# Noise Filtering and Microarray Image Reconstruction Via Chained Fouriers

Karl Fraser[1], Zidong Wang[1], Yongmin Li[1], Paul Kellam[2], and Xiaohui Liu[1]

[1] School of Information Systems, Computing and Mathematics
Brunel University, Uxbridge, Middlesex, UB8 3PH, U.K.
karl.fraser@brunel.ac.uk
[2] Department of Infection
University College London, London, W1T 4JF, U.K.

**Abstract.** Microarrays allow biologists to determine the gene expressions for tens of thousands of genes simultaneously, however due to biological processes, the resulting microarray slides are permeated with noise. During quantification of the gene expressions, there is a need to remove a gene's noise or background for purposes of precision. This paper presents a novel technique for such a background removal process. The technique uses a gene's neighbour regions as representative background pixels and reconstructs the gene region itself such that the region resembles the local background. With use of this new background image, the gene expressions can be calculated more accurately. Experiments are carried out to test the technique against a mainstream and an alternative microarray analysis method. Our process is shown to reduce variability in the final expression results.

**Keywords:** Microarray, Filtering, Reconstruction, Fourier.

## 1 Introduction

The invention of the microarray in the mid-90's dramatically changed the landscape of modern day genetics research. The devices allow simultaneous real time monitoring of expression levels for tens of thousands of genes. One of these so-called "gene chips" contains probes for an organism's entire transcriptome. The different conditions or cell lines render a list of genes with their appropriate activation levels. These gene lists are then analysed with the application of various computational techniques, for example clustering [1], or modelling [2] such that differential expressions are translated into a better understanding of the underlying biological phenomena.

A major challenge with any real-world data analysis process is how to address data quality issues effectively. Although microarray hardware is engineered to very high tolerances, noise (henceforth "noise" and "background" are synonymous) will be introduced into the final output slide. This noise can take many forms, ranging from common artefacts such as; hair, dust and scratches on the slide, to technical errors like; the random variation in scanning laser intensity or the miscalculation of gene expression due to alignment issues. Alongside these

M.R. Berthold, J. Shawe-Taylor, and N. Lavrač (Eds.): IDA 2007, LNCS 4723, pp. 308–319, 2007.

technical errors there exist a host of biological related artefacts; contamination of complementary Deoxyribonucleic Acid (cDNA) solution or inconsistent hybridisation of the multiple samples for example.

Unfortunately, these images are expensive to produce and the current climate is such that “bad” experiment sets must still be analysed, regardless of their quality. Whereas such poor images could simply be discarded in other fields, here, an image must yield some knowledge irrespective of how small. It is common practice throughout then to implement some form of duplication in-situ such that correction tasks can take place during downstream analysis. Much work in the field therefore focuses on post-processing or analysing the gene expression ratios themselves [1,3,4,5,6,7] as rendered from given image sets, which means there is relatively little work directed at pre-processing or improving the original images to begin with [8,9].

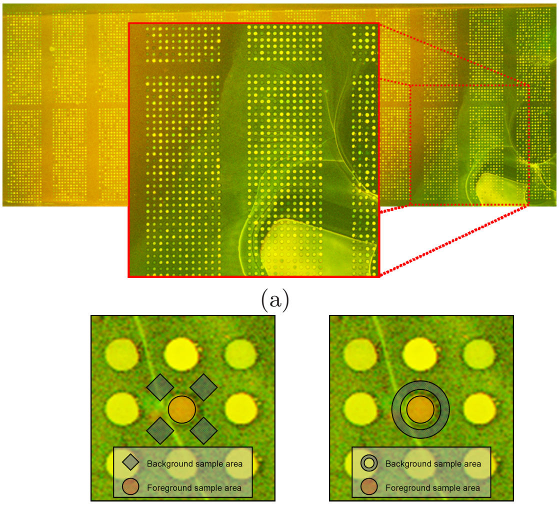
Microarray images are full of background signal that is of no real interest specifically to the experimental process. Nevertheless, these artefacts can have a detrimental effect on the identification of genes as well as their accurate quantification. There are many reasons for this, the most critical of which is due to the similar intensity levels seen between noise and a gene (due to inappropriate DNA binding sites for example). In this paper, we present an algorithm that attempts to remove the biological experiment from the image. In this context, the biological experiment consists of the gene spot regions. Put another way; imagine the image is made up of two separate layers. The bottom layer consists of the glass substrate material upon which the gene spots are deposited onto to begin with. The top layer on the other hand consists of the gene spots. Removal of the biological experiment regions is to clear the top layer such that the hidden regions of the bottom layer can be seen. In effect, this removal process is equivalent to background reconstruction and will therefore produce an image which resembles the “ideal” background more closely in experimental regions. Subtracting this new background from the original image should yield more accurate gene spot regions. The reconstructed expressions are contrasted to those as produced by GenePix [10] (a commercial system commonly used by biologists to analyse images) and O’Neill *et al.* [9] (one of the first reconstruction processes implemented to deal with microarray image data).

