Pattern Recognition Letters 29 (2008) 2129–2136

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# Can graph-cutting improve microarray gene expression reconstructions?

Karl Fraser [a,*], Zidong Wang [a], Yongmin Li [a], Paul Kellam [b], Xiaohui Liu [a]

[a] Centre for Intelligent Data Analysis, School of Information Systems, Computing and Mathematics, Brunel University, Uxbridge, Middlesex UB8 3PH, UK
[b] Department of Infection, University College London, London W1T 4JF, UK

ARTICLE INFO

*Article history:*
Received 23 December 2007
Received in revised form 26 July 2008
Available online 12 August 2008

Communicated by T. Vasilakos

*Keywords:*
cDNA
Two channel
Microarrays
Background reconstruction
Graph-edge cuts

ABSTRACT

Microarrays produce high-resolution image data that are, unfortunately, permeated with a great deal of "noise" that must be removed for precision purposes. This paper presents a technique for such a removal process. On completion of this non-trivial task, a new surface (devoid of gene spots) is subtracted from the original to render more precise gene expressions. The graph-cutting technique as implemented has the benefits that only the most appropriate pixels are replaced and these replacements are replicates rather than estimates. This means the influence of outliers and other artifacts are handled more appropriately (than in previous methods) as well as the variability of the final gene expressions being considerably reduced. Experiments are carried out to test the technique against commercial and previously researched reconstruction methods. Final results show that the graph-cutting inspired identification mechanism has a positive significant impact on reconstruction accuracy.



## 1. Introduction

Although microarray technology (Shalon and Davies, 1995) was invented in the mid–1990s the technology is still widely used in laboratories around the world today. The microarray "gene chip" contains probes for an organism's entire transcriptome where differing cell lines render gene lists with appropriate activation levels. Gene lists can be analysed with application of various computational techniques, be they clustering (Eisen et al., 1998) or modelling (Kellam et al., 2002), for example such that the differential expressions can be translated into a clearer understanding of the underlying biological phenomena present. For a detailed explanation of the microarraying process readers may find references (Coe, 2003; Duyk, 2002; Petricoin et al., 2002; Liu and Kellam, 2003) of interest.

Here, we provide a brief review of this process. Every cell in our body contains an instruction set that is stored in DNA, coding for all functional aspects; from protein synthesis to cell division. To protect the DNA's integrity, a copy is manufactured using RNA and it is this expendable copy that is used throughout the cell. If this RNA is extracted at a given point in time, copies of all the genes in use at that time can be identified, and then quantified by a measure of abundance using receptors tailored for each gene. A microarray is created by printing individual receptors for every gene into specific locations on a specially treated glass slide. This slide is then digitised using a dual laser scanning device, producing a two-channel 16-bit grey-scale image. The gene receptor locations on this image (typically 16–20 pixels in diameter) are identified, their median intensity values measured and then summarised as $\log^2$ ratios across both channels. An example usage of this technology is the comparison between cells for a patient before and after infection by a disease. If particular genes are used more after infection (highly expressed) then it can be surmised that these genes may play an important role in the life cycle of this virus.

Addressing the issue of microarray data quality effectively is a major challenge, particularly when dealing with real-world data, as "cracks" will appear regardless of the design specifications, etc. These cracks can take many forms, ranging from common artifacts such as hair, dust, and scratches on the slide, to technical errors like miscalculation of gene expression due to alignment issues or random variation in scanning laser intensity. Alongside these errors, there exist a host of biological related artifacts such as contamination of the complementary deoxyribonucleic acid (cDNA) solution or inconsistent hybridisation of multiple samples. The focus in the microarray field therefore is on analysing the gene expression ratios themselves (Chen et al., 1997; Eisen et al., 1998; Gasch et al., 2000; Quackenbush, 2001; Kepler et al., 2002; Quackenbush, 2002) as rendered from the image sets. This means there is relatively little work directed at improving the original images (Yang et al., 2002; O'Neill et al., 2003; Fraser et al., 2007a,b) such that final expressions are more realistic.

As noise in the images has a negative effect with respect to the correct identification and quantification of underlying genes, in this paper we present an algorithm that attempts to remove the biological experiment (or gene spots) from the image. In the

* Corresponding author. Tel.: +44 1895266544; fax: +44 1895232806.
*E-mail address:* Karl.Fraser@brunel.ac.uk (K. Fraser).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.07.011

microarray field, it is accepted as part of the analysis methodology that the background domain (non-gene spot pixels) infringes on the gene's valid measure and steps must be taken to remove these inconsistencies. In effect, this removal process is equivalent to background reconstruction and should therefore produce an image which resembles the "ideal" background more closely in experimental (gene spot) regions. Subtracting this new background image from the original should in-turn yield more accurate gene spot expression values. The gene expression results of the proposed reconstruction process are contrasted to those as produced by GenePix (Axon Instruments Inc., 2001) (a commercial system commonly used by biologists to analyse images). Results are also compared with three (3) of the aforementioned reconstruction approaches (O'Neill et al., 2003; Fraser et al., 2007a,b) with respect to like-for-like techniques.

The paper is organised in the following manner. First, we formalise the problem area as it pertains to microarray image data and briefly explain the workings of contemporary approaches in Section 2. Section 3 discusses the fundamental idea of our approach with the appropriate steps involved in the analysis highlighted. We then briefly describe the data used throughout the work and evaluate the tests carried out over both synthetic and real-world data in Section 4. Section 5 summarises our findings and renders some observations into possible future directions.

## 2. Existing techniques

Microarray image analysis techniques require knowledge of a given gene's approximate central pixel and the slide's structural layout; therefore, all analysis techniques have similarities (regardless of their specific implementations). For example, a boundary is defined around the gene – thus marking the foreground region – with any pixels outside a given radius taken to be local background. The median value for this background is then subtracted from the foreground and the result is summarised as a $\log_2$ ratio. Bounding mechanisms include partitioning pixels via their histograms (Chen et al., 1997; GSI Lumonics, 2002), edge-based (Perkins, 1980; Ahuja et al., 1980), region growing (Adams and Bischof, 1994; Ahuja et al., 1980) and clustering (McQueen, 1967; Kaufman and Rousseeuw, 1990) functions, a detailed comparison of the more common approaches can be found in (Yang et al., 2002).

