BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 10 2005, pages 2410–2416
doi:10.1093/bioinformatics/bti341

Gene expression

Background correction for cDNA microarray images using the TV+L^1 model

Wotao Yin[1,3,*], Terrence Chen[2,3], Xiang Sean Zhou[3] and Amit Chakraborty[3]

[1]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA, [2]Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA and [3]Siemens Corporate Research, Inc., Princeton, NJ 08540, USA

Received on November 19, 2004; revised on February 10, 2005; accepted on February 17, 2005
Advance Access publication February 22, 2005

ABSTRACT

Motivation: Background correction is an important preprocess in cDNA microarray data analysis. A variety of methods have been used for this purpose. However, many kinds of backgrounds, especially inhomogeneous ones, cannot be estimated correctly using any of the existing methods. In this paper, we propose the use of the TV+L^1 model, which minimizes the total variation (TV) of the image subject to an L^1-fidelity term, to correct background bias. We demonstrate its advantages over the existing methods by both analytically discussing its properties and numerically comparing it with morphological opening.

Results: Experimental results on both synthetic data and real microarray images demonstrate that the TV+L^1 model gives the restored intensity that is closer to the true data than morphological opening. As a result, this method can serve an important role in the preprocessing of cDNA microarray data.

Contact: wy2002@columbia.edu

1 INTRODUCTION

The cDNA microarrays consist of tens of thousands of individual DNA sequences printed in parallel on a glass microscope slide. They are designed to detect specific genes and to measure their activities in tissue samples by monitoring the differential hybridization of the two DNA or RNA samples to the sequences on the array. The research of cDNA microarrays has greatly contributed to cell biology, human health and disease, drug discovery and other related areas.

From the image analysis perspective, one of the biggest problems of cDNA microarray images is that they are plagued with inhomogeneous backgrounds. On a microarray slide, the measured fluorescence intensity of a spot is a combination of the image background intensity near the spot and the intensity determined by the hybridization level of the mRNA samples with the spotted DNA. Background correction is necessary to estimate the true hybridization level of the cDNA. The existence of inhomogeneous background can make this task very difficult.

In the research community, different methods have been developed to correct microarray background bias. The published methods can be classified into three categories: (1) constant background correction, (2) local background correction and (3) morphological opening (MO). Constant background correction methods use the mean or median intensity of the whole image background as the estimated background intensity, and, consequently are seldom used in real applications with inhomogeneous backgrounds. Local background correction methods calculate background intensity locally using the pixels that are near cDNA spots. These methods give the corrected images by subtracting the mean or median intensity value of local pixels from original images. Figure 1 depicts the different local regions used in ScanAlyze (Eisen and Brown, 1999, http://rana.lbl.gov/EisenSoftware.htm), ImaGene (Medigue *et al*., 1999) and GenePix (Axon Instruments, 1999). One problem of these methods is that the mean or median intensity values of the pixels in a local region of a spot may be higher than the intensity of the spot itself. This happens when the background has big intensity changes near the spot. Consequently, local background correction may give negative spot intensity values, which is wrong. In general, the performance of local correction degrades under the presence of local background artifacts or variation. The third category is MO (Soille, 1999), which estimates background intensity using a non-linear filter. This filter essentially smoothes the entire image, albeit, in a non-uniform way. It removes all local peaks, including both artifacts and spots, and returns a smoothed image as the background estimate. More specifically, MO applies a local minimum filter, which is an erosion process, followed by a local maximum filter, which is a dilation process, to the image. This procedure is used in the software package Spot (Beare and Buckley, 2004, http://spot.cmis.csiro.au/spot/doc/Spot.pdf). MO is considered superior to constant background and local background

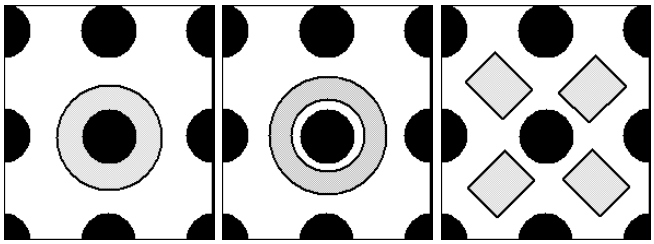


Fig. 1. Different regions (gray) used in local background (white) corrections by ScanAlyze (left), ImaGene (middle) and genePix (right).

*To whom correspondence should be addressed.

corrections due to its robustness against local artifacts and variations. It seldom gives negative spot intensity. Moreover, Yang *et al*. (2000) and Smyth *et al*. (2003) studied MO and claimed that it gave the best results. However, there are some limitations. The most serious one is that it smoothes the edges (i.e. sharp intensity changes) in the background and thus leaves these background edges in the foreground output. This background bias usually has arbitrary sizes and shapes, and is therefore hard to be corrected further by heuristics. Figure 3 (column 4) shows this effect. In addition, MO has the effects of local overerosion or overdilation. The level of these effects depends on the window size of the morphological operator. Smaller window size minimizes these effects, but in the applications to microarray images, the window size must be larger than the spot size.

To overcome these problems in microarray background correction, we propose the use of a total variation (TV)-based regularization method with an L^1-norm fidelity term (Alliney, 1992; Nikolova, 2002; Chan and Esedoglu, 2004, ftp://ftp.math.ucla.edu/pub/camreport/cam04-07.pdf). The preliminary numerical results on both synthetic and real data appear to give significant improvements over MO.