The paper is organised in the follow manner. First, we formalise the problem area as it pertains to real microarray image data sets and briefly explain the workings of contemporary approaches in the next section. Section Three discusses the fundamental idea of our approach with the appropriate steps involved in the analysis. In Section Four, we briefly describe the data used throughout the work then detail and evaluate the tests carried out for the synthetic and real-world data. Section Five summarises our findings, draws out some relevant conclusions and defines future considerations and directions.

## 2 Background

Regardless of the specific techniques used to assist with downstream analysis of microarray image data, all of them have similarities due to the nature of the problem. For example, the techniques require knowledge of a given gene spot's approximate central pixel as well as the slide's structural layout. A boundary is then defined around the gene spot and background pixels with the median of these regions taken to be foreground and background intensities respectively. Then, the background median is subtracted from the foreground and the result is summarised as a $\log_2$ ratio. Other bounding mechanisms include pixel partitioning via histogram [3,11] and region growing [12,13] functions with a detailed comparison of the more common approaches given in [8]. The underlying assumption for these mechanisms is that there is little variation within the gene and background regions.

Unfortunately, this is not always the case as seen in Fig. 1a generally, which depicts a typical test set slide (enhanced to show gene spot locations) with a total of 9216 gene regions on the surface and measuring ∼5000×2000 pixels. A good example of the low-level signal produced in the image can be seen in the close-up sections, where problems such as missing or partial gene spots, shape inconsistencies, and background variation can be seen. Such issues are highlighted in more detail in b and c where the scratch and background illuminations around the presented genes change significantly.



(a)

(b) (c)

**Fig. 1.** Example Images: Typical test set Slide Illustrating Structure and Noise (a) and Sample Gene, Background Locations for GenePix Valleys (b) and ImaGene Circles (c)

What is needed is a more specific background determination process that can account for the inherent variation between the gene and background regions. Texture synthesis represents a possible avenue for such background reconstruction processes. An established reconstruction technique is that as proposed by Efros *et al.* [14] whereby a non-parametric process grows a reconstruction outwards from an initial seed pixel, one pixel at a time via Markov Random Fields. The Bertalmio *et al.* [15] approach on the other hand attempts to mimic techniques as used by professional restorers of paintings and therefore works on the principle of an isotropic diffusion model. Moreover, Chan *et al.* [16] greatly extended the work of [14] and others to propose a curvature model based approach. However, microarray images contain thousands of regions requiring such reconstructions and are therefore computationally expensive to examine with the highlighted techniques. In an attempt to overcome such time restrictions (although not focused at microarray data itself) Oliveira *et al.* [17] aimed to produce similar results to [15] albeit quicker, although as we shall see the approach loses something in translation.

One of the first reconstruction techniques applied specifically to microarray images is that as proposed by O'Neill *et al.* [9] which utilises a simplification of the Efros *et al.* [14] technique. In this context, a gene spot is removed from the surface and recreated by searching a known background region and selecting pixels most similar to the known border. By making the new region most similar to given border intensities it is theorised that local background structures will transition through the new region. However, the best such a process has accomplished in this regard is to maintain a semblance of valid intensities, while the original topological information is lost. The next section describes a technique that attempts to address some of these issues, e.g. retention of topology, process efficiency, and edge definition in a more natural way.