The underlying assumption throughout these mechanisms however is that there is little variation within the gene and background regions. Unfortunately, this is not always the case as can be seen in the example regions of Fig. 1a, which depicts a typical test set slide (enhanced to show gene spot locations) with a total of 9216 gene regions on the surface held within an approximate area of $\sim 5000 \times 2000$ pixels. Note in addition that every image in the test set was created on a so-called two-dye microarray system which means the DNA tagging agents used for the two channels are known as Cyanine 5 (Cy5) and Cyanine 3 (Cy3). The close-up sections provide good examples of the low-level signal produced in the image; problems such as partial or missing gene spots, shape inconsistencies, and background variation are clearly evident. In particular, note how the scratch and background illuminations around the genes change significantly. Note that all figures and diagrams are best viewed in colour.

A background identification process is required such that inherent variations between gene and background regions are handled more appropriately. Texture Synthesis represents one possible avenue for such reconstruction approaches as they deal with a similar problem. For example. Efros and Leung (1999) proposed a non-parametric reconstruction technique that is now well established. The underlying principal of the work was to grow an initial seed pixel (located within a region requiring rebuilding) via Markov Random Fields (MRF).

Bertalmio et al. (2001) took an approach inspired by the techniques as used by professional restorers of paintings; i.e. the principle of isotropic diffusion, to achieve their reconstruction. Inspired by this work, Oliveira et al. (2001) attempted to produce similar

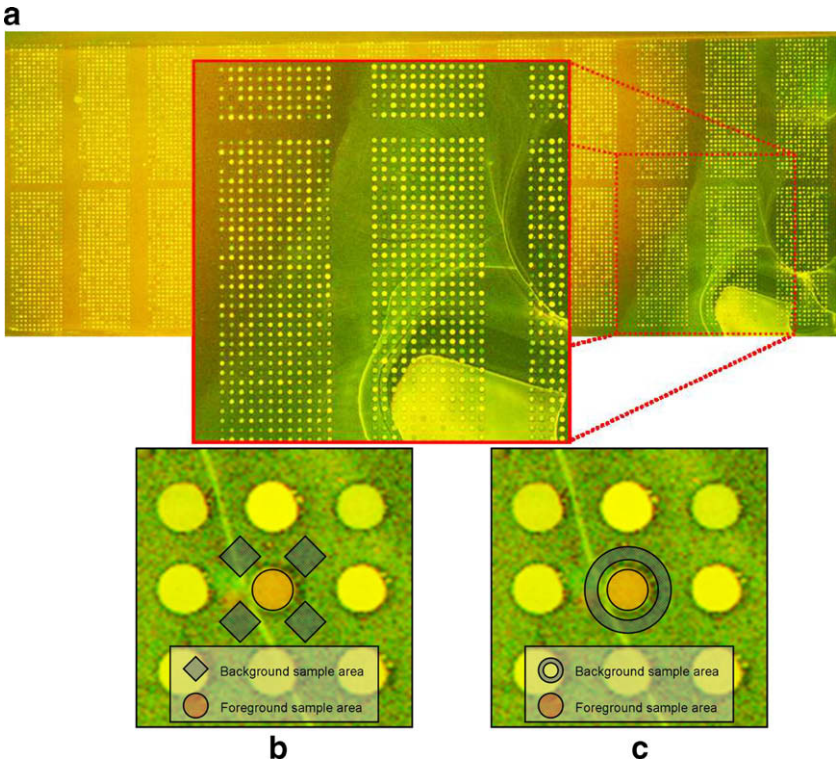


**Fig. 1.** Example images: typical test set slide illustrating structure and noise (a) with sample genes, background locations for GenePix (Axon Instruments Inc., 2001) Valleys (b) and ImaGene (BioDiscovery, 2002) circles (c).

results, albeit much faster. Indeed, Oliveira traded accuracy for speed and succeeded in reducing computation complexity somewhat with their results being of a similar level. Chan et al. (2002) then extended these works along with other related techniques and proposed an elastic curvature model approach that combined amongst others Bertalmio's transportation mechanism with the authors earlier Curvature Driven Diffusion models to produce yet more accurate reconstructions. Sun et al. (2005) proposed an interactive inpainting method with missing strong visual structures by propagating structures (similar to the Bertalmio approach) according to user-specified curves, such an approach does improve on previous methods somewhat, but the interactivity clause would be inappropriate in the context of microarray data (due to the number of objects that need to be rebuilt therein).

In 2003, O'Neill et al. (2003) attempted to address the issue of applying inpainting methodologies into object removal (for microarray imagery specifically) by harnessing ideas from the Efros et al. technique. Specifically, O'Neill et al. remove gene spots from the surface by searching known background regions and selecting pixels most similar to the reconstruction border. By making the new region most similar to given border intensities it is theorised that local background structures 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. This is not to say however that such an approach is void of merit as the resulting surface reconstructions do significantly improve on prior methods (Yang et al. (2002) and Axon Instruments Inc. (2001) for example). In addition, although O'Neill et al. moved away from the inherent aesthetic bounds of the parent techniques to generate more accurate background pixel estimates, it is perhaps unsurprising that they found a blur process to be beneficial for their final results.

An alternative approach to the inpainting mechanism that could be of great benefit in this medical image context is that as held in the Graphing community. Specifically, graph-cutting type processes as used in the general field of image-editing could have potential with respect to such a reconstruction problem. For example, Perez et al. (2003) proposed a Poisson Image Editing technique to compute optimal boundaries between source and target images, while Agarwala et al. (2004) created an interactive Digital Photomontage system that combined parts of a set of photographs into a composite picture. Kwatra et al. (2003) proposed a system that attempted to smooth the edge between different target and source images and is perhaps closest to our current ideology. Other Graph related methods can be seen with Barrett and Cheney (2007), Rother et al. (2004) and Nielsen and Nock (2005), for example. The critical point with respect to these approaches is that they contain an interactive element. Clearly, the interactive element lends itself to an aesthetic resultant very well as such aesthetic improvements are subjective in nature.

