Reducing the Variability in cDNA Microarray Image Processing by Bayesian Inference

Neil D. Lawrence†, Marta Milo†, Mahesan Niranjan†,
Penny Rashbass‡and Stephan Soullier‡
†Department of Computer Science,
Regent Court, 211 Portobello Road, Sheffield, S1 4DP, U.K.
‡Centre for Developmental Genetics,
University of Sheffield School of Medicine and Biomedical Science,
Firth Court, Western Bank, Sheffield, S10 2TN, U.K.

15th August 2003

Abstract

Motivation: Gene expression levels are obtained from microarray experiments through the extraction of pixel intensities from a scanned image of the slide. It is widely acknowledged that variabilities can occur in expression levels extracted from the same images by different users with the same software packages. These inconsistencies arise due to differences in the refinement of the placement of the microarray 'grids'. We introduce a novel automated approach to the refinement of grid placements that is based upon the use of Bayesian inference for determining the size, shape and positioning of the microarray 'spots', capturing uncertainty that can be passed to downstream analysis.

Results: Our experiments demonstrate that variability between users can be significantly reduced using the approach. The automated nature of the approach also saves hours of researchers' time normally spent in refining the grid placement.

Availability: A MATLAB implementation of the algorithm and tiff images of the slides used in our experiments, as well as the code necessary to recreate them are available for non-commercial use from `http://www.dcs.shef.ac.uk/~neil/VIS`.

Contact: `neil@dcs.shef.ac.uk`

1 Introduction

The basis of DNA microarray technology is the construction of high density arrays of spots on glass slides that are hybridised with fluorescently-labelled mRNA-derived targets [Eisen & Brown, 1999, Schena, 2000]. The pattern of hybridisation to the elements is visualised by fluorescent imaging and the raw microarray data images are subsequently transformed into gene expression matrices [Brazma & Vilo, 2000]. Several different software packages are available to determine the position of each element by fitting grids over the image to extract the level of the sample's expression [Yang *et al.*, 2000].

The recent increase in microarray experiments has led the academic community to an international effort to adopt standards for DNA-array experiment annotations and data representation[1]. The Minimum Information About a Microarray Experiment (MIAME) [Brazma *et al.*, 2001] standard, for example, aims to establish gene expression data repositories that will allow consistent comparison of the data collected from different sources. It is therefore imperative that the processing of microarray image data by different researchers leads to consistent results. Our experience and that of other users of microarray technology suggests this is not the case; we have found that even when two researchers use the same software package to process the same images, results are not consistent. The inconsistencies we focus on are not associated with normalisation of the slides, but with differences arising from misplacement of the circles which specify the spots.

In the following experiment, labelled cDNA was prepared from mRNA obtained from optic primordia dissected from E11.5 wild-type and *aphakia*

[1]Microarray Gene Expression Database Group coordinated by the EBI, `http://www.mged.org/`.

mouse embryos [Varnum & Stevens, 1968]. A microarray slide was hybridised with 2 μg of Cy3-labelled cDNA and $2\mu g$ of Cy5-labelled cDNA. Scans were obtained at 10 pixels$/\mu m$ resolution using a Genomic Solutions GeneTac LSIV scanner. Figure 1 compares log ratios of expression levels obtained through analysis by two different users of this slide using the ScanAlyze[2] software package. The slide contains 6144 PCR products printed in duplicate (12,288 spots) from plates 1-16 of the National Institute of Aging $15K$ cDNA Clone Set (NIA $15K$)[Tanaka *et al.*, 2000] obtained from the MRC UK Human Genome Mapping Project. The ScanAlyze package relies on information provided by the user, including the layout of the grids, size and shape[3] of each spot. Each user provided the software with the same information (grid layout), but then placed the grids independently and subsequently refined the spot size and position to allow for discrepancies in the array manufacture. Figure 1 illustrates that image analysis by different researchers leads to large discrepancies in the gene expression level values obtained from a single array. Note that these differences are *not* confined to a single source error, *i.e.* they arise from discrepancies in both the specification of centers and widths, and they are not limited to low intensity spots which are potentially removable. Figure 1 shows that there is up to a two-fold variation in the ratios purely arising from grid placement errors. Interestingly, when performing the same experiment across different slides other researchers [Hughes *et al.*, 2000] have discovered variations of a similar magnitude, but assumed that they resulted from 'biological noise'. Whilst slide image data and grid placement information is not publicly available for this work, it seems plausible that a portion of this variation was due to grid placement error.

In this paper we present a novel way of processing microarray images based upon Bayesian inference. The method reduces variation between different researchers' grid placements and, since it is automated, saves researchers' time[4]. In Section 2, the concept of Bayesian inference, upon which the method is based, the likelihood function, and the prior distribution will be introduced. Section 2.3 describes the posterior distribution that can then be used to extract expression levels. Because the resulting posterior density does not have a simple functional form, it cannot be computed exactly and in Appendix A we discuss approximations. Finally in Section 3 we repeat the analysis of the slide used in Figure 1 with the automated approach demonstrating greater consistency, comparisons with an alternative, automated spot finder and using an additional slide are also included.

