Chapter 5

Kernel Regression of SNP Data to Predict Breeding Value

5.1 Introduction

In recent times there has been a dramatic increase in the amount of genomic data generated, with some studies incorporating whole genome scans with millions of markers (see for example Wong et al. (2004); Zimdahl et al. (2004)). It is anticipated that these data will be used to predict the genetic value of animals more accurately at an earlier age than is currently being achieved (Schaeffer, 2006; Meuwissen et al., 2001).

Locating quantitative trait loci (QTL) has been the subject of considerable research, but recently the focus has moved towards using information from the whole genome to predict genetic value. Meuwissen et al. (2001) compared the accuracy of prediction of best linear unbiased prediction (BLUP) and Bayesian analysis when relatively few individuals were genotyped for many markers from a whole genome scan. A hierarchical model was suggested by Gianola et al. (2003) and the model was incorporated into BLUP and Bayesian analyses to utilize marker information to infer phenotype marker associations. Similarly, Xu (2003) used a Bayesian method whereby each marker was assigned a normal prior with the prior variance to be determined from the data to simultaneously evaluate marker effects from the entire genome. Woolaston et al. (2007) (Chapter 3) used principal components regression on single nucleotide polymorphisms (SNPs) from the whole genome to predict molecular breeding value in a dairy population.

All of the aforementioned methods can be broadly classified as parametric techniques, whereby the modeler determines the function relating the phenotypic response variable and genotypic explanatory variables. It is difficult to model the inherently complicated nature of the genome with these parametric models. Hence, the use of non-parametric methods may be sensible, whereby the data determine the shape of the relationship between phenotypes and genotypes. Gianola et al. (2006) demonstrated that non-parametric techniques such as kernel regression could be incorporated to predict genotypic merit via reproducing kernel Hilbert space (RKHS) regression. In the RKHS model presented by Gianola et al. (2006) the intrinsically discrete marker data was treated as a continuous variable. However, the effect of treating intrinsically discrete marker data as continuous is unclear.

In this chapter kernel regression is applied to simulated and real data from a dairy cattle population. The aim of the kernel regression is to predict the molecular breeding value (MBV) of an animal, using only its genotypic information and both the phenotypic and genotypic information of other animals. The effect of treating the discrete genotypic information as a continuous trait is examined by comparing the predictive performance of different forms of kernel regression that accommodate discrete and continuous data.

5.2 Materials

5.2.1 Real Data

These are the data which were the subject of Chapters 3 and 4. The details are restated here for context of the new analysis. These data are comprised of 15,380 SNPS recorded from $n_{an} = 1,546$ dairy sires. However, 298 SNPs showed no variation and were removed from these data, so $n_s = 15,082$ SNPs are retained. These data are arranged into a matrix, $\mathbf{X}_{n_{an} \times n_s}$, where

$$X_{ij} = \begin{cases} 0 \text{ if the } j\text{th SNP for the } i\text{th animal is aa} \\ 1 \text{ if the } j\text{th SNP for the } i\text{th animal is aA (unordered)} \\ 2 \text{ if the } j\text{th SNP for the } i\text{th animal is AA .} \end{cases} \tag{5.1}$$

A total of 6.89% of all the SNPs were missing values and these were replaced with 1's to be consistent with Mendel's first law. Table 5.1 gives a breakdown of the percentage of SNPs missing per animal.

Table 5.1: Distribution of the percentage of missing SNP values per animal.

% of missing SNP values	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Number of animals	1447	85	7	3	1	1	2

The dairy sires were born between 1955 and 2001 and estimated breeding values (EBVs) for milk protein percent were estimated as a part of the Australian Dairy

Herd Improvement Scheme. Table 5.2 gives a breakdown by years of the EBVs. These EBVs are highly reliable.

Table 5.2: Summary of the EBVs of the dairy sires.

Year of Birth	Before 1972	1972-1981	1982-1991	1992-2001	All animals
Number of animals	10	82	457	997	1546
Mean of EBVs (%)	-0.176	-0.079	-0.038	0.020	-0.004
sd(EBV)	0.120	0.108	0.108	0.123	0.123
Mean Reliability of EBVs	0.778	0.940	0.918	0.874	0.890

5.2.2 Simulated Data

These simulated data are similar to the data simulated in Chapter 3 but only 200 chromosomes are used to generate the base population and the number of known chromosomes is varied here. A population comprising 1,204 animals was simulated. Animals consisted of two copies of one chromosome of length 20 million base pairs. The simulation process can be divided into 5 steps, which are explained below. The algorithm is displayed in Appendix A.1.

(i) Location of SNPs and probability of mutations at the SNP loci SNPs were placed on the chromosome, with their base pair positions randomly sampled from the integers between 1 and 20 million without replacement. Either (i) $n_s = 10$, (ii) $n_s = 100$ or (iii) $n_s = 1{,}000$ SNPs were assumed to be known. Of these SNPs, (a) $n_a = 10$, (b) $n_a = 100$ and (c) $n_a = 1{,}000$ ($n_a \leq n_s$) were simulated to have an additive effect and these effects were sampled from a Gamma distribution with shape parameter 0.59 and scale parameter 7.1 (Hayes and Goddard, 2001). Each effect was randomly assigned to be positive or negative with probability 0.5 (Meuwissen et al., 2001). The minor allele frequeny at the *ith* site, p$_i$, was randomly sampled between 0 and 0.5.

(ii) Base population of chromosomes In order to simulate the effect of Linkage Disequilibrium (LD), a small number of chromosomes (200) were created in order to generate the base population. The haplotype values for the jth chromosome in the *ith* position is given by:

$$B_{ij} = \begin{cases} 0 \text{ with probability } 1 - p_i \\ 1 \text{ with probability } p_i. \end{cases} \tag{5.2}$$

(iii) Base population The top 60 of the rows of the matrix **B** were paired to form 30 males and the remaining 140 paired up to form 70 females. Random mating was performed to produce the first generation of 500 individuals. The distance between cross-overs in the breeding process was sampled from a Poisson distribution with parameter 1 million, so that each chromosome is 20 Morgans long. No mutation was simulated.

(iv) Subsequent generations The population structure was intended to be a simplified representation of the breeding structure in place in the dairy industry in Australia. The initial population of 500 animals was split into 40 males and 460 females and random breeding was once again simulated to form 395 new animals. Ten of these animals were assigned to be male and 385 were female. Thirty of the males and 75 of the females were retained from the previous generation and were added to the current population of 10 males and 385 females to form the next generation. This process was repeated for 10 generations and the last three generations were stored.

(v) Calculate MBVs and phenotypes The MBV for each animal stored in the last three generations was calculated as:

$$MBV = \sum_{i=1}^{j=1000} q_i a_i, \tag{5.3}$$

and the phenotypic values calculated as:

$$y = MBV + \epsilon, \tag{5.4}$$

where q_i is the number of minor alleles (0, 1 or 2) at SNP position i, a_i is the allelic substitution effect of the ith polymorphism and ϵ is sampled from a N(0,σ_e^2) distribution. The allele effects were additive. The predefined heritability ($h^2 = 0.1, 0.4$ and 0.7) and the additive genetic variance (σ_a^2) determined σ_e^2 via the equation $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$.

5.3 Kernel Regression

The aim is to predict the MBV of the ith animal, MBV_i, from the SNP values of that animal, $\mathbf{x}_i$, with some regression function, $g(.)$:

$$M\hat{B}V_i = g(\mathbf{x}_i).$$

Using kernel regression the data determines the shape of the function, $g(.)$, by assigning weights to each data point according to its proximity to the focal point, $\mathbf{x}_i$. The kernel regression function can be written as:

$$g(\mathbf{x}_i) = \sum_{j \neq i} w_{ij} y_j, \tag{5.5}$$

where w_{ij} is the weight for the jth animal in the regression and y_j is the EBV or phenotypic record of the jth animal.

A common method of determining the weights, w_{ij}, was proposed by Nadaraya (1964) and Watson (1964):

$$w_{ij} = \frac{K(d[\mathbf{x_i}, \mathbf{x}_j]/k)}{\sum_{j \neq i} K(d[\mathbf{x}_i, \mathbf{x}_j]/k))}, \tag{5.6}$$

where $K(.)$ is a kernel function, $d[.,.]$ is a (pseudo-quasi) metric and k is a smoothing parameter.

Upon substitution of the w_{ij} term in equation (5.5) with equation (5.6), the Nadaraya-Watson estimator of the regression function is obtained:

$$g(\mathbf{x}_i) = \frac{\sum_{j \neq i} K(d[x_i, x_j]/k) y_j}{\sum_{j \neq i} K(d[x_i, x_j]/k)}.$$

The important choices here are that of the metric, $d[.,.]$ and kernel function, $K(.)$. Three different combinations of the kernel function and metric are examined here.

In the following sections, the term focal genotype is used to refer to the genotype of the animal whose MBV is to be estimated. The observed genotype, $\mathbf{x}_i$, refers to the ith individual in the training set.

5.3.1 Binomial Kernel

Gianola et al. (2006) suggested that the number of disagreements between the focal genotype, $\mathbf{x}$, and observed genotype, $\mathbf{x}_i$, be used as the metric. That is, the metric $d[.,.]$ is defined to be:

$$d[\mathbf{x},\mathbf{x}_i] = 2 \times \sum_{j=1}^{n_s} \mathbf{U}_{x(j)}(x_i(j)), \tag{5.7}$$

where

$$\mathbf{U}_{x(j)}(x_i(j)) = \begin{cases} 0 \text{ if } x(j) = x_i(j) \\ 1 \text{ if } x(j) \neq x_i(j). \end{cases}$$

We note that the scaling factor of 2 makes no difference when using the metric in equation 5.7, but this factor is included to be consistent with the method described by Gianola et al. (2006). The kernel function is defined to be:

$$K(\mathbf{x},\mathbf{x}_i,k) = k^{p-d[\mathbf{x},\mathbf{x}_i]}(1-k)^{d[\mathbf{x},\mathbf{x}_i]},$$

with $\frac{1}{2} \leq k \leq 1$ and p is the maximum value of $d[\mathbf{x},\mathbf{x}_i]$.

The Nadaraya-Watson estimator of the MBV at the focal point, $\mathbf{x}$ (which may or may not be equal to any $\mathbf{x}_i, i = 1, 2...n$) is given by (Kernel 1):

$$g(\mathbf{x}) = \frac{\sum_{i=1}^{n} k^{p-d[\mathbf{x},\mathbf{x}_i]}(1-k)^{d[\mathbf{x},\mathbf{x}_i]}.y_i}{\sum_{i=1}^{n} k^{p-d[\mathbf{x},\mathbf{x}_i]}(1-k)^{d[\mathbf{x},\mathbf{x}_i]}}. \tag{5.8}$$

5.3.1.1 Missing Data

The metric defined in equation (5.7) is not robust to missing values in the data. The approach of assuming all missing values to be heterozygotes tends to shrink the 'genetic distance' between animals from their true values. Hence, the quasi-pseudo metric $d^*[.,.]$ is used:

$$d^*[\mathbf{x}, \mathbf{x}_i] = \frac{n_s d[\mathbf{x}^*, \mathbf{x}_i^*]}{n_s^*},$$

where $\mathbf{x}_i^*$ and $\mathbf{x}^*$ contain the common known SNPs in $\mathbf{x}_i$ and $\mathbf{x}$ and n_s^* is the number of such SNPs. The idea is to find the 'distance' between animals for the known SNPs and scale this distance according to the total amount of SNPs. The $d^*[.,.]$ metric could easily be replaced with the $d[.,.]$ metric, with $\mathbf{x}$ and $\mathbf{x}_i$ replaced by their respective conditional expectations $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}_i$ (Gianola et al., 2006). However, in this study we do not assume a known pedigree, or that the animals are related.