The rest of the paper is organized as follows. We describe the TV+L^1 model in Section 2 and give its properties in Section 3, including the choice of the only parameter λ. Section 4 discusses implementation aspects. Finally, Section 5 demonstrates the advantages of the TV+L^1 model over MO using numerical examples.

2 THE TV+L^1 MODEL

We begin with the discussion of a generalized signal regularization framework, which solves a variational problem. In this framework, f is modeled as the sum of image cartoon u and texture v, where f, u and v are functions with bounded supports. Cartoon contains background hues and important boundaries. The rest of the image, which is texture, is characterized by small-scale oscillating patterns. If present, noises and small artifacts are included in v. Since cartoon u is more regular than texture v, we can obtain u from input f by letting u to be the solution that minimizes an irregularity measure while being close to f with respect to a fidelity (distance) measure. The choice of the irregularity measure and the fidelity measure is application dependent.

A popular method for image processing in this framework is given by Rudin *et al*. (1992). They proposed using TV $\int |\nabla u|$ of u as the irregularity measure and the L^2-norm of $(f-u)$ as the fidelity measure for noisy image restoration, where u is defined in the space of functions with bounded variations (the BV space). This method removes noises while maintaining the sharp edges, which are important in most applications. Instead of using the L^2-norm, Alliney (1992), Nikolova (2002) and Chan and Esedoglu (2004) proposed and analyzed the TV model using the L^1-norm. We borrow the idea of image decomposition from Chan and Esedoglu (2004) and apply this model to background correction. Formally, this TV+L^1 model is formulated as:

$$\min_{u\in \mathrm{BV}(\Omega)} \int_\Omega |\nabla u(x)|\,\mathrm{d}x \quad s.t.\ \|f(x)-u(x)\|_{L^1} \le \sigma, \tag{1}$$

where Ω is the image domain and functions f and u are defined on Ω. Since (1) is a convex optimization problem, it can be reformulated as

$$\min_{u\in \mathrm{BV}(\Omega)} \int_\Omega |\nabla u(x)| + \lambda |f(x)-u(x)|\,\mathrm{d}x, \tag{2}$$

where λ is the Lagrange multiplier. This equivalence is covered in most textbooks on convex analysis (Rockafellar, 1996). When we apply this model to cDNA microarray image background correction, f is the microarray image input. Since u is the cartoon of f, it is the corrected background output, and therefore, $f-u$ is the signal of cDNA spots. Noting that the solution of Equation (2) depends on scalar λ, we often write u_λ and $v_\lambda = f - u_\lambda$ as the background and spot output.

In order to explain why this model can correctly extract cDNA spots from the input with inhomogeneous backgrounds, we describe its analytical properties and derive λ in the following section.

3 ANALYTICAL PROPERTIES

In this section, we discuss the important properties of the TV+L^1 model that help to understand its good performance on microarray background correction. To avoid too much analysis on the boundary of Ω, we assume $\Omega = \mathbb{R}^2$. The following lemma bridges the gap between the variational problem (2), which is easy to solve, and its geometrical equivalent, which is easy to analyze.

LEMMA 1 (Chan and Esedoglu, 2004). *Solving Equation (2) with $\Omega = \mathbb{R}^2$ is equivalent to solving the following level-set-based problem:*

$$\min_{u\in \mathrm{BV}} \int_{-\infty}^{+\infty} \mathrm{Per}\left(\{x : u(x) > \mu\}\right) + \lambda \mathrm{Vol}\left(\{x : u(x) > \mu\} \oplus \{x : f(x) > \mu\}\right)\,\mathrm{d}\mu, \tag{3}$$

where Per *is the perimeter function,* Vol *is the volume function and $S_1 \oplus S_2 := (S_1 \setminus S_2) \cup (S_2 \setminus S_1)$ for the given sets S_1 and S_2.*

According to this lemma the TV+L^1 model is operated on the level sets $\{x : u(x) > \mu\}$ and $\{x : f(x) > \mu\}$, for $\mu \in (-\infty, \infty)$, and it minimizes a geometric problem in them. This paves the way to the rest of the analysis.

Since cDNA spots are almost round, we use disk signal to approximate them. Using Lemma 1, we can analytically derive the exact solution $v_\lambda = f - u_\lambda$ in Equation (2):

(1) In the case when the input has the background with intensity c_0 and a cDNA spot with intensity $c_0 + c_1$, we have the input $f = c_0 + c_1 1_{B_r(y)}(x)$, i.e. f is a function with the value $c_0 + c_1$ in the disk centered at y and with radius r, and the value c_0 anywhere else. (Chan and Esedoglu, 2004)

$$v_\lambda = \begin{cases} c_1 1_{B_r(y)}(x) & 0 < \lambda < 2/r, \\ \{s 1_{B_r(y)}(x) : 0 \le s \le c_1\} & \lambda = 2/r, \\ 0 & \lambda > 2/r. \end{cases} \tag{4}$$

PROOF. Without loss of generality, we assume $c_1 > 0$. Clearly, solution $u(x)$ of Equation (2) is bounded between c_0 and $c_0 + c_1$. It follows that Equation (3) is simplified to:

$$\min_{u\in \mathrm{BV}} \int_{c_0}^{c_0+c_1} \mathrm{Per}\left(\{x : u(x) > \mu\}\right) + \lambda \mathrm{Vol}\left(\{x : u(x) > \mu\} \oplus \{x : f(x) > \mu\}\right)\,\mathrm{d}\mu. \tag{5}$$