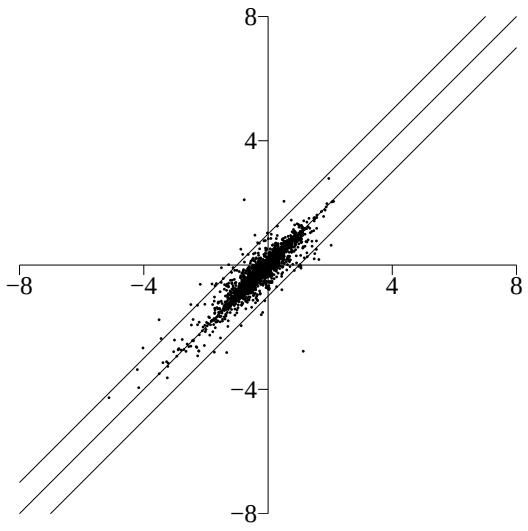

Figure 1: The $\log_2$ ratios extracted from the same microarray image by two different researchers using ScanAlyze software with the same grid parameters. Ideally points should lie along the central line. The outer lines lie along points for which there is a two-fold variation in the ratios. Whilst there were 6,144 PCR products on the slide, we only plot points for those spots which had hybridised and were not contaminated by dust leading to 4,900 separate points.

2 Methods

One possible cause of the inconsistencies we have observed is *uncertainty*. The correct location and shape of the spot is not a deterministic concept. When such uncertainties exist a popular way of making progress is through Bayesian Inference. Bayesian inference [Bernardo *et al.*, 1980, Cox, 1946] is a mathematical way of formulating

[2] ScanAlyze software is free for academic use and downloadable from `http://rana.lbl.gov/EisenSoftware.htm`.

[3] In this paper, we shall generally use shape to refer to the x-radius and y-radius of an ellipsoid. However it should be noted that some packages allow spots to be of arbitrary shape.

[4] Fine manual placement of the grids, such as that undertaken for the experiment shown in Figure 1, may take 2-6 hours depending on factors such as the quality of the slide. Our approach, in common with others, requires the initial rough grid layout to be specified manually, for a previously unused microarray slide this may take about 15 minutes. For the slide we considered our automated refinement algorithm then takes about half an hour to execute. The run, however, may take place unattended.

the process of extracting useful information from data. For the problem we are considering, the information of interest is the position and shape of the spots on our microarray slide. Before observing the microarray we hold a 'belief' about how the spots are laid out, this belief might involve an assumption that the spots are laid out in grids. We may know from the manufacturer that these grids contain, for example, sixteen rows and sixteen columns. We may also believe that the spots are circular with a particular radius. Finally we may be given some information about the spacing between the spots. Once we view the image, we soon realise that whilst our prior belief was roughly true, many of the spots do not lie exactly on the grid and they also, due to inconsistencies in the manufacture of the slide, may vary in shape. Therefore, having observed the image, we can update our belief about the size and positioning of the microarray spots. Bayesian inference provides us with a mathematical technique for this updating of beliefs in the light of observed data.

As an example of the inference process we will first consider the positioning of the spot centres. Our initial belief is that the centres are laid out on a grid. This can be combined with an estimate of the uncertainty of these positions and incorporated in a *prior probability distribution*. This distribution, $p(\mathbf{c}|G)$, gives the probability of the centre positions, $\mathbf{c}=\begin{pmatrix} c_x \\ c_y \end{pmatrix}$, given the parameters of the grid, G. What we are really interested in obtaining is the *posterior distribution*, $p(\mathbf{c}|G,\mathcal{I})$. It is our updated belief about the position of the spot centres given the grid parameters *and* the intensity of the pixels in the image. The relationship between the prior and the posterior is given by *Bayes' rule*,

$$p(\mathbf{c}|G,\mathcal{I}) = \frac{p(\mathcal{I}|\mathbf{c})\,p(\mathbf{c}|G)}{p(\mathcal{I}|G)}, \qquad (1)$$

where we have also introduced the *likelihood function*, $p(\mathcal{I}|\mathbf{c})$, which represents a probabilistic model of the observed image given that there is a spot centre at $\mathbf{c}$. The normalisation constant, $p(\mathcal{I}|G)$, may be computed in terms of the likelihood function and the prior distribution by the integral

$$p(\mathcal{I}|G) = \int p(\mathcal{I}|\mathbf{c})\,p(\mathbf{c}|G)\,d\mathbf{c}. \qquad (2)$$

For this paper we assume that the spots are made up of axis aligned oval shapes, so as well as their centres we must also consider a radius in the x-direction and a radius in the y-direction. We pair these as a vector and denote it as $\mathbf{r} = \begin{pmatrix} r_x \\ r_y \end{pmatrix}$. More complex assumptions are of course possible (*e.g.,* involving rotated ellipses or a more general class of shapes) and extending the algorithms proposed here to such shapes is straightforward. However, it should be borne in mind that, in implementation, evaluation of the likelihood function dominates CPU-time requirements, so any increase in complexity will be accompanied by a corresponding increase in CPU-time. Our use of axis constrained ellipses mimics the approach of the ScanAlyze software which has become a *de facto* standard.

To undertake the inference process we must

- Specify our prior distribution, $p(\mathbf{c},\mathbf{r}|G)$.
- Specify a likelihood function $p(\mathcal{I}|\mathbf{c},\mathbf{r})$.
- Obtain our posterior distribution, $p(\mathbf{c},\mathbf{r}|G,\mathcal{I})$, through taking the product of these two distributions and normalising it:

  $$p(\mathbf{c},\mathbf{r}|\mathcal{I}) = \frac{p(\mathcal{I}|\mathbf{c},\mathbf{r})\,p(\mathbf{c},\mathbf{r})}{p(\mathcal{I})}, \qquad (3)$$

  where we have dropped the dependence of the grid parameters, G, to avoid cluttering the notation.
- Use the posterior for estimating gene expression levels, and their variances, from the image.