## 3 A New Analysis Technique

In this work, we have proposed Chained Fourier Image Reconstruction (*CFIR*), a novel technique that removes gene spot regions from a microarray image surface. Although this may seem counter-intuitive (the gene spots are the elements of value in a microarray after all), the successful removal of these regions leads to more accurate or natural looking background surfaces, which can be used to yield yet more accurate gene spot intensities. Techniques such as O'Neill work in the spatial domain exclusively and essentially compare all gene border pixels to those of the local background to produce appropriate pixel mappings. Although this works well, such brute force methods are typically expensive with respect to execution time. However, if we harness the frequency domain along with more traditional spatial ideas we can render a reconstruction that inherently deals with the issues (illumination, shading consistencies etc) more efficiently.

Taking the diagram of Fig. 1b as our reference, let us now detail the CFIR process outlined in Table 1. Initially due to the nature of the microarraying process, gene spots can be rendered with different shapes and dimensions, both individually and through the channels. Therefore, a generic window centred at the gene (as determined by GenePix) can be used to capture all pixels $p_{xy}$ within a specified square distance from this centre, where *(x,y)* are the relative coordinates of the pixels in the window centred at pixel *p*. Window size is calculated directly from an analysis of the underlying image along with resolution meta-data if needed. The window can then be used to determine the appropriate *srcList* and *trgList* pixel lists (foreground and background) accordingly. Note that in the current implementation, the background region resembles a square as defined by the outer edges of the diamonds in Fig. 1b.

With the two lists (*srcList*, *trgList*) in place a Fast Fourier Transform (FFT) is applied to both lists independently (as highlighted in lines 2∼4). If *f(x,y)* for

**Table 1.** Pseudo-Code of Chained Fourier Transform Reconstruction Function

```
Input
      srcList: List of gene spot region pixels
      trgList: List of sample region pixels
Output
      outList: srcList pixels recalibrated into trgList range

Function fftEstimation(srcList,trgList):outList
 1. For each gene
 2.   srcMask = fourier transform srcList members
 3.   trgSample = fourier transform trgList members
 4.   recon = srcMask * trgSample // to generate initial reconstructed surface
 5.   While doneIterate = 0
 6.     recon = fourier transform initial reconstructed surface
 7.     reconPhase = phase elements of reconstructed surface
 8.     minimum recon element = smallest element in the trgSample surface
 9.     recon = inverse fourier transform merged recon and reconPhase surfaces
          // such that subtle characteristics are retained
10.     recon elements $\leq 0$ = smallest element in srcMask
11.     recon elements $\geq 65535$ = largest element in srcMask
12.     reset non-gene pixels in recon = trgList
13.     if difference between recon and trgSample ¡tolerance
14.         doneIterate = 1
15.     End If
16.   End While
17.   outList = reconstructed region
18. End For
End Function
```

*x;y=0,1,...,M-1;N-1* respectively denote the $M \times N$ image region, the digital FFT for *F(u,v)* can be defined as

$$F(u,v) = \sum_{M-1}^{x=0} \sum_{N-1}^{y=0} f(x,y) e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{1}$$

where *(u,v)* represent the frequency coordinates of their spatial *(x,y)* equivalents. Note the inverse transform is computed in much the same way. The real $R$, imaginary $I$ and phase $\phi$ components of the resulting FFT spectrum can then be broken up according to

$$|F(u,v)| = \left[R^2(u,v) + I^2(u,v)\right]^{\frac{1}{2}}\text{, and} \tag{2}$$

$$\phi(u,v) = tan^{-1}\left[\frac{I(u,v)}{R(u,v)}\right]\text{, respectively} \tag{3}$$