Note that graph cutting in the computer science context is derived from graph theory, a study of graphs such that mathematical structures are used to model pair wise relations between objects from a given collection. The so-called graph then refers to that collection of vertices that are related by connected edges. In this way such connected edges could be undirected, meaning there is no distinction between any two vertices associated with each edge. Or alternatively, directed which means the relation between the edges may be directed from one vertex to another. There are in fact several variations, but in essence graphs can be broken up into two groups. The term graph cut then refers to partitioning the vertices such that they themselves fall into two distinct sets. In the context of the paper such set partitioning would be the gene spot and background domains, respectively.

However, microarray images (and indeed medical imagery generally) contain tens of thousands of regions requiring such reconstructions and are therefore either computationally expensive to examine with the aforementioned techniques (not practical) or their interactivity clause renders them to be of limited use (with respect to the number of objects to process). Also, let us not forget that such methods are focused at aesthetic reconstructions. Medical images by their very nature demand reconstruction processes that go beyond such aesthetic considerations.

What is needed is a technique that generates true pixel replacements for an area needing rebuilt rather than the estimates as returned by current approaches. It would also be beneficial if the technique only rebuilt regions that required it (meaning the bounding box around a gene did not include unnecessary pixels as per O'Neill). The following section describes an approach that attempts to address some of the issues related to object removal by using a graph detection mechanism in an automatic and natural way.

## 3. Proposed technique

In this work, we propose graphS-Cut Image Reconstruction (*SCIR*), a technique that removes gene spots from a microarray image surface such that they are indistinguishable from the surrounding regions. Removal of these regions leads to more accurate gene spot intensities. Our previous work in this domain examined the effects of Recalibration (*HIR*) and Fourier Chaining (*CFIR*). Fraser et al. (2007a,b), respectively, techniques. Although *CFIR* dealt with shading and illumination issues more appropriately than *HIR*, *HIR* produced similar results significantly faster. However, both techniques can produce poorer reconstructions in regions dominated by strong artifacts (a saturated gene surrounded with similar level artefact, for example). This work therefore attempts to improve on this issue; while at the same time generating exact pixel values.

### 3.1. Description

The technique is designed to replace gene spot pixels with their most appropriate background neighbour. For example, a scratch on a photograph could be removed such that it is unidentifiable after reconstruction. In the context of this work, a scratch is equivalent to the gene spot region itself. Therefore, removal of this "scratch" should yield the underlying background region in the gene spot area. However, due to the nature of the microarraying process, gene spots can be rendered with different shapes and dimensions, individually and through the channel surfaces.

Therefore, we use a window centred at a target gene (as determined by GenePix) to capture all pixels $p_{x,y}$ within a specified square distance from this centre. Note $x$ and $y$ are the relative coordinates of the pixels in the window centred at pixel $p$. The Window size is calculated directly from an analysis of the underlying image along with resolution meta-data. The window can then be used to determine the appropriate *srcList* (list of gene spot region pixels) and *trgList* (list of sample region pixels) pixel lists, respectively.

The gene spot pixels list can be defined via this windowed region as, $G^p = \Omega^w(g_{x,y})$, with $\Omega^w$ representing pixels falling into the windowed region and $(g_{x,y})$ meaning those pixels falling into the gene spot. The second list $B^p = \Omega^w(\bar{g}_{x,y})$ denotes those pixels within the same window that are not held in gene list $G^p$ (and must therefore be representative of local background pixels).

The graph-cutting process then uses the *srcList* to determine those neighbouring pixels that have the strongest intensity through the surface. While *trgList* is used to determine the weakest neighbouring background intensities, respectively. In the general sense, if we let image $\mathbf{I}$ be a $n \times m$ surface, and if $x,y = 0, 1, \ldots, M-1$; $N-1$ parses said image, a vertical graph cut $Gv$ through the two lists could be defined as

$$Gv = \{Gv_v\}_{x=1}^{n} = \{(x, Gv(x))\}_{x=1}^{n} \quad \forall x.|Gv(x) - Gv(x-1)| \leqslant m, \quad (1)$$

Note that the use of the term *parses said image* means that every pixel within image surface **I** is examined during the reconstruction event.

The vertical graph therefore is an 8-way connected neighbouring set of pixels in the image from top-to-bottom with one pixel per row. Initially, the image is parsed such that cumulative energy for all possible connected pixel sets should be at a minimum for each *x*,*y* pairing through the surface.

In essence a mapping of this nature means that strong foreground pixels are replaced with their appropriate weak background equivalents. Such a replacement policy guarantees that the new foreground surface is not artificially biased to a particular intensity range. Indeed, if anything the new regions will consist of slightly lower intensity than perhaps is necessary meaning therefore a built-in buffer is also applied presently.

### 3.2. Pseudo-code and example

A pseudo-code implementation of the SCIR algorithm can be found in Table 1. For clarity, the implementation is based on processing target window regions, which each contain a distinct set of pixels that are separated into gene spot and background sets.

Initially, the SCIR process creates two distinct lists for a given gene spot location. The source list represents gene spot pixels as demarcated within the square window centred at the gene, while the target list consists of the remaining pixels in the window. Eq. 1 is executed on the lists with the local background taken as the source region and the gene pixels the region to be reconstructed. Essentially, the approach tries to create a chain (or neighbouring set) of pixels through the region that have (in some sense) a maximal/minimal intensity, respectively. This can be thought of as a gradient function that searches for high-contrast (or edge) pixels within the gene spot region and low-contrast pixels within the local background region.

Fig. 2 presents a sample-reconstructed region from Fig. 1a image as processed by the documented techniques.

Note in particular how the SCIR surface looks sharper than that of O'Neill. This is due in part to the O'Neill surface being blurred such that resulting outliers, etc., are suppressed. The SCIR technique on the other hand generates absolute surfaces without this blur stage.

