Focused Calibration of Microarray Images

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

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

Abstract. Due to the nature of a microarray experiment, gene expression levels across and through a slides channels can experience up to 10^3 fold change differences in intensity. Such variance in the data is caused by various 'noise' elements, which can influence final expressions. This paper proposes a technique whereby a fast, texture synthesis inspired process is applied to reduce noise artefacts significantly. Akin to a magic eraser, the technique attempts to blend pixels associated with gene spots. Identification of pixels is relatively straightforward, but blending them with appropriate values is non-trivial. Once the replacement values are determined, the image should be a very good approximation original microarray surface. Then, by subtracting this new surface from the original, the gene spot regions would be more accurate. Experiments were designed and carried out with the results compared against a mainstream analysis process "GenePix" and one of the first microarray reconstruction assemblies "O'Neill". Not only was our process shown to be significantly quicker in execution time, it also reduced final expression results while at the same time typically generating less variation within gene spot's.

1 Introduction

With application of complementary Deoxyribonucleic Acid (cDNA) microarray technologies, biologists are able to study all of the genes within an organism to obtain a global view of gene interactions and regulations. This technology has huge potential with respect to obtaining a greater understanding of biological processes. However, the technology is still in the early stages of development with improvements required at key phases of the microarraying process: spotting, hybridisation, and scanning. Because of the multi-layer nature of the process, the digitised microarray image data is often permeated with 'noise', which would propagate through all later phases of analysis. To realise the true potential of such technology it is therefore crucial to obtain high quality image data that reflects the underlying biology as closely as possible.

Although recently there has been much work focused on how to detect and eliminate various artefacts and other such errors from natural image data, progress in this field has been slow. Indeed, improvement of microarray image data itself has received relatively little attention [1–3] when compared with the advances made in post image analysis work (Normalisation [4–6], Modelling [7] and Clustering [8–10] for example).

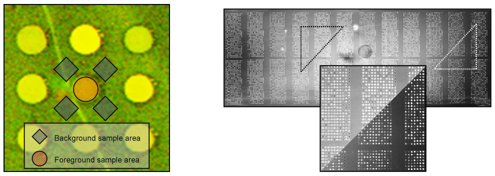
In essence, this paper attempts to rebuild a microarray's background such that the biological experiment regions are removed from the image and we are left with a very accurate background. The new background can then be subtracted from the original image to yield ever more accurate gene spots. The reconstructed expressions as rendered from these new gene spot regions are compared to those as produced by a commonly used commercial system, GenePix [11] and as proposed by O'Neill *et al.* [2].

The paper is organised as follows. We formalise the problem as it pertains to microarray image data and briefly explain the workings of contemporary approaches in the next section. Then, Section three discusses the approach and highlights the steps involved for analysis. In Section four, we start by describing the data used throughout the work and then detail and evaluate the tests carried out for synthetic and real-world data. Section five summarises our findings and defines future directions.

2 Background

Microarray image analysis techniques typically require knowledge of a given gene spot's approximate central pixel as well as the slide's structural layout. A boundary is defined around the gene spot and background pixels, with the medians of these regions taken to be foreground and background intensities respectively. With the subtraction of the background medians from those of the foreground, the result can then be summarised as a $\log_2$ ratio. An example of this can be seen with GenePix which uses a circle of varying diameter, while other techniques use partitioning of pixels by use of a histogram [4, 12] or growing a region from the centre [13, 14] for example. For an extensive comparison of these techniques and more details about their implementation, see Yang *et al.* [1].

To assess the slide background relative to the spot intensities, most methods calculate the difference between the spot intensity and the background noise, often by sampling the gene spots surrounding area. This works well when the assumption holds that there is little local variation between what is behind the gene spot and the surrounding area or in situations where the background is evenly distributed over the slide. However, in many situations this is not the case as highlighted in Fig. 1b, and shown more generally in 1a, where the artefacts can change significantly over the surface.



(a) (b)

Fig. 1. Sample Gene and Background Locations for GenePix Valleys (a) and a Typical Slide Structures Variance (b)

What is needed is a more specific background determination process that can account for the inherent variation between the gene spot and background regions. Indeed the data suffers from an abundance of problems as could be expected from real-world analysis and would include such issues as noise from missing, inconsistent, and outlier elements. No matter how much care is taken when carrying out an experiment, it is certain that there will be errors within the slide. These errors take many forms, but generally, they can be technical, like the random variation in scanning laser intensity, inaccurate measurement of gene expressions and a wide range of artefacts such as hairs or dust on the slide. Alternatively, the errors can be of a more physical nature, inconsistent hybridisation, or washing processes for example.