Since $\{x : f(x) > \mu\} \equiv B_r(y)$ for $\mu \in (c_0, c_0 + c_1)$, $S(\mu) := \{x : u(x) > \mu\}$ must solve the following geometrical problem:

$$\min_S \mathrm{Per}\left(S(\mu)\right) + \lambda \mathrm{Vol}\left(S(\mu) \oplus B_r(y)\right), \tag{6}$$

for all $\mu \in (c_0, c_0+c_1)$. First, $S(\mu) \subseteq B_r(y)$ holds; otherwise, $S(\mu) \cap B_r(y)$ achieves a lower objective value of Equation (6) than $S(\mu)$. Then, it follows that

$$\text{Vol}\,(S(\mu) \oplus B_r(y)) = \text{Vol}\,(B_r(y) \backslash S(\mu))\,. \tag{7}$$

Therefore, to minimize Equation (6) is to minimize the perimeter of S while maximizing its volume. According to the Isoperimetric Theorem (Siegel, 2003), $S(\mu)$ must be either $\emptyset$ or a disk. Let r_S denote the radius of S, it is easy to see from the optimality of Equation (6) that $r_S = r$ if $\lambda > 2/r$, $r_S = 0$ if $0 < \lambda < 2/r$ and $r_S \in \{0, r\}$ if $\lambda = 2/r$. Equation (4) follows from relationship $v_\lambda = f - u_\lambda$.

Note that λ, which determines whether v_λ is $c_1 1_{B_r(y)}(x)$ or 0, depends only on the disk radius r but not on the values c_0 and c_1 and the disk center y! When $\lambda = 2/r$, Equation (2) gives multiple solutions. Generally, the analytical solutions of v_λ are not unique for at most a countable number of λs (Chan and Esedoglu, 2004). Therefore, we can omit these values in the forthcoming analysis and in the numerical tests.

(2) In the case when the cDNA spot has inhomogeneous intensity and its signal resembles an annulus, we have $f = c_0 + c_1 1_{A_{r_1,r_2}(y)}(x)$, where $0 < r_2 < r_1$ and A_{r_1,r_2} represents the annulus lying between two concentric circles with radii r_1 and r_2. In other words, f takes the value $c_0 + c_1$ between the two circles and the value c_0 elsewhere. Then

$$v_\lambda = \begin{cases} c_1 1_{A_{r_1,r_2}(y)}(x) & 0 < \lambda < \min\left\{\frac{2r_1}{r_1^2 - 2r_2^2}, \frac{2}{r_1 - r_2}\right\}, \\ -c_1 1_{B_{r_2}(y)}(x) & \frac{2}{r_1 - r_2} < \lambda < \frac{2}{r_2}, \\ 0 & \lambda > \max\left\{\frac{2}{r_1 - r_2}, \frac{2}{r_2}\right\}. \end{cases} \tag{8}$$

As in the previous case, scalar λ, which determines whether the entire spot signal is given in the output, only depends on size parameters r_1 and r_2. The following two properties describe the analytical solutions of two more complicated cases:

(3) Suppose $f = c_0 + c_1 1_{B_{r_1}(y_1)}(x) + c_2 1_{B_{r_2}(y_2)}(x)$, where $c_1, c_2 > 0$, $0 < r_2 < r_1$ and $B_{r_2}(y_2) \subset B_{r_1}(y_1)$

$$v(\lambda) = \begin{cases} (c_1 1_{B_{r_1}(y_1)} + c_2 1_{B_{r_2}(y_2)})(x) & 0 < \lambda < \dfrac{2}{r_1}, \\ c_2 1_{B_{r_2}(y_2)}(x) & \dfrac{2}{r_1} < \lambda < \dfrac{2}{r_2}, \\ 0 & \lambda > \dfrac{2}{r_2}. \end{cases} \tag{9}$$

(4) Assume the same as in the above property except $-c_1 < c_2 < 0$ and $y_1 = y_2 := y$, then

$$v_\lambda = \begin{cases} c_1 1_{B_{r_1}(y)} + c_2 1_{B_{r_2}(y)} & 0 < \lambda < \frac{2}{r_1}, \\ -c_2 1_{A_{r_1,r_2}(y)}(x) & \frac{2}{r_1} < \lambda < \min\left\{\frac{2r_1}{r_1^2 - 2r_2^2}, \frac{2}{r_1 - r_2}\right\}, \\ c_2 1_{B_{r_2}(y)}(x) & \frac{2r_1}{r_1^2 - 2r_2^2} < \lambda < \frac{2}{r_2}, \\ 0 & \lambda > \max\left\{\frac{2}{r_1 - r_2}, \frac{2}{r_2}\right\}. \end{cases} \tag{10}$$

Equation (4) shows that, when both the background and the spot have homogeneous intensities, any $\lambda < 2/r$, where r is the spot radius, makes the TV+L^1 model to return the exact spot intensity and the correct background. The other three properties infer that this is also true when the spots are not homogeneous, which is often the case in reality. Similar properties can be further extended to more general cases as long as the feature level sets have smooth boundaries (Chan and Esedoglu, 2004). In general, the decomposition using the TV+L^1 model is only scale dependent, and this property together with the edge preserving property (Strong and Chan, 2003) explain why the model is suitable for extracting small-scale signal under large-scale inhomogeneous background.