Global features of the image regions (repeating patterns, overall region intensity etc) thus become localised within the frequency spectrum, while non-repeating structures become scattered. Retaining this phase information in the reconstructed region is crucial as this has the effect of aligning global features (much the same as the isotropic diffusion approach of [15] does for example) and as such presents subtle surface characteristics (illumination and shading features etc). In order to capture this subtle intensity information within the background (*trgList*) region and allow the gene spot (*srcList*) area to inherit it, a simple minimisation function is used (as per lines 7∼9). More complicated criteria could be computed in this regard but after critical testing it was found that the minimum of the region produces good results and is thus used at present

$$R(u,v) = minimum\,|srcList(R), trgList(R)| \tag{4}$$

The final stage of the algorithm (lines 10∼12) replaces modified background (*trgList*) pixels within the gene spot (*srcList*) area with their original values. Recall, the FFT function disregards spatial information, which means subsequent modifications (like the minimiser function) could well change inappropriate pixels with respect to the reversed FFT of line 9. Therefore, the original non-gene spot pixels must be copied back into the modified regions such that erroneous allocations are not propagated through the reconstructed region during the next cycle.

The actual convergent criterion as highlighted in line 13 can be determined by the mean squared error (MSE) or other such correlation methods between the reconstructed and background regions. Generally though, the MSE falls rapidly over the initial iterations and thenceforth slows until a minimum is reached. Conversely, the correlation coefficient approach would be expected to rise rapidly over the initial iterations and then slow as convergence approaches. Regardless however, the tolerance criterion guarantees termination of the reconstruction process when iterative changes are at a minimum. In practice, tolerance is calculated such that

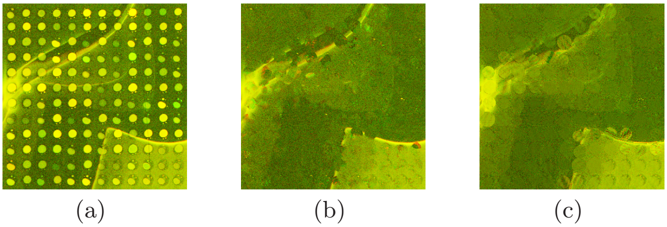


**Fig. 2.** Chained Fourier Transform Example: Original Image (a), Reconstructed O'Neill (b) and CFIR (c) Regions

the absolute difference (for the gene spot (*srcList*) pixels specifically) between all original and reconstructed pairs (for an individual region) are monotonically decreasing. Such monotonicity helps with retention of illumination and tone information that would otherwise be lost. Fig. 2 presents a sample-reconstructed region from the Fig. 1a image as processed by the techniques.

Application of frequency and spatial methods when applied separately to such problems can work well (see Fig. 2b for example) but there are better ways to carry out such processes. The formulation as described for CFIR allows us to inherently combine advantages from both the frequency and spatial domains such that reconstructed regions not only retain implicit domain information but, are processed faster than contemporary methods. Related to this implicit domain information (and suggested in Fig. 2bc) is the problem of correct edge classification. Generally however, CFIR improves handling and production of results accordingly as will be seen in the next section.

## 4 Experiments

This section details the results of numerous experiments that were designed to test empirically the performance and differences between the O'Neill and CFIR algorithms with respect to GenePix. Although there are many ways that such performance characteristics can be distilled, for this work the focus is on the resulting median expression intensities. These intensities become the raw gene expressions (as used in post-analysis [1,3,4,5,6,7] work for example) and therefore render overall insight into the reconstruction event. In addition, these values allow us to drill down into a particular gene spots repeat set and as such help clarify reconstruction quality.

As it is not possible to determine the optimal background for a gene spot region, the best validation in this context would seem to be to compare against rebuilt background regions. Such a comparison renders a clearer understanding of the reconstruction characteristics. To aid in this, 64 synthetic gene spots (SGS) were created and placed into existing background regions of the Fig. 1a image. If the reconstruction processes were to perform in an ideal manner, the SGSs would be removed from the slide such that these SGS regions are indistinguishable from their local surrounding region.

With the potential errors inherent in the GenePix background generation thus highlighted, the next stage of testing involved determining how these errors translate onto the reconstruction of real gene spots. Experiment two therefore examines the median intensities of the control gene spots for all blocks across the Fig. 1a image and the entire test set. Finally, experiment three carries out an explicit comparison between the techniques and thus yields the gained improvements (or not) of a particular technique as compared with GenePix.