There is an unfortunate complication in the simple picture of Bayesian inference that we have portrayed thus far; the normalisations of the type shown in eqn (2) will very often be intractable. Under these circumstances we will not be able to obtain an analytic representation of the posterior distribution and we will be forced to turn to approximations to make progress. We will deal with this problem in Appendix A, for now we disregard such considerations and turn to our choice of likelihood function and prior.

2.1 Likelihood Function

The likelihood function represents what we expect to see in the image given a particular centre and radius of oval, we would expect to find the pixels within the oval are from the spot and those outside the oval are from the background. For the sake of simplicity it is common, though rarely correct, to assume that the probability of each pixel coming from the background is independent of the origin of other, perhaps neighbouring, pixels. In this 'first stab' at Bayesian inference for processing of

microarray slides, we will use a fairly simple likelihood function that models this scenario. Similar likelihood models are widely used in computer vision, in particular in video based tracking; see, *e.g.* [Isard & MacCormick, 2001].

We consider a 'box' of pixels with associated intensities, $\mathcal{I}_{\text{box}}$, surrounding the proposed position of our microarray spot. The positioning and size of the oval we place in this box constitutes a hypothesis, $\mathcal{H}$, about the location and shape of the spot (see Figure 2). Consider the intensity of an individual pixel from the box, I_i. If the likelihood model is based on individual likelihoods for each pixel, we can model the probability of the intensity given that the pixel is part of a spot, $p(I_i|S)$, and the probability of the intensity given that a pixel is not part of a spot, $p(I_i| \sim S)$. A particular hypothesis for our oval's position and size partitions the pixels in the box into a set which comes from the spot, S, and a set that is not from the spot, $\sim S$ as shown in Figure 2. The likelihood function, due to our independence assumption, may then be written as the product of each individual pixel's likelihood

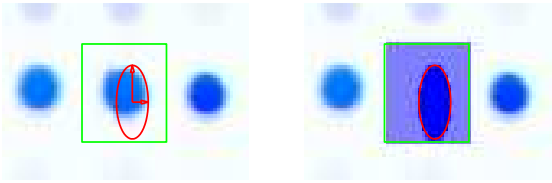

Figure 2: *Left:* A hypothesised oval with its radii shown as arrows over a spot from a microarray image. The box shows pixels which belong to the set $\mathcal{I}_{\text{box}}$. *Right:* Pixels which are allocated to the set S are shaded dark, those allocated to the set $\sim S$ are shaded lighter.

$$p(\mathcal{I}_{\text{box}}|\mathbf{c},\mathbf{r}) = \prod_{i\in S} p(I_i|S) \prod_{i\in\sim S} p(I_i| \sim S).$$

To complete our likelihood function, we are required to estimate the individual pixels' likelihoods $p(I_i|S)$ and $p(I_i| \sim S)$. We built histograms representing the intensities of the foreground and background pixels for both the red and green channels. These histograms[5] were based on the initial rough grid localisation. The likelihood of each pixel was then taken to be the product of the likelihoods from the red and green histograms. More complex likelihood functions may also be straightforwardly implemented.

[5]The histograms are generated separately for each sub-grid in the microarray slide to account for changes in background levels across the slide.

2.2 Prior Distribution

The remaining part of the model is the prior distribution, our belief about the centre, $\mathbf{c}$, and the radii, $\mathbf{r}$, of the oval. The user specifies a grid that represented the approximate centres of the spots, $\mathbf{m_c}$, and radius values representing the expected size of the spots, $\mathbf{m_r}$. A simple prior which can incorporate this information is a Gaussian distribution[6]. For simplicity we consider covariance matrices which lead to circular Gaussian priors so for the prior governing the radii we take $\mathbf{\Sigma_r} = \alpha_{\mathbf{r}}^{-1}$, and for that governing the centres we have $\mathbf{\Sigma_c} = \alpha_{\mathbf{c}}^{-1}$, where is the 2×2 identity matrix and $\alpha_{\mathbf{c}}$ is a scalar. Each α is then a measure of the 'precision': the higher its value, the more confident we are about our prior belief. Finally, we assume independence in the prior between the radii and the centres, giving,

$$p(\mathbf{c},\mathbf{r}) = p(\mathbf{c})\, p(\mathbf{r}),$$

where

$$p(\mathbf{c}) = \mathcal{N}\left(\mathbf{c}|\mathbf{m_c},\alpha_{\mathbf{c}}^{-1}\right),\ p(\mathbf{r}) = \mathcal{N}\left(\mathbf{r}|\mathbf{m_r},\alpha_{\mathbf{r}}^{-1}\right)$$

and we have dropped the dependence of G to avoid cluttering the notation. Figure 3 shows some samples from this type of prior distribution. Reasonable parameters may vary from slide to slide[7], we found that standard deviations of the order of two pixels worked well across a range of slides. Our definition of the prior and the likelihood give us the elements we need to compute the posterior.

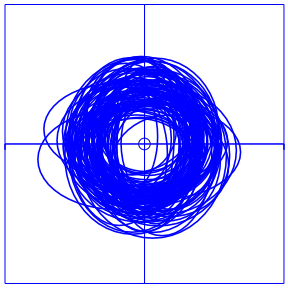

Figure 3: The ovals in this figure represent samples from the prior distribution, $p(\mathbf{c},\mathbf{r})$. For this example $\mathbf{m_c} = [0\ 0]^{\mathrm{T}}$, $\alpha_{\mathbf{c}} = 1$, $\mathbf{m_r} = [5\ 5]^{\mathrm{T}}$ and $\alpha_{\mathbf{r}} = 1$.