**Table 1**
Graphs-cut reconstruction functions pseudo-code

| | |
|---|---|
| | **Input** |
| | *srcList*: List of gene spot region pixels |
| | *trgList*: List of sample region pixels |
| | **Output** |
| | *outList*: srcList pixels recalibrated into trgList range |
| | **Function graphsCut(srcList,trgList):outList** |
| 1. | For each gene |
| 2. | geneRadius=radius of current gene spot |
| 3. | While geneRadius $\geqslant 0$ |
| 4. | fgEnergy=max pixel surface from srcList members |
| 5. | bgEnergy=min pixel surface from trgList members |
| 6. | fgChain=Parse fgEnergy to determine max-neighbour pixel chain |
| 7. | bgChain=Parse bgEnergy to determine min-neighbour pixel chain |
| 8. | remove fgChain from fgEnergy |
| 9. | remove bgChain from bgEnergy |
| 10. | copy bgChain pixels into srcList locations |
| 11. | geneRadius-=1 |
| 12. | *outList*=srcList |
| 13. | End While |
| 14. | End For |
| | **End Function** |

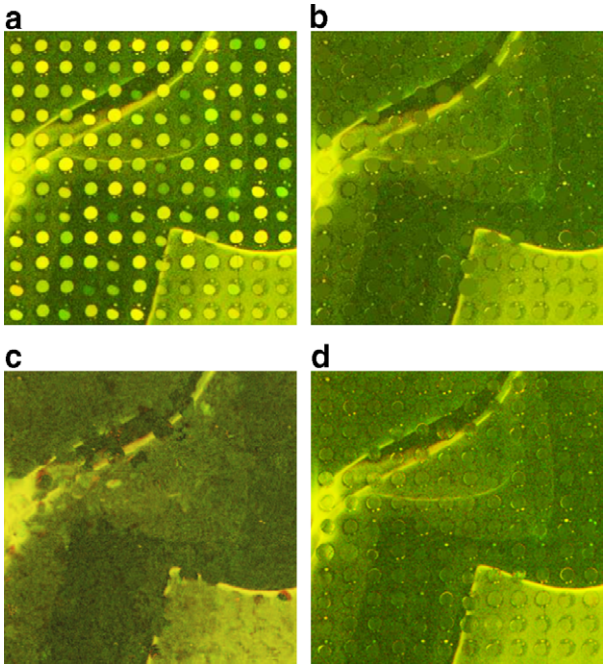


**Fig. 2.** Reconstruction examples: original image (a), reconstructed GenePix (b) O'Neill (c) and SCIR (d) Regions.

## 4. Experiments and results

This section details numerous experiments that were designed to empirically test the performance characteristics of the reconstruction methods. Median expression intensities are utilised in the comparisons as these values are in fact the raw gene expressions (as used in post-analysis, Chen et al. (1997), Eisen et al. (1998), Gasch et al. (2000), Quackenbush (2001), Kepler et al. (2002) and Quackenbush (2002) work, for example). These values help provide clearer understanding of a gene spots repeat set and as such assist with clarification of the reconstruction quality itself.

### 4.1. Data set characteristics

The images used in this paper are derived from the human gen1 clone set (http://www.hgmp.mrc.ac.uk/Research/Microarray/HGMP-RC_Microarrays/description_of_array.jsp) data. These experiments were designed to contrast the effects of two cancer-inhibiting drugs (PolyIC and LPS) over two different cell lines. One cell line represents the control (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. Each slide consists of 24 gene blocks with each block containing 32 columns and 12 rows of gene spots. The gene spots in the first row of each odd-numbered block are known as the Lucidea ScoreCard (Samartzidou et al., 2001; Stability studies, 2003) and consist of a set of 32 pre-defined genes that can be used to test various experiment characteristics. The remaining 11 rows of the odd-numbered blocks contain the human genes themselves. The even-numbered blocks are repeats of their odd-numbered counterparts. This means that each slide has 24 repeats of the 32 ScoreCard genes and 4224 repeats of the human genes, respectively. Note that it is generally accepted that extreme pixel values should be ignored as these values could go beyond the scanning hardware's capabilities.

*4.2. Synthetic data*

The guiding principle of the technique is the feasibility that replacing gene spot pixels with pixels from neighbouring regions will result in a reconstructed area that is indistinguishable from the neighbouring region. Put another way, the gene spots should simply vanish from the surface which means that their new texture has to be very similar to the neighbouring region. Note that