The Nadaraya-Watson for the $d^*[.,.]$ metric with the binomial kernel becomes (Kernel 1*):

$$g(\mathbf{x}) = \frac{\sum_{i=1}^{n} k^{p-d^*[\mathbf{x},\mathbf{x}_i]}(1-k)^{d^*[\mathbf{x},\mathbf{x}_i]}.y_i}{\sum_{i=1}^{n} k^{p-d^*[\mathbf{x},\mathbf{x}_i]}(1-k)^{d^*[\mathbf{x},\mathbf{x}_i]}}. \tag{5.9}$$

5.3.2 Gaussian Kernel

Under a Gaussian framework, the vectors of SNP values are assumed to be continuously valued, although strictly, they are discretely valued. The Gaussian kernel is of the form (Silverman, 1986):

$$K(\frac{\mathbf{x}_i - \mathbf{x}}{k}) = \frac{1}{(2\pi)^{\frac{p}{2}}} e^{[-\frac{1}{2}(\frac{\mathbf{x}_i - \mathbf{x}}{k})'(\frac{\mathbf{x}_i - \mathbf{x}}{k})]}. \tag{5.10}$$

The Nadaraya-Watson estimator of the Gaussian kernel function for a genotype $\mathbf{x}$ is (Kernel 2):

$$g(\mathbf{x}) = \frac{\sum\limits_{i=1}^{n} e^{[-\frac{1}{2}(\frac{\mathbf{x}_i - \mathbf{x}}{k})'(\frac{\mathbf{x}_i - \mathbf{x}}{k})]}.y_i}{\sum\limits_{i=1}^{n} e^{[-\frac{1}{2}(\frac{\mathbf{x}_i - \mathbf{x}}{k})'(\frac{\mathbf{x}_i - \mathbf{x}}{k})]}}.$$

5.3.3 Predicting MBV of Each Animal From All Other Animals

Each of the 1,546 animals in these real data have their MBVs estimated from the remaining 1,545 animals using the 3 different forms of the Nadaraya-Watson estimator introduced above (equations (5.8), (5.9) and (5.10)).

The smoothing parameter, k, is varied and the accuracy of prediction is calculated as:

$$\text{accuracy} = \rho(\mathbf{g}(\mathbf{x}_i), \mathbf{EBV}),$$

where ρ denotes the correlation, $\mathbf{g}(\mathbf{x}_i)$ is the vector of estimated MBVs and $\mathbf{EBV}$ is the vector of EBVs, or in the case of the simulated data, true MBVs.

5.3.4 Predicting a Group of Animals From All Other Animals

In practice, the optimal value of the smoothing parameter, k, is not known and needs to be calculated from a training set. The animals are partitioned into those with EBVs, K, and those to have their MBVs estimated, U. The smoothing parameter is found by performing a search on the parameter space using animals in the set K and choosing the k which minimizes the mean square error of prediction (Gianola et al., 2006). The MBVs of animals in the set, U, are then estimated by kernel regression using the animals in K and the estimate of the smoothing parameter obtained from the training set. This procedure is repeated for the different (pseudo-quasi) metrics and kernel functions introduced above.

In the case of these real data, $n_u = 150$ animals are randomly assigned to form the set of animals, U. A different set of animals is randomly chosen 100 times for cross validation. For these simulated data, the youngest generation comprising of 396 animals form the set of unknown animals, U and the remaining 808 animals are the training set.

5.4 Results

5.4.1 Simulated Data

Table 5.3 displays the accuracy for Kernels 1 and 2 for predicting MBV, with the number of recorded SNPs, number of SNPs with an additive effect and at different heritabilities. Kernel 2 (kernel regression with a Gaussian kernel) is superior to Kernel 1 in all instances.

Of the variables studied, heritability has the largest impact on the accuracy of both methods of kernel regression. If all other factors are held constant, a low heritability is associated with low accuracy of prediction and high heritability is coupled with high accuracy. This is hardly surprising, since if all animals are unrelated, the accuracy of the Nadaraya-Watson estimator is proportional to the square root of heritability (see Appendix A.2). There are small departures from this relationship here because the animals are related.

The number of known SNPs also influences the accuracy of kernel regression. A smaller number of known SNPs is associated with better accuracy and a larger amount of known SNPs is associated with lower accuracy, all other factors held constant.

The number of SNPs with an additive effect has minimal impact on the accuracy of kernel regression in estimating MBV if the number of SNPs and heritability are held constant. For example when there are 1,000 known SNPs and a heritability of 0.4, the respective accuracies for Kernel 1 are 0.461, 0.467 and 0.467 when 10, 100 and 1,000 SNPS have an additive effect.

The optimal value for the smoothing parameter in Kernel 1, k, is influenced by the heritability and the number of known SNPs. Lower values of k are associated with low heritabilities and larger numbers of known SNPs.

Similarly, the optimal value for the smoothing parameter in Kernel 2 is also influenced by the heritability and the number of known SNPs. Lower values of k are associated with high heritabilities and smaller numbers of known SNPs.

Table 5.3: Mean accuracies (standard error) of prediction and mean optimal smoothing parameters for binomial kernel regression (Kernel 1) and Gaussian kernel regression (Kernel 2) for these **simulated** data (100 replications for each simulation).

			Kernel 1		Kernel 2	
n_s	n_a	h^2	$\bar{Accuracy}$	$\bar{k}$	$\bar{Accuracy}$	$\bar{k}$
10	10	0.1	0.245 (0.005)	0.656	0.262 (0.005)	0.71
		0.4	0.548 (0.005)	0.724	0.572 (0.005)	0.56
		0.7	0.770 (0.004)	0.796	0.792 (0.003)	0.48
100	10	0.1	0.197 (0.005)	0.517	0.216 (0.006)	2.19
		0.4	0.469 (0.006)	0.526	0.510 (0.006)	1.79
		0.7	0.679 (0.007)	0.534	0.707 (0.006)	1.58
100	100	0.1	0.192 (0.006)	0.517	0.221 (0.005)	2.18
		0.4	0.461 (0.005)	0.526	0.502 (0.006)	1.82
		0.7	0.667 (0.005)	0.546	0.724 (0.004)	1.59
1000	10	0.1	0.188 (0.005)	0.502	0.203 (0.006)	7.00
		0.4	0.461 (0.005)	0.503	0.510 (0.005)	5.90
		0.7	0.670 (0.006)	0.504	0.702 (0.005)	5.00
1000	100	0.1	0.196 (0.006)	0.502	0.231 (0.005)	7.05
		0.4	0.467 (0.005)	0.503	0.519 (0.005)	5.91
		0.7	0.669 (0.005)	0.504	0.705 (0.004)	5.00
1000	1000	0.1	0.198 (0.006)	0.502	0.220 (0.005)	7.09
		0.4	0.467 (0.006)	0.503	0.507 (0.005)	5.89
		0.7	0.673 (0.004)	0.504	0.705 (0.004)	5.00

n_s is the number of known SNPS, n_a is the number of SNPs with an additive effect, h^2 is the heritability and k is the smoothing parameter.

5.4.2 Real Data

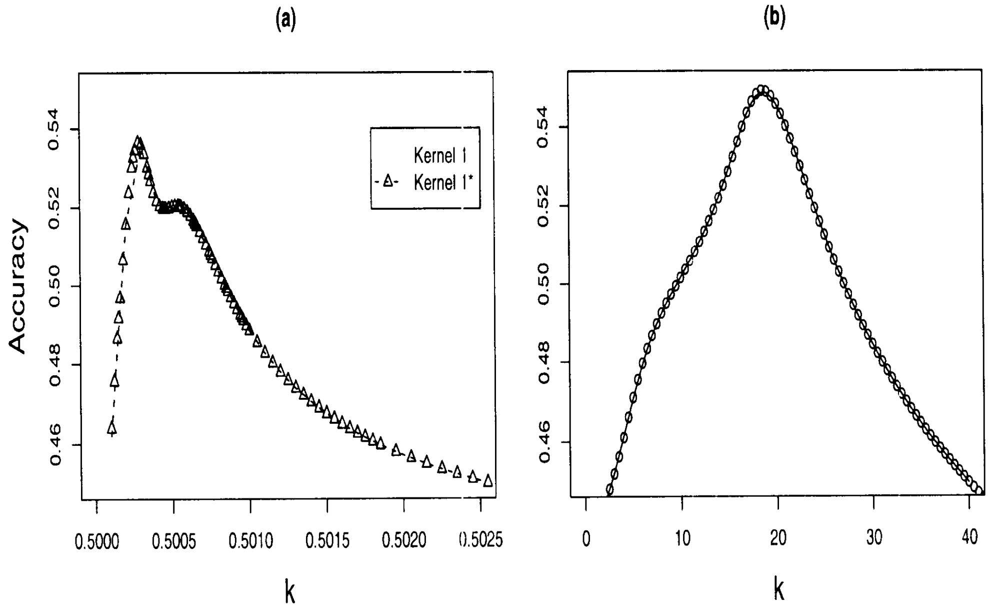
5.4.2.1 Predicting the MBV for Each Animal

Figure 5.1 displays the accuracies of the 3 kernels as each of the respective smoothing parameters vary. The best accuracy obtained is 0.546 when Kernel 2 is used with a smoothing parameter of 18 and the optimal value of the smoothing parameter is clearly identified for this kernel from the plot (Figure 5.1(b)).

Figure 5.1(a) displays the accuracies for Kernel 1 and 1*. A global maximum accuracy of 0.519 is observed when a smoothing parameter of 0.5006 is used for Kernel 1. However, there is another local maximum observed when the smoothing parameter is 0.5013 for Kernel 1. Similarly, there are 2 local maximums in accuracy for Kernel 1* that almost occur in the same locations as for Kernel 1. However the peak observed when the smoothing parameter is 0.5013 is less pronounced. The maximum accuracy for Kernel 1* (0.535) is higher than the maximum obtained by Kernel 1.

For Kernel 1, when the smoothing parameter is greater than 0.6, the accuracy approaches 0.459. When k in Kernel 1* is between 0.6 and 1, the accuracy is 0.417 and for small k ($k < 0.5$) the accuracy of Kernel 2 is 0.438. In all of the abovementioned cases, the estimates of MBV for each animal approach the EBV of the 'genetically closest' animal, where 'closest' is defined by the metric used in each respective kernel function.

Figure 5.1: Accuracy of prediction for **real** data for varying smoothing parameter, k. (a) Binomial kernels (b) Gaussian kernel.



5.4.2.2 Predicting the MBV for a Group of Animals

Table 5.4 displays the accuracy of prediction and the mean optimal smoothing parameter for the three kernel regression methods when the MBVs of a group of 150 animals are predicted from the remaining 1,396 animals. The standard deviations for the mean optimal smoothing parameter in Kernels 1 and 1* are less than 0.0001 and less than 0.1 for Kernel 2. The mean accuracies for Kernel 1 and 2 are the same as when each animal had its MBV estimated from all other animals. However, the mean accuracy for Kernel 1* is significantly lower for predicting a group of animals MBVs compared to predicting each MBV individually. This would suggest that this kernel is not robust to a drop in the number of animals in the training set.

Table 5.4: Accuracy of predicting the MBV of 150 animals randomly sampled from the remaining 1,396 animals (100 replicates).

	Kernel 1	Kernel 1*	Kernel 2
Accuracy	0.519 (0.005)	0.518 (0.005)	0.542 (0.005)
k	0.5006	0.5005	18.0

Kernel 1 is binomial kernel regression, Kernel 1* is binomial kernel regression adjusted for missing values and Kernel 2 is Gaussian kernel regression.

5.5 Discussion

The smoothing parameter, k, has to be close to 0.5 in Kernels 1 and 1* when there are a large number of SNPs, or the estimate at a focal point tends towards the EBV (or phenotypic value) of the 'genetically closest' animal. To demonstrate this, equation (5.8) can be re-written as:

$$\frac{\sum_{i=1}^{n}(\frac{1-k}{k})^{d[\mathbf{x},\mathbf{x}_i]}y_i}{\sum_{i=1}^{n}(\frac{1-k}{k})^{d[\mathbf{x},\mathbf{x}_i]}} = \frac{y_m + \sum_{i\neq m}^{n}(\frac{1-k}{k})^{d[\mathbf{x},\mathbf{x}_i]-m_d}.y_i}{1+\sum_{i\neq m}^{n}(\frac{1-k}{k})^{d[\mathbf{x},\mathbf{x}_i]-m_d}}, \tag{5.11}$$

where m_d is the minimum value of $d[\mathbf{x}, \mathbf{x}_i]$ for $i = 1, 2, 3..., n$, y_m is the EBV of the corresponding animal and $\frac{1}{2} \leq k \leq 1$. It has been observed that $d[\mathbf{x}, \mathbf{x}_i] >> 0$ for $i \neq j$, which implies that equation (5.11) tends towards y_m when k is only fractionally larger than 0.5. Similar calculations can be performed to show that small values of k in Kernel 2 lead to estimates of MBV approaching that of the 'genetically closest' animal. Consequently, an examination of the accuracy of prediction for large values of k in Kernels 1 and 1* and for small values of k in Kernel 2 gives further indication as to which notion of 'genetic distance' is most plausible. Surprisingly, the $d[,.,]$ metric gives the best correlation between an animals EBV and that of its 'genetically closest' animal.

