Animal (2010), **4:2**, pp 165–172 © The Animal Consortium 2009
doi:10.1017/S1751731109991054

animal

Integrating biological information into the statistical analysis and design of microarray experiments*

G. J. M. Rosa† and A. I. Vazquez

Department of Dairy Science, University of Wisconsin, Madison, WI 53706, USA

(Received 4 February 2009; Accepted 8 August 2009; First published online 5 October 2009)

Microarray technology is a powerful tool for animal functional genomics studies, with applications spanning from gene identification and mapping, to function and control of gene expression. Microarray assays, however, are complex and costly, and hence generally performed with relatively small number of animals. Nevertheless, they generate data sets of unprecedented complexity and dimensionality. Therefore, such trials require careful planning and experimental design, in addition to tailored statistical and computational tools for their appropriate data mining. In this review, we discuss experimental design and data analysis strategies, which incorporate prior genomic and biological knowledge, such as genotypes and gene function and pathway membership. We focus the discussion on the design of genetical genomics studies, and on significance testing for detection of differential expression. It is shown that the use of prior biological information can improve the efficiency of microarray experiments.

Keywords: microarray, experimental design, differential expression, biological knowledge

Implications

Because of their high costs, microarray gene expression studies are generally performed with relatively small number of animals. Nonetheless, they generate huge data sets, as virtually all the genome is assessed for each subject in a single experiment. Therefore, careful experimental planning and data analysis tools are needed for a successful microarray study. Here, we discuss how prior genomic and biological knowledge can be incorporated into the design and data mining steps of such experiments to improve their efficiency and statistical power.

Introduction

Microarray technology has been increasingly used in animal functional genomics studies to further our understanding related to gene functions and interactions. Applications of microarray technology in livestock span a broad range of research endeavors, from investigation on gene identification and mapping to function and control of gene expression. Such knowledge can ultimately lead to breakthroughs in farm animal production and welfare, enhancing profitability and trade of animal agriculture and its products (e.g. Hiendleder *et al.*, 2005; Cassar-Malek *et al.*, 2008).

Microarray gene expression assays, however, are extremely complex, involving a number of laboratorial procedures such as sample collection, RNA extraction, reverse transcription and labeling, and the final hybridization. Given the high costs and intense labor requirement, microarray experiments are generally performed with relatively small numbers of samples (animals). Hence, such trials require a great deal of planning and careful experimental design. Moreover, after microarray experiments are conducted, researchers are faced with data sets of unprecedented complexity and dimensionality, requiring tailored statistical and computational tools for an appropriate data mining.

The statistical analysis of such data sets generally involves a data preprocessing step, which refers to the summarization of signal intensities into expression indexes, as well as data transformation and normalization, for removing non-biological systematic effects contributing to variation on expression measurements (Jaffrezic *et al.*, 2007). After the microarray data has been normalized, the statistical analysis may involve, e.g., clustering techniques, discriminant analysis, or differential expression significance test (see, e.g., Jaffrezic *et al.*, 2007; Sorensen *et al.*, 2007). Subsequently, results are generally complemented with some sort of validation or interpretation using biological

*This paper has been presented at the session "Genomics selection and bioinformatics" of the 59th Annual meeting of the European Association for Animal Production held in Vilnius (Lithuania), 24 to 27 August 2008. Dr A. Maki-Tanila acted as a guest editor.

† E-mail: grosa@wisc.edu

https://doi.org/10.1017/S1751731109991054 Published online by Cambridge University Press

knowledge such as gene networks and functional categories, e.g., for assessing enrichment of gene sets (e.g. Beissbarth and Speed, 2004; Subramanian *et al*., 2005; Sorensen *et al*., 2007). On a different approach, rather than using biological information on a follow-up step only, some authors have suggested the integration of external genomic information into the statistical analysis itself, as well as in the design of microarray trials.

In this paper, we review experimental design and data analysis strategies, which take advantage of prior genomic and biological knowledge to improve the efficiency of microarray experiments. Regarding the design component of the manuscript, we concentrate the discussion on genetical genomics studies, which make use of information on genotypes and genetic relationships among the animals available for the expression profiling. In regard to the data analysis module, we focus on the analysis of differential expression, which use genomic information such as gene function and pathway membership, to improve the power of detecting differential expression, as well as the reliability of fold change estimates. This paper is organized as follows: In the first section, we provide some examples to support the notion of the relationship between gene sequence and transcriptional activity, including coexpression of genes. In the second section, we present experimental design alternatives proposed in the literature on how molecular maker information can be used to improve the efficiency of genetical genomics studies. The following section discusses the use of genomic information, such as gene networks, on the statistical analysis of microarray data and lastly some concluding remarks are provided.