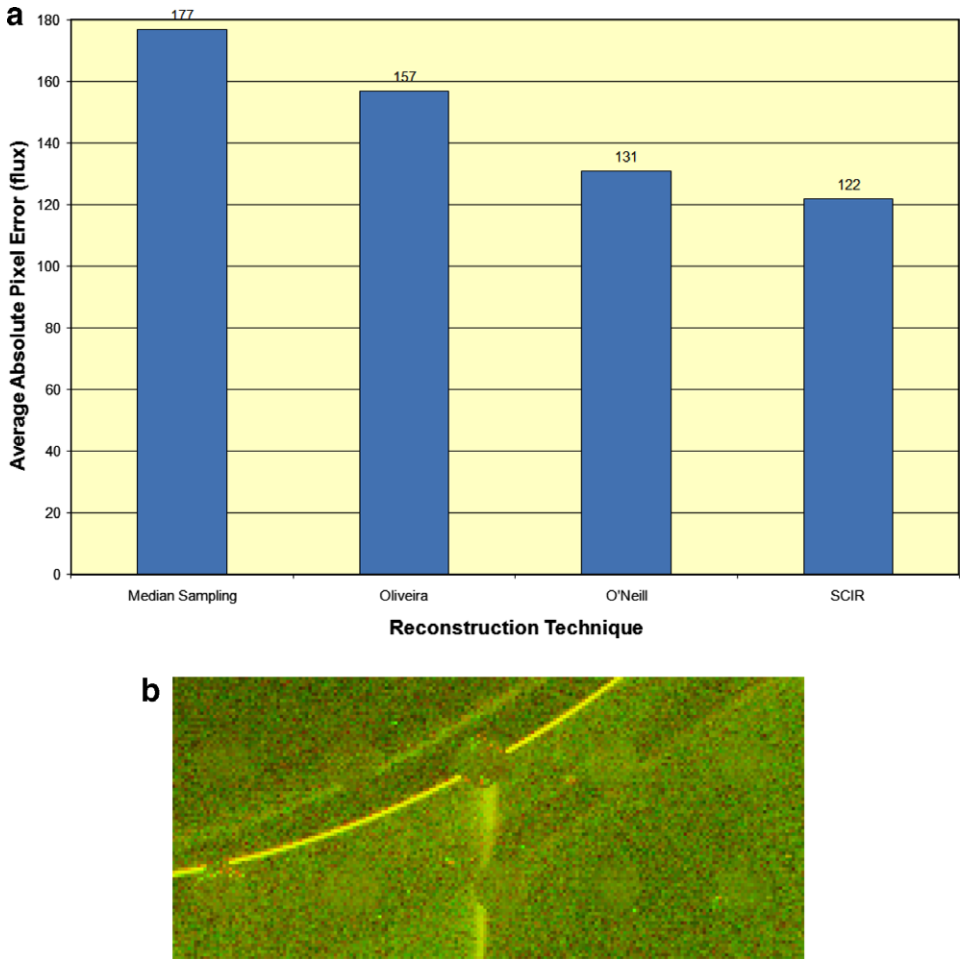


**Fig. 3.** Synthetic gene spots: average absolute pixel error (a) and close-up of a Fig. 1a region with 10 synthetic spots (b).

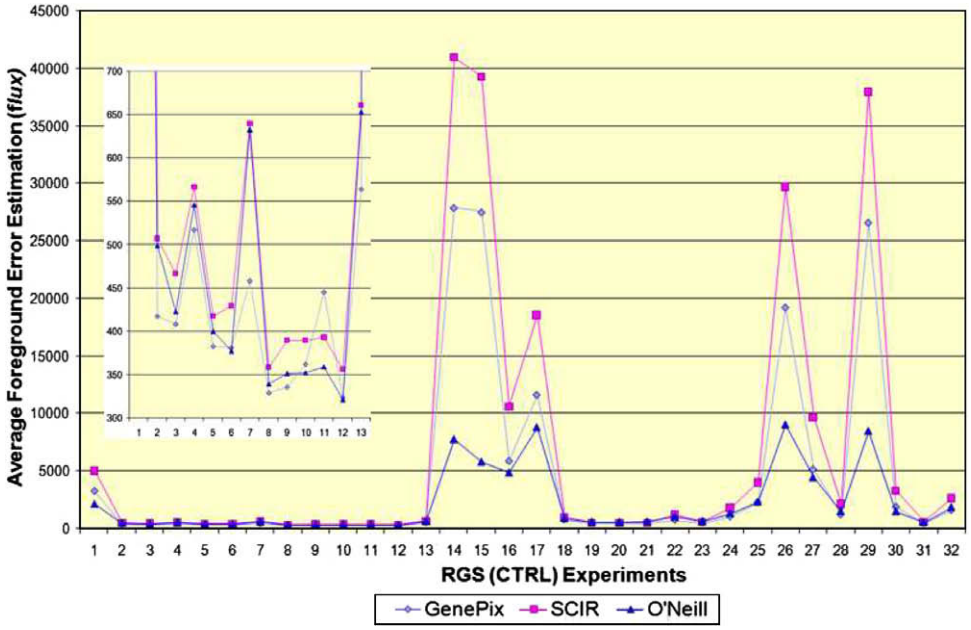


**Fig. 4.** Real gene spots: overview of scorecard gene standard deviations.

regions with strong and sharp intensity differences (an artefact edge, for example) will be harder to "blend" successfully. In order to verify that the principle is at least valid, one would need to rebuild an obscured known region and compare before and after surfaces for accuracy. However, as the gene spot sits above the optimal background surface it is not possible to determine optimal rebuild pixels. In order to validate rebuild feasibility therefore, we use the Synthetic Gene Spot (SGS) creation process as outlined in Fraser et al. (2007b).

The first experiment is focused at answering "how well the SCIR process removes synthetic gene spots from the image"? Sixty-four (64) realistic SGS's were placed into existing background regions of Fig. 1a images Cy5 and Cy3 surfaces. These synthetic genes were then reconstructed with the before and after surfaces compared for similarity. Note that as the artefact region itself could be considered gene spot similar, our reconstruction processes also attempt to build the region such that the artefact pixels are removed. This process yields a ball-park-figure for the potential distillation errors generated by the various background reconstruction techniques. Such potentials as rendered from test imagery can be seen in Fig. 3a, while Fig. 3b highlights a close-up sample region of the aforementioned SGS's.

The graph presents the potential intensity flux error (PIFE) for the reconstruction techniques. On average, the GenePix advocated median sampling approach yields a PIFE of 177 per pixel per SGS region while the other techniques yield decreasing values (our process value of 122 represents a ~30% reduction over GenePix). Such a finding reiterates that downstream analysis when based on GenePix (specifically the BackGround Correction (BGC) stage) estimates directly; produce more erroneous gene expressions than perhaps appreciated. It should be noted the word flux is used in the traditional sense such that it represents the sum of all the pixel