A possible avenue for such background reconstruction processes is that as represented by the Texture Synthesis community. Efros *et al.* [15] proposed a non-parametric reconstruction technique that is now well established. Efros and Co. worked from the principal of growing an initial seed pixel (located within a region requiring rebuilding) via Markov Random Fields (MRF). Although this works well, the nature of the MRF is such that speed is sacrificed for accuracy. Chan *et al.* [16] greatly extended this work along with other related techniques and proposed a curvature model based approach which is accurate but relatively slow. Bertalmio *et al.* [17] on the other hand took an approach relying on the techniques as used by professional restorers of paintings and therefore worked with the principle of an isotropic diffusion model. Note that all of these techniques were designed to work with natural imagery and therefore provide aesthetic reconstructions. Oliveira *et al.* [18] tried to produce similar results to [17] albeit at a much faster pace. Alas, microarray images on the other hand contain tens of thousands of regions requiring such reconstructions and are therefore computationally expensive to examine with the highlighted techniques. As is typically the case however, something has to give such that significant speed increases can be found and the Oliveira approach is no exception, as we shall see.

With these considerations in mind, O'Neill *et al.* [2] attempted to harness ideas from the highlighted techniques and improve background prediction results while also reducing computation time somewhat. Specifically, O'Neill utilises a simplification of the Efros technique in which gene spots are removed from the surface and recreated by searching known background regions and selecting pixels most similar to the reconstruction border. As the rebuilt region is derived from given border intensities, it is hoped that local background structures will be retained due to an implementation feature of the technique however, only those regions that straddle the centre of an artefact are built most appropriately. The next section describes a technique that attempts to address some of the problem areas as mentioned.

3 Proposed Solution

Calibrated Image Reconstruction (CIR) is a simple technique designed to blend gene spot pixels in a microarray image surface as natural and quickly as possible. Although it is the gene spots we are ultimately interested in, their successful blending at this stage will yield more accurate background regions once this new surface is subtracted from the original. CIR therefore, utilises histogram calibration at the gene level across the two or more microarray slide channels to determine an accurate estimation of the given gene spots background. Due to the nature of the micraorraying process, genes are typically rendered with different shapes and dimensions across channels. Therefore, CIR uses a generic square window Ω^w centred at the gene (as determined by GenePix), to capture all pixels $p_{x,y}$ within a specified distance from this centre.

A gene spot list *srcList* can be defined as *srcList*$=\Omega^w(g_{x,y})$ with Ω^w representing pixels falling into the windowed region only and $(g_{x,y})$ meaning such pixels belong to the gene spot, while list *trgList* denotes background pixels (those not held in *srcList*) and is defined *trgList*$=\Omega^w(\bar{g}_{x,y})$. Although such a blending process is unlikely to create perfect background renditions, it should be possible to improve upon existing template median approach as advocated by packages like GenePix and brute force methods such as O'Neill.

Specifically, the pixel lists as defined are blended by histogram calibration or matching. In simple notation a calibration task can be defined

$$s = T(r) \tag{1}$$

where s and r represent the images pixels with grey levels $\in[0\sim1]$ for the processed and original images over an (x,y) range after some transformation T function. We assume at this point that function T

1. is continuous and r has been normalised to the interval $[0\sim1]$
2. that $T(r)$ is single valued and monotonically increasing in the interval $0 \leq r \leq 1$
3. and $0 \leq T(r) \leq 1 \mapsto 0 \leq r \leq 1$

Point 2 guarantees that an inverse mapping exists and the intensity range is non-invertible, while point 3 holds that the transformed image will exist in the same dynamic range as the original image.

Taking probability density function (PDF) of s, one can transform r into a uniform distribution by using the cumulative density function (CDF) as the transforming function.

As a digital image consists (strictly speaking) of discrete values the probability of grey level r_k occurring within an image can be approximated by the summation

$$r_k = \frac{n_k}{n}, k = 0, 1, 2, ..., L\text{ - }1 \tag{2}$$

where n denotes total number of input pixels, n_k the number of pixels having grey level r_k and L the total number of possible grey levels in the image. This means discrete PDF and CDF functions can be written

$$s_k = T(r_k) = \sum_{j=0}^{k} r(r_j) = \sum_{j=0}^{k} \frac{n_j}{n}, k = 0, 1, 2, ..., L\text{ - }1 \tag{3}$$

$$G(z_k) = \sum_{i=0}^{k} z(z_i) = s_k, k = 0, 1, 2, ..., L - 1 \tag{4}$$

$$z_k = G^{-1}[T(r)], k = 0, 1, 2, ..., L - 1 \tag{5}$$

such that z values will be uniform and independent of r. Briefly, Equ. 3 maps the original intensity levels into their appropriate levels as based on the original histogram. Equ. 4 computes transform function G from the histogram as defined by z and Equ. 5 renders an approximation of the pixel levels for the processed image via the inverse cumulative density function (iCDF) of r. Put another way, calibration is performed by making a frequency count of pixels falling into a given bin location when divided by total pixels (the PDF). The CDF is generated by tracking the accumulation of successive bin contents while the iCDF is created by performing a linear interpolation of the standard CDF function such that samples are evenly spaced in the [0∼1] range. High level pseudo-code for the CIR process is given in Table 1.