The increase in the optimal smoothing parameter value in the binomial kernel with heritability is due to a decrease in noise. The left hand side of equation (5.11) is useful to determine why this is the case. For lower heritabilities, there is an increase in noise, so animals genetically 'further' from the focal animal require a higher weighting in the regression, i.e. the smoothing parameter has to be closer to 0.5. Conversely, for a highly heritable trait, there is less noise in the regression so that the MBV of an animal can be estimated with a high weighting on the animals genetically 'close' to it. That is, the smoothing parameter can be higher than 0.5 to place more emphasis on the genetically 'close' animals when there is less noise in the system.

The existence of 2 local maximums in accuracy for Kernels 1 and 1* when these real data are used (Figure 5.1(a)) is probably an artifact of missing SNP values. The use of the $d^*[.,.]$ metric as opposed to the $d[.,.]$ metric in the binomial kernel function increases the accuracy of the global maximum and does little to the accuracy of prediction at the local maximum on the right. This would imply that when the smoothing parameter is 0.5013, the local maximum occurs because assuming missing SNP values are heterozygotes tends to deflate the genetic distance between animals, so that a higher value of k is locally optimal.

The superiority of Kernel 2 in all comparisons examined here indicates that assuming SNPs are continuous rather than discrete is acceptable. The accuracy of prediction of the binomial kernel when the $d^*[.,.]$ metric is used is significantly higher than when the $d[.,.]$ metric is used, indicating that there are considerable improvements to be made by adequately modeling the missing SNP values in kernel regression.

Chapter 6

Spatial Effects in cDNA Microarray Slides

6.1 Introduction

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 (Nguyen et al., 2002).

There are multiple possible causes for spatial trends on a slide within each step of a microarray experiment, not all of which are fully understood. However, the problem is similar to issues encountered in geostatistics and field experiments in that the data are arranged in a regular grid. In geostatistics a common approach to modeling spatial variation is a two sweep procedure (Cressie, 1993). The large scale trend is removed in the first sweep and the small scale vibrations are removed in the second sweep. In agricultural field experiments it is suggested that the large scale trends be modeled with splines and that time series models be used to remove vibrations (Cullis and Gleeson, 1991; Gilmour et al., 1997). A first order autoregressive (AR1) correlation is suggested as a suitable time series model for both rows and columns, with the overall correlation structure being the Kronecker product of the row and column correlation matrices (AR1 × AR1). An advantage of this approach is that it fits within the restricted maximum likelihood (REML) framework and estimates of the parameters of interest can be gained simultaneously with estimates of the nuisance spatial effect. Burgueño et al. (2005) use an AR1 × AR1 correlation structure with some success in spatial prediction for microarray slides. Baird et al. (2004) use AR1 × AR1 with splines in spatial linear mixed models to account for the nuisance spatial trends that arise in microarray experiments.

Many of the methods for spatial corrections in cDNA microarrays are done using the log ratio of red (R) to green (G) intensities. Yang et al. (2002) propose a lowess adjustment within each printing block group to account for intensity bias and spatial bias. The implicit spatial correction is removing the printing block effect. This approach is extended by Cui et al. (2003) by fitting a lowess curve for the ratio as a function of the mean intensities, row position and column position. They combine spatial and intensity adjustments in a lowess function. Spatial detrending is done by fitting a smooth surface to the intensity, $I = (log(R)+log(G))/2$ over the grid of spots defined by rows and columns. In fitting a surface, Cui et al. (2003) recognise that the spatial effect be represented as a smooth trend over the slide and that there be a local correction due to spot position. This comparison is analogous to comparing the spatial models of Cullis and Gleeson (1991), Gilmour et al. (1997) and Baird et al. (2004) with classical incomplete block analysis. A semi-parametric model is proposed by Fan et al. (2005) to account for spatial trends. This too relies on the assumption that hidden spatial effects can be represented as a smooth surface. All of the above studies apply to intensity ratios. Other authors (e.g. Jin et al. (2001) and Kerr and Churchill (2001b)) caution against complete reliance upon the ratios as they may mask important information. Jin et al. (2001) use the individual log intensities to allow direct comparisons of the different sources of variation. Kerr and Churchill (2001b) point out that not all relevant information can be captured by the ratio of intensities at a spot. This would imply that spatial adjustments on the ratios may not be ideal to capture all relevant information.

The sources of variation in a microarray slide may not act continuously, so may be fractal in dimension. Methods assuming separability and smoothness such as AR1 × AR1, splines and lowess adjustments may not be as efficient as other methods of removing spatial dependencies such as wavelet decomposition when the data are of fractal dimension. Huang and Cressie (2000) use a thresholding method of the discrete wavelet transform (DWT) to recover signal from noisy data, but little has been done with wavelets involving microarrays. A Gaussian filter of the Fourier transform of microarray images is used by Shai et al. (2003) to eliminate spatial trend, but wavelets have an advantage over the Fourier transform where there are discontinuities (Huang and Cressie, 2000).

In this chapter, a correction of spatial effects using the wavelet transformation is proposed. The multiresolutional nature of wavelets permits the estimation of both large and fine scale spatial effects simultaneously and accommodates the fractal nature of the data. After intensities are corrected for spatial effects, a set of over-expressed genes is identified. A comparison of different methods of spatial correction,

(i) wavelets and (ii) splines plus AR1 × AR1, is given and the sets of differentially expressed genes for each method are compared. Control genes are a set of genes that are not expected to be differentially expressed in an experiment. Thus, control genes allow us to estimate spatial variability within a slide. Since the choice of control genes will influence estimates of error and spatial trends, two schemes for selecting a set of control genes are explored. The first approach is to take a pre-determined set of genes to be control genes and the second approach is to use genes whose relative expression is approximately constant throughout the experiment as control genes.

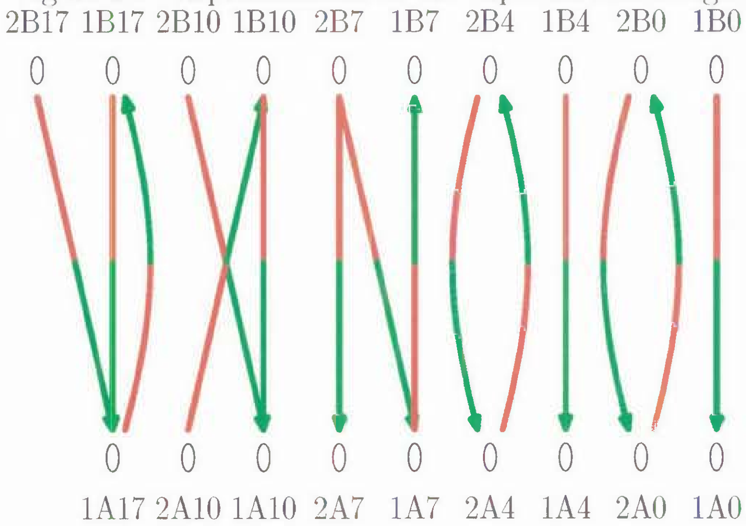
6.2 Materials and Methods

6.2.1 Data

6.2.1.1 First Phase of the Experiment

The data used in this study came from RNA extracted from mice livers in experiments conducted by Harry Noyes at The University of Liverpool. The treatments that were applied to the microarray were determined by strain of mice, challenge, replicate and time. Two strains of mice were used, AJ (A) and C57BL6 (B). The mice were further divided into challenges, challenge 1 and challenge 2. From combinations of mouse strain and challenge, two independent biological replicates were taken. 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 (each is an interaction between strain of mouse, challenge and time) applied to each, one coloured red and the other green. Figure 6.1 is a graphical representation of the experimental design (Kerr and Churchill, 2001a), where the nodes represent the cDNA and an edge between nodes indicates a slide. The edges are directed from the samples that were associated with the red dye, to the samples that were associated with the green dye. The nodes are labeled in the order challenge, mouse type and time point, so that for example 1A7 comes from challenge one, strain AJ and time 7 days.

Figure 6.1: Representation of the experimental design.



6.2.1.2 Second Phase of the Experiment

Each slide was divided into four metacolumns and twelve metarows, forming 48 printing blocks. 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, however there were 2,040 missing readings. The total number of genes in the experiment was 11,218, so clearly not all genes were printed on each array. However 14 of the 15 slides had an identical pattern of printed genes. In all, 73 pre-determined control genes were used. In this experiment the pre-determined control genes were printed in the first and last row of each metarow and the third row of each metarow contained some empty cells, which served the same function as control genes.

6.2.2 Statistical Models

In this section we outline two statistical models for interpreting the microarray data. The spatial mixed linear model has been proposed by Baird et al. (2004) drawing on the work from Wolfinger et al. (2001). We propose the wavelets model as an alternative. The response to be analyzed is $y = log_2(intensity)$.

The techniques used are variations on those that have been successfully applied in the analogous models for field experiments. These techniques seek an efficient estimate of the random components associated with correlated data so that inferences about the fixed effects are reliable and efficient. Efficiency is essential in gene expression data lest informative genes be unrecognised.

A hierarchical approach developed in the context of field experiments is to first fit the systematic part of the model and examine the residual components through variograms (Gilmour et al., 1997). The variogram identifies spatial components and the observations corrected for spatial positions. Thus the estimates of the true error components are improved by the removal of the nuisance spatial effects.

The first model considered is in the style explained by Gilmour et al. (1997) and Baird et al. (2004) in that spatial effects are estimated using splines. This model assumes that the spatial correlation can be represented as a smooth, separable process. That is, the correlation matrix $\mathbf{R}$ can be written in the form:

$$\mathbf{R} = \mathbf{R}_c \otimes \mathbf{R}_r,$$

where $\mathbf{R}_c$ and $\mathbf{R}_r$ are the correlation matrices for the smooth underlying one dimensional column and row processes. The second model does not assume that the spatial correlation can be expressed as the Kronecker product of two smooth processes.

6.2.2.1 Spatial Mixed Linear Model

A spatial mixed linear model (Wolfinger et al., 2001; Baird et al., 2004) is fitted for each colour on each slide. The model for $\mathbf{y}$ is:

$$\mathbf{y} = \left[\begin{array}{c|c|c} \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \end{array}\right] \left[\begin{array}{c} \tau_1 \\ \hline \tau_2 \\ \hline \tau_3 \end{array}\right] + \left[\begin{array}{c|c} \mathbf{Z}_1 & \mathbf{Z}_2 \end{array}\right] \left[\begin{array}{c} \mathbf{u}_1 \\ \hline \mathbf{u}_2 \end{array}\right] + \mathbf{E}$$

where

- τ_1 contains the fixed site effects.
- τ_2 contains the linear row and column effects.
- τ_3 contains the effects of the control genes.
- $\mathbf{u}_1$ is a random vector which contains the test gene effects and is a sample from a normal distribution.
- $\mathbf{u}_2$ contains spline terms.
- $\mathbf{E}$ is the grand error term, which can be decomposed as $\mathbf{E} = \xi + \epsilon$, where ξ is a spatially dependent error term and ϵ is the pure error term.

Furthermore:

$$\begin{pmatrix} \mathbf{u_1} \\ \mathbf{u_2} \\ \xi \\ \epsilon \end{pmatrix} \sim \mathbf{N} \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{G_1}(\gamma_1) & 0 & 0 & 0 \\ 0 & \mathbf{G_2}(\gamma_2) & 0 & 0 \\ 0 & 0 & \mathbf{R}(\phi) & 0 \\ 0 & 0 & 0 & \theta\mathbf{I} \end{pmatrix} \right].$$

The model is fitted with:

(i) no spatial correlation assumed (Splines), so that:

$$\mathbf{R}(\phi) = \mathbf{I}$$

and

$$\mathbf{G_1}(\gamma_1) = \lambda\mathbf{I},$$

(ii) auto-correlated rows and auto-correlated columns (AR1 × Splines), so that:

$$\mathbf{R}(\phi) = \mathbf{R}_c \otimes \mathbf{R}_r$$

and

$$\mathbf{G_1}(\gamma_1) = \lambda\mathbf{I},$$

with the symmetric matrix:

$$\mathbf{R}_c = \begin{pmatrix} 1 & & & & \\ \rho_c & 1 & & & \\ \rho_c^2 & \rho_c & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \rho_c^{n_c-1} & \rho_{n_c}^{n_c-2} & \rho_c^{n_c-3} & \dots & 1 \end{pmatrix}$$

and $\mathbf{R}_r$ is similarly defined for correlation between rows. That is, there is a one step correlation between rows and a one step correlation between columns. Fixed linear row and column effects are also fitted in the AR1 × Splines model.

