ANALYSIS OF MICROARRAYS INCORPORATING ADJUSTMENTS FOR SPATIAL EFFECTS

A.F. Woolaston[1], R. Murison[2] and B. Tier[1]

[1]Animal Genetics and Breeding Unit[*], University of New England,
Armidale, New South Wales, 2351.
[2]School of Mathematics, Statistics and Computing Science and The Institute for Genetics and Bioinformatics, University of New England, Armidale, New South Wales, 2351.

SUMMARY
Various models were used to extract spatial effects from microarray data. Large discrepancies between the rankings of genes for the different methods were found, due to the roughness of the signal. Models assuming separability and autocorrelation did not perform as well as wavelets because the data were fractal in dimension, so assumptions underlying those models were violated.
Keywords: Microarray, spatial effects, wavelet, fractal.

INTRODUCTION

Microarray technology is becoming increasingly available to animal scientists to estimate expression levels of genes in biological processes. Intensity bias, dye bias and spatial bias can all influence the estimated expression level of genes if they are not included in the model or considered in the design of the experiment. This may result in inefficient allocation of resources in subsequent studies. This paper is concerned with spatial bias. Spatial trends can accumulate in the various stages of a microarray experiment. The final intensity reading of each spot is the result of a complex process involving array fabrication, sample preparation, cDNA synthesis and labeling, hybridization and microarray quantification. There are many possible causes for spatial trends on a slide within each step of a microarray experiment, not all of which are fully understood. The sources of variation in a microarray slide may not act continuously, so may be fractal and discontinuous in dimension.

Several authors have considered models to account for spatial bias involving autocorrelation (Burgueno *et al*. 2005, Baird *et al.* 2004). Models involving autocorrelation and splines assume separability (Adler 1981). Such methods rely on the data having an integer fractal dimension. When this condition is violated they may not be as efficient as other methods of removing spatial dependencies such as wavelet decomposition. This paper compares the efficiency of four methods in removing spatial dependencies from a murine microarray experiment.

MATERIALS AND METHODS

Data. The data used in this study came from cDNA extracted from mice livers in experiments conducted by Harry Noyes at The University of Liverpool. The treatments applied to the microarray were determined by strain of mice, challenge, replicate and time and these factors were unbalanced. Two types of mice were used, AJ and C57BL6. The mice were further divided into two challenges.

[*] AGBU is a joint unit of NSW Primary Industries and the University of New England.

From each mouse challenge two replicates were taken and the biological samples were applied to the microarrays at time points of 0, 4, 7, 10 and 17 days. Fifteen slides were used in total, with two biological treatments applied to each, one coloured red and the other green. The design of the first stage of the experiment involving the mice is not relevant to this paper.

Each slide was divided into four metacolumns and twelve metarows, forming 48 printing blocks on each slide. There were twelve rows within each metarow and sixteen columns within each metacolumn yielding 9,216 spots per array. Thus with 15 slides, 2 colours and 9,216 spots there were 276,480 data points. 11,218 genes were spotted on the arrays in all, so clearly not all genes were printed on each array. Housekeeping genes were used on all arrays, and included empty cells. They were printed in a very similar pattern on each array, with the first and last row of each metarow typically filled with housekeeping genes and the third row of each metarow usually containing some empty cells.

Statistical models. A mixed model of the form $\mathbf{y = X\beta + Zu + E}$, $\mathbf{E \sim N(0,\Sigma)}$, was fitted for each colour on each array with $\log_2$ of intensity as the response variable. Thus, dye bias was modeled as a part of the systematic model. The effects of housekeeping genes, $\boldsymbol{\beta}$, were treated as fixed effects and the effects of test genes, $\mathbf{u}$, were assumed to have come from a normally distributed ($\mathbf{N(0,\sigma^2_u)}$) population. The model was fitted with no spatial correlation assumed and then separable first order auto-correlated rows and first order auto-correlated columns (AR1×AR1). The grand error term, $\mathbf{E}$, was partitioned into mutually orthogonal components, a spatial component, $\boldsymbol{\xi}$, which modeled the spatial bias and remaining error term**,** $\boldsymbol{\varepsilon}$. The spatial component was estimated using wavelets, splines and AR1×AR1 with splines. A model was also fitted with no consideration for spatial effects as a benchmark. Once $\boldsymbol{\xi}$ was found, this term was subtracted from the original intensities to give adjusted intensities, $\mathbf{y^*}$. As before a model of the form $\mathbf{y^* = X\beta + Zu + \varepsilon}$ was fitted and the 100 gene by treatment interactions with the highest absolute expression from each method were compared.