Table 1. Pseudo-Code of CIR Function

```
Input
    nGenes: Total number of gene spots to process
    srcList: List of given gene spot pixels
    trgList: List of given genes background pixels
Output
    outList: srcList pixels recalibrated into trgList pixel range

Function HistogramEstimation(srcList,trgList):outList
For each nGenes
    For each srcList pixel in nGenes
        set nBin to number of pixels in trgList
        calculate trgLists histogram characteristics and set histX
        determine srcLists histogram characteristics and set thisX
        set thisCDF to thiscount accumulation
        set histCDF to histCount accumulation
        scale histogram bins of thisX into histX range
        linear interpolate new set of bins from underlying histBin
        set out to new bin content
    End For
End For
End Function
```

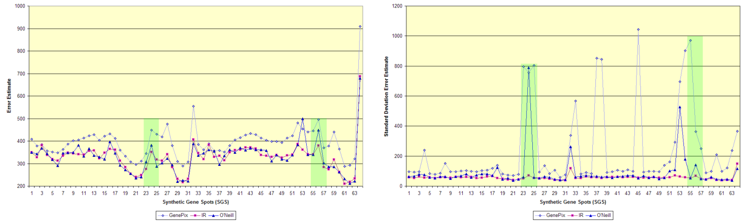
4 Experiments

All of the images used in the paper were derived from two experiments conducted using human gen1 clone set [19] data. The human gen1 experiments were designed to contrast the effects of two cancer-inhibiting drugs over different cell lines. The first, **PolyIC**, is classed as the control or normal cell line while the other **Hela** is known as the treatment line. In total, there are 47 distinct slides (with 9216 genes per slide) generated from a series of several specific time points. The gene spots are divided up into distinct blocks such that their demarcation during later analysis is easier for the biologist to perform. Along with the raw microarray slides, there are also the corresponding GenePix processed results.

In order to gain a clear understanding of a given reconstruction events characteristics over the data, we focus on the absolute error of the median expression intensities as calculated from a given gene spots repeat set. However, such expressions are first calculated over a set of 64 synthetic gene spots (SGS) regions, so that performance differences between the proposed (CIR) and comparison (GenePix and O'Neill) techniques can be gleaned more effectively.

4.1 Synthetic gene spot Experiments

Altogether 64 synthetic genes were created in known background areas of a slide with their locations chosen to encompass a range of possible artefacts and 'more regular' regions as per true genes. Plotting the differences between GenePix calculated backgrounds and the newly reconstructed surfaces yields an understanding of the absolute error per gene (AEPG) for these SGS regions. The Fig. 2a plot presents this AEPG information for the 64 SGS regions.



(a) (b)

Fig. 2. Synthetic Gene Spot Curves: Absolute Median (a) and Standard Deviation of (b) Regions

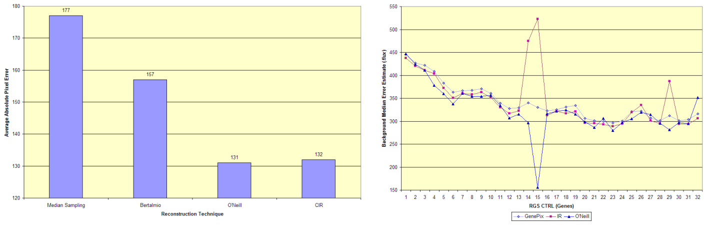
Note the significant drop in background estimation values via our approach over GenePix (due to GenePix relying on median values). The results of our approach and those of O'Neill are highly similar in their intensity estimates, the differences being related to the handling of artefact regions. The overall mean of the individual process curves are 399, 337, and 339 flux for GenePix, CIR and O'Neill respectively.