Once the spatial trends are found, they are subtracted from the original data to give adjusted intensities $\mathbf{y}^{\mathbf{a}}_{\mathbf{Splines}}$ and $\mathbf{y}^{\mathbf{a}}_{\mathbf{AR1\times Splines}}$.

A mixed linear model that ignores spatial effects is also fitted as a benchmark. That is, a model of the form $\mathbf{y} = \mathbf{X}\tau + \mathbf{Zu} + \epsilon$, is fitted where τ is the vector of the fixed effects site and control genes and u contains the test gene effects. As before, $\epsilon \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, $u \sim \mathrm{N}(\mathbf{0}, \lambda\mathbf{I})$ and $\mathbf{u}$ and ϵ are orthogonal.

6.2.2.2 Wavelets

The idea here is to perform a discrete wavelet transform (DWT) to the data and threshold the coefficients in the wavelet domain to remove the spatial trend. The Daubechies wavelet basis function (Daubechies, 1988) is chosen for the DWT. The multiresolutional attributes of wavelets allow decomposition of the spatial components in the frequency domain (Huang and Cressie, 2000). Wavelet coefficients below a threshold correspond to noise and the larger scale wavelet coefficients correspond to the spatial trend. Thus the larger wavelet coefficients corresponding to spatial trends could be removed in the wavelet domain and the remaining coefficients transformed back to the original scale. That is, fit a model of the form:

$$\mathbf{Dy} = \mathbf{DX}\tau + \mathbf{DZu} + \mathbf{D}\xi + \mathbf{D}\epsilon,$$

where τ is the vector of the fixed effects due to site and control gene, $\mathbf{u}$ is the vector of test genes random effects, ξ is the spatial trend and ϵ is the pure error term. The matrix representation of the discrete wavelet transform is $\mathbf{D}$, which is a two dimensional version of the pyramid decomposition algorithm (Mallat, 1989; Meyer, 1993). The best linear unbiased estimator (BLUE) of τ, the best linear unbiased predictor (BLUP) of $\mathbf{u}$ and the spatial components are to be found. However, the structures of $\mathbf{DX}\tau$, $\mathbf{DZu}$ and $\mathbf{D}\xi$ cannot be simultaneously determined in the wavelet domain with certainty (Huang and Cressie, 2000). Hence, the BLUE for τ, $\hat{\tau}$ and the BLUP $\tilde{\mathbf{u}}$ for $\mathbf{u}$ are found on the original scale assuming no spatial dependency and subtracted from the original data, to yield the grand error term, $\mathbf{E}$.

$$\mathbf{E} = \mathbf{y} - \mathbf{X}\hat{\tau} - \mathbf{Z}\tilde{\mathbf{u}}.$$

The DWT is applied to $\mathbf{E}$ and thresholding of the wavelet coefficients is performed for each colour on each slide to separate the spatial trend from random noise.

$$\mathbf{DE} = \mathbf{D}\xi + \mathbf{D}\epsilon.$$

A hard thresholding method is used. If $\mathbf{E}^*$ are the wavelet coefficients of $\mathbf{E}$ after the DWT is applied, then the hard thresholding value is given by:

$$T(E^*_{j,k_1,k_2}) = \begin{cases} E^*_{j,k_1,k_2}; & \text{if } |E^*_{j,k_1,k_2}| > \lambda; \\ 0; & \text{if } |E^*_{j,k_1,k_2}| \leq \lambda. \end{cases}$$

The choice of thresholding parameter, λ is the universal threshold parameter given by Donoho and Johnstone (1994), $\sigma\sqrt{2\log n}$, where σ is the noise parameter and n is total number of points in the image. Wavelet coefficients above the threshold are considered to correspond to a spatial trend and smaller coefficients to noise. The correction of the nuisance spatial effects is done by thresholding the wavelet coefficients, applying the inverse DWT and subtracting that entity from the raw data:

$$\mathbf{y}^{\mathbf{a}}_{\mathbf{wavelet}} = \mathbf{y} - \mathbf{D}^{-1}(T(E^*_{j,k_1,k_2})).$$

6.2.3 The Sample Variogram

A sample variogram of ϵ is computed and plotted for each model to compare the models. The variogram ordinates are g.ven by $\gamma_{ij} = \frac{1}{2}[\epsilon_i - \epsilon_j]^2$ (Cressie, 1993). The sample variogram is the triple $(|\tau_i^r - \tau_j^r|, |\tau_i^c - \tau_j^c|, v_{ij})$, where τ_i^r is the row position of the ith point, τ_i^c is the column position of the ith point and v_{ij} is the mean of the variogram ordinates with a distance of $(|\tau_i^r - \tau_j^r|, |\tau_i^c - \tau_j^c|)$, between them. A flat variogram with a low plateau illustrates that the model has efficiently removed spatial trends. The sample variogram is also calculated as a function of the absolute distance between points, $v(|\tau|)$.

6.2.4 Control Genes

The model fitted so far has used pre-specified control genes to indicate the spatial variation. It could be argued that the control genes should be determined by the data rather than a priori (Schadt et al., 2001; Tseng et al., 2001; Speed, 2003). In this section the robustness of spatial adjustments using the Wavelets and AR1 × Splines models is investigated when the control genes are (a) chosen a priori and (b) determined from the data as being stable with rank invariant selection. Under the rank invariant selection scheme, the rank for each gene is found independently for each sample on a slide. If the gene is not differentially expressed, then the rank of the gene should be relatively unchanged between samples. Thus, genes whose rank varies by less than a threshold value, d, on a given slide are taken and the intersection of all such genes on all slides are the new control genes.

6.2.5 Analysis of Variance

Before an analysis of variance is conducted the intensity dependent dye bias is removed. This is done for each array by fitting a lowess curve to the RI plot which is a plot of the $\log_2$ intensity ratio of red to green ('R') against the mean log intensity ('I') (Yang et al., 2002). The lowess curve is added to the $\log_2$ of the green intensities so that the intensity dependent dye bias is accounted for.

A typical ANOVA model for microarray data is (Kerr and Churchill, 2001a,b):

$$\mathbf{y^a_{ijkl}} = \mu + \mathbf{A_i} + \mathbf{C_j} + \mathbf{T_k} + \mathbf{G_l} + \mathbf{(AG)_{il}} + \mathbf{(CG)_{jl}} + \mathbf{(TG)_{kl}} + \epsilon_{\mathbf{ijkl}},$$

where A is the array effect, C is the colour (dye) effect, G is the gene effect and T is the treatment effect. The gene by treatment interaction is the effect of interest. It is assumed here that genes do not interact with colour or array. The treatment effect is confounded with the colour by array effect, so the treatment effect is also dropped from the model.

The model fitted is:

$$\mathbf{y^a_{ijkl}} = \mu + \mathbf{A_i} + \mathbf{C_j} + \mathbf{G_l} + \mathbf{(AC)_{ij}} + \mathbf{(TG)_{kl}} + \epsilon_{\mathbf{ijkl}}.$$

The ANOVA is conducted on the data adjusted by the Wavelets and the AR1 × Splines models.

6.2.6 Separability and Fractals

For a separable field, the covariance matrix, $\mathbf{\Sigma}$ can be expressed as $\mathbf{\Sigma} = \mathbf{\Sigma_x} \otimes \mathbf{\Sigma_y}$, where $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_y}$ are the variance matrices of the underlying one-dimensional processes (Adler, 1981; Scaccia and Martin, 2005). Such a separation ensures that the determinant and inverse of $\mathbf{\Sigma}$ can be easily calculated because of the properties of the Kronecker product. Thus, the assumption of separability makes the iterative maximum likelihood calculations relatively fast.

The fractal dimension gives an indication of the roughness of the signal (Mandelbrot and Van Wess, 1968; Adler, 1981; Constantine and Hall, 1994). If the signal is smooth, the fractal dimension is an integer. In the case of the microarray data at hand, a fractal dimension of 2 would indicate a smooth surface and any significant departure from 2 would indicate that the data are fractal in nature.

A stationary Gaussian field, $f(x)$, is assumed, with a position independent correlation structure, ρ, so that the correlation between two points is a function of the distance (τ) between the points:

$$\rho(\tau) = E[f(x)f(x+\tau)]. \tag{6.1}$$

For such surfaces, the correlation function behaves like $1-\rho(\tau) = |\tau|^{\alpha}$, for small τ, $0 < \alpha \leq 2$ and the fractal dimension of the surface is given by (Adler, 1981; Gneiting and Schlather, 2004):

$$D = 3 - \frac{\alpha}{2}.$$

If the signal is rough, separability may not be appropriate for the data. The correlation function in equation (6.1) can be modeled as:

$$\rho(\tau) = e^{-|\tau|^{\alpha}},$$

with $0 < \alpha \leq 2$. If a separable model is to be used, it is required that:

$$\rho(\tau_r, \tau_c) = \rho(\tau_r) \times \rho(\tau_c),$$

where τ_r and τ_c are the row and column displacements respectively. That is:

$$e^{-\sqrt{|\tau_r|^2+|\tau_c|^2}^{\alpha}} = e^{-(|\tau_r|^{\alpha}+|\tau_c|^{\alpha})}.$$

These correlation functions cannot be equal unless $\alpha = 2$, or equivalently unless the Gaussian field to be modeled is smooth, with fractal dimension 2.

6.2.7 Calculating Fractal Dimension

The fractal dimension of all 15 slides is calculated for both the red and green treatments by box counting, a variogram-based method and using the `fdim` (de Pison Ascacibar et al., 2000) package in R (R Development Core Team, 2004).

6.2.7.1 Box Counting Estimator

If a set, S, can be completely covered with a minimum number of cubes, $N_\tau(S)$ of side length τ, then the box counting fractal dimension is given by (Falconer, 1990):

$$D_{BC} = \lim_{\tau\to 0} \frac{ln(N_\tau(S))}{-ln(\tau)}.$$

In order to calculate the box counting fractal dimension of each microarray image, a triple of each row position, column position and intensity is formed. Each microarray slide is represented as an image in $\mathbb{R}^3$. The minimum number of cubes to completely cover the image is found for a variety of cube sizes and the logarithm of the cube size is plotted against the logarithm of the number of cubes required. The slope of the plots for small τ gives the estimates of the box counting fractal dimension, $\hat{D}_{BC}$. However Hall and Wood (1993) have shown the box counting estimator to be biased.

6.2.7.2 Variogram-based Estimator

The sample variogram, $\gamma|d|$, can be used to estimate the exponent α in the function for fractal dimension (Constantine and Hall, 1994). Assuming that $\gamma|d| \propto |d|^{\alpha}$ as $d \rightarrow 0$, then,

$$ln(\gamma|d|) = \text{constant } + \alpha ln(d) + \epsilon.$$

Thus, the slope of the plot of $ln(d)$ and $ln(\gamma|d|)$ for small d is used to find an estimate of α and then an estimate of D is given by:

$$\hat{D}_{VB} = 3 - \frac{\hat{\alpha}}{2}.$$

6.3 Results

6.3.1 Results of Spatial Models

Figure 6.2 shows the spatial component and pure error term for the Splines, AR1 × Splines and Wavelets models for slide 15. This slide has one of the more apparent spatial trends. It is seen that the Wavelets method accounts for more of the systematic spatial trend than the Splines and AR1 × Splines methods. In particular, the AR1 × Splines and Splines models do a poor job of accounting for the irregularity in the printing block centered at row 13 and column 8. The sample variograms in Figure 6.3 confirm that the pure error term for the Wavelets model has the least spatial trends. These variograms are typical of these data.

Figure 6.2: Decomposition of the grand error term, **E**, for red, slide 15 with the (a) Splines, (b) AR1 × Splines and (c) Wavelets models.

