td>2185.66</td>
<td>2433.91</td>
<td>2092.47</td>
</tr>
<tr>
<td>Green channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPTY</td>
<td>Cholera</td>
<td>386.57</td>
<td>468.70</td>
<td>349.54</td>
</tr>
<tr>
<td></td>
<td>Heat shock</td>
<td>132.81</td>
<td>337.85</td>
<td>84.35</td>
</tr>
<tr>
<td></td>
<td>Diauxic</td>
<td>3207.59</td>
<td>3923.36</td>
<td>2865.08</td>
</tr>
<tr>
<td></td>
<td>BAT3</td>
<td>1239.39</td>
<td>1288.85</td>
<td>1275.53</td>
</tr>
<tr>
<td>SSDNA</td>
<td>Cholera</td>
<td>1054.40</td>
<td>1078.04</td>
<td>970.51</td>
</tr>
<tr>
<td></td>
<td>AFGC</td>
<td>1901.29</td>
<td>2196.84</td>
<td>1782.69</td>
</tr>
</tbody>
</table>

Empirical assessment of channel specific variance in quantification of negative control spots for two different types of such spots (blank and SSDNA). These are computed for each of the three background assessment strategies assessed, assuming normal distribution for the distributions of the spot intensities. The smallest (best performing) algorithm for each channel and dataset is indicated with bold face.

<table>
<thead>
<tr>
<th>Negative control spot</th>
<th>Variance measure</th>
<th>Algorithm</th>
<th>Red channel</th>
<th>Green channel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>EM/MAXENT</td>
<td>NONE</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPTY</td>
<td>Dataset</td>
<td>Cholera</td>
<td>42.09</td>
<td>258.77</td>
</tr>
<tr>
<td></td>
<td>Heat shock</td>
<td>304.88</td>
<td>409.88</td>
<td>318.57</td>
</tr>
<tr>
<td></td>
<td>Diauxic</td>
<td>747.39</td>
<td>2083.43</td>
<td>863.96</td>
</tr>
<tr>
<td></td>
<td>BAT3</td>
<td>25.01</td>
<td>153.66</td>
<td>53.93</td>
</tr>
<tr>
<td>SSDNA</td>
<td>Cholera</td>
<td>270.26</td>
<td>433.32</td>
<td>224.48</td>
</tr>
<tr>
<td></td>
<td>AFGC</td>
<td>445.07</td>
<td>1111.31</td>
<td>627.64</td>
</tr>
<tr>
<td>Green channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMPTY</td>
<td>Cholera</td>
<td>32.67</td>
<td>105.83</td>
<td>34.57</td>
</tr>
<tr>
<td></td>
<td>Heat shock</td>
<td>16.30</td>
<td>50.92</td>
<td>32.17</td>
</tr>
<tr>
<td></td>
<td>Diauxic</td>
<td>678.82</td>
<td>1185.67</td>
<td>823.67</td>
</tr>
<tr>
<td></td>
<td>BAT3</td>
<td>38.10</td>
<td>92.74</td>
<td>58.12</td>
</tr>
<tr>
<td>SSDNA</td>
<td>Cholera</td>
<td>191.42</td>
<td>412.24</td>
<td>206.62</td>
</tr>
<tr>
<td></td>
<td>AFGC</td>
<td>385.74</td>
<td>1181.21</td>
<td>792.13</td>
</tr>
</tbody>
</table>

Empirical assessment of channel specific variance in quantification of negative control spots for two different types of such spots (blank and SSDNA). These are computed for each of the three background assessment strategies assessed, assuming normal distribution for the distributions of the spot intensities.