Note that all of the images used in this paper were derived from two experiments conducted using the human gen1 clone set data. These two experiments were designed to contrast the effects of two cancer inhibiting drugs (PolyIC and LPS) onto two different cell lines, one being normal (control, or untreated) and the other the treatment (HeLa) line over a series of several time points. In total, there are 47 distinct slides with the corresponding GenePix results present.

The first experiment is designed to determine how well the reconstruction process can remove a synthetic gene from the image. When removed, the new region can be compared to the original with the difference calculated explicitly. Fig. 3 distils this difference information into a clear plot by calculating the average absolute pixel error for the SGS regions, as determined by the reconstruction techniques.

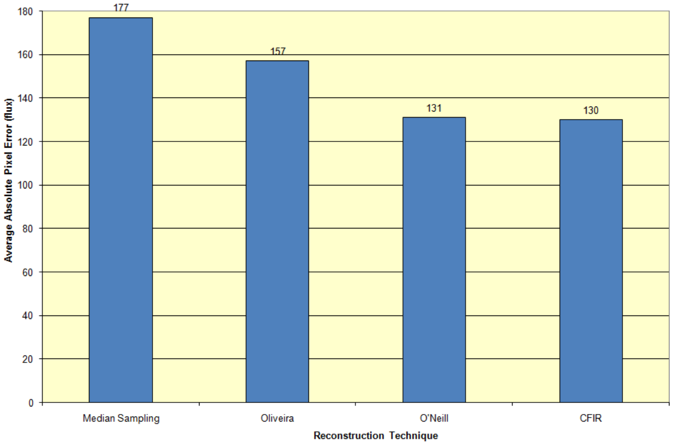
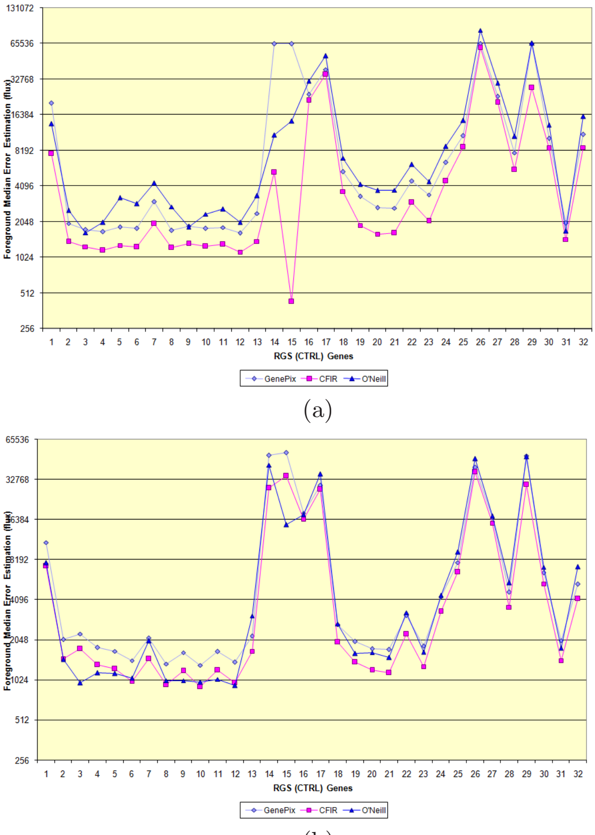


**Fig. 3.** Synthetic Gene Spots: Average Absolute Pixel Error

In effect, the graph shows that on average, GenePix's median sampling approach to background classification yields a potential intensity error of 177 flux per pixel for an SGS region, while the other techniques yield smaller error estimates. A consequence of this is that downstream analysis based on GenePix results directly must produce more erroneous gene expressions than realised.

The second experiment (with results as shown in Fig. 4) conducts the performance evaluation of CFIR, O'Neill and GenePix using true gene spots. This experiment is particularly focused on the relationships between expression measurements for all control repeat genes in the same slide (Fig. 4a) and across all slides in the test set (Fig. 4b). In all cases the underlying assumption for repeated genes is that they (the genes) should have highly similar intensity values (ideally) for a given time point, regardless of their location on the slide surface. Although we would perhaps expect to see some differences in the values as the time points increase over the duration of the biological experiments.