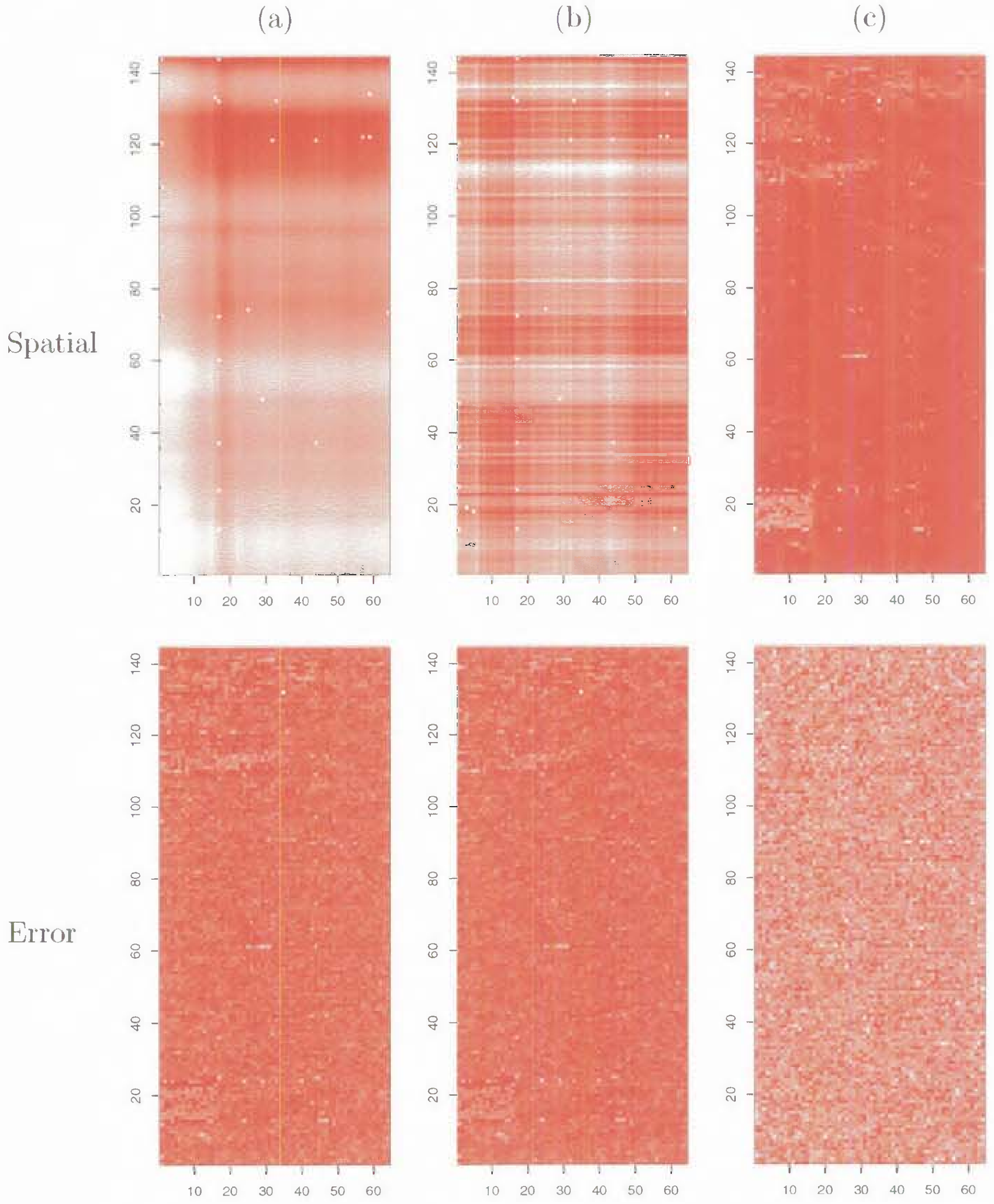
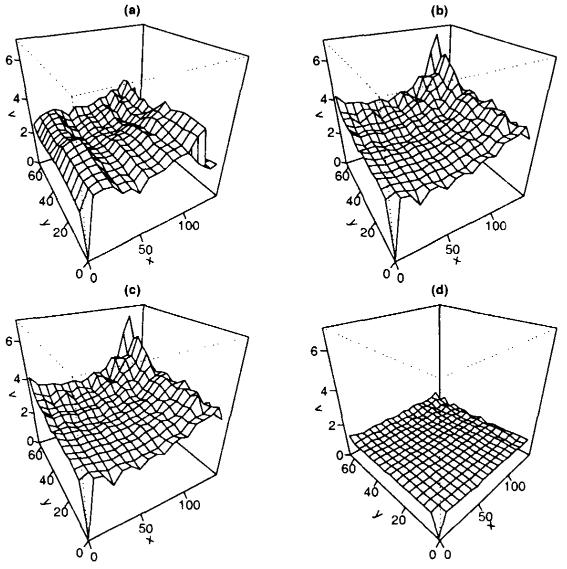


Figure 6.3: Sample variograms for red, slide 15 with (a) No spatial trend modeled. (b) Splines. (c) AR1 × Splines. (d) Wavelets.

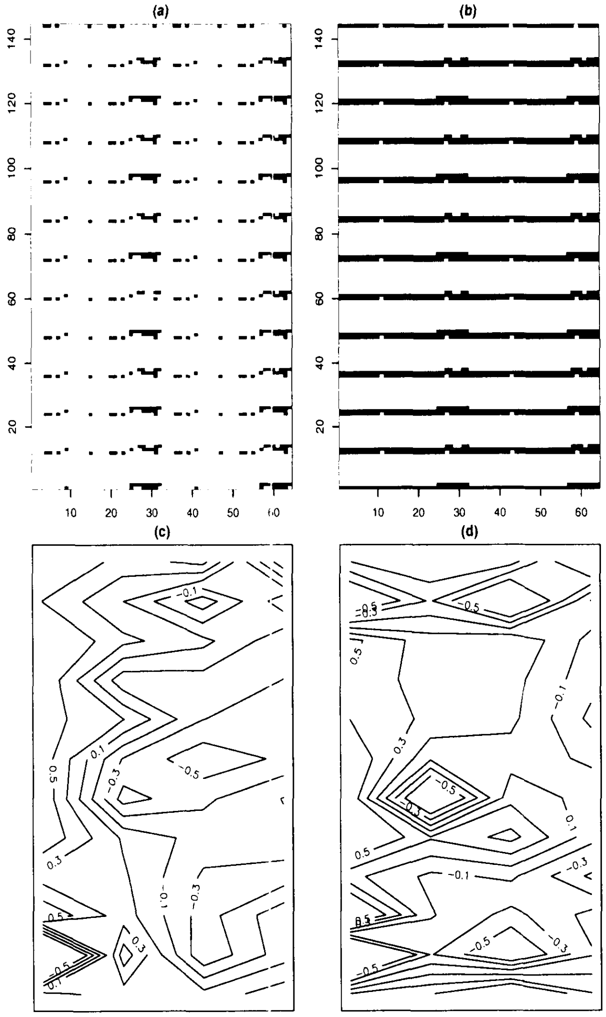


6.3.2 Control Genes

The rank invariant selection scheme approach finds 11 and 13 control genes when the Wavelet model and AR1 × Splines models are used respectively. All of these genes are in the set of 73 pre-determined control genes and nine of these genes are in both sets found by the rank invariant selection scheme.

Figures 6.4(a) and (b) show where the control genes are spotted on slide 15 when these genes are determined from the data with both the Wavelets model and a priori methods of determining housekeeping genes. There are significantly fewer data-determined control genes than pre-specified control genes and this is reflected in the contour plots in Figures 6.4(c) and (d). The two contour plots are similar in shape, but the contour plot for the pre-specified control genes shows a much more sloping surface to the contour plot for the data-determined control genes.

Figure 6.4: (a) Location of control genes found with the Wavelets model. (b) Location of the pre-specified control genes. (c) Contour plot of the spatial trend for red, slide 15 using the control genes identified using the Wavelets model. (d) Contour plot of the spatial trend for red, slide 15 using the pre-determined control genes.



6.3.3 Fractal Dimension

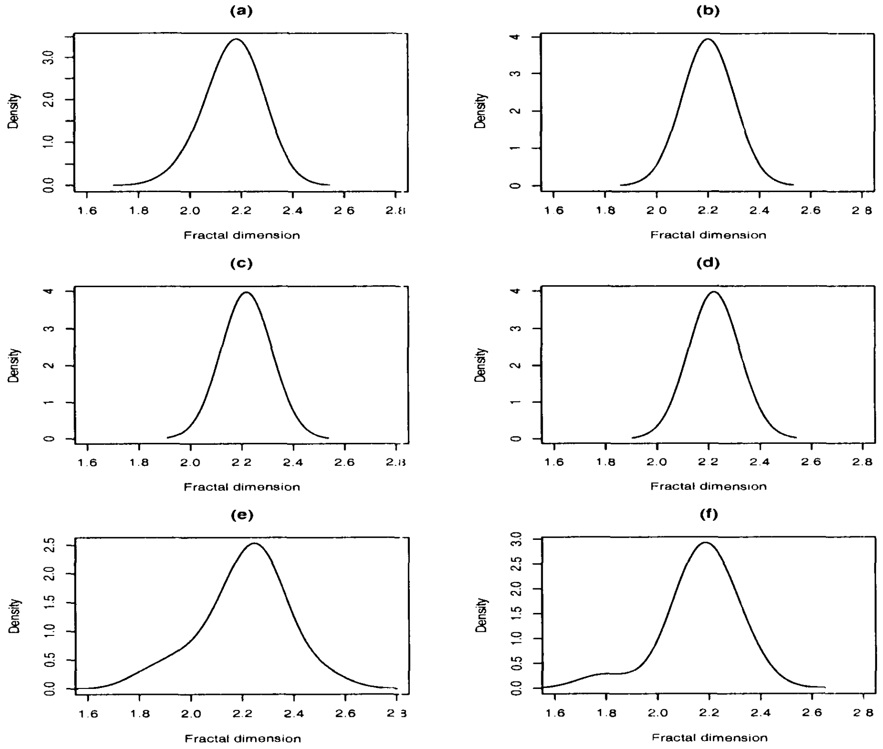
The distribution of estimated fractal dimensions for each colour and each method is shown in Figure 6.5. The mean fractal dimension of the red (green) images for the fdim estimator, Box counting method and variogram-based estimator are 2.17 (2.20) ± 0.05, 2.22 (2.22) ± 0.01 and 2.21 (2.17) ± 0.14 respectively. There is a clear departure from 2 in fractal dimension for the means of all the methods and the means for all of the methods are tightly grouped. The estimates for the variogram-based estimator have a high variance because there were few points in the variogram with a distance small enough for the slope near zero to be calculated.

6.3.4 Analysis of Variance

Intensity dependent colour bias is evident in these data. Figure 6.6(a) illustrates that for slide 1 the 'R' and 'I' values are not independent, with the lowess curve being non-linear. The shape of this lowess curve is typical of these data. Figure 6.6(b) shows the data after the lowess correction.

Genes that are ranked in the top 10 gene by treatment interactions for each treatment are taken as genes of interest. This results in 96 genes being taken in the set of genes of interest for the Wavelets model and 80 genes being taken for the AR1 × Splines model. There are 64 genes contained in both sets. Systematic effects of arrays, genes, treatments and colour account for 62% of the total variation in the ANOVA model with spatial variance adjusted by wavelets. The counterpart ANOVA of AR1 × spline corrected spatial effects attributes 63% of the variation to these systematic effects.

Figure 6.5: Distributions of fractal dimension estimates of all 15 slides using (a) The fdim package for the red images. (b) The fdim package for the green images. (c) Box counting, red images. (d) Box counting, green images. (e) Variogram-based estimator, red images. (f) Variogram-based estimator, green images.



The way in which control genes are determined influences the estimate of gene expression level. Figure 6.7 shows the expression levels of two genes under different treatments when the different methods of removing spatial trend and selecting control genes are used. In both genes there is little difference in estimated expression for the AR1 × Splines method when the different sets of control genes are used. However the Wavelets method is somewhat sensitive to a change in control genes.

Figure 6.6: RI plots for slide 1 (a) before the lowess correction and (b) after the lowess correction.