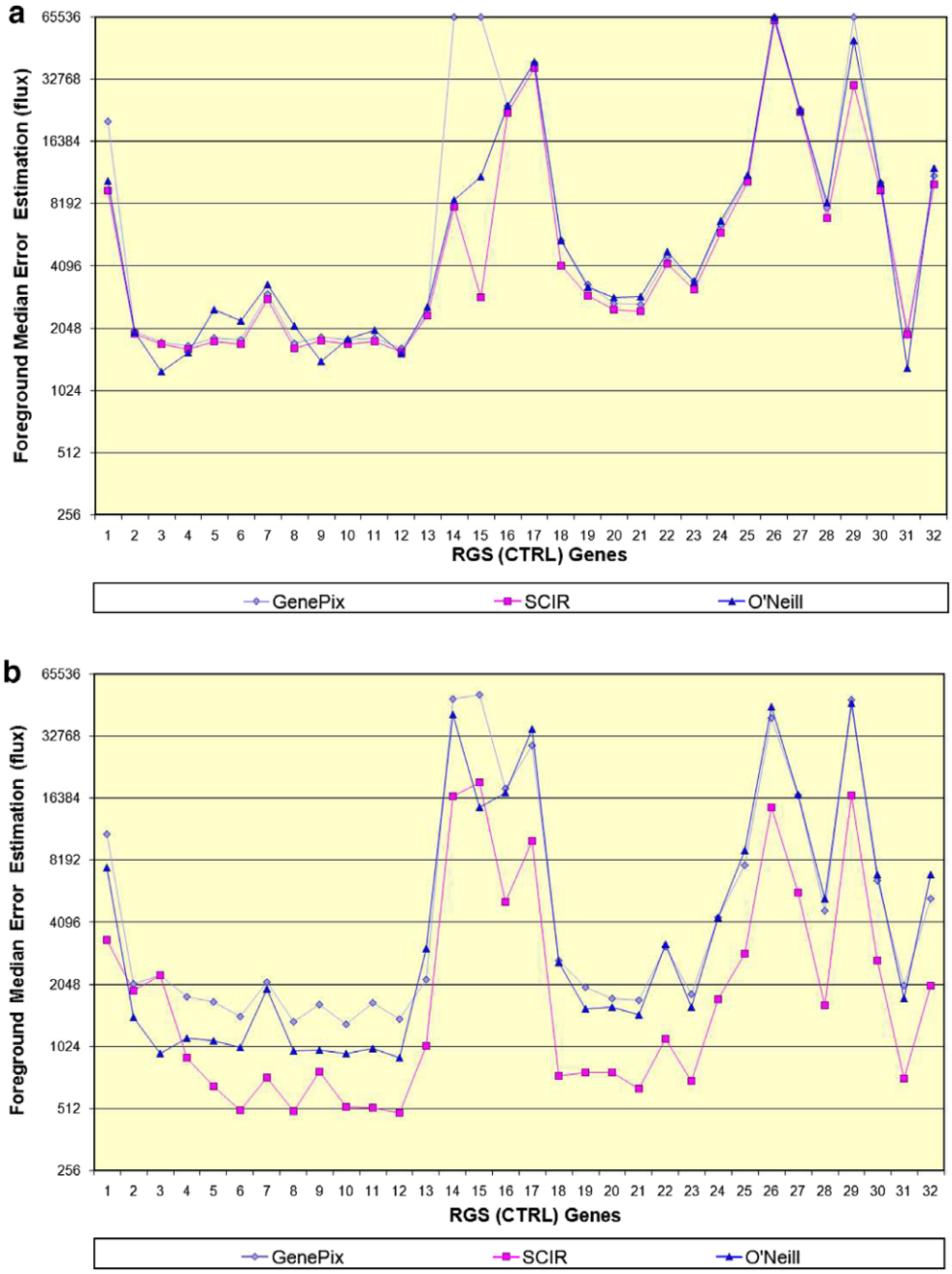


**Fig. 5.** Real gene spots: absolute medians for 32 genes over Fig. 1a (a) and test set (b) regions.

intensities emitted by a gene spot. In other words, flux is a measure of the intensity difference between the original and reconstructed gene spot regions and represents a score value for a reconstruction event.

It should be noted the word flux is used in the traditional sense such that it represents the sum of all the pixel intensities emitted by a gene spot. In other words, flux is a measure of the intensity difference between the original and reconstructed gene spot regions and represents a score value for a reconstruction event.

The panel b surface highlights a sample of the SGS region with a large artefact running through two (2) gene spot regions. Also note we can see that the strong artefact edge has been successfully replaced with appropriate background substitutions. Note however that such strong edges can cause greater challenges within real data as shall be seen.

### 4.3. Real data

With our confidence in the reconstruction techniques abilities enhanced by the synthetic results, the next stage is to understand how such reconstructions fare with real data. In particular, "how badly do strong artefact edges interfere with a reconstruction event"?

Experiment two only uses the ScoreCard control genes for all blocks across the test images. Recall, the composition of the test imagery is such that we have more technical repeats of the control genes than the human ones. Also, the control genes are completely independent of the biological experiment which means ideally they should fluoresce in exactly the same way across the images regardless of environmental conditions (in principle).

Fig. 4 plot presents the tracking of the standard deviations (STD) for the 32 ScoreCard genes over the 24 repeat locations. Note however that due to the way in which O'Neill calculates a given gene spots region, their STD's are somewhat lower than expected. However, the plot still imparts general characteristics for the given reconstruction techniques.

If we disregard saturated gene spots for a moment and examine the close-up section of the plot we see the profile residuals follow each other fairly well. This means the processes do reduce STDs at least in a partial sense.

Critically then, this leads to a need to understand the reconstruction techniques performance characteristics more closely. Specifically, the relationships between expression measurements for all ScoreCard genes in the same slide (Fig. 5a) and across all slides in the test set (Fig. 5b) are compared. Note that it is expected that some intensity differences will appear as the experimental time point's increase as required through the biological processes.

These plots show the bound absolute foreground median values for the multiple image channels for the documented techniques. From Fig. 5a it can be seen that SCIR and O'Neill performed in a similar vein with very little difference amongst them. However, the saturated gene spots – 15 in this case has caused a blip in the profile plot for SCIR. Recall, that by the very nature of a saturated gene spot, the surface is close to a constant value and obviously artificially high. But in this instance the gene in question also has a strong artefact intercepting it. During reconstruction, the constant type value of the gene is not a major challenge to rectify; more problematic is how to deal with the strong intercepting artifacts appropriately. Note that the replacement pixel sets as derived during reconstruction actually do a fair job overall. For this image, the saturated gene spots did not affect the outcome of the final quantification stage greatly.