A disadvantage is that the model always leaves small areas where spots are located with constant intensity values in the estimated background. In the above properties, if the true background intensity is a ramp ranging from c_1^l to c_1^h in the area under the spot, then $u_\lambda = c_1^h$ when $\lambda < 2/r$. This happens because TV+L^1 decomposes f based on the scale of its level sets and the μ level set of f, for $\mu \in [c_1^l, c_1^h)$, is of a much larger scale than the c_1^h level set—the spot support. However, the intensity of large-scale inhomogeneous backgrounds changes relatively slowly except which crossing sharp boundaries; hence the background intensity in the small area under a spot can be regarded homogeneous (constant) in most cases. Consequently, we do not observe this effect in corrected backgrounds of real cDNA microarray images (Fig. 5) although we do so in the synthetic results (Fig. 3) with steep ramps in backgrounds (Fig. 2).

4 IMPLEMENTATION: PDE AND SOCP APPROACHES

In this section, we introduce two independent approaches to solve the discretized version of the TV+L^1 model. The microarray images can be represented as two-dimensional (2D) $m \times n$ matrices in $\mathbb{R}^{m\times n}$. Let matrix $f \in \mathbb{R}^{m\times n}$ denote the input microarray image. f contains inhomogeneous background $u \in \mathbb{R}^{m\times n}$, which the background correction process should identify. Let $v(= f - u)$ denote the rest of the image. Since v contains microarray spot signal, which is in small-scale, v is treated as the texture part in the TV+L^1 model.

The existing partial differential equation (PDE) approach (Chan and Esedoglu, 2004) requires the TV+L^1 model to be represented in the relaxed form [Equation (2)] and the existence of its first-order optimal condition: the Euler–Lagrange equation

$$\nabla \cdot \left(\frac{\nabla u}{|\nabla u|}\right) + \lambda \frac{f - u}{|f - u|} = 0. \tag{11}$$

In practice, to avoid division by zero, a small value $\epsilon > 0$ is added to $|\nabla u|$ and $|u - f|$. This change also makes the method strongly convex and ensures uniqueness. To solve Equation (2), the PDE approach uses an artificial time stepping method (an evolving method solving heat PDE with heat source) to find the solution of Equation (11). The evolving formula used in Chan and Esedoglu (2004) is

$$u_{i,j}^{n+1} = u_{i,j}^n + \delta_t \ell\, \partial_x^- \left(\frac{\partial_x^+ u_{i,j}^n}{M}\right) + \delta_t \ell\, \partial_y^- \left(\frac{\partial_y^+ u_{i,j}^n}{M}\right) + \delta_t \ell\, \lambda \frac{f_{i,j}^n - u_{i,j}^n}{\sqrt{(f_{i,j}^n - u_{i,j}^n)^2 + \epsilon}}, \tag{12}$$

where

$$M := \sqrt{(\partial_x^+ u_{i,j}^n)^2 + (\partial_y^+ u_{i,j}^n)^2 + \epsilon}, \quad u_{i,j}^0 = f_{i,j},$$

$\delta_t \ell$ is the time step, and ∂^+ and ∂^- are forward and backward partial finite differences, respectively. Clearly, this approach has cheap per-iteration cost as it only computes the first-order discrete derivatives in the LHS of Equation (11) in each iteration. Numerical experiments, however, show that this advantage is offset by the large number of iterations needed to reach a steady state u. Since the last term in Equation (12) is very sensitive to the sign of $f_{i,j}^n - u_{i,j}^n$, especially when the system is close to a steady state, $\delta_t \ell$ has to be very small for the update formula (12) to catch all important changes in the flow field u. As a result, the PDE approach must reduce $\delta_t \ell$ with the increase of n. We cannot reduce $\delta_t \ell$ arbitrarily close to zero because too small $\delta_t \ell$ causes numerical problems and increases the execution time. On the other hand, the second-order cone programming (SOCP) approach (Goldfarb and Yin, 2004) described in the following paragraph neither relies on the existence of Equation (11) nor uses any time steps, and hence, avoids the problems of the PDE method in practice.

The SOCP (Alizadeh and Goldfarb, 2003) is an extension to linear programming (LP). The vector inequality constraint in the form of $a \geq b$ in LP is extended to $a - b \in \mathcal{K}$ in SOCP, where $\mathcal{K}$ is one or a Cartesian product of second-order cones $\{(s_0; \bar{s}) : s_0 \geq \sqrt{\bar{s}'\bar{s}}\}$. In the 1D case, a second-order cone reduces to $\{s_0 : s_0 \geq 0\}$. In the 3D space, $\{(s_0; s_1, s_2) : s_0 \geq \sqrt{s_1^2 + s_2^2}\}$ looks like an ice cream cone. SOCPs can be solved by modern interior-point algorithms that are efficient both in theory and in practice. Goldfarb and Yin (2004) introduced SOCP to a group of TV-based image regularization models, including Equations (1) and (2). The TV term $\int_\Omega |\nabla u| \, \mathrm{d}x$ is handled discretely by $\sum_{i,j} \ell_{i,j}$ subject to $\ell_{i,j}^2 \geq (\partial_x u_{i,j})^2 + (\partial_y u_{i,j})^2$ and the L^1 term is handled discretely by $\sum_{i,j} w_{i,j}$ subject to $w_{i,j} \geq f_{i,j} - u_{i,j}$ and $w_{i,j} \geq u_{i,j} - f_{i,j}$. In addition, by defining operators ∂_x and ∂_y as forward, backward or centered finite differences that are linear in u, Equations (1) and (2) can be formulated as SOCPs. Each of the SOCPs minimizes a linear function subject to a set of linear equality and inequality constraints and second-order constraints $\ell_{i,j}^2 \geq (\partial_x u_{i,j})^2 + (\partial_y u_{i,j})^2$, for each i, j. In SOCP interior-point methods, the per-iteration cost of solving the SOCP is much higher than calculating Equation (12), but on the other hand they usually take only 10–30 iterations to return an accurate solution in practice, and this number of iterations does not increase with image size. Moreover, Goldfarb and Yin (2004) demonstrated the application of domain decomposition in the TV+L^2 model, which further cuts the computation time and memory usage. This decomposition technique can be also applied to the TV+L^1 model.