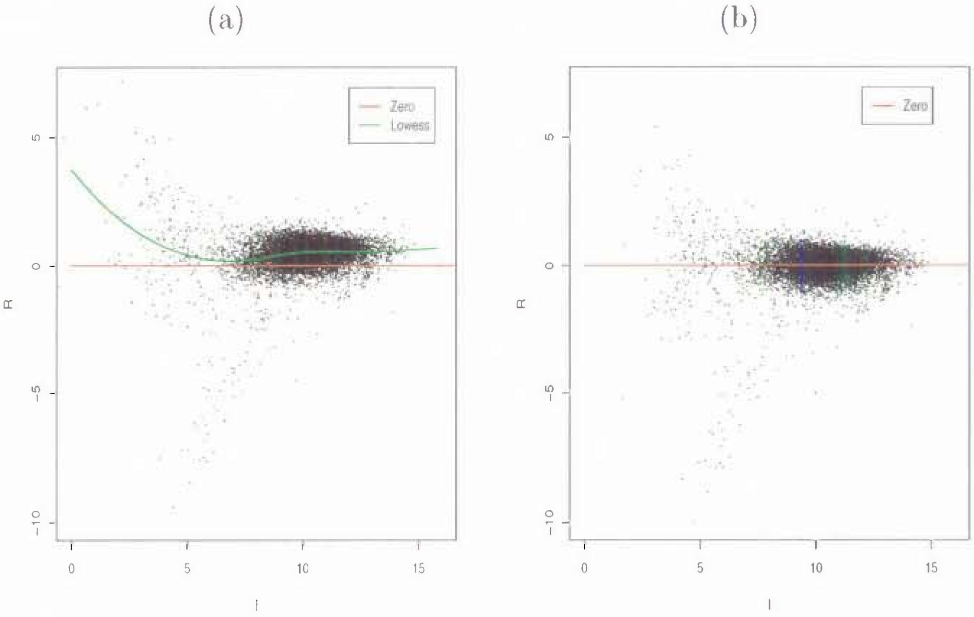
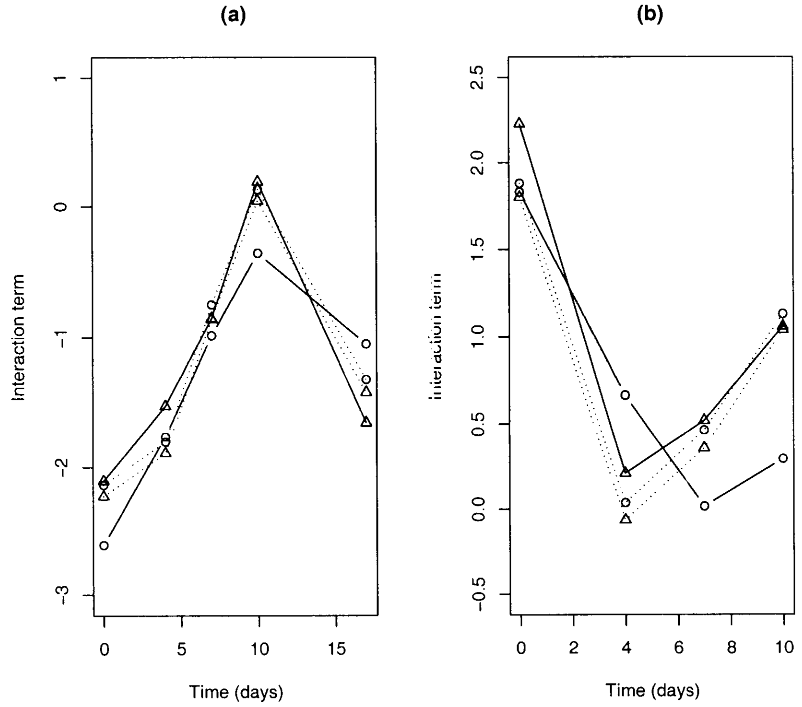


Figure 6.8 shows which of the 64 genes that are identified as genes of interest from both the Wavelets and AR1 × Splines models are estimated to be over- or under-expressed for each treatment after the ANOVA model is fitted. There are clear differences between the two models, with different genes identified as being differentially expressed. In particular, the AR1 × Splines model identifies many genes for time 4 days, breed B, strain 2 that are not identified after the Wavelets model is used. Similarly, the Wavelets model identifies genes at time 0 days, breed B, strain 1 that are not found to be differentially expressed when the AR1 × Splines model is used to remove spatial bias. There are very few treatments where both the Wavelets and the AR1 × Splines models identify the same differentially expressed genes.

Figure 6.7: △ Pre-determined control genes. ○ Data-determined control genes. ··· AR1 × Splines. —— Wavelets.

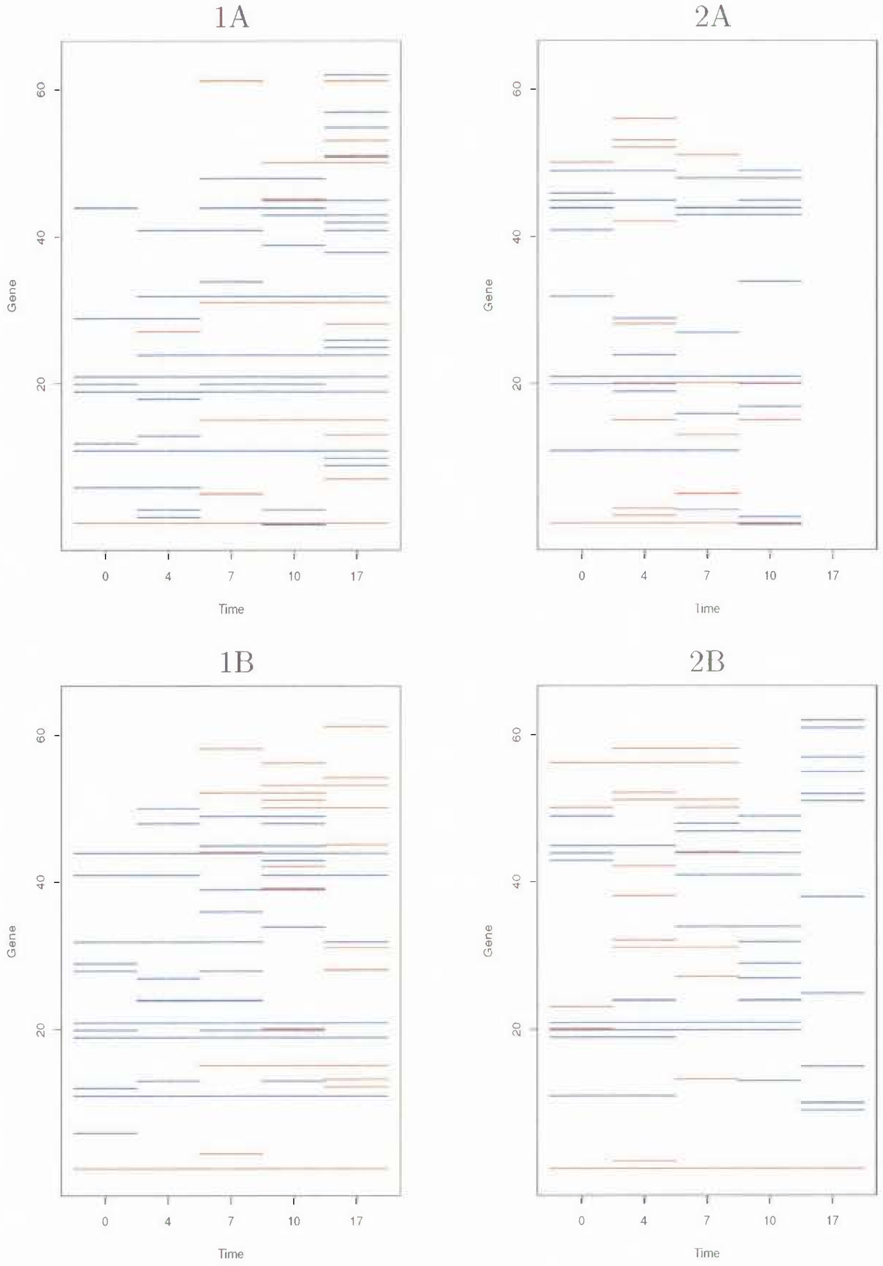
(a) Interaction of Gene NM_007414 with challenge one, strain AJ. (b) Interaction of Gene AB041540 with challenge two , strain C57BL6.



6.4 Discussion

The difference in the state space diagram and descrepancies between estimates of gene expression level between the methods used to model spatial trends 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 6.3, it can be seen that wavelets are much more successful in removing the spatial trend than the other methods for these data. The AR1 × Splines and Splines models are very similar or marginally better than fitting no spatial trend.

Figure 6.8: State space diagram for each challenge (1 and 2) and strain (A and B) combination for the 64 genes that are identified as genes of interest after the — Wavelets and — AR1 × Splines models are used to remove spatial trends. A line indicates that the gene is over- or under-expressed at the corresponding time point.



The selection of control genes is important when removing spatial trends in microarray data. In these data it is found that the AR1 × Splines model is reasonably robust against the method of selecting control genes, whereas there are large changes in estimated expression levels when the Wavelets method is used.

There are large differences in the resulting analysis of variances when the data are adjusted by the AR1 × Splines model and the Wavelets model. The way that spatial trend is modeled has a large impact on the genes identified as genes of interest for subsequent studies.

Mathematically, the wavelet analysis is not disparate from the spatial models proposed by Baird et al. (2004). In both cases, the underlying spatial regression variables are tensor products of basis functions, i.e. $F_{row} \otimes F_{columns}$. Whereas cubic splines basis functions are associated with smooth functions, Daubechies wavelet bases accommodate fractal and discontinuous trends. The representation of small scale 'vibrations' by AR1 models play the same role as the high frequency component of the wavelet analysis. 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. These properties of the wavelet transform will be further examined in the following chapter using simulated data.

Chapter 7

A Simulation Study of Spatial Effects in cDNA Microarray Slides

7.1 Introduction

There are many sources of variation in microarray studies that need to be accounted for so that gene expression estimates are accurate. One such source of variation is spatial trends that can accumulate on the microarray slides. This spatial variation may arise from disparities between printing blocks, pin effects and uneven washing of the solution over the slide. These spatial trends can significantly change the biological conclusions of an experiment depending on how they are modeled (see Chapter 6, for example). Hence, it is important to test methods of removing spatial bias in microarray slides so that a method can be recommended for a particular slide depending on its properties. Complexities in the procedures and high costs of microarrays mean that performing real microarray experiments to test methods of removing bias is not always viable. Simulation is a practical way to test potential strategies in processing microarray experiments. Previous simulation studies consider sources of variation such as background noise, expression signals, spot location, spot shape and irregularities on the slide (Wierling et al., 2002; Balagurunathan et al., 2002). Simulation of spatial trends in the single channel background adjusted signal will be considered here.

In this chapter, results from a previous murine study (Chapter 6; Woolaston et al., 2005) are used to generate data to assess the effectiveness of three methods of modeling spatial noise in microarray slides. Spatial effects of varying roughness or fractal dimension are also simulated. The methods to be compared are (i) No spatial modeling (Basic), (ii) Splines with autocorrelated rows and autocorrelated columns (AR1 × Splines) and (iii) A discrete wavelets threshold model (Wavelets).

7.2 Spatial Models to be Compared

A summary of the 3 spatial models to be compared is given here. More details of these spatial models can be found in Chapter 6.

7.2.1 Basic Model

The basic model is a mixed model of the form $\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \epsilon$, $\epsilon \sim (0, \sigma^2)$, $E(u.\epsilon) = 0$, $u \sim N(0, \sigma_u^2)$, containing no spatial terms with only the control gene and test gene effects fitted. Control genes are fitted as fixed effects and test genes as random effects. ASReml (Gilmour et al., 2000) was used to fit the model.

7.2.2 AR1 × Splines

The AR1 × Splines model is similar to the basic model, with additional terms added. The fixed effects are due to the control genes, printing blocks, rows and columns. The random component contains the test gene effects and spline terms with 5 knot points in each direction to model the long term spatial variation. As in the basic model the test gene effects are assumed to have come from a $N(0, \sigma_u^2)$ distribution. First order autocorrelated rows and first order autocorrelated columns (AR1 × AR1) are also fitted to account for small scale vibrations in the spatial trend. Once again, ASReml (Gilmour et al., 2000) was used to fit the model.