[6]A K-dimensional Gaussian distribution is specified by its mean, $\mathbf{m}$, and its covariance matrix, $\mathbf{\Sigma}$: $\mathcal{N}(\mathbf{x}|\mathbf{m},\mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{K}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right)$.

[7]For large standard deviations the algorithm becomes less sensitive to the prior's means, for small standard deviations the algorithm becomes more dependent on the prior's means.

2.3 Posterior Distribution

In microarray image processing the objective is to obtain the level of gene expression from each spot image. The expression level for a particular spot, E, is a function of the pixel intensities, the spot position and shape,

$$E = f(I, \mathbf{c}, \mathbf{r}). \quad (4)$$

We are given the image intensities, I, so if $\mathbf{c}$ and $\mathbf{r}$ are also precisely specified then the expression level may be computed. In the case of Bayesian inference though, $\mathbf{c}$ and $\mathbf{r}$ are not specified exactly, our knowledge about them is encapsulated in a posterior distribution, $p(\mathbf{c}, \mathbf{r}|G, I)$. Therefore, instead of simply computing the function in eqn (4), we now compute its expectation under the posterior distribution,

$$\langle E \rangle_{p(\mathbf{c},\mathbf{r}|I)} = \int f(I, \mathbf{c}, \mathbf{r})\, p(\mathbf{c}, \mathbf{r}|G, I)\, \partial\mathbf{c}\partial\mathbf{r}. \quad (5)$$

It is possible to compute other quantities of interest such as a 'error bars' which may be obtained through the expression level's variance $\mathrm{var}(E) = \langle E^2 \rangle - \langle E \rangle^2$, where we have dropped the subscript on the expectations, which are assumed to be taken (as in eqn (5)) over the posterior distribution.

The posterior distribution is, in theory, now fully specified; in practice though we find that it is difficult to recover. As noted before the principal obstacle is the normalisation which makes up the denominator of eqn (3). For the prior and likelihood given in the preceding sections it is not obvious how the relevant integral may be computed. This problem is common to many implementations of Bayesian inference; to resolve it we must look to approximations. The approximations we used are detailed in Appendix A and we refer to the algorithm which uses them as the variational importance sampler (VIS). They allow us to approximate expectations under the true posterior.

Note that the use of a probabilistic model within the VIS algorithm allows us to, in a very straightforward manner, compute the probability that there is no spot at the location. We simply need to evaluate the likelihood of no spot being present and compare it with the likelihood of our proposed spot locations. Accordingly we may choose to flag the location as not containing a spot. These flags were used to identify locations where no spots were present for all the results above. For non-probabilistic approaches decision making of this type generally relies on arbitrarily chosen thresholds.

3 Results and Discussion

To demonstrate the utility of the variational importance sampler (VIS) algorithm the grid layouts as identified by the two different researchers using ScanAlyze software[8] for Figure 1, were used to initialise our algorithm. In Figure 4 we compare the expectations of the $\log_2$ ratios as computed by our algorithm. The computation of the $\log_2$ ratios was as suggested in the manual of the ScanAlyze software package, and are therefore directly comparable to those presented in Figure 1. Spots that were identified as not being present, corrupted by artifacts and those with an associated variance of greater than $\frac{1}{16}$ are not presented.

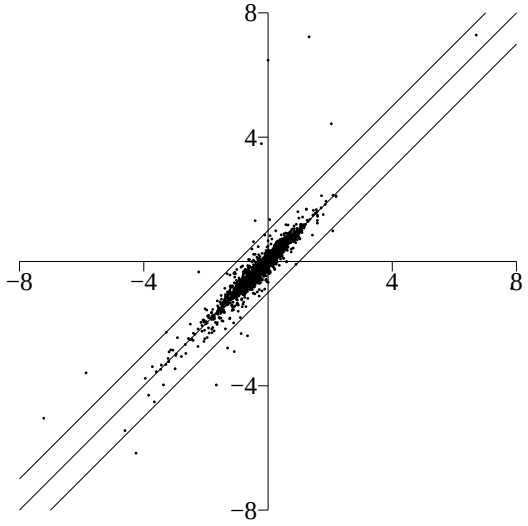

Figure 4: The $\log_2$ ratios extracted from the same microarray image using our Bayesian inference algorithm. The initial grid positions were given by two different researchers, initialising the algorithm in different places. The x-axis shows the $\log_2$ ratio given by initialising with Researcher 1's grid, and the y-axis shows the $\log_2$ ratio obtained through initialising with Researcher 2's grid.

Quantitatively the mean squared error between the points was calculated as 6.0×10^{-2} which compares to the mean squared error as calculated for the manual grids of 8.23×10^{-2}. In other words our automated approach leads to a 37% increase in consistency. The approach also affords additional information in the form of variance estimates as-

[8] The positions as identified by the two researchers were used for initialisation here to demonstrate that from inconsistent starting points we converge on a more consistent solution. Of course in practice, one would initialise the algorithm with only a rough placement of the grid.

sociated with the $\log_2$ ratios. In Figure 5 we plot the base 10 logarithm of these variance estimates against $\log_{10}$ 'intensity', where intensity has been defined [Yang *et al.*, 2001] as the product of the red and green channel. The ratio between the two channels is more sensitive to small changes in the intensity of the channels when the intensity is low. As a result, we might expect that the variances associated with the $\log_2$ ratios to be high when the intensity is low. This hypothesis is confirmed by the results depicted in Figure 5 which, while indicating some spread for a given intensity, generally exhibit a downward trend as intensity is increased.