Splines. A simple model was fitted for each slide of the form $\mathbf{E = printing\ block + R_{spl} + C_{spl} + \varepsilon}$, where $\mathbf{R_{spl}}$ and $\mathbf{C_{spl}}$ are the spline trends due to rows and columns respectively. The spatial component, $\boldsymbol{\xi} = \mathbf{printing\ block + R_{spl} + C_{spl}}$, was subtracted from the original data.

Wavelets. The Discrete Wavelet Transform (DWT) was applied to the grand error term, $\mathbf{E}$, and then thresholded in the wavelet domain to remove the spatial trend, $\boldsymbol{\xi}$. Wavelet coefficients below a threshold are considered to correspond to noise and the larger scale wavelet coefficients correspond to the spatial trend (Huang and Cressie 1997). Thus the larger wavelet coefficients corresponding to spatial trends were transformed back into the spatial domain before subtraction from the original intensities.

A sample variogram of $\boldsymbol{\varepsilon}$ was computed and plotted for each model to compare the models. The variogram ordinates are given by $\gamma_{ij} = \frac{1}{2}\,[\varepsilon_i - \varepsilon_j]^2$. The sample variogram is the triple $(|\tau^r_i - \tau^r_j|, |\tau^c_i - \tau^c_j|, v_{ij})$, where τ^r_i is the row position of the ith point, τ^c_i is the column position of the ith point and v_{ij} is the mean of the variogram ordinate with a distance of $(|\tau^r_i - \tau^r_j|, |\tau^c_i - \tau^c_j|)$ between them. A flat variogram with a low plateau illustrates that the model has efficiently removed spatial trends. The sample variogram was also calculated as a function of the absolute distance between points, $v(|\boldsymbol{\tau}|)$.

Calculating Fractal Dimension. The fractal dimension of all 15 slides was calculated for both the red and green treatments by both box counting and a variogram based method.

Box counting. If a set **S** can be completely covered with a minimum number of cubes N_δ**S** of side length δ, then the box counting fractal dimension is given by $D_{BC} = \lim_{\delta \to 0} \frac{\ln(N_\delta(S))}{\ln(\delta)}$. In order to calculate the box counting fractal dimension of a microarray image, a triple of each row position, column position and intensity was formed. Now each microarray slide was represented as an image in R^3. The minimum number of cubes to completely cover the image was found for a variety of cube sizes and the logarithm of the cube size was plotted against the logarithm of the number of cubes required. The slope of the plot for small δ gave the estimate of the box counting fractal dimension, D_{BC}.