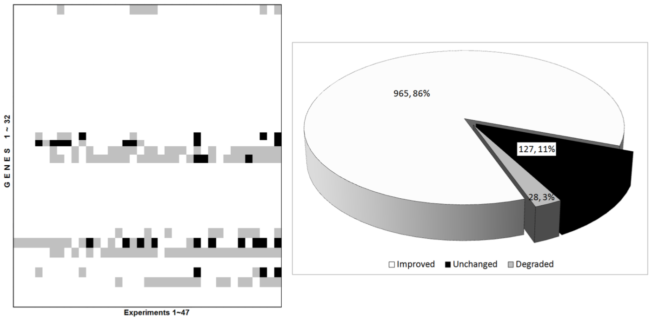
(a)

(b)

**Fig. 4.** Real Gene Spot Curves: Absolute Medians for 32 genes over Fig. 1a (a) and the Test set (b) Regions

The plots represent the absolute foreground median from both image channels for the tested techniques. It is clear from Fig. 4a that CFIR outperformed the other methods comfortably as far as the reduction of individual gene intensities is concerned. However, for the saturated gene spots (gene 15 for example) the estimation difference increases, although it is still closer than GenePix. Generally, for this particular image, CFIR outperformed both O'Neill and GenePix in relation to reconstruction with the specific residuals at 14514, 7749 and 13468 flux for GenePix, CFIR and O'Neill respectively. Indeed, as far as control data is concerned the CFIR process reduced noise by ∼46%.

Plotting the entire 47-slide test set produces a cross-sectional average through the data. The O'Neill and CFIR processes are much closer overall as compared to the sample slide; however, there are subtle handling differences in the saturated gene regions. Essentially, the intensity jump relationship for the saturated genes has flipped, meaning that CFIR seems to cope with such issues in a smoother way overall. The respective residuals for the entire test set are 10374, 7677 and 9213 flux respectively, which indicates that although not perfect, CFIR tends to produce lower repeat scores in general. To be fair the O'Neill technique has improved the overall score somewhat with respect to the individual image surface. It is clear that reconstructing the true gene spot regions does have a positive effect on the final expression results but, not so obvious, are the ramifications this reconstruction is having over the entire test set. Fig. 5a therefore plots a comparison chart which shows explicitly the improvement (or not) of a particular reconstruction method as compared to the original GenePix expressions. In addition, execution time plays a critical role in the reconstruction task, as techniques need to run as fast as possible given the number of gene spots that must be processed. Therefore, Fig. 5b presents a brief breakdown of the timings required for the techniques to parse a small percentage of the entire test set.



(a)

| | *Time on Xeon 3.4GHz* (hh:mm:ss) | | |
|---|---|---|---|
| *Image* | *GenePix* | *O'Neill* | *CFIR* |
| Hela PolyIC t00*(1a)* | 4:00:00+ | 1:53:00 | 00:52:39 |
| Hela PolyIC t05 | 4:00:00+ | 1:58:00 | 00:58:15 |
| Hela PolyIC t18 | 4:00:00+ | 1:52:00 | 00:54:44 |

(b)

**Fig. 5.** Final Results Comparison: Matrix for test set showing difference in repeat expression fluctuations (a); GenePix, CFIR and Both techniques are assigned the colours black, white and grey (∼10% difference) respectively and Sample Timing Chart (b)

The distinct banding occurring in gene regions 3∼8 and 16∼19 of Fig. 5a are associated with saturated (or near background) intensities as created by the scanner hardware and suggest more work is needed with respect to these genes. The non-banded genes on the other hand are indicative of the individual reconstruction techniques being able to account more appropriately for gene intensity replacement. Table 5b highlights the significant speed increase gained by applying the CFIR process to reconstruction rather than the O'Neill and GenePix methods.

## 5 Conclusions

The paper looked at the effects of applying both existing and new texture synthesis inspired reconstruction techniques to real-world microarray image data. It has been shown that the use of existing methods (which have typically focused on aesthetic reconstructions) to medical image applications can be highly effective, however their output quality and processing time's need to be significantly improved. As for microarray-focused reconstruction, pixel fill order as applied by the O'Neill technique plays a crucial role and should therefore be crafted with greater care.