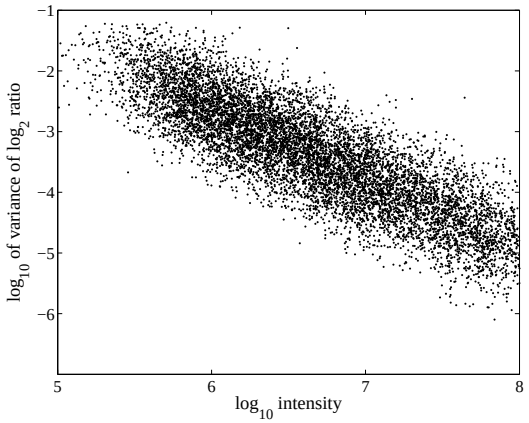

Figure 5: Plot of the $\log_{10}$ intensity, the against the $\log_{10}$ of the variance of the $\log_2$ ratio. As expected the variances tend to reduce as intensity increases.

Estimates of variance can be made further use of in downstream analysis, for example, they may be used to weight the $\log_2$ ratios for each of the replicates when averaging them to obtain a representative value for a single gene [Lawrence *et al.*, 2003].

In the next section we compare the $\log_2$ ratios extracted by the VIS with a set extracted by another automated software package known as 'Spot[9]'.

3.1 Brief Comparison with Spot

One alternative approach to extracting $\log_2$ ratios involves allowing each spot to take on any general shape which may be determined using 'seeded region growing'. This is the approach taken by the Spot software package [Yang *et al.*, 2000]. In this package the location of each spot is specified with a given pixel. Given this 'seed' pixel, the algorithm is deterministic in that it will allocate the same pixels as foreground and background every time it is run. However, for different seeds the regions will also differ.

To assess the quality of the algorithms, we took advantage of the replicates which were present on the slide. Each replicate is attempting to measure the same gene's expression ratio. Their measurements differ due to various noise processes as well as the inconsistencies we identified above. By assessing the mean squared error between these values we are able to assess the consistency of the different algorithms across different measurements associated with the same gene. As far as the practical utility of the methods, this is an important performance measure. Unfortunately it does not allow us to fully evaluate performance as any algorithm which is consistently incorrect[10] will still perform well under this measure. However a true assessment of the quality of the $\log_2$ ratios would require extensive PCR experiments which are beyond the scope of this paper.

We therefore compared the $\log_2$ ratios associated with each of the two replicates for 2,077 probes[11]. In Figure 6 we have depicted the $\log_2$ ratios for each of the these replicates on two different axes. The NIA $15K$ is a 'whole genome' slide, therefore only ten per cent of the probes were hybridised. We therefore also considered a slide developed specifically for the mouse eye (ME) by [Soullier *et al.*, 2003]. This slide contained 5,760 PCR products printed in duplicate of which 5,492 had hybridised and were uncontaminated. The mean squared errors between the estimates of the $\log_2$ ratios for both slides are summarised in Table 1 Recall that here a portion of each error is due to noise processes which are not associated with image processing so the improvement in consistency over spot (which initially appears to be 347% for NIA and 13% for the eye specific slide)

Slide	Manual	VIS	Spot
NIA	0.270	0.232	0.806
ME	·	4.16×10^{-2}	4.71×10^{-2}

Table 1: Mean squared errors for $\log_2$ ratios from the manual grid refinement, the VIS algorithm and the Spot algorithm.

[9] See `http://experimental.act.cmis.csiro.au/Spot/` and [Yang *et al.*, 2000].

[10] Consider, for example an algorithm which simply assigns a $\log_2$ ratio of 0. This algorithm will perform extremely well under our measure but will be useless in practice.

[11] Whilst there were 6,144 PCR products on the slide, we only evaluated the error for those which had hybridised and were not contaminated by dust leading to 2,077 separate points.

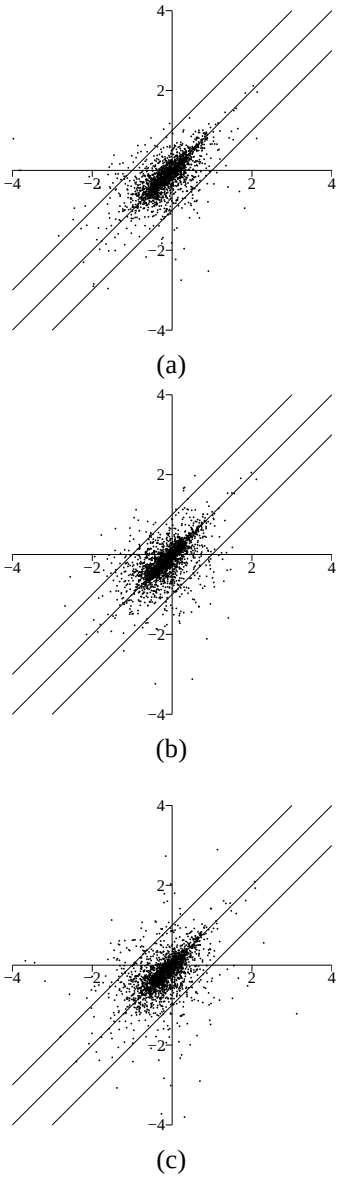

Figure 6: For these plots the axes represent the $\log_2$ ratios from the two replicas of each spot that are present on the slide. There are 2,077 points on the plots representing the 4,154 spots on the slide out of 12,288 which were present and not corrupted by artifacts. (a) Results from manual grid refinement. (b) Results from refinement with VIS. (c) Results from the Spot algorithm.

is likely to be higher. To demonstrate that there is no bias between the Spot algorithm and the VIS approach we have have plotted the extracted log ratios for each algorithm against each other (Figure 7). Also included in the plot are error bars derived from the variances of the VIS log ratios. Note that often these error bars are larger when the extracted values from the two approaches are inconsistent.