The relationship between gene sequence and transcriptional activity

With the accumulation of more microarray data and the availability of complete sequences for several species provided by sequencing projects, the relationship between the expression and the sequence has been demonstrated by many studies. For example, Cohen *et al*. (2000) and Kruglyak and Tang (2000) independently found that adjacent pairs of genes in yeast (*Saccharomyces cerevisiae*) have correlated expression patterns. The same phenomenon has been observed in human (Lercher *et al*., 2002), worm (Roy *et al*., 2002), fruit flies (Spellman and Rubin, 2002), and *Arabidopsis thaliana* (Williams and Bowles, 2004). Fukuoka *et al*. (2004) performed whole genome transcriptional profiling in six eukaryotic genomes, including yeast, fruit fly, worm, rat, mouse and human, and also found that neighboring genes are likely to be coexpressed.

In addition, other studies showed that genes with common sequence elements in their upstream regions (i.e. regulatory motifs) often share the similar expression patterns (e.g. Bussemaker *et al*., 2001). Coexpression patterns have been used then to identify regulatory motifs and also for inferring the cause of such coexpressions, i.e., shared transcription factor (TF)-binding sites in the promoters of coregulated genes (Bussemaker *et al*., 2001; Conlon *et al*., 2003). By studying the relationship between gene sequence and transcriptional activity, Zhang *et al*. (2009) also provided evidence that sequence similarity, gene neighborhood, as well as pathway membership contribute to observed coexpression patterns.

To take advantage of the relationship between gene sequence and expression, a number of recent publications in the microarray literature propose the use of different kinds of biological information into the design and analysis of such experiments. In this review, we discuss two specific applications of this concept, namely (i) the design of genetical genomics studies and (ii) the significance test for detection of differential expressed genes across experimental or observational groups.

Planning of microarray experiments using molecular marker information

The planning of a gene expression study entails a variety of aspects, including the choice of the microarray platform that best suits the specific experiment goals, as well as careful decision regarding the cell type(s), tissue(s), and the developmental stage(s) to be assayed. Various commercial and homemade microarray platforms are available for gene expression profiling in different livestock species, including the so-called two-color (using either cDNA or long oligo probes) and single-channel (generally with short oligo probes) technologies. The choice of microarray platform is generally guided by the availability of alternative technologies for the organism being studied. For example, cDNA microarrays are available only for species from which expressed sequence tags (ESTs) were obtained from cDNA libraries. Likewise, slides using oligonucleotide probes can be generated only if DNA sequence is available, and an optimal probe set can be designed only for those species with completely sequenced genomes. Hence, a broader range of commercial and homemade array platforms is generally available for research in cattle, pigs and chickens as opposed to other livestock species. Besides economics, the choice of microarray platform depends also on the nature of the research question proposed. For example, traditional platforms may not be optimal for exon level expression profiling to uncover alternative splicing events.

In addition to the choice of array platform, another important consideration before execution of microarray experiments is the complexity of the tissue of interest, and whether to use RNA from whole tissue *v.* individual purified cell populations of interest. For example, when whole tissue RNA is utilized in microarray experiments, changes in gene expression of interest may be diluted by contribution of transcripts from other cell types and tissue components or may be masked by potential divergent gene regulation within individual cell populations (Smith and Rosa, 2007).

Another interesting and extremely important component of microarray experiment planning refers to the choice of the experiment layout regarding the allocation of mRNA

https://doi.org/10.1017/S1751731109991054 Published online by Cambridge University Press

samples within assay batches, slides, dye labeling and other local control factors (Rosa *et al*., 2006). With single-channel platforms, the experiment setup is usually straightforward, as each sample is assayed in an independent slide. With two-color microarrays, on the other hand, there are always numerous ways of pairing samples and assigning dyes. The simplest alternative refers to the so-called reference design, though the loop structure is shown to be generally more efficient (Kerr and Churchill, 2001; Yang and Speed, 2002). In addition, variations of the classical reference and loop structures may be obtained with different combinations of biological and technical replication (Rosa *et al*., 2005), i.e., number of subjects within each experimental group and number of replicated arrays or spots for each biological sample, respectively. In a more general context, optimal design methodologies may be used to search for more efficient experiments for specific goals of the experiment (Bueno Filho *et al*., 2006).

It is clear then that prior biological knowledge and the specific experiment goal(s) play a crucial role in the planning phase of a microarray assay, including the choice of array platform, the tissue/cells to be assayed and the overall experiment layout. In addition, prior knowledge may help also on sample size and power calculations, which require input in terms of effects size values, i.e., expected fold change and variability of the system. In this manuscript, however, we focus the discussion on the design of experiments for genetical genomics studies (Jansen and Nap, 2001), which combine gene expression with information on phenotypic traits, family (or pedigree) structure and molecular markers. Genetical genomics experiments are used, e.g., to search for polymorphisms associated with gene expression variability, denoted expression quantitative trait locus (eQTL), and to estimate heritability of mRNA transcript abundances, among other applications. A comprehensive review of alternative design strategies for genetical genomics studies was presented by Rosa *et al*. (2006). Here, we briefly discuss the main approaches available in the literature, as well as their extension to outbred populations, as generally encountered in livestock research.