4.4. Model based evaluation of bias and variance of background adjustment methods

In order to better characterize the bias–variance performance of the three methods, we applied mixed-effects models to replicate measurements of genes whose "true" expression ratio was known either because of the experimental design employed or because RT-PCR verification had been performed by the original investigators. In the self–self hybridizations in the AFGC dataset, the true log-expression ratio is zero, whereas the bacterial deletion mutants for the sigma factor (rpoS) in the Cholera dataset are expected to have a negative log-expression ratio compared to the wild type. In the Cholera dataset, bacteria growing in the stationary phase should have a substantially higher log-expression ratio than cultures in the log phase. Finally in the BAT3 dataset [38] HeLa cells stably transfected with expression vectors for sense and antisense EST of the BAT3 locus had been compared to parental cell cultures that did not express such ESTs by both microarray and RT-PCR. According to the latter, the BAT3 gene is down-regulated when compared to parental cell cultures included zero for the self–self hybridization experiments or whether the self–self hybridization experiments in which the true log-expression ratio was known to be equal to zero, the EM/MAXENT method yielded expression ratios that were consistently within 20% of the true value. On the other hand, application of the SBC method yielded expression ratios that were consistently within 20% of the true value. On the other hand, application of the SBC method yielded expression ratios that could be >30% (−0.33 in log scale) away from the expected ratio.

To analyze the model based estimates of bias, the linear mixed model was applied to all three methods and the confidence intervals for the log-expression ratio were classified according to the inclusion of the expected log-expression ratio i.e. whether they included zero for the self–self hybridization experiments or whether they did not cross zero for down-regulated probes. The odds ratios of correct classification of the log-expression ratio were 1.79 (no correction), 1.65 (EM/MAXENT) and 1.28 (SBC) for datasets normalized by the global median method (ANOVA p value testing the equality of these three odds ratios was 0.23). When the datasets were normalized with the lowess method, the corresponding...
odds ratios were 2.07, 1.89 and 2.07 (p-value for equality 0.89). Taken together, these results indicate (Fig. 7) that the performance of the three methods is roughly equivalent as far as the introduction of bias is concerned.

In 9/12 assessments the SBC resulted in measurements that were substantially more variable than the original data and met the criteria for statistical significance. In no case was the SBC superior to using the original data i.e. without background adjustment. In contrast, application of the EM/MAXENT resulted in measurements that were substantially less variable than the original data in 7/12 cases. These results achieved statistical significance at the 0.05 level without or with a Bonferroni adjustment for the number of comparisons (24). In 4/12 comparisons the variance of the EM/MAXENT method was not statistically different from the variance of the original data and only one case was it worse. When the relative variance was examined as a dichotomous outcome (smaller/same versus larger than the original data), the EM/MAXENT method resulted in log gene expression ratios that had the same or smaller variability than the SBC method (p-value 0.003, Fisher’s exact test). Similar results were obtained when the datasets were normalized with the lowess method; 30% of all EM/MAXENT assessed genes yielded variances >1, whereas all probes subjected
to the SBC method had a ratio >1 (p-value 0.002 Fisher’s exact test).
Expressed as a continuous outcome, the median variance of the EM/MAXENT method relative to the original data was 0.86 and 0.88 when the datasets were normalized with the global median and the tip-based lowess, respectively. In contrast, the median variance of the SBC method was 1.77 and 1.84 times, respectively.

5. Discussion

5.1. The role of background adjustment in microarray analysis

Ever since the introduction of microarrays in high throughput gene expression profiling, background correction has been
considered a necessary step in the experimental pipeline [44], which may in fact have a significantly higher bearing on the final results than subsequent analytic steps i.e. normalization [11,14]. A simple model is one which notes that in any given finite area of the microarray surface there are both specific and non-specific hybridization sources of fluorescence, which combine additively. Such considerations lead to the additive background corruption model that is endorsed by the traditional (SBC) and other alternative approaches [9,10]. Yet, unless the adjustment algorithm is properly constrained by extraneous pre-data information about the measurement scale and the relative magnitude of the background to the signal, its consistency in repeated applications is likely to suffer. This has indeed been the experience with previous evaluations of the default SBC method for background adjustment [11,14,45], a finding that we also observed in the five datasets we examined in this report.