Figure 7: Plot of the results from the VIS algorithm against those from the Spot algorithm for the ME array. Error bars at one standard deviation are shown for the VIS results.

3.2 Discussion

We have shown how Bayesian inference can be used in the extraction of gene expression levels from microarray data. The framework outlined allows refinement of prior information about grid layouts using information in the image intensities. It circumvents a problem with a 'vanilla' importance sampling technique, where a naive choice of proposal distribution, namely the prior distribution, is often made, by introducing a novel, hybrid, Bayesian approximation, whose details are given in the appendix. The resulting algorithm showed how inconsistencies between different researchers could be reduced.

In theory the true Bayesian posterior can be multimodal whilst our proposal distribution is always unimodal. In practice this could mean that the sampling procedure focuses on one of the multiple modes which may not be that with the largest probabilistic mass. For our experiments this was not found to be a problem.

One advantage of our approach is the facility to place error bars on our estimates of the gene expression levels. This is possible because our Bayesian

approach gives us samples from a posterior distribution over the position and and sizes of the spots, giving us an estimate of the uncertainty. This uncertainty can be propagated through the downstream analysis.

The algorithm is stochastic, and variations in initial random seeds do result in variations in extracted $\log_2$ ratios. However, even though Spot, for example, uses a theoretically deterministic algorithm we found that similar levels of variation resulted for Spot when initial rough location of the grids was changed by only a little as one pixel. The robustness, in terms of repeatability, of the VIS algorithm to different initialisations was illustrated in Figure 4.

The algorithm led to a four and a half fold improvement in consistency over manual refinement of grids and at least a two fold improvement in consistency over another automated spot finding algorithm.

The algorithm was implemented in the MATLAB programming environment. The processing of the NIA $15K$ slide, containing 12,288 spots, took approximately 62 minutes on a Pentium III 1.2 GHz computer, or in other words about 0.3 of a second per spot. Note, however, that the researcher need not be in attendance during this processing time.

A pre-release MATLAB version of our software is available for academic use from `http://www.dcs.shef.ac.uk/~neil/VIS`.

Acknowledgments

PR is supported by a fellowship from the Lister Institute of Preventative Medicine and SS is a Fight for Sight research fellow.

A Posterior Approximations

Our options for approximating the posterior are twofold; firstly, we could look for a functional form which approximates our intractable posterior distribution, expectations under our approximation could then be substituted for the true expectations. We shall return to this approach in Section A.2, for the moment we focus on the alternative: Monte Carlo methods [Gilks *et al.*, 1996] and specifically importance sampling. Whilst we may not be able to normalise our posterior distribution, there are methods which allow us to obtain samples from it. Then, instead of obtaining expectations of interest exactly, we may obtain their sample based approximations.

Importance sampling involves the following steps; consider we wish to approximate an expectation, $\langle \phi(x) \rangle_{p(x)}$, over a distribution, $p(x) = \frac{p^*(x)}{Z}$, which is known subject to a constant Z. If we obtain a set of samples $\{x_s\}$ from a further *proposal* distribution, $R(x)$, we may estimate the expectation as $\langle \phi(x) \rangle_{p(x)} \approx \sum_{s=1}^{N} \hat{w}_s \phi(x_s)$, where the parameters $\hat{w}_s$, known as *importance weights*, are given by $w_s = \frac{p^*(x_s)}{R(x_s)}$ with $\hat{w}_s = \frac{w_s}{\sum_{s=1}^{S} w_s}$. A key component of importance sampling is the proposal distribution: for the method to be effective it is important that the proposal distribution is similar to the distribution of interest. In the context of Bayesian inference, a candidate for the proposal distribution is the prior: if the prior belief is accurate, it might be hoped that the prior is similar to the posterior. In this case, the unnormalised importance weights are simply given by the likelihood of each sample. This is one approach taken by, for example, *particle filters* [Doucet *et al.*, 2001]. One disadvantage of this approach is that if the prior distribution is not well matched to the posterior, the importance sampler's estimate will be dominated by a small number of samples with very large weights; to finesse this problem, we will introduce a technique which combines variational inference [Jordan *et al.*, 1998] with importance sampling. The first step though will be to introduce a 'hierarchical prior'.

A.1 Hierarchical Prior

In the main text we focused on a Gaussian prior for the centres and radii of the spots. In practice we used an alternative form of prior. It was still based around a Gaussian distribution, but one whose parameters, $\mathbf{m_c}$ and $\alpha_\mathbf{c}$ are governed by a so-called *hyper-prior*. In the following text we demonstrate this procedure for the values of the centres, $\mathbf{c}$, however exactly the same reasoning holds for the radii, $\mathbf{r}$,

$$p(\mathbf{c}) = \int p(\mathbf{c}|\mathbf{m_c}, \alpha_\mathbf{c})\, p(\alpha_\mathbf{c})\, p(\mathbf{m_c})\, \partial \mathbf{m_c} \partial \alpha_\mathbf{c}$$

$$\begin{aligned} p(\mathbf{c}|\mathbf{m_c}, \alpha_\mathbf{c}) &= \mathcal{N}\left(\mathbf{c}|\mathbf{m_c}, \alpha_\mathbf{c}^{-1}\right), \\ p(\mathbf{m_c}) &= \mathcal{N}\left(\mathbf{m_c}|\boldsymbol{\mu}_\mathbf{c}, \beta_\mathbf{c}^{-1}\right), \\ p(\alpha_\mathbf{c}) &= \text{gam}\left(\alpha_\mathbf{c}|a_\mathbf{c}, b_\mathbf{c}\right), \end{aligned}$$

where $\text{gam}(\cdot)$ is the gamma distribution defined by $\text{gam}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$, in which $\Gamma(\cdot)$ is the gamma function. This is a convenient choice of prior as it is *conjugate* to the Gaussian precision

$\alpha_{\mathbf{c}}$. The parameters of the gamma hyper-prior govern our confidence in $\alpha_{\mathbf{c}}$. The mean of the distribution is at $a_{\mathbf{c}}/b_{\mathbf{c}}$ and its variance decreases as $a_{\mathbf{c}}$ and $b_{\mathbf{c}} \to \infty$.