Whereas Fig. 5a plot represents a specific image surface, it does not render a given reconstruction techniques abilities to deal with a range of image modalities. The panel b plot therefore shows the same information across the entire 47-slide test set. This should allow us to see how exactly a reduction manifests itself onto the final gene metrics. Clearly, the SCIR process has reduced the technical repeats to a greater extent than perceivable from the sample image alone. The respective profile values for the test set are 10,374, 3742 and 9213 flux, respectively.

Clearly, reconstruction of gene spot's does have a positive effect on the final expression results but, not so obvious, are the ramifications that the reconstruction has over the test set. Fig. 6a therefore is a comparison chart showing explicitly the improvement (or not) of a particular reconstruction technique against the original GenePix expressions.

The general banding region of genes 16–17 and partial banding of gene 30 as seen in Fig. 6 are associated with aforementioned saturated (or near background) gene intensities as created by the Axon (Axon Instruments Inc., 2001) scanner hardware and are suggestive of more work needed. The non-banded genes on the other hand are indicative of the individual reconstruction techniques

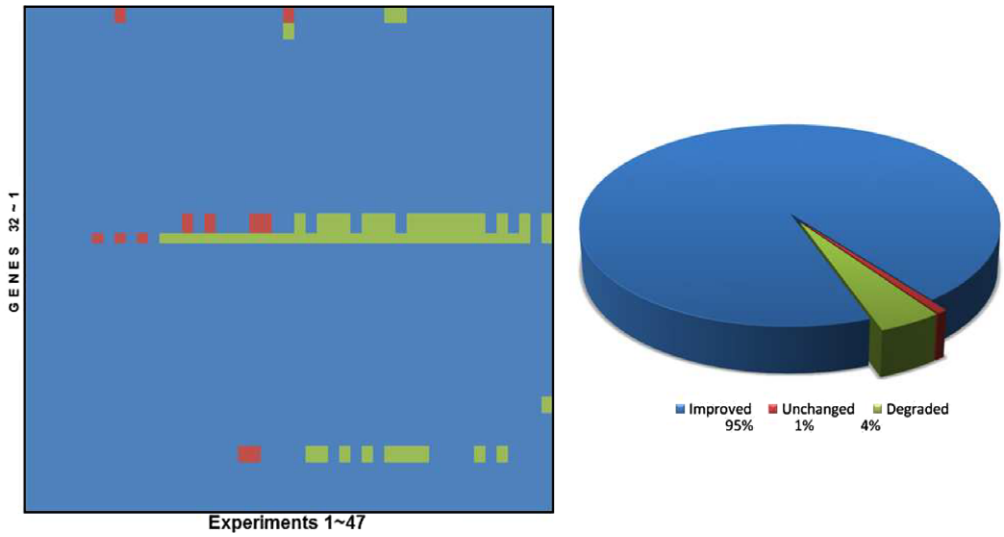


**Fig. 6.** Final results comparison: matrix for test set showing difference in repeat expression fluctuations; the GenePix, SCIR and O'Neill techniques are assigned the colours blue (darkest), red and green (lightest) (~10% difference), respectively. (For interpretation of the references in colour in this figure legend, the reader is referred to the web version of this article.)

being able to account more appropriately for gene intensity replacement.

## 5. Conclusions

The paper looks at the effects of applying current and new texture synthesis inspired reconstruction techniques to real-world microarray image data. In particular, we propose an approach to reconstructing a gene's underlying background by attempting to focus on problematic pixels only. Previously, we took a purely local constraint based approach to the problem and, although the reconstructions were better than those of contemporary approaches, clear areas of improvement existed. Later work relaxes such local constraints and tries to use a level of "localness" to guide the harnessing of an image's global knowledge more closely. Although such a holistic process was shown to be highly effective and a great improvement, again the approach still had weaknesses. The primary weakness is that as related to the nature of the reconstruction task as, in order to rebuild a region appropriately a local area must be sampled in some way. Such a sampling must be larger than the actual region to be rebuilt as local gene spot extended edge discontinuities need to be taken into account. However, not only does such a consideration increase computation complexity but also, one could argue that rebuilding the extended edges causes an artificial increase in flux error which is propagated through to final expressions. Note an extended edge type problem exists in a gene spot's internal regions also.

In this work, we were specifically interested in addressing issues related to extended edge problems of gene spot reconstruction. The proposed approach therefore utilises a graph theory inspired pixel identification mechanism to select those pixels that are most similar to their direct neighbours within a pre-defined region. The highlighted pixel chains are then replaced according to their closest background region representative. The results show that the new method makes a significant improvement in gene expression reduction both directly and when compared with technical repeat variances.

Although in future we would like to be able to broaden our analyses with respect to contrasting our reconstruction processes with other mainstream methods like Spot (Yang et al., 2002), ScanAlyze (Eisen Labs., 2002), ImaGene (BioDiscovery, 2002) and QuantArray (GSI Lumonics, 2002), for instance, it is difficult to acquire appropriate final results as our collaborators use GenePix exclusively. In addition, a critical element of such a critique would require the internal result workings of the mentioned methods.

It is quite probable that a hybrid reconstruction system (able to classify to some extent a gene region) will be of great benefit to this analysis task. Such a hybrid system would use what is deemed to be the most appropriate reconstruction technique for a given gene. As we have now developed several separate reconstruction techniques, shown to be highly effective at their task, it is our belief that such a hybrid system can now be tackled appropriately as several reconstruction specific component parts are in place.

## Acknowledgements

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

## References

Adams, R., Bischof, L., 1994. Seeded region growing. IEEE Trans. Pattern Anal. Mach. Intell. 16, 641–647.

Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, s., Colburn, A., Curless, B., Salesin, D., Cohen, M., 2004. Interactive digital photomontage. ACM Trans. Graph. 23 (3), 294–302.

Ahuja, N., Rosenfeld, A., Haralick, R.M., 1980. Neighbour gray levels as features in pixel classification. Pattern Recognition 12, 251–260.

Axon Instruments Inc., 2001. GenePix Pro Array Analysis Software.