5.2. The rationale for Maximum Entropy methods in microarray image analysis

In order to make the most out of the initial “objective” information about the microarray measurement process, we used the method of Maximum Entropy [46,47] to render such information into probability distributions. Our rationale for using this approach does not follow from the assumption that the data from microarray experiments come from such distributions. On the contrary, we feel that the complexity of the microarray measurement process is so high, that the search for the “true” distribution of the microarray data is a very difficult task and an approximation is needed instead. However, our decision to express the sought after approximation in terms of Maximum Entropy distributions is motivated by both theoretical and practical features that such distributions possess. From the Bayesian viewpoint (a view that was not adopted in this paper) their major theoretic attraction is their least informative nature. Stated in other terms, such distributions are the ones most compatible with the “objective” pre-data information at our disposal while being maximally noncommittal about the missing information. To a frequentist, the most appealing features would most likely be the attainment of the lower bound of the Cramer-Rao inequality and the unbiased-ness of their moment estimators [48]. Stated in other terms, such distributions are expected to achieve the minimum error (bias and variance) out of all distributions compatible with the initial information and the experimental data at hand. Last but certainly not least, MaxEnt distributions are endowed with computationally efficient procedures for estimation and learning based on sufficient statistics. This aspect is important considering the large size of microarray images, which render alternative approaches (e.g. Monte Carlo or resampling methods) computationally prohibitive for batch work. By embedding the Maximum Entropy distributions into the Expectation–Maximization algorithm we derived a computationally efficient probabilistic algorithm (EM/MAXENT) for the adjustment of microarray images.

5.3. Discussion and Interpretation of evaluation results

The technical innovation of the EM/MAXENT algorithm provides a partial answer to the following important questions: (a) should we even attempt to correct gene expression measures for the microarray background? (b) If so, what is the best method to achieve this? We approached both questions by calculating empirical measures of bias and variance from multiple categories of negative control and non-control spots from five datasets generated over the course of a decade. These measures of bias and variance were then utilized to compare the proposed method to the standard, subtractive approach, using the uncorrected data as reference. Our results indicate that the subtractive approach is characterized by a considerable magnitude of empiric bias and variance in the case of spots that are expressed close to background. On the other hand the proposed method does not appear to suffer from these disadvantages and the method can filter out spot
measures with intensities that are numerically close to the background level. An alternative semi-quantitative view of the EM/MAXENT algorithm is that it implements a thresholded filter with a non-linear activation function. Such a filter would reject spots with intensities that are indistinguishable from the background of a given image but would correct all the others using a statistically consistent estimator for the level of the noise. From this perspective the improvement noted in the measures of variance in the model based assessments can be traced to the efficiency of the entropic distributions in approximating the unknown distribution of the background with the least error. The implications for the analysis of other microarray datasets can be understood by referring to the variance of the EM/MAXENT and SBC method relative to the original data. Using the model based estimates (Section 4.4) as an index of method performance, the former method results in a reduction in the standard deviation by \(1 - \sqrt{0.86} \approx 3.3\%\) compared to the original data. On the other hand application of the SBC method will increase the standard deviation by \(\sqrt{1.77} - 1 \approx 33\%\). Assuming that the signal to noise ratio of microarray measurements falls by the square root of the sample size, application of the EM/MAXENT method may reduce the required sample size by 14% for a given level of an experimental noise and signal to noise ratio. On the other hand, the SBC method will increase the required sample size by 77% relative to the original uncorrected data in line with previous reports.

5.4. Outlook and directions for future work

A few limitations of the proposed methodology and open methodological relations should be noted when evaluating its merits for routine work. These relate to the optimality of a local (rather than described herein global) background adjustment procedure, the need to derive a more realistic additive/multiplicative inverse measurement error model that could link background adjustment and normalization. We postulate that such a derivation may be facilitated by the adoption of a Bayesian perspective, a direction that we are actively pursuing. Furthermore, the estimated reductions in variance should serve as rough initial estimates which could be improved upon by applying the proposed EM/MAXENT method to additional datasets. Ideally these datasets should include independent quantification of the expression ratio of multiple genes with non array based techniques, in order to facilitate meaningful comparisons of bias and variance. Finally, we note that one natural extension of the proposed statistical framework/background correction method involves oligo-nucleotide arrays (e.g. Affymetrix). Even though in such arrays, an explicit segmentation of mismatch probes and correction of perfect match probes (Affymetrix). Even though in such arrays, an explicit segmentation of mismatch probes and correction of perfect match probes

6. Conclusion

Background correction has been considered a necessary step in the experimental pipeline of high throughput gene expression profiling. The proposed method for background adjustment based on Maximum Entropy reduces standardized measures of log expression variability across replicates in situations of differential and non-differential gene expression without increasing the bias. It incorporates computationally efficient procedures for estimation and learning based on sufficient statistics and can filter out spot measures with intensities that are numerically close to the background level.

References