If we examine this information while relating results to what is known about the original region, we begin to appreciate the fundamental differences between the approaches. Fig. 2b plots standard deviations of the regions rather than median intensities directly and gives an idea of the smoothness of an SGS area. Overall, there is very little difference between the quality of O'Neill and CIR results (as hinted at by 2a). Notable exceptions are the way in which artefacts have been dealt with during execution. Recall, O'Neill attempts to rebuild the local background and hence remove associated artefact intensity from the region. We, on the other hand take the view that such intensities should be merged with local regions as such artefacts are themselves by-products of DNA binding. Due to their very nature, their removal typically renders the intersecting genes unusable; therefore attempting to filter the artefact will typically help to retain information. The average standard deviation residuals are 100, 55.5, and 62.03 flux for the GenePix, CIR and O'Neill processes respectively which means CIR reduced inter-gene spot standard deviations by half with O'Neill just lagging this result. Crucially, the CIR surface is inherently more accurate as the topology has been retained in a more natural state.

Fig. 3a distils the information from the two previous graphs into their differences by calculating the average absolute pixel error in the SGS regions. In effect, this shows that GenePix generates a potential intensity error of 177 flux per pixel for an SGS region, while the other techniques render smaller estimates. This highlights that downstream analysis (of GenePix results) are usually performed on erroneous gene expressions that background reconstruction can help to improve upon.

4.2 Real gene spot Experiments

With such idealised results detailed and our confidence in CIR enhanced, let us examine characteristics with respect to real gene spot regions. Note that only control genes will be used during this analysis, as there are 768 (24 rows × 32 columns) of them. The first experiment examines the median intensity of gene spots in one block across a slide to give a feel for typical variances. Fig. 3b presents plots for two reconstruction techniques when compared to the equivalent GenePix performance curve. The reconstruction processes have made subtle changes to the underlying background regions through the image. Other than spike points in the curves the two result sets fluctuate around each other relatively close. Gene 14 and 15 curve spikes are due to these genes becoming saturated during the digitisation process. Note how the two techniques have processed this saturated data differently. O'Neill essentially resets gene 15s background intensity, which translates into subtracting little from the original gene during background correction (BGC). By comparison, CIR estimates a larger background value and would therefore remove more gene intensity during the BGC stage. Clearly, the techniques have different strengths and weaknesses, but the saturation issue reveals a particular inconsistency with the O'Neill process. Genes 14 and 15 were both saturated to a **very similar** intensity level originally

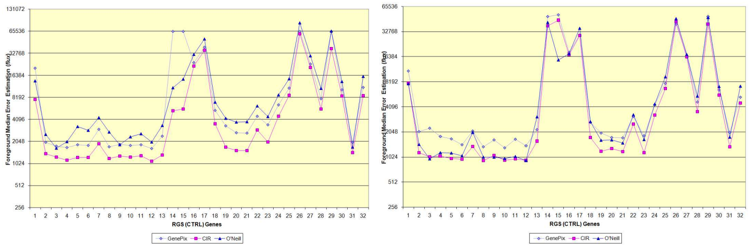


(a) (b)

Fig. 3. Gene Spot Curves: SGS Average Absolute Pixel Error (a), Absolute Medians for 32 genes over block one of a typical slide (b)

and should therefore be processed into a similar result. However, it is clear that O'Neill estimates gene 14s intensity to be much larger than 15 compared with the same genes of CIR, which are more appropriate (their residues alone are three times smaller). Keep in mind however, that as these genes are saturated their topology plateaus and their accuracy should ultimately be questioned regardless. That said CIR has handled these saturated genes more consistently.

Now that we have a feel for general image characteristics, "How do the median characteristics change with respect to a full slide and consequently all slides in the data set?" The second experiment attempts to answer these two points by examining the relationships between expression measurements for all control repeat genes in the one slide initially and then across the entire test set. In all cases the underlying assumption for repeated genes is that they should have highly similar intensity values for a given time point, irrespective of their location on the slide surface. We would expect to see differences as the time point's increase over the biological experiments duration.



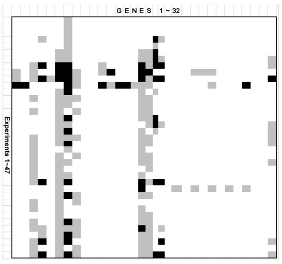
(a) (b)

Fig. 4. Gene Spot Curves: Absolute Medians for 32 genes over typical slide (a) and Entire Test set (b) Regions

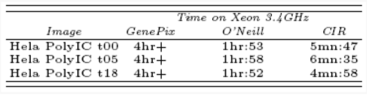
The Fig. 4 plots represent the absolute foreground medians from both channels for the techniques. Generally, for this particular image, CIR outperformed the other methods comfortably. In addition, we would expect gene intensities to drop with removal of background, as peer background estimation methods tend to over specify this intensity.