Comparing these two approaches, we find that the PDE approach is more suitable to process large cDNA microarray images in a batch mode while the SOCP approach is more useful to process smaller cDNA microarray images with higher accuracy.

5 NUMERICAL EXPERIMENTS

In this section, we compare our proposed method against MO on both synthetic data and real cDNA microarray images.

5.1 Parameter selection

As shown in Section 3, the choice of parameter λ in the TV+L^1 model only depends on the scale of the signal to be extracted. To return entire spot signal in v_λ, the model can use any $\lambda < 2/r_{\max}$, where $r_{\max}$ denotes the largest spot radius in a microarray image. Since too small λ may cause numerical inaccuracy, it is better to choose λ that is slightly $<2/r_{\max}$. For fair comparison, we adjusted the window sizes when applying MO. We found that the use of oversmall and overlarge window sizes causes the estimated background to be contaminated by spot intensity and affected by overerosion and overdilation, respectively, hence we chose to use the window sizes that minimized these effects and showed the best results. Following these guidelines, for all the six synthetic images with $r_{\max} = 5$, we used $\lambda = 0.35$ in the TV+L^1 model and 8×8 window size in MO. For the real cDNA microarray images, we used λs varying between 0.3 and 0.8 and MO window sizes varying between 5×5 and 10×10 since these images are obtained from different sources and thus in different zooms. In practice, if spot sizes of a batch microarray images are fixed (e.g. in the microarray images produced by the same device) then a single λ is good for all the microarray images.

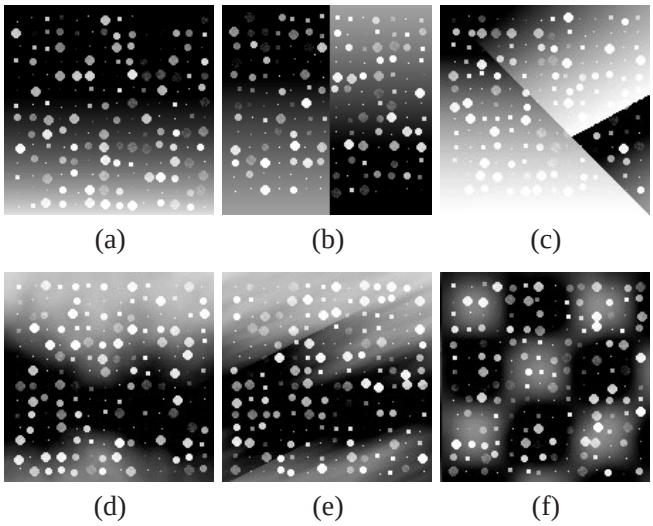


Fig. 2. Generated synthetic data

5.2 Codes and running times

We developed our PDE code in C++ and used commercial optimization package Mosek (called in Matlab) as our SOCP solver. We also implemented our MO code in C++. The average running times to process a 150×150 grayscale image are 18.34 s by the PDE approach, 12.25 s by the SOCP approach and 4.14 s by the MO approach on a Pentium IV-2.8 GHz Windows workstation with 1GB RAM. We recommend Goldfarb and Yin (2004) to readers for optimizing the SOCP solver and for a comprehensive comparison between the PDE and the SOCP approaches.

5.3 Synthetic data generation

A synthetic microarray image f^s is the sum of a foreground image v^s of spots and a background image u^s, which are generated on the 256 gray level scale. The intensity of each spot in v^s is uniformly distributed between 15 and 150. The radius of each spot is uniformly distributed between 0 (not visible) and 5 pixels. They are located along an $n \times n$ grid but their centers are subject to small Gaussian disturbance. These spots simulate cDNA microarray spots. Different backgrounds have been generated for the test. They are backgrounds with linear intensity changes (a), multilinear