A number of suggestions for taking advantage of prior knowledge (in terms of molecular marker information or genetic relationships between individuals) are available in the literature, which can be split into two main categories. The first group relates to the subset selection of individuals to be assayed, when a larger panel of individuals is available. This strategy is generally referred to as selective phenotyping, or selective transcriptional profiling. A second group of design techniques refer to optimal strategies to allocate mRNA samples across slides and dye labeling, if two-color microarrays are being used.

Selective phenotyping is advised when the number of animals available is larger than the number of mRNA samples to be assayed. In such cases, instead of randomly choosing individuals from the available panel, one may use molecular marker information to try to maximize genetic dissimilarity among selected individuals to increase the power of detecting QTL additive effects (Jin *et al*., 2004). Such rationale can be extended also if interest includes QTL with dominant effects (Piepho, 2005; Bueno Filho *et al*., 2006). Another cleaver strategy for selective phenotyping in genetical genomics, whose goal is an increased efficiency of localizing eQTL, is a subset selection based on the number of recombination events observed in each individual (Jannink, 2005; Xu *et al*., 2005). Such approach may aim not only an increased number of recombination events in the subsample, but also an even distribution of recombinations across the genome. On a somewhat different genetical genomics research goal, if interest refers to inferences regarding variance components and heritability of transcriptional activity, then the selective phenotyping criterion should take into account the genetic relatedness among the available subjects, and the optimal designs are shown to be very specific for each pedigree structure (Bueno Filho *et al*., 2006).

After the individuals to be assayed have been selected, an optimal strategy to allocate mRNA samples across slides and dye labeling should be sought. Again, the choice of experimental layout should be intimately connected with the research goal. Some alternative strategies have been proposed in the literature for specific goals of genetical genomics studies. For example, Fu and Jansen (2006) proposed a design strategy for allocating pairs of samples to two-color microarray slides in situations when only two genotypes are possible for each locus, such as with recombinant inbred lines, backcross and double haploid. Their approach, called distant pair design, maximizes the ratio of within- over between-slides genotypic dissimilarity to increase the power of detection of additive effects. Similar rationale was considered by Piepho (2005), for situations where three genotypes are possible (e.g. F_2 population), and the main interest refers to the estimation of dominant effects. A generalization of the distant pairing concept proposed by Fu and Jansen (2006) for finite loci was presented by Bueno Filho *et al*. (2006) for experiments aiming at the estimation of variance components and heritability of gene expression. It is shown that given a sample of subjects with a certain relatedness structure (pedigree), the optimal designs tend to pair less related individuals in each slide.

Specifically for genetical genomics with livestock, Cardoso *et al*. (2008) extended and compared a number of selective phenotyping strategies in the context of an outbred F_2 population, and Lam *et al*. (2008) propose an extension of the distant pairing strategy for outbred experimental crosses. It is shown that, with limited resources or a fixed number of microarray slides, an optimal selective phenotyping strategy and experimental layout can substantially improve the efficiency of genetical genomics studies.

Integrating genomic information into the statistical analysis of microarrays

General aspects

Microarray assays involve multiple sources of systematic effects (e.g. differences among slides and dye labeling and

https://doi.org/10.1017/S1751731109991054 Published online by Cambridge University Press

spatial variation within slides), such that their statistical analysis generally requires a data preprocessing step, usually referred to as normalization, for removing non-biological effects contributing to variation on expression measures. A review and comparison of normalization techniques is provided, e.g., in Jaffrezic *et al*. (2007).

Normalization approaches can be based either on control genes, or on data-driven selected genes seemingly equally expressed across experimental groups, or most commonly, on all genes represented in the slides. In this latter case, robust statistical techniques are used, with the assumption that the majority of the genes are not differentially expressed. In some cases, e.g., with boutique arrays developed specifically for studies with certain tissue(s) and developmental stage(s), this assumption does not hold and alternative normalization techniques are required (Oshlack *et al*., 2007). Hence, it is apparent that biological knowledge should be used to guide researchers to choose an appropriate data normalization strategy for their experiments.

Other examples of use of prior biological information in preprocessing of microarray data can be found, among others, in Tuikkala *et al*. (2006) who used information on functional similarities, assessed by semantic correspondences originating from gene ontology (GO) annotations, to improve the selection of relevant genes for missing data imputation, and in Kim *et al*. (2007) who investigated the incorporation of genetic relationships among animals for the normalization of expression data in genetical genomics studies.

After this first step of data preprocessing, a higher-level statistical analysis is performed. Depending on the objective(s) of the study such analysis may involve, e.g., clustering, classification techniques or hypothesis testing for detection of differential expression (for a comparison of different approaches, see, e.g., Jaffrezic *et al*., 2007 and Sorensen *et al*., 2007). In this manuscript, we will specifically discuss the comparison of expression profiles across groups or populations using ANOVA models. Such models (including those with random effects) are the most popular in the microarray literature (see, e.g., Jaffrezic *et al*., 2007), due to their flexibility and ease of use as well as the availability of software for their implementation, e.g. MAANOVA and LIMMA, both developed in R language and available at Bioconductor website (http://www.bioconductor.org/).