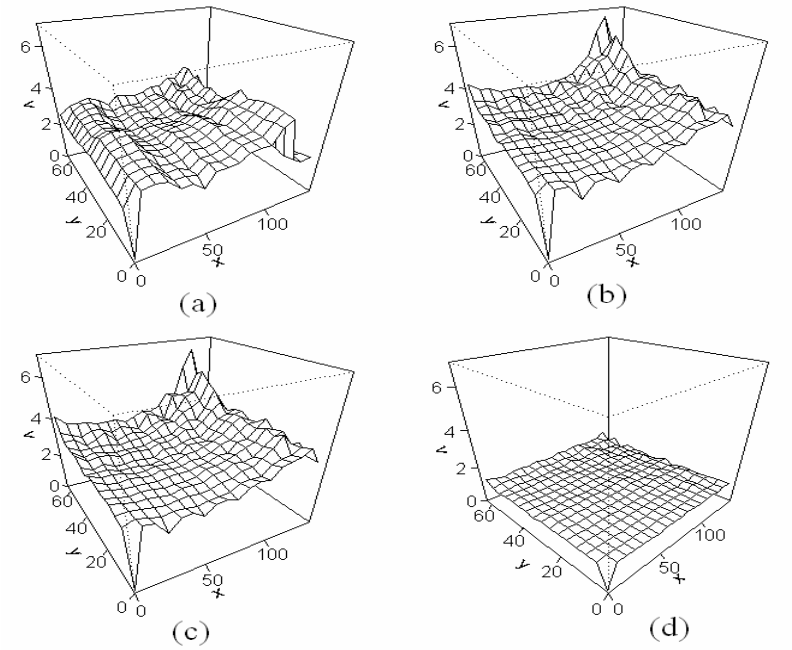

Figure 1. Sample variograms for red, slide 15 with (a) No spatial trend modeled. (b) Splines. (c) AR1×AR1 with splines. (d) Wavelets.

Variogram Based Estimator. The sample variogram, v(|**τ**|), was used to estimate the fractal dimension (Constantine and Hall 1994). Assuming that v(|**τ**|)=c| **τ** |$^{\alpha}$ as τ approaches 0, then log(v(|**τ** |))= constant + αlog(**τ**) + error. The slope of the plot of log(**τ**) and log(v(| **τ** |)) for small **τ** was used to find an estimate of α and then an estimate of the fractal dimension was given by D=3-α/2.

RESULTS

The model used to adjust for spatial trend affected the ranking of the expression levels of the gene by treatment interaction. Each of the models resulted in the same gene (<K01391-a) being the most highly ranked in absolute expression level for a given treatment, but overall only 30 of the gene by treatment interactions were ranked in the top 100 interactions for all four methods. Only 31 interactions were ranked in the top 100 for each of the three models correcting for spatial bias. The majority of the genes in the 100 most highly ranked gene by treatment interactions were spotted once per array, but some were spotted up to 40 times per array. Between three and nine of the genes in the 100 most highly ranked gene by treatment interactions for each method were only printed on one array. The Pearson's rank correlation between methods ranged between 0.355 and 0.365 for all pair wise comparisons of all ranked gene by treatment interactions. The sample variograms for the four methods are displayed in Figure 1 for the red dye, slide 15. The shape of the variograms for this slide and colour was typical of the data.

The mean fractal dimension for the red (green) treatment of all 15 slides was 2.22 (2.22) ± 0.01 and 2.21 (2.17) ± 0.13 for the box counting and variogram based methods respectively.

DISCUSSION

The low number of genes ranked in the top 100 for all four methods illustrates the large difference between methods and that care should be taken when choosing methods for removing spatial effects from data. From the variograms in Figure 1, it can be seen that wavelets were much more successful in removing the spatial trend than the other methods. The splines and AR1×AR1 with splines models were marginally better than fitting no spatial trend. Their variograms had a flatter plateau than if no spatial trend were modeled.

Departure from 2 in the fractal dimension of the data, suggests that the data are insufficiently smooth for the separable autocorrelation model to be efficient. The wavelet transform for correcting spatial trends in microarrays is a generic method for capturing the spatial information irrespective of fractal dimension. The location-frequency resolution of wavelets allows identification of broad spatial trends and fine scale vibrations.

REFERENCES

Adler, R.J. (1981) "The Geometry of Random Fields". New York: Wiley.

Baird, D. , Johnstone P.(2004), and Wilson T., *Bioinformatics* **20**:3196.

Burgueno, J., Crossa, J., Grimanelli, D., Leblanc, O. and Autran, D.(2005), *Crop. Sci* **45***(2)*:748.

Constantine, A.G. and Hall, P. (1994), J Roy. Stat. Soc B. Met. **56**:97.

Huang, H.C. and Cressie, N. (1997), *Statistical Laboratory Preprint No. 97-23*, Iowa State University, Ames.