7.2.3 Wavelets

The Wavelets model is a two step process. The residuals from the basic model are transformed to the wavelet domain via the discrete wavelet transform (See for example Liò, 2003). Large wavelet coefficients are considered to correspond to spatial trend and small coefficients to noise (Huang and Cressie, 2000). Hence the transformed data is thresholded in the wavelet domain to find the spatial trend. The discrete wavelet transform and thresholding was performed using the *wavethresh* package (Nason, 2004) as part of the statistical program R (R Development Core Team, 2004). The threshold value used is $\sigma\sqrt{2\log(n)}$ (Donoho and Johnstone, 1994), where n is the number of data points and σ is the noise parameter. The spatial trend found by this approach is transformed back to the original domain and subtracted from the original intensities to give intensities adjusted for spatial trend. The basic model is then fitted to these adjusted data.

7.3 Methods of Simulation

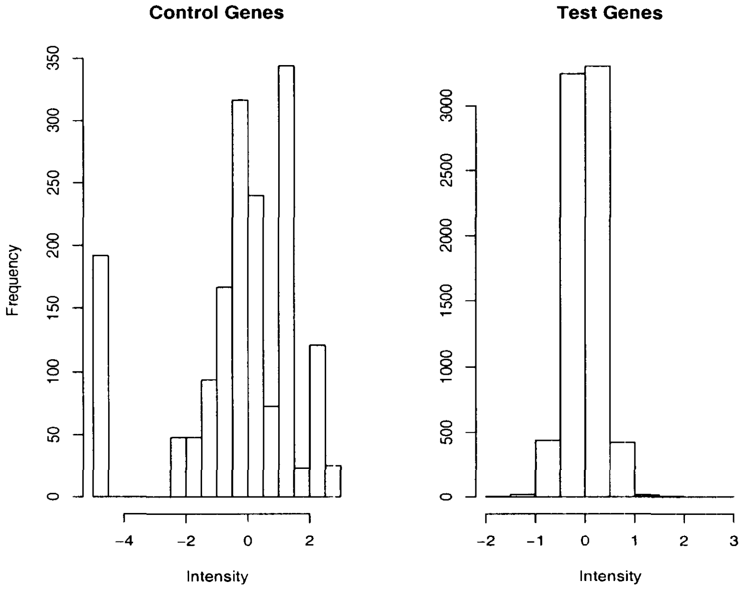
The simulation was a two step procedure, with all values assumed to have undergone a $\log_2$ transformation. First the gene effects are simulated and these values are randomly applied to an array. Secondly, the spatial noise is simulated and applied to the array. The simulation of spatial noise was simulated in two ways: (i) spatial noise taken to be estimated spatial noise from a real experiment (ii) spatial noise entirely simulated.

7.3.1 Simulation of Gene Effects

The simulated control and test gene effects were the estimated effects from a real microarray experiment. In the murine study, 2 strains of mice were infected with 2 challenges of a disease. Biological samples were applied to the arrays at time points 0, 4, 7, 10 and 14 days. There were 14 slides taken from the experiment. Each slide had 7,444 test genes spotted once per array and 45 control genes spotted between 23 and 192 times per array. The frequency of the simulated intensities that were added to the simulated spatial noise is displayed in Figure 7.1. The most under-expressed simulated test gene had an intensity of -4.57 and the most over-expressed test gene had a simulated intensity of 2.95. All other test genes have a simulated value between -2.1 and 1.9.

The majority of the simulated control genes had an intensity between -2 and 2, however there was one gene spotted 192 times per array with an intensity of -5. This corresponded to empty cells in the original experiment.

Figure 7.1: Simulated intensities of the genes.



7.3.2 Spatial Noise Taken From Real Dataset

The results from a previous study were used to obtain the simulated background noise and the test and control gene effects. Wavelets and the AR1 × Splines approach were used separately in the previous study to remove the spatial trends. Since using the spatial trend found in one model from the results of the previous study in this simulation would give that model an advantage in removing spatial bias, both sets of results were used. For each of the 14 analyzed slides, the test and control gene effects were randomly arranged over the spatial noise found by the respective model for accounting for spatial patterns in the arrays. Each slide was then analyzed with Wavelets, AR1 × Splines and the Basic models. The random application of effects and analysis was repeated 30 times per slide.

7.3.3 Simulation of Spatial Noise

A combination of systematic effects and real valued random Gaussian fields were used to simulate the spatial variation in a microarray slide.

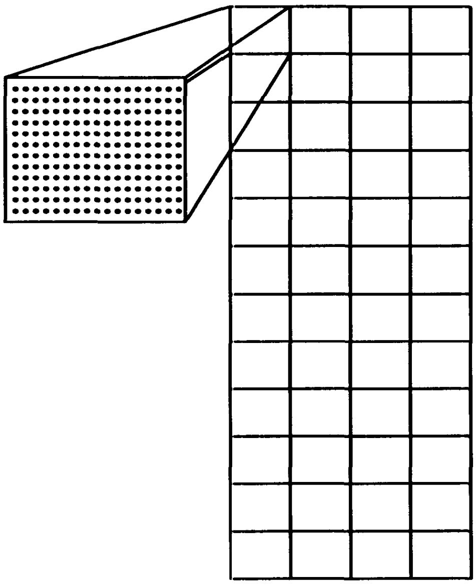
7.3.3.1 Printing Block and Pin Effects

The design of the arrays is illustrated in Figure 7.2. There were 48 printing blocks arranged in a 4 × 12 grid. Within each printing block there were spots arranged in a 16 × 12 grid. Each of these spots was associated with the corresponding pin in the printing block. Printing block effects were modeled as $s_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, 3, ...48$, where μ_i was sampled from a $N(0, \sigma^2)$ distribution and σ_i^2 was sampled from an Inverse Gamma distribution with shape parameter ν and scale parameter s. Pin effects were modeled as $p \sim N(0, \sigma_p^2)$. The values of ν, s, σ^2 and σ_p^2 were estimated from the real dataset and were and 20, 0.083, 0.5 and 0.4 respectively.

7.3.3.2 Spatial Gaussian Effects

A stationary Gaussian process, $Z(s)$, was assumed with a variety of covariance structures used to simulate spatial noise found in microarray slides that was not accounted for by printing block effects or pin effects. A Gaussian random process, $Z : D \rightarrow \mathbb{R}, D \subset \mathbb{R}^2$, was used to generate realizations of the Gaussian process, $Z(\mathrm{s}), s \in (1, 2, ..., 144) \times (1, 2, ..., 64)$, on a 144 × 64 grid.

Figure 7.2: The layout of printing blocks and printed spots for the simulated array.



$$Z(\mathrm{s}) = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,63} & z_{1,64} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,63} & z_{2,64} \\ \vdots & \vdots & & \vdots & \vdots \\ z_{143,1} & z_{143,2} & \cdots & z_{143,63} & z_{143,64} \\ z_{144,1} & z_{144,2} & \cdots & z_{144,63} & z_{144,64} \end{pmatrix}$$

The matrix $Z(\mathbf{s})$ can arranged into a vector $Z_v(\mathbf{s})$ with the *vec* operator so that the appropriate 9216×9216 covariance matrix, $\mathbf{C}$, can be formed for $Z_v(\mathbf{s})$.

$$Z_v(\mathbf{s}) = [vec\{Z(\mathbf{s})\}] = \begin{pmatrix} z_{1,1} \\ z_{2,1} \\ \vdots \\ z_{144,1} \\ z_{1,2} \\ z_{2,2} \\ \cdots \\ z_{144,2} \\ \vdots \\ z_{143,64} \\ z_{144,64} \end{pmatrix} .$$

Since a stationary Gaussian field is assumed, the covariance function, c, depends only on the distance, τ, between points:

$$c(z_{ij}, z_{kl}) = c(\sqrt{(i-k)^2 + (j-l)^2}) = c(|\tau|)$$

The covariance matrix can be written as:

$$\mathbf{C} = E(Z_v(\mathbf{s})Z_v(\mathbf{s})^T) = \begin{pmatrix} c(0) & c(1) & \cdots & c(63) & c(1) & c(\sqrt{2}) & \cdots & c(\sqrt{63^2+143^2}) \\ c(1) & c(0) & \cdots & c(62) & c(\sqrt{2}) & c(1) & \cdots & c(\sqrt{62^2+143^2}) \\ \vdots & & \ddots & & & & & \vdots \\ c(63) & & & \ddots & & & & \vdots \\ c(1) & & & & \ddots & & & \vdots \\ c(\sqrt{2}) & & & & & \ddots & & \vdots \\ \vdots & & & & & & \ddots & \vdots \\ c(\sqrt{63^2+143^2}) & & \cdots & \cdots & \cdots & \cdots & \cdots & c(0) \end{pmatrix} .$$

Once the covariance matrix is formed it can be decomposed via the Cholesky decomposition:

$$\mathbf{C} = \mathbf{L}\mathbf{L}^T.$$

Then $Z(\mathbf{s})$ is generated by:

$$Z(\mathrm{s}) = \mathbf{L}\mathbf{X},$$

where $\mathbf{X}$ is a vector from a $N(0, \mathbf{I})$ distribution. It is easily shown that $Z(\mathbf{s})$ has the desired properties (i.e. $E(Z(s)) = 0$ and $var(Z(s)) = \mathbf{C}$) and all that is required is to choose the covariance function of the spatial Gaussian noise.

Two different correlation functions were used separately to generate the spatial Gaussian noise, (a) the stable correlation with the Cauchy correlation function to separate the short term and long term dependency and (b) the stable correlation function to model short term dependency only.

(a) Separate consideration of short term and long term behaviour. The simulated spatial noise in a microarray slide was decomposed into short term and long term behaviour:

$$\rho(\tau) = \rho_s + \rho_l.$$

The short term dependency was modeled with the stable correlation function:

$$\rho_s(\tau) = e^{-|\tau|^\alpha}, \tag{7.1}$$

with $0 < \alpha \leq 2$ and a variance of 1. The fractal dimension of a realization of such a field can be calculated as $D = 3 - \frac{\alpha}{2}$. The long range dependency was modeled with a stationary Gaussian field with a Cauchy correlation structure

$$\rho_l(\tau) = (1 + |\tau|^2)^{-\beta}, \beta > 0,$$

and a variance of 0.5. The parameter β controls the long range dependency.

(b) Short term dependency only. It may be appropriate to consider only short term dependency in the correlation function, since the site effects are accounting for some long range dependency. Once again the stable correlation function (equation (7.1)) is used to model the short term dependency, with a variance of 1.5.

For both of the correlation functions the parameter α took values between 1 (rough surface) and 2 (smooth surface) in increments of 0.1. Consequently, the simulated fractal dimension took values between 2 and 2.5 in increments of 0.05. Woolaston et al. (2005) found the mean fractal dimension for the data used in this study to be 2.2 (see also Chapter 6). There were 30 replicates for each value of α. For approach (a) the parameter β was set at 0.4, which was the mean value calculated from the real slides.

7.4 Results

7.4.1 Spatial Noise Taken From Real Dataset

There are significant differences between the three methods of removing spatial trend. Table 7.1 shows the mean of the standard deviation between the simulated and estimated intensity readings. The simulated intensities for genes and background noise in Table 7.1(a) are from the estimates from the real data with the AR1 × Splines model and Table 7.1(b) are the estimates obtained with the Wavelets model. In both cases it can be seen that the Wavelets method clearly outperforms the AR1 × Splines and Basic methods for slides with a high standard deviation between simulated and estimated intensities (for example slides 2,4,5,10 and 13). The corresponding real slides have particularly non-smooth surfaces.

In Table 7.1(a) it is seen AR1 × Splines model has more success than the Wavelets model for slides 1 and 3. These are particularly smooth slides. The Basic model does not predict the simulated intensity of genes as successfully as the Wavelets model on most slides, however in some cases the Basic model does outperform the Wavelets and/or AR1 × splines models.

Table 7.1: Standard deviation of the difference between simulated and estimated intensity reading. The simulated background noise is found with (a) the AR1 × Splines model and (b) the Wavelets method from a previous experiment. All standard errors were 0.01 or less.