The ANOVA approach for analyzing microarray data was first presented by Kerr and Churchill (2001), with a model described as:

$$y_{adgvr} = \mu + A_a + D_d + (AD)_{ad} + G_g + (AG)_{ag} + (DG)_{dg} + (VG)_{vg} + \varepsilon_{adgvr}, \quad (1)$$

where y_{adgvr} represents the expression intensities on the log scale; μ is an overall constant; A_a, D_d and $(AD)_{ad}$ are 'global factors' that account for variation between arrays and dyes; G_g is the gene effect; $(AG)_{ag}$ represents array by gene interactions; $(DG)_{dg}$ is gene-specific dye effects; $(VG)_{vg}$ is the quantity of interest, related to differential expression of gene g specifically attributable to variety (treatment) v; and ε_{adgvr} is random residual term with variance σ_ε^2. Wolfinger *et al*. (2001) extended this approach to allow for random effects as well for gene-specific variance components. For computational convenience, a two-step methodology was proposed. In the first stage, a model including only global factors is fitted to all data simultaneously, which can be written as:

$$y_{adgvr} = \mu + A_a + D_d + (AD)_{ad} + e_{adgvr}, \quad (2)$$

where the terms are defined as in model (1). The second stage uses the estimated residuals $\hat{e}_{adgvr}$ from model (2), and fits a series of gene specific models, expressed as:

$$\hat{e}_{adgvr} = \mu_g + A_{ag} + D_{dg} + V_{vg} + \varepsilon_{adgvr}, \quad (3)$$

where μ_g, A_{ag}, D_{dg} and V_{vg} are gene-specific constant, array effects, dye effects and treatment effects, respectively.

The mixed effects modeling approach of Wolfinger *et al*. (2001) relaxes the assumption of a common variance for all genes, as in model (1), but at the expense of performing independent analyses for each gene, one at a time. This means that genes are treated as independent entities, ignoring the natural covariance among their expression levels due to coregulation processes. The main drawback of this approach is that, because sample sizes for each gene are usually relatively small, the resulting statistical tests have low power and fold change estimates have low precision. Most of the strategies available in the literature to overcome this situation refer to some sort of shrinkage estimation of variance components, to borrow information across genes (e.g. Smyth, 2004 and Cui *et al*., 2005). However, those approaches are still based on the assumption of independence among genes, and they do not take advantage of prior genomic or biological knowledge. Exogenous information is generally used only after a list of genes has been selected, e.g., to study gene set enrichment (Huang *et al*., 2009).

Modeling covariance between gene expression

Some recent publications have suggested the integration of prior biological knowledge into the statistical analysis of microarray data. For example, Imoto *et al*. (2003) and Werhli and Husmeier (2007) exploit multiple sources of prior knowledge for gene regulatory network reconstruction; and Huang and Pan (2006), Pan (2006a) and Tseng (2007) propose biologically informed gene clustering techniques. However, not much research has been published regarding the incorporation of prior knowledge into the analysis of differential expression, one exception being Pan (2006b), who uses gene functional annotation for improving detection of differentially expressed genes. Here, we discuss some strategies for integrating prior biological information into the statistical analysis of differential expression. The main idea is to build parsimonious multivariate ANOVA models, as an alternative to the shrinkage estimation approaches usually found in the literature.

https://doi.org/10.1017/S1751731109991054 Published online by Cambridge University Press

We start by combining the gene-specific models into a single multigene model (as discussed by Hoeschele and Li, 2005), which can be expressed in matrix notation as:

$$y = X\beta + Zu + \varepsilon$$

where $\boldsymbol{y}$ is a vector with expression intensities for all genes and samples; $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are vectors with fixed (such as treatment and dye) and random (such as slide and animal) effects, respectively; $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices; and $\boldsymbol{\varepsilon}$ is a vector of residual terms, assumed to be independent from each other. For modeling the dependence across genes, a multivariate version of the mixed model described above could be considered, with an unstructured covariance matrix for modeling the residuals. In this case, the variance of ε could be represented as $\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma}_{(\mathrm{k}\times\mathrm{k})} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & \sigma_k^2 \end{bmatrix}$$

is a matrix describing the variances and covariances between genes, k is the number of genes and $\boldsymbol{I}_n$ is an identity matrix of order n (n is the number of samples or subjects in the experiment).