Barrett, W.A., Cheney, A.S., 2007. Object-based image editing. In: Proc. 29th Conf. on Computer Graphics and Interactive Techniques, San Antonio, TX, pp. 777–784.

Bertalmio, M., Bertozzi, A., Sapiro, G., 2001. Navier–stokes, fluid dynamics, and image and video inpainting. In: IEEE Computer Vision and Pattern Recognition.

BioDiscovery Inc., 2002. ImaGene Array Analysis Software.

Chan, T., Kang, S., Shen, J., 2002. Euler's elastica and curvature based inpaintings. J. Appl. Math. 63 (2), 564–592.

Chen, Y., Dougherty, E.R., Bittner, M.L., 1997. Ratio-based decisions and the quantitative analysis of cdna microarray images. J. Biomed. Opt. 2, 364–374.

Coe, B., 2003. You want ketchup with your DNA chips? An overview of expression microarrays. BioTeach Online J. 1 (1), 89–94.

Duyk, G.M., 2002. Sharper tools and simpler methods. Nat. Genet. 32, 474–479.

Efros, A.A., Leung, T.K., 1999. Texture synthesis by non-parametric sampling. In: IEEE Internat. Conf. on Computer Vision, pp. 1033–1038.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. In: Proc. National Academy of Sciences, USA, pp. 14863–14868.

Eisen Labs., 2002. ScanAlyze Array Analysis Software.

Fraser, K., Wang, Z., Li, Y., Kellam, P., Liu, X., 2007a. Improving microarray expressions with recalibration. In: American Institute of Physics, AIP, Utrecht, The Netherlands, pp. 3–16.

Fraser, K., Wang, Z., Li, Y., Kellam, P., Liu, X., 2007b. Noise filtering and microarray reconstruction via chained fouriers. Adv. Intell. Data Anal. VII 7 (1), 308–319.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., Brown, P.O., 2000. Genomic expression program in the response of yeast cells to environmental changes. Mol. Biology Cell 11, 4241–4257.

GSI Lumonics., 2002. QuantArray Analysis Software.

H.G.M. Project. Human gen1 clone set array. Available from: <http://www.hgmp.mrc.ac.uk/Research/Microarray/HGMP-RC_Microarrays/description_of_arrays.jsp>.

Kaufman, L., Rousseeuw, J.P., 1990. Finding groups in data: An introduction to cluster analysis. In: Kaufman, L., Rousseeuw, J.P. (Eds.), Clustering Large Applications (Program CLARA), vol. 3. John Wiley and Sons, New York, pp. 126–163.

Kellam, P., Liu, X., Martin, N., Orengo, C.A., Swift, S., Tucker, A., 2002. A framework for modelling virus gene expression data. J. Intell. Data Anal. 6 (3), 265–280.

Kepler, B.M., Crosby, L., Morgan, T.K., 2002. Normalization and analysis of dna microarray data by self-consistency and local regression. Genome Biology 3 (7) (research0037.1).

Kwatra, V., Schödl, A., Essa, I., Turk, G., Bobick, A., 2003. Graphcut textures: Image and video synthesis using graph cuts. In: Proc. 2003 ACM SIGGRAPH Conf., San Diego, California, pp. 277–286.

Liu, X., Kellam, P., 2003. Mining gene expression data. In: Orengo, C.A., Jones, D.T., Thornton, J.M. (Eds.), Bioinformatics: Genes, first ed., Proteins and Computers, vol. 15. BIOS Scientific Publishers, Oxford, pp. 229–244.

McQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Le Cams, L., Neyman, S. (Eds.), Proc 5th Berkeley Symposium on Mathematics Statistics and Probability, Berkeley, CA, pp. 281–297.

Nielsen, F., Nock, R., 2005. Clickremoval: Interactive pinpoint image object removal, Hilton, Singapore, pp. 315–318.

Oliveira, M.M., Bowen, B., McKenna, R., Chang, Y.S., 2001. Fast digital image inpainting. In: Proc. Visualization, Imaging and Image Processing, Marbella, Spain, pp. 261–266.

O'Neill, P., Magoulas, G.D., Liu, 2003. Improved processing of microarray data using image reconstruction techniques. IEEE Trans. Nanobiosci. 2 (4), 176–183.

Perez, P., Gangnet, M., Blake, A., 2003. Poisson image editing. ACM Trans. Graph. 22 (3), 313–318.

Perkins, W.A., 1980. Area segmentation of images using edge points. IEEE Trans. Pattern Recognition Machine Intell. 2 (1), 8–15.

Petricoin III, E.F., Hackett, J.L., Lesko, L.J., Puri, R.K., Gutman, S.I., Chumakov, K., Woodcock, J., Feigal Jr., D.W., Zoon, C.K., Sistare, D.F., 2002. Medical applications of microarray technologies: A regulatory science perspective. Nat. Genet. 32, 474–479.

Quackenbush, J., 2001. Computational analysis of microarray analysis. Nat. Rev. Genet. 2 (6), 418–427.

Quackenbush, J., 2002. Microarray data normalization and transformation. Nat. Genet. 32, 490–495.

Rother, C., Kolmogorov, V., Blake, A., 2004. Grabcut – Interactive foreground extraction using iterated graph cuts. In: Proc. 2004 ACM SIGGRAPH Conf., Los Angeles, California, pp. 309–314.

Samartzidou, H., Turner, L., Houts, T., Frome, J., Worley, M., Albertsen, H., 2001. Lucidea microarray scorecard: An integrated analysis tool for microarray experiments. Life Sci. News 7 (13), 1–10.

Shalon, D., Davies, R., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 250 (5235), 467–470.

Stability studies of dyes in microarray applications, Technical Report 1.

Sun, J., Yuan, L., Jia, J., Shum, H. 2005. Image completion with structure propagation, in: Proceedings of the 2005 ACM SIGGRAPH conference, pp. 861–868.

Yang, Y.H., Buckley, M.J., Dudoit, S., Speed, T.P., 2002. Comparison of methods for image analysis on cdna microarray data. J. Comput. Graph. Statist. 11, 108–136.