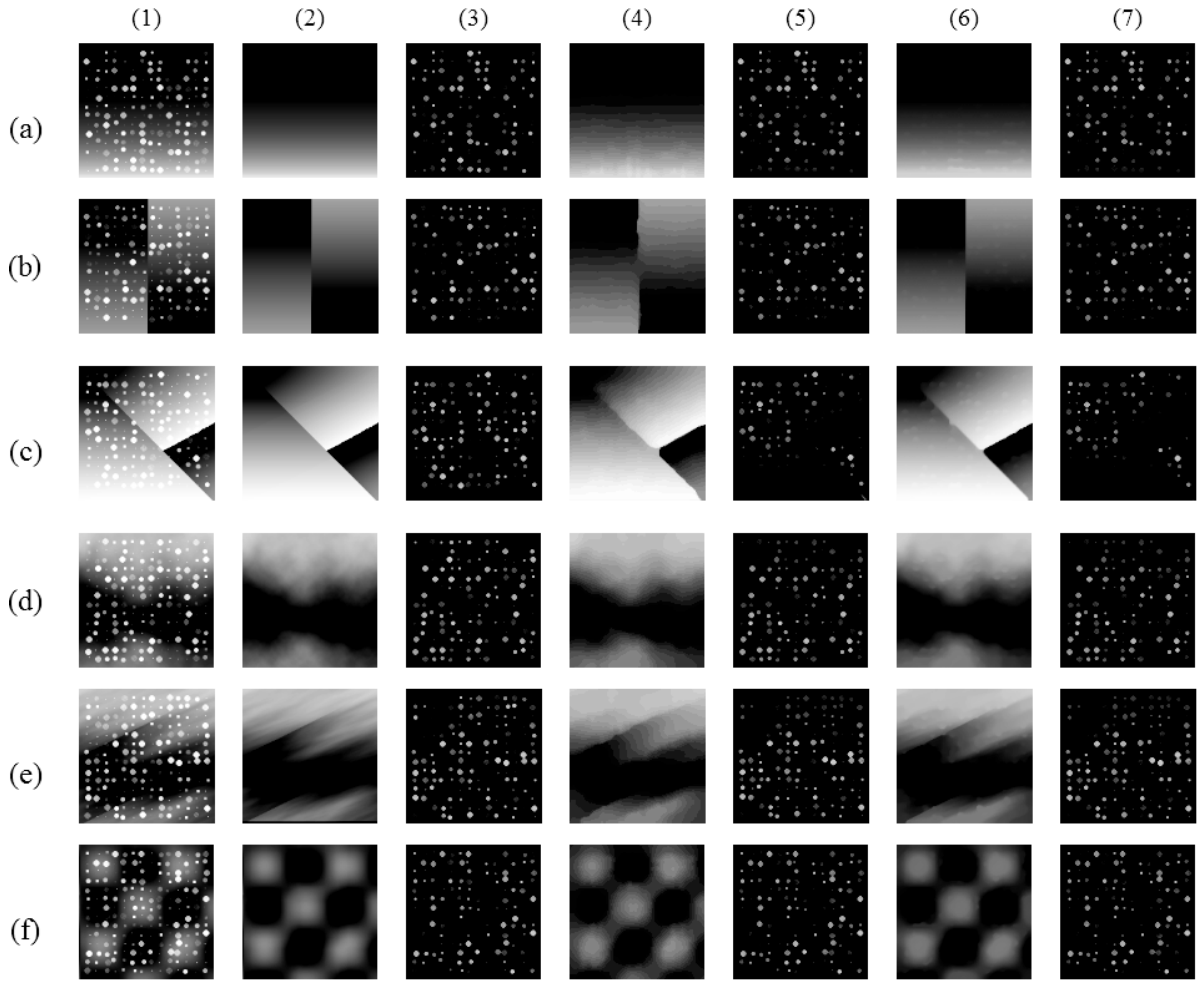
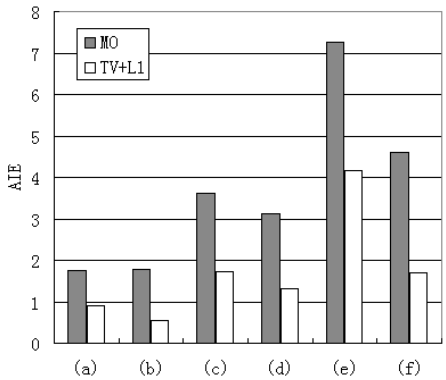


Fig. 3. Background removal results of synthetic microarray images (**a**)–(**f**). Each row shows one synthetic example. Column: (1) Original images, (2) ground truth background, (3) ground truth signal, (4) background estimation using MO, (5) Restored foreground using MO, (6) background estimation using the TV+L^1 model and (7) restored foreground using the TV+L^1 model.

gradient backgrounds (b) and (c), mildly inhomogeneous background (d), mildly inhomogeneous background with sharp edges (e) and large-scale Gaussian background (f), which are depicted in Figure 2. We use these examples to simulate real cDNA microarray images.

5.4 Synthetic data test results

Figure 3 shows the background correction results given by MO and TV+L^1. In all our tests with synthetic data, the TV+L^1 model gave more accurate backgrounds as compared with the originals than MO. While TV+L^1 preserves the edges in the backgrounds in tests (b), (c) and (e), MO creates the obvious edge distortions in these tests. In the TV+L^1 result (c6), we also observe the constant intensity left in the small areas where the spots are located. The MO results of tests (a), (b) and (e) show oversmoothed edges but stair-cased ramps in the background, and the MO results of test (d) and (f) show over-erosion and dilation effects. Moreover, in test (c), we can see some background intensity is left in the foreground output of MO (c5). To quantize the differences, we defined the average intensity errors



	(a)	(b)	(c)	(d)	(e)	(f)
MO	1.76	1.79	3.62	3.13	7.26	4.61
TV+L^1	0.91	0.55	1.73	1.32	4.17	1.70

Fig. 4. The AIEs of restored intensities.