The covariance matrix $\boldsymbol{\Sigma}$, however, has $\mathrm{k}(\mathrm{k}+2)/2$ distinct elements that need to be estimated (k variances and $\mathrm{k}(\mathrm{k}-1)/2$ covariances), which makes fitting such a model unfeasible given the small number of samples relative to the number of genes, i.e. $n \ll \mathrm{k}$. For example, with an experiment involving $\mathrm{k} = 10\,000$ genes, there would be over 50 million (co)variance parameters to be estimated. Nonetheless, currently microarray experiments in livestock do not involve more than a few hundreds of slides.

To overcome this problem, biological information could be used, e.g., to split the genes into subsets of potentially coexpressed genes such that multivariate models of much lower dimensions could be fitted within each set of genes. Biologically meaningful gene groupings, such as regulatory networks and pathways, can be assessed using freely available databases such as the GO Consortium (GO – http://www.geneontology.org/), the Kyoto Encyclopedia of Genes and Genomes (KEGG – http://www.genome.jp/kegg/) among others. In this context, the assumption is that genes in different groups are independent but that genes within each group can have any pattern of coexpression. Under these circumstances, the new covariance matrix $\boldsymbol{\Sigma}$ is block diagonal, and can be described as:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{B}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B}_\mathrm{b} \end{bmatrix},$$

where b is the number of gene subsets, and

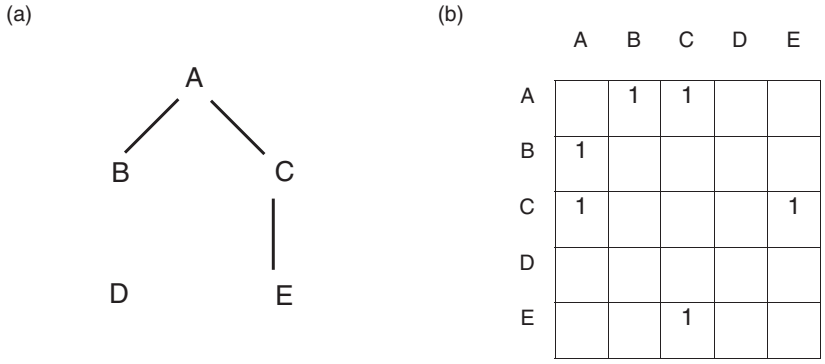
$$B_\mathrm{i} = \begin{bmatrix} \sigma^2_{1(\mathrm{i})} & \sigma_{12(\mathrm{i})} & \cdots & \sigma_{1\mathrm{k_i}(\mathrm{i})} \\ \sigma_{12(\mathrm{i})} & \sigma^2_{2(\mathrm{i})} & \cdots & \sigma_{2\mathrm{k_i}(\mathrm{i})} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1\mathrm{k_i}(\mathrm{i})} & \sigma_{2\mathrm{k_i}(\mathrm{i})} & \cdots & \sigma^2_{\mathrm{k_i}(\mathrm{i})} \end{bmatrix}_{(\mathrm{k_i}\times\mathrm{k_i})}.$$

The number of non-zero parameters in $\boldsymbol{\Sigma}$ is now $p^* = \sum_{i=1}^{\mathrm{b}} \frac{\mathrm{k_i}(\mathrm{k_i}+1)}{2}$, which is much smaller than the original $p = \frac{\mathrm{k}(\mathrm{k}+1)}{2}$ parameters, where $k = \sum_{i=1}^{\mathrm{b}} \mathrm{k_i}$. Moreover, the analysis under such settings can be implemented independently for each gene set, making model dimensions even smaller and computations less cumbersome. Nonetheless, small samples sizes can still hinder the gene set models to be fitted, especially for larger groups of genes. In this case, parsimonious structures for each covariance matrix $\boldsymbol{B}_\mathrm{i}$ can be built using biological information, e.g., by describing covariances between genes as a function of their number of shared elements, their positions in the genome, as well as information on biological processes (Vazquez *et al.*, 2009).

A standard way of summarizing current knowledge about the biological process is via a regulatory network. Wei and Pan (2008), e.g., used biological information summarized in a network with a model in which the joint prior probability of gene effects is multivariate normal, and the conditional prior probabilities for any gene depends only on the effects of neighboring genes. The mean of the conditional distribution is the mean of the effects of neighboring genes, and its variance decrease with edge degree. A more general approach may be obtained by assigning a multivariate normal prior on gene effects with a general (co)variance structure derived from the network, in the form of a kernel.