	Slide													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	(a) AR1 × Splines													
AR1×Spl	0.11	0.38	0.16	0.36	0.47	0.31	0.32	0.12	0.15	0.40	0.36	0.18	0.52	0.35
Wavelets	0.14	0.35	0.17	0.33	0.43	0.30	0.30	0.13	0.15	0.38	0.34	0.18	0.46	0.32
Basic	0.13	0.37	0.17	0.36	0.47	0.30	0.31	0.12	0.15	0.40	0.35	0.18	0.51	0.35
	(b) Wavelets													
AR1×Spl	0.16	0.40	0.19	0.34	0.39	0.26	0.23	0.11	0.16	0.38	0.25	0.17	0.40	0.43
Wavelets	0.19	0.38	0.18	0.31	0.37	0.25	0.22	0.10	0.16	0.37	0.24	0.16	0.37	0.38
Basic	0.18	0.39	0.19	0.34	0.39	0.25	0.23	0.10	0.16	0.38	0.25	0.16	0.41	0.42

7.4.2 Simulation of Spatial Noise

Figure 7.3 compares the three methods of removing spatial noise by displaying the standard deviation of the difference between the simulated and estimated intensity readings for each correlation function used to generate to spatial Gaussian noise.

Figure 7.3(a) shows that when the correlation function is decomposed into long term and short term behaviour, as the fractal dimension is increased in the stable correlation function, the estimates from the AR1 × Splines model become further from the simulated values and the standard deviation of the difference increases. The estimates from the Wavelets and Basic models have similar deviations from the simulated values, with the Wavelets model slightly better for all fractal dimensions. The standard deviation of the difference does not increase with the fractal dimension for the Wavelets or Basic models.

Similarly, Figure 7.3(b) shows that when only the short term correlation is modeled, as the fractal dimension is increased the estimates that arise from the AR1 × Splines model become further from the simulated values. The AR1 × Splines model is comparable the Wavelets model for smoother images, but the Wavelets model outperforms the AR1 × Splines model for more jagged images. Once again, the Wavelets and Basic models produce similar results to each other, with the Wavelets model slightly better, and the accuracies of prediction are not dependent on the fractal dimension of the spatial surface.

Figure 7.3: Standard deviation of the difference between simulated and estimated intensity reading. The spatial noise is simulated with Gaussian fields with (a) stable and Cauchy correlation functions and (b) stable correlation function.

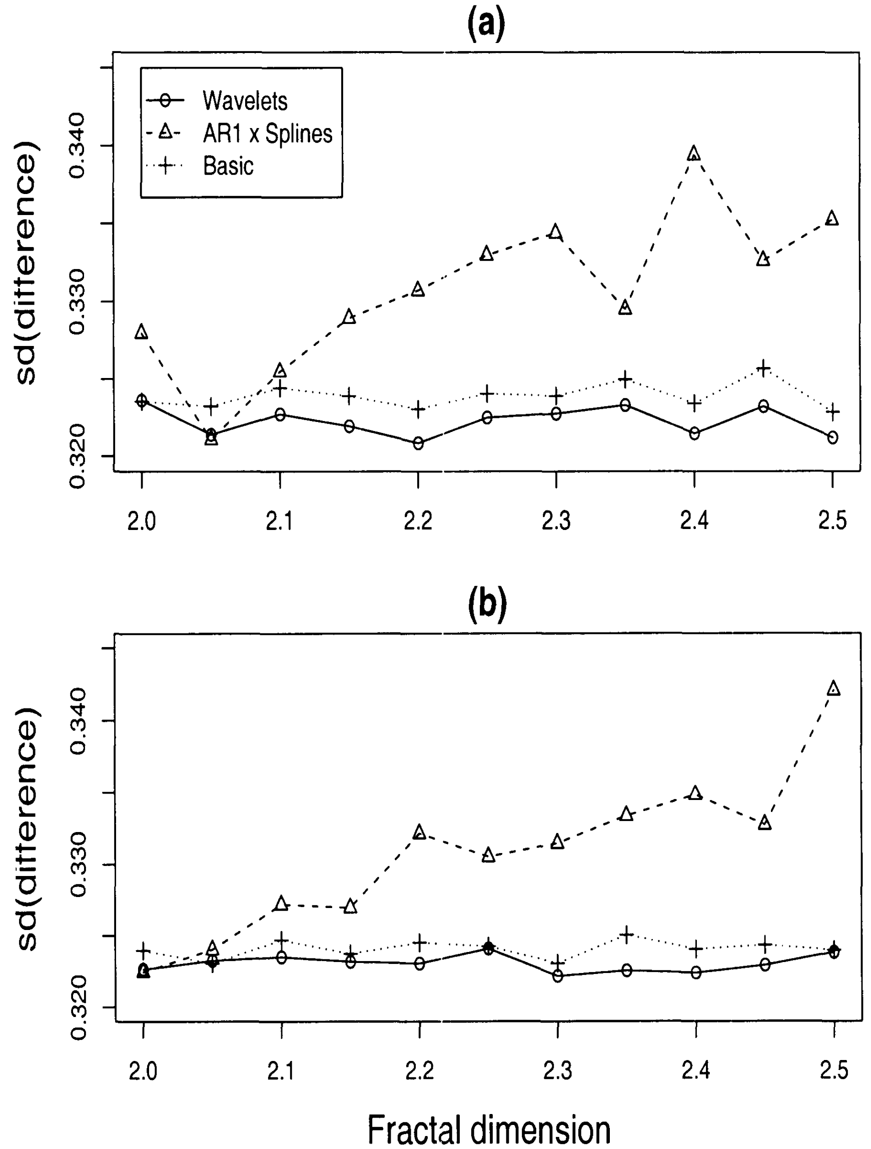


Figure 7.4: Empirical cumulative distribution function of the estimated value minus the simulated value for a simulation with stable and Cauchy correlation functions and with fractal dimension 2.2.

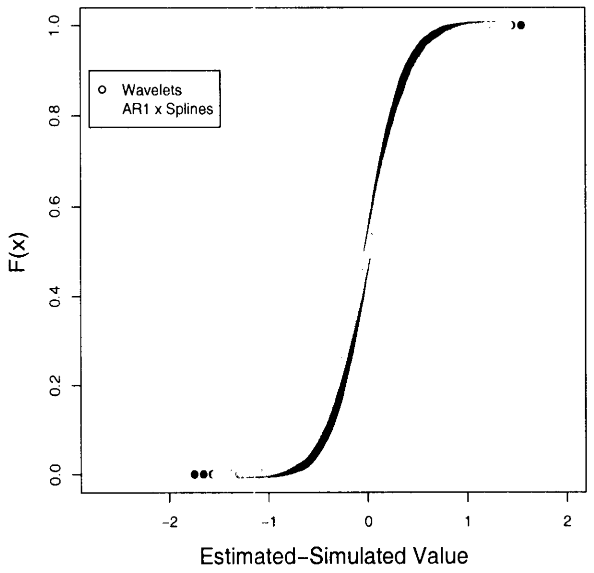


Figure 7.4 displays a typical empirical cumulative distribution function of the difference between the simulated and estimated values of the test genes for the Wavelets and AR1 × Splines models. The spatial noise was generated with the stable and Cauchy correlation structure with a fractal dimension of 2.2. The plot shows that the Wavelets model predicts more test gene effects accurately than the AR1 × Splines model. It is also seen that the AR1 × Splines model severely over-estimates the expression level of one of the genes, with the difference of the simulated log_2 intensity and the estimated log_2 intensity being 2.5.

7.4.3 Computation Time

Table 7.2 shows the mean computation time for one slide. The AR1 × Splines model is the slowest and the Basic model is the fastest under all simulations. The Wavelets model is significantly faster than the AR1 × Splines model, but slightly slower than the Basic model in all circumstances.

Table 7.2: Mean computation time (seconds) for fitting a spatial model for one slide. The times given for the simulated spatial noise are for Gaussian fields with a fractal dimension of 2.2.

	Real Background Noise		Simulated Spatial Noise	
Model	AR1 × Splines	Wavelets	Stable	Stable + Cauchy
AR1 × Splines	174.95 ± 47.09	178.99 ± 56.16	41.62 ± 9.00	48.66 ± 9.28
Wavelets	10.47 ± 0.27	10.49 ± 0.64	6.90 ± 0.23	6.88 ± 0.22
Basic	1.02 ± 0.01	1.03 ± 0.11	1.12 ± 0.06	1.12 ± 0.07

7.5 Discussion

The AR1 × Splines model may not be as successful as the Wavelets model for rougher images because this model assumes smoothness in the spatial trend. Hence, the jagged images may cause under-estimation of the spatial trends and this is reflected in the estimate of the test gene effects. Similarly for smoother images, the Wavelets model may estimate the spatial trend to be too jagged, which would also reduce the precision of the estimated test gene effects. This could be remedied with a more appropriate choice of basis function for the discrete wavelet transform.

If the spatial noise in a microarray slide is smooth, the AR1 × Splines model may be appropriate to account for this spatial bias. However, care should be taken for more fractal images.

It has been demonstrated here that not accounting for the spatial surface at all may be better than fitting a model that assumes smoothness, such as the AR1 × Splines model. The multiresolutional nature of the wavelet function means that this method is well suited at removing spatial bias regardless of the fractal dimension of the nuisance spatial effects. Furthermore, because of the simple threshold approach in the frequency domain, the wavelets approach has another advantage of computational speed.

Chapter 8

General Discussion

There has been a huge investment globally on genotyping and related laboratory tools, which promise to increase the rate of genetic gain and aid in the understanding of regulatory pathways. It is well accepted that DNA-based selection aids will impact mostly on traits that are costly or difficult to measure. As the technology becomes more accessible, it is likely that even larger amounts of genotypic data will be generated for more traits. This highlights the necessity of statistical methods to at least keep pace with the development of genetic technologies lest information gathered be less than optimally utilized.

Some useful statistical techniques to make use of the vast amounts of data available to geneticists have been demonstrated in this thesis. The major points to be taken are:

- New technologies, such as single nucleotide polymorphism (SNP) chips and microarrays, give rise to large amounts of data which require appropriate statistical techniques, so that useful conclusions can be drawn from studies involving such methods.

- Data reduction techniques, such as principal component analysis, can be effective in reducing the dimension of the space in which SNP values are contained. The resultant principal components can be used in multiple linear regression to give relatively accurate molecular breeding values for young animals with no phenotypic records or known pedigree.

- Non-parametric models can be useful in genetic studies, particularly when there are far more explanatory variables than response variables. However, care should be taken when applying methods that assume a continuous structure to marker data. It has been shown empirically that using an intrinsically continuous Gaussian kernel is more accurate than using the discrete binomial kernel in a Nadaraya-Watson estimator for predicting molecular breeding value from SNP data. However, it is important to check the repercussions of treating discrete data as continuous.

- Care should be taken when correcting microarray data for spatial noise. The way in which the spatial noise is accounted for can alter the outcomes of a microarray experiment. The wavelets transform is a useful tool for accommodating spatial noise into the modeling process. Due to the multiresolutional nature of wavelets, broad and fine scale trends can be modeled simultaneously, making the method well suited for fractal spatial surfaces in microarray slides.

The use of marker data from the whole genome when selecting young animals for breeding is an improvement on traditional BLUP methods. The BLUP method predicts genetic merit as the mean of the parents breeding values plus a Mendelian sampling term. Marker assisted selection (MAS) allows prediction of breeding value with the segments of chromosome that are inherited from each parent known, resulting in better specified estimated breeding values. The inclusion of markers in selection may also allow greater control of genetic variance in future generations so that further genetic improvements can be made, whilst managing the level of inbreeding.

Further issues arise when MAS is implemented. First, the question of which animals to genotype and phenotype so that resources are efficiently employed needs to be addressed. Second, the transferability of results from one population to another needs to be assessed. For example, can herds be used from different regions or countries in the same study; and can animals at different time points in each herd be treated as equal when estimating breeding values using markers?

Since the chapters in this thesis regarding spatial artefacts in microarray slides have been completed, the technology has improved, with spatial trends less prevalent on slides. However, the removal of spatial noise is still a crucial part of the normalization of microarray data. Hence, the discrete wavelet transform is a useful method for microarray normalization. Furthermore, the tendency toward high throughput methods to collect genotypic data may mean that further technologies are developed that stimulate fractal responses. Wavelets are an ideal method for accommodating such fractal data.

Currently, many of the SNPs used in genotypic studies are located in non-coding regions of the chromosome, so that it is difficult to combine marker and microarray experiments. It is envisaged that more SNPs from exonic regions of the chromosome will be able to be genotyped, allowing studies to readily incorporate expression and marker data simultaneously. Such studies have been conducted for mice, allowing insight into the biological processes underlying phenotypic observations (Schadt et al., 2003). This methodology could be used extensively within the livestock industries to aid understanding of biological processes and consequently, further accelerate the rate of genetic improvement.