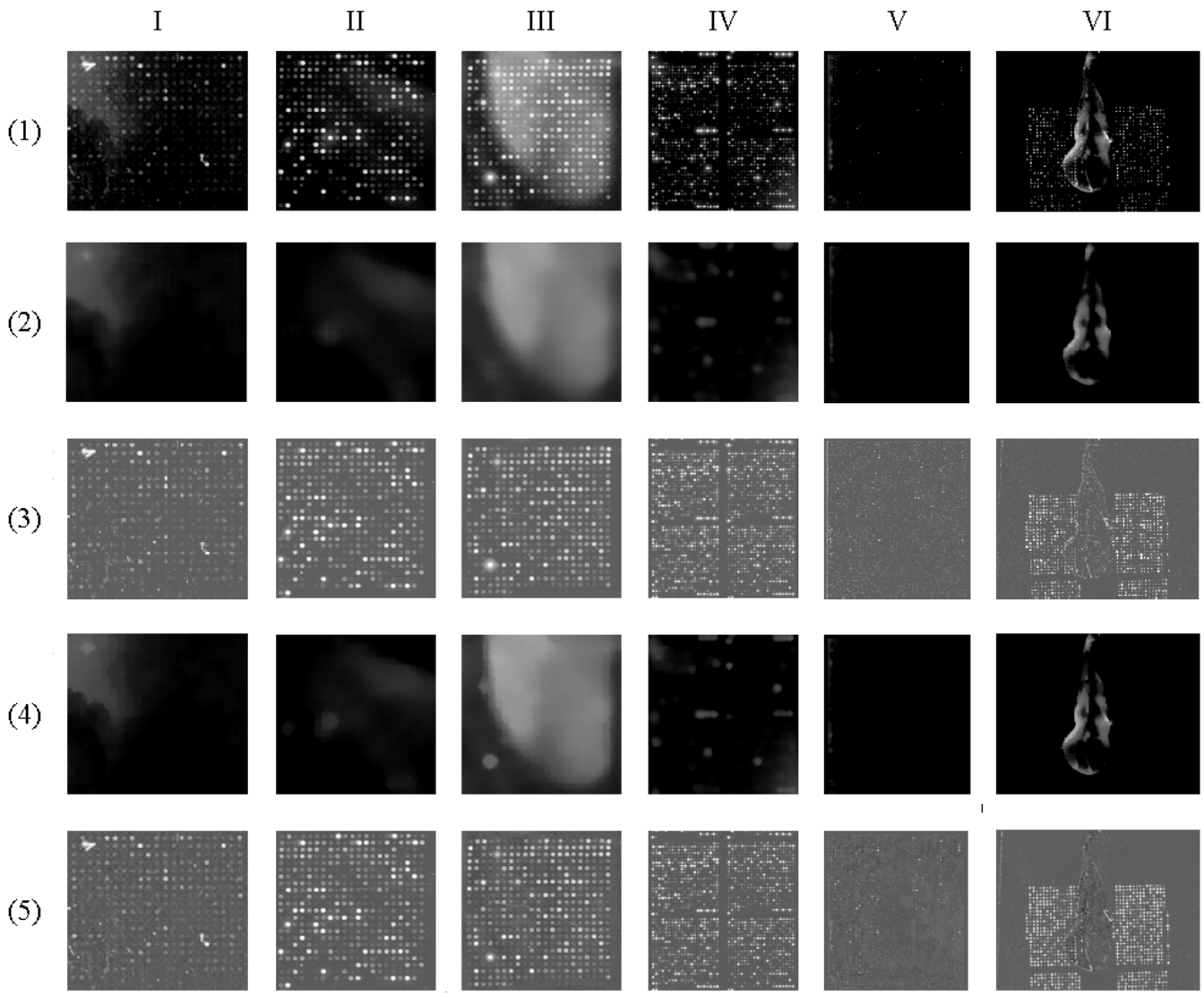


Fig. 5. Background removal results of real microarray images *I*–*VI*. Rows: (1) Original images, (2) background estimation using MO (3) restored foreground intensity using MO (+100 for better visualization), (4) background estimation using the TV+L^1 model and (5) restored foreground intensity using the TV+L^1 model (+100 for better visualization).

(AIEs) (Fig. 4) of these results as:

$$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}|u_{i,j}^{\text{true}}-u_{i,j}|}{m\times n},$$

where u^{true} and u are true and extracted backgrounds in form of $m\times n$ matrices, respectively. Figure 4 clearly shows that the AIEs of the results obtained by applying the TV+L^1 model are much smaller than those obtained by applying MO.

5.5 Real data test results

In this subsection, we continue to demonstrate the effectiveness of the TV+L^1 model by comparing it with MO on six problematic but representative real cases of cDNA microarray images. Background correction is often separately applied to the red and the green channels during the generation of real cDNA microarrays. However, since our testing microarray images have the similar inhomogeneous backgrounds in both channels, we directly apply the MO and TV+L^1 models to the images with combined red and green channels and a reduced color depth of 256 grays. Figure 5 depicts the results.

- *Case I* (Yang, 2004, http://www.biostat.ucsf.edu/jean/Presentation/ShareMAF/SMFQualityV3.pdf). This case demonstrates the importance of the edge preserving property of the TV+L^1 model. In the MO correction [Fig. 5I-2] of the water stain near the upper left corner [Fig. 5I-1], the edge of the stain is smoothed. The TV+L^1 model, however, keeps the edge of the stain perfectly intact [Fig. 5I-4]. This renders the restored signal closer to the original.
- *Cases II, III and IV* (Yang and Barczak, 2003, http://arrays.ucsf.edu/presentations/SandlerLabMeeting.2003.07.07.ppt).

These three cases demonstrate the weakness of MO. As we have mentioned in the introduction, MO using a large window size will smooth away smaller stains while MO using a small window size will affect the spot intensity values. This is why some small stains are not seen in the MO background correction. On the other hand, the TV+L^1 model does not have this problem as we can see in these three cases that all the large stains and the small stains are well kept.