The key element in this approach is the form of $\boldsymbol{\Sigma}$, which is a function of a gene network (or graph). Figure 1 illustrates how to express an undirected unweighted graph into an adjacency matrix, $\boldsymbol{E}$, whose entries e_{ij} are equal to 1 if genes (or nodes) i and j have a common edge connecting them (e.g. indicating (co)expression), or 0 otherwise. As an example, in Figure 1a, genes A and B are connected with an edge, and D is an isolated gene. Kernel candidates are functions of $\boldsymbol{E}$, such as the exponential $\exp(-d_{ij})$, or the inverse $1/d_{ij}$, where d_{ij} is the length of the path between nodes i and j, with an infinite distance between no connected genes (e.g. genes A and D in Figure 1a). A particular function of $\boldsymbol{E}$ is the diffusion kernel (Kondor and Lafferty, 2002), which has been previously used in microarrays in a classification problem with a support vector machine approach (Vert and Kanehisa, 2003). The diffusion kernel matrix has larger entry values between nodes more closely connected and smaller values between two nodes with a longer path, with the strength of the correlations controlled by a bandwidth parameter h. On Figure 1, two different values of the bandwidth parameter were used to construct diffusion kernels based on the simple network depicted ($\mathrm{h} = 0.2$ and 0.5 for Panels c and d, respectively).

https://doi.org/10.1017/S1751731109991054 Published online by Cambridge University Press



(c)

	A	B	C	D	E
A	1.000	0.197	0.196	0.000	0.019
B	0.197	1.000	0.019	0.000	0.001
C	0.196	0.019	1.000	0.000	0.197
D	0.000	0.000	0.000	1.000	0.000
E	0.019	0.001	0.197	0.000	1.000

(d)

	A	B	C	D	E
A	1.000	0.456	0.443	0.000	0.103
B	0.456	1.000	0.103	0.000	0.015
C	0.443	0.103	1.000	0.000	0.456
D	0.000	0.000	0.000	1.000	0.000
E	0.103	0.015	0.456	0.000	1.000

Figure 1 Example of an undirected unweighted graph representing a 5-node network (panel a), and its corresponding adjacent matrix (panel b), and two possible diffusion kernels (panels c and d).

As an illustration of the potential of using prior biological knowledge in the analysis of microarray data, we consider a simple example with the data set and network used by Wei and Pan (2008). Briefly, log-ratio expression of 4609 genes of *S. cerevisiae* was measured in absence and presence of a TF, using ChIP-on-chip technology. A weighted average of three independent replicates in each group (i.e. with and without TF) were analyzed, incorporating or not the information provided by a probabilistic functional gene network obtained from an independent study (Lee *et al*., 2004).

The following model was used to incorporate the gene network information:

$$y_k = m + g_k + e_k,$$

where y_k is the average (over the three replicates) of the log-ratio intensity values for gene k, m is an overall mean, g_k is the differential expression (fold change in the log scale) relative to gene k, and e_k is a residual term associated with the observation y_k. It is assumed that $e_k \overset{iid}{\sim} N(0, \sigma^2_e)$, and that $\mathbf{g} = [g_1, g_2, \ldots, g_K]' \sim N(\mathbf{0}, \boldsymbol{K}_h)$. A diffusion kernel with bandwidth parameter $h = 0.2$ ($\boldsymbol{K}_{0.2}$) was then generated from the network. A model that ignores the gene network information was also considered, in which case $\boldsymbol{g} = [g_1, g_2, \ldots, g_K]' \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2_g)$, i.e. $g_k \overset{iid}{\sim} N(0, \sigma^2_g)$. These models were evaluated in terms of predictive ability using a leave-one-out cross-validation (LOOCV) approach. In the model with the gene network ($\boldsymbol{K}_{0.2}$ model), prediction of ratio values were obtained by $\hat{y}_k = \hat{m} + \hat{g}_k$, whereas in the model ignoring the gene network, because of the confounding between g_k and e_k, predictions were equal to $\hat{y}_k = \hat{m}$ for all k.

With the $\mathbf{K}_{0.2}$ model, the estimate of variance due to the gene structure was 0.006, whereas the residual variance was 0.045, such that the gene structure explained about 12% of the total variability. When ignoring the network, however, it was not possible to separate both variances, since the data set does not have replications. The network allowed not only the estimation of variances, but also the prediction of gene expression differences (fold changes). The correlation between predicted and observed log-ratio values was 0.76, indicating that the network produced considerable information borrowing across genes. The mean squared error of the prediction (MSEP) evaluated via LOOCV was smaller in $\boldsymbol{K}_{0.2}$ compared with the model ignoring the network, being 0.0503 and 0.0545, respectively (i.e. an improvement of 7.6%).

For successful modeling of covariance structures for microarray data using prior biological knowledge, the choice of an adequate kernel is critical. Building adequate kernels requires two main components: (i) an informative input set and (ii) an appropriate function to generate the kernel from the adjacent matrix. Informative input sets can be gathered from trustable and validated biological knowledge from public databases, whereas an adequate kernel can be obtained, e.g., by comparing candidate functions using cross-validation techniques. In this case, the incorporation of prior knowledge into the statistical analysis of microarray data should yield lower MSEP, i.e. more accurate and or less biased predictions.

https://doi.org/10.1017/S1751731109991054 Published online by Cambridge University Press