To overcome timing and accuracy issues, we proposed a novel approach to reconstructing a gene's background by attempting to harness an image's global information more intently along with the gene's neighbour pixels. The proposed technique takes advantage of the grouping concept of the frequency domain and characterises global entities. At the same time, we use local spatial knowledge of a gene to help restrict a constructed regions spread. Results obtained from several experiments showed great improvement over a commonly used package (GenePix) and a brute force approach (O'Neill). Specifically, not only was the gene repeat variance reduced from slides in the test set, but in addition the construction time was decreased $\sim$50% in comparison to O'Neill's technique.

In future studies we wish to investigate gene spots that straddle strong artefact edges along with general transition issues as they are subject to the sub-allocation of replacement pixel(s). A transition edge will have a sharp yet convoluted evolution that can influence the accuracy of a gene's background. Such inconsistencies render themselves in the surface as halos and it would be beneficial if such halos were removed more appropriately. The artefact sub-allocation problem on the other hand requires a subtle approach to correction and we believe a weighted transition map would be more appropriate than the current bivalence approach.

## Acknowledgment

This work is in part supported by EPSRC grant (EP/C524586/1) and an International Joint Project sponsored by the Royal Society of the U.K. and the National Natural Science Foundation of China. The authors would also like to thank the anonymous reviewers for their comments.

## References

1. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. In: Proceedings of the National Academy of Sciences, USA, December 1998, pp. 14863–14868 (1998)
2. Kellam, P., Liu, X., Martin, N., Orengo, C.A., Swift, S., Tucker, A.: A framework for modelling virus gene expression data. Journal of Intelligent Data Analysis 6(3), 265–280 (2002)
3. Chen, Y., Dougherty, E.R., Bittner, M.L.: Ratio-based decisions and the quantitative analysis of cdna microarray images. Journal of Biomedical Optics 2, 364–374 (1997)
4. Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., Brown, P.O.: Genomic expression program in the response of yeast cells to environmental changes. Molecular biology of the cell 11, 4241–4257 (2000)
5. Quackenbush, J.: Computational analysis of microarray analysis. Nature Reviews Genetics 2(6), 418–427 (2001)
6. Kepler, B.M., Crosby, L., Morgan, T.K.: Normalization and analysis of dna microarray data by self-consistency and local regression. Genome Biology 3(7) (2002)
7. Quackenbush, J.: Microarray data normalization and transformation. Nature Genetics 32, 490–495 (2002)
8. Yang, Y.H., Buckley, M.J., Dudoit, S., Speed, T.P.: Comparison of methods for image analysis on cdna microarray data. Journal of Computational and Graphical Statistics 11, 108–136 (2002)
9. O'Neill, P., Magoulas, G.D., Liu, X.: Improved processing of microarray data using image reconstruction techniques. IEEE Transactions on Nanobioscience 2(4) (2003)
10. Anonymous: GenePix Pro Array analysis software. Axon Instruments Inc.
11. Anonymous: QuantArray analysis software. GSI Lumonics
12. Adams, R., Bischof, L.: Seeded region growing. IEEE Transactions on Pattern Analysis and Machine Intelligence 16, 641–647 (1994)
13. Ahuja, N., Rosenfeld, A., Haralick, R.M.: Neighbour gray levels as features in pixel classification. Pattern Recognition 12, 251–260 (1980)
14. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: IEEE International Conference on Computer Vision, pp. 1033–1038. IEEE Computer Society Press, Los Alamitos (1999)
15. Bertalmio, M., Bertozzi, A., Sapiro, G.: Navier-stokes, fluid dynamics, and image and video inpainting. In: IEEE Computer Vision and Pattern Recognition, IEEE Computer Society Press, Los Alamitos (2001)
16. Chan, T., Kang, S., Shen, J.: Euler's elastica and curvature based inpaintings. Journal of Applied Mathematics 63(2), 564–592 (2002)
17. Oliveira, M.M., Bowen, B., McKenna, R., Chang, Y.S.: Fast digital image inpainting. In: Proceedings of the Visualization, Imaging and Image Processing, Marbella, Spain, pp. 261–266 (September 2001)