- *Case V* Bibeau *et al*. (2005), http://www.corning.com/lifesciences/technical_information/techdocs/troubleshootingUltraGAPS_ProntoReagents.asp). In this case, the left side of the image has water stain overlapped with the signals. This stain is not completely kept in the background using MO as we can see a vertical line near the left edge of the estimated foreground (Fig. 5V-4). The TV+L^1 model successfully keeps the stain in the background.
- *Case VI* (Bibeau *et al*., 2005). In this extreme example, both methods give satisfying results. The advantage of the TV+L^1 model over MO is demonstrated by the sharper edges and more details of the water drop in the estimated background and the clearer spot expressions under the water stain in the estimated foreground.

In cases I–IV where the spot sizes are similar, λs were set to 0.3. In cases V and VI, λs were set to 0.8 as the spot sizes are much smaller than those in the previous cases. We also adjusted the window sizes of MO and selected the smallest window size that keeps the spots in the estimated foregrounds. We note that, in the estimated foregrounds of the TV+L^1 model, only spot signals (and small-scale artifacts) have positive intensity values. The remaining area has zero intensity uniformly. Clearly, this property is very useful in spot finding (Saeed *et al*., 2003; Jain *et al*., 2002; MicroDiscovery, 2004, http://www.microdiscovery.de; Koada Technology, 2004, http://www.koadarray.com), another important process on cDNA microarray images.

6 DISCUSSION AND CONCLUSION

In this paper, we propose the use of the optimization model of minimizing the TV and an L^1-norm fidelity term for correcting background intensity inhomogeneities. This model decomposes the input into a large-scale background part and a small-scale signal part. It is suitable for background correction because the decomposition is independent of the feature intensity and is controlled simply by a scalar parameter λ. Moreover, the correct λ can be easily calculated. We generate synthetic data with various background bias and measure the accuracy of restored signal. The numerical results show that the method performs better than the prevailing method—MO. This is further supported by experimental results on six real microarray images. One disadvantage of the proposed method is that it leaves small areas where spots are located with constant intensity values in the estimated background. In future we will develop finer correction methods to remove this defect.

In conclusion, we believe that the proposed work will contribute to the field of cDNA microarray data analysis on account of a more accurate restoration of the original intensities of the hybridized spots.

ACKNOWLEDGEMENTS

The authors want to thank Dorin Comaniciu, Luyong Wang and Diana Luca for valuable advice and proofreading of the manuscript. We also thank the referees for suggestions on the improvement of the paper.

REFERENCES

Alizadeh,F. and Goldfarb,D. (2003) Second-order cone programming. *Math. Program.*, **95**, 3–51.

Alliney,S. (1992) Digital filters as absolute norm regularizers. *IEEE Trans. Signal Process.*, **40**, 1548–1562.

Axon Instruments, Inc. (1999) *GenePix 4000A User's Guide*. Union City, CA.

Beare,R. and Buckley,M. (2004) *Spot: cDNA Microarray Image Analysis—User's Guide*. CSIRO Mathematical and Information Sciences.

Bibeau,C. *et al*. (2005) Troubleshooting guide for Corning UltraGAPS slides and Corning Pronto! universal hybridization kits and reagents, Figures 16 and 18, Corning Incorporated, Corning NY.

Chan,T.F. and Esedoglu,S. (2004) Aspects of total variation regularized L^1 function approximation. *UCLA CAM Report 04-07*, UCLA, Los Angeles, CA.

Eisen,M.B. and Brown,P.O. (1999) ScanAlyze.

Goldfarb,D. and Yin,W. (2004) Second-order cone programming methods for total variation-based image restoration, *Columbia University CORC Report TR-2004-05*, Department of Industrial Engineering and Operations Research, Columbia University, NY.

Jain,A.N. *et al*. (2002) Fully automatic quantification of microarray image data. *Genome Res.*, **12**, 325–332.

Koada Technology (2004) Koadarray.

Medigue,C. *et al*. (1999) Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics*, **15**, 2–15.

MicroDiscovery GmbH (2004) GeneSpotter.

Nikolova,M. (2002) Minimizers of cost-functions involving nonsmooth data-fidelity terms, *SIAM J. Numer. Anal.*, **40**, 965–994.

Rochafellar,R.T. (1996) *Convex Analysis (Princeton Landmarks in Mathematics and Physics)*. Princeton University Press, Reprint Edn.

Rudin,L. *et al*. (1992) Nonlinear total variation based noise removal algorithms. *Physica D*, **60**, 259–268.

Saeed,A.I. *et al*. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.

Siegel,A. (2003) An isoperimetric theorem in plane geometry. *J. Discr. Comput. Geometry*, **29**, 239–255.

Smyth,G.K. *et al*. (2003) Statistical issues in cDNA microarray data analysis. In Brownstein,M.J. and Khodursky,A.B. (eds), *Funcational Genomics: Methods and Protocols*, pp. 111–136.

Strong,D. and Chan,T. (2003) Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Problems*, **19**, 165–187.

Soille,P. (1999) *Morphological Image Analysis: Principles and Applications*. Springer, NY.

Yang,Y.H. (2004) Assessing microarray data quality, Slide 33 (Original data provided by Christina Lewis, UCSF).

Yang,Y.H. and Barczak,A. (2003) Standards for assessing microarray data quality, Slides 19 and 20.

Yang,Y.H. *et al*. (2000) Comparison of methods for image analysis on cDNA microarray data. *UC Berkeley Technical Report No. 584*, Statues Department, UC Berkeley, CA.