Concluding Remarks

In this manuscript, we discuss the incorporation of genomic and biological knowledge into the experimental design and data analysis of microarray gene expression experiments. Specifically on the design of microarray experiments, we review strategies for subset selection of animals and for pairing of RNA samples for each competitive hybridization in genetical genomics studies. It is shown that the use of molecular marker information and genetic relationships among animals can increase the power of eQTL detection, and also improve the accuracy of estimates of transcriptional activity heritability. The incorporation of biological information, such as gene function and pathway membership, is also shown to improve results of microarray data mining tools. More specifically, it is shown to considerably enhance performance of clustering techniques. Here, we discuss how statistical models for detection of differential expression can incorporate biological information, such as gene networks. The key element of the approach discussed is the form with which the network graph is translated to a covariance structure describing gene coexpression. This is an extremely contemporary and promising area of research, which should enhance the efficiency of microarray experiments and lead their analysis to a more integrated, systems biology approach. It is worthy noting that, as technology develops, in the near future gene expression will probably be measured using resequencing techniques (e.g. Morozova and Marra, 2008; Torres *et al.*, 2008). Nonetheless, most of the design and data analysis strategies discussed here will still apply.

Acknowledgments

The authors acknowledge support from the Wisconsin Agriculture Experiment Station.

References

Beissbarth T and Speed TP 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20, 1464–1465.

Bueno Filho JSS, Gilmour SG and Rosa GJM 2006. Design of microarray experiments for genetical genomics studies. Genetics 174, 945–957.

Bussemaker HJ, Li H and Siggia ED 2001. Regulatory element detection using correlation with expression. Nature Genetics 27, 167–171.

Cardoso FF, Rosa GJM, Steibel JP, Ernst CW, Bates RO and Tempelman RJ 2008. Selective transcriptional profiling and data analysis strategies for eQTL mapping in outbred F2 populations. Genetics 180, 1679–1690.

Cassar-Malek I, Picard B, Bernard C and Hocquette J-F 2008. Application of gene expression studies in livestock production systems: a European perspective. Australian Journal of Experimental Agriculture 48, 701–710.

Cohen BA, Mitra RD, Hughes JD and Church GM 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nature Genetics 26, 183–186.

Conlon EM, Liu XS, Lieb JD and Liu JS 2003. Integrating regulatory motif discovery and genome-wide expression analysis. Proceedings of the National Academy of sciences of the United States of America 100, 3339–3344.

Cui XG, Hwang JTG, Qiu J, Blades NJ and Churchill GA 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics 6, 59–75.

Fu J and Jansen RC 2006. Optimal design and analysis of genetic studies on gene expression. Genetics 172, 1993–1999.

Fukuoka Y, Inaoka H and Kohane IS 2004. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics 5, 4.

Hiendleder S, Bauersachs S, Boulesteix A, Blum H, Arnold GJ, Frohlich T and Wolf E 2005. Functional genomics: tools for improving farm animal health and welfare. Revue Scientifique Et Technique-Office International Des Epizooties 24, 355–377.

Hoeschele I and Li H 2005. A note on joint versus gene-specific mixed model analysis of microarray gene expression data. Biostatistics 6, 183–186.

Huang DS and Pan W 2006. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics 22, 1259–1268.

Huang DW, Sherman BT and Lempicki RA 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research 37, 1–13.

Imoto S, Higuchi T, Goto T, Kuhara S, Miyano S and Tashiro T 2003. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. Proceedings IEEE Computer Society Bioinformatics Conference 104–113.

Jaffrezic F, De Koning DJ, Boettcher PJ, Bonnet A, Buitenhuis B, Closset R, Déjean S, Delmas C, Detilleux JC, Dovc P, Duval M, Foulley JL, Hedegaard J, Hornshoj H, Hulsegge IB, Janss L, Jensen K, Jiang L, Lavric M, Lê Cao KA, Lund MS, Malinverni R, Marot G, Nie H, Petzl W, Pool MH, Robert-Granié C, SanCristobal M, van Schothorst EM, Schuberth HJ, Sorensen P, Stella A, Tosser-Klopp G, Waddington D, Watson M, Yang W, Zerbe H and Seyfert HM 2007. Analysis of the real EADGENE data set: comparison of methods and guidelines for data normalisation and selection of differentially expressed genes. Genetics Selection Evolution 39, 633–650.

Jannink JL 2005. Selective phenotyping to accurately map quantitative trait loci. Crop Science 45, 901–908.

Jansen RC and Nap J 2001. Genetical genomics: the added value from segregation. Trends in Genetics 17, 388–391.

Jin CF, Lan H, Attie AD, Churchill GA, Bulutuglo D and Yandell BS 2004. Selective phenotyping for increased efficiency in genetic mapping studies. Genetics 168, 2285–2293.

Kerr MK and Churchill GA 2001. Statistical design and the analysis of gene expression microarray data. Genetical Research 77, 123–128.

Kim Y, Doan BQ, Duggal P and Bailey-Wilson JE 2007. Normalization of microarray expression data using within-pedigree pool and its effect on linkage analysis. BMC Proceedings 1, S152.