We could proceed with our new prior and once again use an importance sampler to approximate expectations of interest; but this would not necessarily improve the match between the proposal distribution (our new prior) and the distribution of interest (the posterior). Instead we turn to an alternative method of approximate inference known as variational inference.

A.2 Variational Inference

In variational inference [Jordan *et al.*, 1998] an intractable posterior, $p(x)$, is approximated by an alternative tractable distribution, $q(x)$. The approximation is found by minimising an information theoretic measure of the discrepancy between two distributions known as the Kullback-Leibler (KL) divergence,

$$\mathrm{KL}\left(q(x) \,||\, p(x)\right) = -\int q(x) \ln \frac{p(x)}{q(x)} dx.$$

If no constraints are imposed upon the form of $q(x)$, minimising the KL-divergence simply recovers $q(x) = p(x)$, to discover a tractable approximation we must impose constraints on the functional form of $q(x)$. For our model we choose to force the posterior approximation, $q(\mathbf{c}, \mathbf{m}_{\mathbf{c}}, \alpha_{\mathbf{c}})$, to be separable across its component variables,

$$q(\mathbf{c}, \mathbf{m}_{\mathbf{c}}, \alpha_{\mathbf{c}}) = q(\mathbf{c})\, q(\mathbf{m}_{\mathbf{c}})\, q(\alpha_{\mathbf{c}}).$$

Given this constraint it is possible to show, through a free-form minimisation of the KL divergence [Waterhouse *et al.*, 1996], that

$$\begin{aligned} q(\mathbf{m}_{\mathbf{c}}) &= \mathcal{N}(\mathbf{m}_{\mathbf{c}} | \bar{\mathbf{m}}_{\mathbf{c}}, \Sigma_{\mathbf{m}}), \\ q(\alpha_{\mathbf{c}}) &= \mathrm{gam}(\alpha_{\mathbf{c}} | \bar{a}_{\mathbf{c}}, \bar{b}_{\mathbf{c}}), \end{aligned}$$

where we have made use of

$$\begin{aligned} \bar{\mathbf{m}}_{\mathbf{c}} &= \Sigma_{\mathbf{m}} \left(\langle \alpha_{\mathbf{c}} \rangle \langle \mathbf{c} \rangle + \beta_{\mathbf{c}} \mu_{\mathbf{c}} \right), \\ \Sigma_{\mathbf{m}} &= \left(\langle \alpha_{\mathbf{c}} \rangle + \beta_{\mathbf{c}} \right)^{-1} \\ \bar{a}_{\mathbf{c}} &= a_{\mathbf{c}} + 1, \\ \bar{b}_{\mathbf{c}} &= b_{\mathbf{c}} + \left\langle \mathbf{c}^{\mathrm{T}} \mathbf{c} \right\rangle \\ &\quad - 2 \langle \mathbf{c} \rangle^{\mathrm{T}} \langle \mathbf{m}_{\mathbf{c}} \rangle + \left\langle \mathbf{m}_{\mathbf{c}}^{\mathrm{T}} \mathbf{m}_{\mathbf{c}} \right\rangle. \end{aligned}$$

Additionally it is possible to obtain some of the necessary expectations:

$$\begin{aligned} \langle \mathbf{m}_{\mathbf{c}} \rangle = \bar{\mathbf{m}}_{\mathbf{c}}, \qquad & \langle \alpha_{\mathbf{c}} \rangle = \bar{a}_{\mathbf{c}} / \bar{b}_{\mathbf{c}}, \\ \left\langle \mathbf{m}_{\mathbf{c}}^{\mathrm{T}} \mathbf{m}_{\mathbf{c}} \right\rangle &= \bar{\mathbf{m}}_{\mathbf{c}}^{\mathrm{T}} \bar{\mathbf{m}}_{\mathbf{c}} + \mathrm{Tr}\left(\Sigma_{\mathbf{m}}\right). \end{aligned}$$

Unfortunately no such simple functional form can be found for $q(\mathbf{c})$. This is one weakness of the variational approach: the class of models to which it can be applied is restricted to those which are in the conjugate exponential family [Ghahramani & Beal, 2001], the likelihood function, $p(\mathcal{I}|\mathbf{c})$, does not fall within this family and we must therefore return to importance sampling to obtain the expectations we require under $q(\mathbf{c})$.