Fig. 4b presents a cross-sectional average of the 47-slides individual data plots. Although O'Neill outperformed the other reconstruction methods, note the subtleties of the saturated regions. Clearly, there is a large jump in overall intensity between genes 14 and 15 (again) for O'Neill as opposed to other techniques, but, these genes have very similar intensity (the difference is related to the local region for the gene in question) and should therefore be more similar. Overall, the entire test set yields 10374, 8874, and 9213 flux (for GenePix, CIR and O'Neill).

Clearly this fast matching technique is having a positive effect on the resulting median values, but, "How does this reflect over the test set as a whole?" Fig. 5a attempts to clarify this issue by comparing the improvement (or not) of a particular reconstruction to the original GenePix rendered expressions. In addition, as execution time plays a critical role in reconstruction 5b highlights technique timings for three microarray images. The distinct banding occurring in gene regions 3∼8 and 16∼19 of 5a are associated with saturated (or near) intensities as created by the scanner hardware. Note that these particular gene groupings represent saturated control genes (they should be saturated). In some cases however, one technique or the other has reduced the repeat variance more, clearly meaning that CIR has more difficulty in this regard.



(a)



	Time on Xeon 3.4GHz		
Image	*GenePix*	*O'Neill*	*CIR*
Hela PolyIC t00	4hr+	1hr:53	5mn:47
Hela PolyIC t05	4hr+	1hr:58	6mn:35
Hela PolyIC t18	4hr+	1hr:52	4mn:58

(b)

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

5 Conclusions

The paper looked at the effects of applying texture synthesis techniques to real-world microarray image data. It has been shown that the use of such methods in medical image applications can be highly effective. We take a different approach to reconstruction with respect to surface topology and provide highly appropriate results, while crucially, reducing computation time for typical imagery significantly.

The proposed recalibration technique offers a viable process to background reconstruction and has shown that relatively simple formulations can have a dramatic improvement on background regions, although there is stillroom for improvement. One thing that became quite apparent is the possible advantages of a hybrid process. All gene spots in the images were processed by the given techniques exclusively. However, there will undoubtedly be gene spots that would benefit more from different assimilation methods. Therefore, determining a genes classification could lead to the ability to select the most appropriate reconstruction technique for a given gene. For example, CIR may be most appropriate in a first pass capacity with time intensive techniques better utilised in highly inconsistent regions that fail to satisfy some quality mechanism. Another interesting point is that the current implementation also applies the reconstruction across the whole region equally. Although this clearly works well, in gene spots that cross an artefact rich area, such global consistency may be inappropriate.

Acknowledgment

This work is in part supported by EPSRC grant (EP/C524586/1)

References

1. 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** (2002) 108–136
2. O'Neill, P., Magoulas, G.D., Liu, X.: Improved processing of microarray data using image reconstruction techniques. IEEE Transactions on Nanobioscience **2**(4) (2003)
3. Bengtsson, A., Bengtsson, H.: Microarray image analysis: background estimation using quantile and morphological filters. BMC Bioinformatics **7**(96) (2006)
4. Chen, Y., Dougherty, E.R., Bittner, M.L.: Ratio-based decisions and the quantitative analysis of cdna microarray images. Journal of Biomedical Optics **2** (1997) 364–374
5. 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) research0037.1
6. Quackenbush, J.: Microarray data normalization and transformation. Nature Genetics **32** (2002) 490–495
7. 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) (2002) 265–280
8. 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) 14863–14868
9. 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** (2000) 4241–4257
10. Quackenbush, J.: Computational analysis of microarray analysis. Nature Reviews Genetics **2**(6) (June 2001) 418–427
11. Anonymous: GenePix Pro Array analysis software. Axon Instruments Inc.
12. Anonymous: QuantArray analysis software. GSI Lumonics.
13. Adams, R., Bischof, L.: Seeded region growing. IEEE Transactions on Pattern Analysis and Machine Intelligence **16** (1994) 641–647
14. Ahuja, N., Rosenfeld, A., Haralick, R.M.: Neighbour gray levels as features in pixel classification. Pattern Recognition **12** (1980) 251–260
15. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: IEEE International Conference on Computer Vision. (1999) 1033–1038
16. Chan, T., Kang, S., Shen, J.: Euler's elastica and curvature based inpaintings. Journal of Applied Mathematics **63**(2) (2002) 564–592
17. Bertalmio, M., Bertozzi, A., Sapiro, G.: Navier-stokes, fluid dynamics, and image and video inpainting. In: IEEE Computer Vision and Pattern Recognition. (December 2001)
18. 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 (September 2001) 261–266
19. Project., H.G.M.: Human gen1 clone set array