Kondor RI and Lafferty J 2002. Diffusion Kernels on Graphs and Other Discrete Input Spaces. Proceedings of the 19th International Conference on Machine Learning, pp. 315–322.

Kruglyak S and Tang H 2000. Regulation of adjacent yeast genes. Trends in Genetics 16, 109–111.

Lam AC, Fu JY, Jansen RC, Haley CS and de Koning DJ 2008. Optimal design of genetic studies of gene expression with two-color microarrays in outbred crosses. Genetics 180, 1691–1698.

Lee I, Date SV, Adai AT and Marcotte EM 2004. A probabilistic functional network of yeast genes. Science 360, 1555–1558.

Lercher MJ, Urrutia AO and Hurst LD 2002. Clustering of housekeeping genes provides a unified model of gene order in human genome. Nature Genetics 31, 180–183.

Morozova O and Marra MA 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92, 255–264.

Oshlack A, Emslie D, Corcoran LM and Smyth GK 2007. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. Genome Biology 8, R2.

Pan W 2006a. Incorporating gene functional annotations in detecting differential gene expression. Applied Statistics 55, 301–316.

Pan W 2006b. Incorporating gene functional annotations in detecting differential gene expression. Journal of the Royal Statistical Society Series C 55, 301–316.

Piepho H-P 2005. Optimal allocation in designs for assessing heterosis from cDNA gene expression data. Genetics 171, 359–364.

Rosa GJM, De Leon N and Rosa AJM 2006. Review of microarray experimental design strategies for genetical genomics studies. Physiological Genomics 28, 15–23.

Rosa GJM, Steibel JP and Tempelman RJ 2005. Reassessing design and analysis of two-colour microarray experiments using mixed effects models. Comparative and Functional Genomics 6, 123–131.

https://doi.org/10.1017/S1751731109991054 Published online by Cambridge University Press

Roy PJ, Stuart JM, Lund J and Kim SK 2002. Chromosomal clustering of muscle-expressed genes in *Canenorhabditis elegans*. Nature 418, 975–979.

Smith GW and Rosa GJM 2007. Interpretation of microarray data: trudging out of the abyss towards elucidation of biological significance. Journal of Animal Science 85(suppl. E), E20–E23.

Smyth GK 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3(1), Article 3.

Sorensen P, Bonnet A, Buitenhuis B, Closset R, Déjean S, Delmas C, Duval M, Tosser-Klopp G, Hedegaard J, Hornshoj H, Hulsegge IB, Jaffrézic F, Jensen K, Jiang L, de Koning DJ, Lê Cao KA, Nie H, Petzl W, Pool MH, Robert-Granié C, San Cristobal M, van Schothorst EM, Schuberth HJ, Seyfert HM, Waddington D, Watson M, Yang W, Zerbe H and Lund MS 2007. Analysis of the real EADGENE data set: Multivariate approaches and post analysis. Genetics Selection Evolution 39, 651–668.

Spellman PT and Rubin GM 2002. Evidence for large domains of similarly expressed genes in *Drosophila* genome. Journal of Biology 1, 5.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102, 15545–15550.

Torres TT, Metta M, Ottenwälder B and Schlötterer C 2008. Gene expression profiling by massively parallel sequencing. Genome Research 18, 172–177.

Tseng GC 2007. Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. Bioinformatics 23, 2247–2255.

Tuikkala J, Elo L, Nevalainen OS and Aittokallio T 2006. Improving missing value estimation in microarray data with gene ontology. Bioinformatics 22, 566–572.

Vazquez AI, Rosa GJM, de los Campos G, Weigel KA and Gianola D 2009. Biologically informed models for microarray gene expression data analysis. In symposium on statistical genetics of livestock for the post-genomic era, Madison, WI, USA, 4 to 6 May 2009. (http://dysci.wisc.edu/sglpge/posters.html)

Vert J-P and Kanehisa M 2003. Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. In Advances in Neural Information Processing Systems 15 (ed. S Becker, S Thrun and K Obermayer), pp. 1425–1432. MIT Press, Cambridge, MA.

Wei P and Pan W 2008. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. Bioinformatics 24, 404–411.

Werhli AV and Husmeier D 2007. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. Statistical Applications in Genetics and Molecular Biology 6, 15.

Williams EJB and Bowles DJ 2004. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Research 14, 1060–1067.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C and Paules RS 2001. Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology 8, 625–637.

Xu Z, Zou F and Vision TJ 2005. Improving QTL mapping resolution in experimental crosses by the use of genotypically selected samples. Genetics 170, 401–408.

Yang YH and Speed T 2002. Design issues for cDNA microarray experiments. Nature Reviews Genetics 3, 579–588.

Zhang K, Lefkowitz ER, Wei H, Lorraine A, Page G and Allison DB 2009 Does sequence similarity predict expression similarity – A case study in *Arabidopsis thaliana* (in preparation).

https://doi.org/10.1017/S1751731109991054 Published online by Cambridge University Press