A.3 Variational Importance Sampler

Using a free-form optimisation may re-write $q(\mathbf{c})$ as

$$q(\mathbf{c}) = \frac{1}{Z'_{\mathbf{c}}} p(\mathcal{I}|\mathbf{c})\, \mathcal{N}\left(\mathbf{c} | \langle \mathbf{m}_{\mathbf{c}} \rangle, \langle \alpha_{\mathbf{c}} \rangle^{-1}\right)$$

which means that we can obtain expectations under $q(\mathbf{c})$ using importance sampling with a proposal distribution, $R(\mathbf{c}) = \mathcal{N}\left(\mathbf{c} | \langle \mathbf{m}_{\mathbf{c}} \rangle, \langle \alpha_{\mathbf{c}} \rangle^{-1}\right)$. This importance sampler is then used to obtain estimates for $\langle \mathbf{c} \rangle$ and $\left\langle \mathbf{c}^{\mathrm{T}} \mathbf{c} \right\rangle$. These estimates are then substituted into the equations in Section A.2 to obtain new values $\langle \mathbf{m}_{\mathbf{c}} \rangle$ and $\langle \alpha_{\mathbf{c}} \rangle$ giving a new proposal distribution from which new estimates for $\langle \mathbf{c} \rangle$ and $\left\langle \mathbf{c}^{\mathrm{T}} \mathbf{c} \right\rangle$ can be found. This process is repeated until convergence. Details of the initialisation of the algorithm and one possible ordering of the updates can be found in [Lawrence *et al.*, 2003].

The advantage of combining the importance sampler with the variational approach is that the proposal distribution becomes focused on an area of high probability density in the posterior. This should alleviate a common problem in importance sampling where the sum used to approximate an expectation is dominated by a few large values of $\hat{w}_s$.

References

[Bernardo *et al.*, 1980] Bernardo, J. M., DeGroot, M. H., Lindley, D. V. & Smith, A. F. M., eds (1980) *Bayesian Statistics*. Valencia University Press, Valencia.

[Brazma *et al.*, 2001] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. & Vingron, M. (2001) Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nature Genetics*, **29**, 365–371.

[Brazma & Vilo, 2000] Brazma, A. & Vilo, L. (2000) Gene expression data analysis. *FEBS Letters*, **480**, 17–24.

[Cox, 1946] Cox, R. T. (1946) Probability, frequency and reasonable expectation. *American Journal of Physics,* **14** (1), 1–13.

[Doucet *et al.*, 2001] Doucet, A., de Freitas, J. F. G. & Gordon, N. J., eds (2001) *Sequential Monte Carlo Methods in Practice*. Springer-Verlag.

[Eisen & Brown, 1999] Eisen, M. B. & Brown, P. O. (1999) DNA arrays for analysis of gene expression. *Methods in Enzymology,* **303**.

[Ghahramani & Beal, 2001] Ghahramani, Z. & Beal, M. J. (2001) Graphical models and variational methods. In *Advanced Mean Field Methods — Theory and Practice,* (Opper, M. & Saad, D., eds),. MIT Press.

[Gilks *et al.*, 1996] Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

[Hughes *et al.*, 2000] Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakrabutty, K., Simon, J., Bard, M. & Friend, S. H. (2000) Functional discovery via a compendium of expression profiles. *Cell,* **102**, 109–126.

[Isard & MacCormick, 2001] Isard, M. & MacCormick, J. (2001) BraMBLe: a Bayesian multiple-blob tracker. In *Proc. 8th Int. Conf. Computer Vision* IEEE.

[Jordan *et al.*, 1998] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1998) An introduction to variational methods for graphical models. In *Learning in Graphical Models*, (Jordan, M. I., ed.), vol. 89, of *Series D: Behavioural and Social Sciences* pp. 105–162 Kluwer, Dordrecht, The Netherlands.

[Lawrence *et al.*, 2003] Lawrence, N. D., Milo, M., Niranjan, M., Rashbass, P. & Soullier, S. Bayesian processing of microarray images. To appear at *NNSP*.

[Schena, 2000] Schena, M., ed. (2000) *DNA Microarray, A Practical Approach*. Oxford University Press.

[Soullier *et al.*, 2003] Soullier, S., Milo, M., Moss, J., Lawrence, N. D., Williams, D., Smith, L., van Heyningen, V., Freeman, T., Greenfield, A., Niranjan, M. & Rashbass, P. The role of pitx3 in mouse eye development dissected using microarray analysis. Submitted.

[Tanaka *et al.*, 2000] Tanaka, T. S., Jaradat, S. A., Lim, M. K., Wang, G. J. K. X., Grahovac, M. J., Pantano, S., Sano, Y., Piao, Y., Nagaraja, R., Doi, H., 3rd, W. H. W., Becker, K. G. & Ko, M. S. H. (2000) Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc. Natl. Acad. Sci. USA,* **97**, 9127–9132.

[Varnum & Stevens, 1968] Varnum, D. S. & Stevens, L. C. (1968) Aphakia, a new mutation in the mouse. *Journal of Heredity,* **59** (2), 147–150.

[Waterhouse *et al.*, 1996] Waterhouse, S., MacKay, D. J. C. & Robinson, T. (1996) Bayesian methods for mixtures of experts. In *Advances in Neural Information Processing Systems*, (Touretzky, D. S., Mozer, M. C. & Hasselmo, M. E., eds), vol. 8, pp. 351–357 MIT Press, Cambridge, MA.

[Yang *et al.*, 2000] Yang, Y. H., Buckley, M. J., Dudoit, S. & Speed, T. P. (2000). Comparison of methods for image analysis on cDNA microarray data. Technical report Department of Statistics, University of California, Berkeley. Available from `http://www.stat.Berkeley.edu/~terry`.

[Yang *et al.*, 2001] Yang, Y. H., Dudoit, S., Luu, P. & Speed, T. P. (2001). Normalization for cDNA microarray data. Technical report Department of Statistics, University of California, Berkeley. Available from `http://www.stat.Berkeley.edu/~terry`.