ESTIMATING BAYESIAN NETWORKS FOR HIGH-DIMENSIONAL DATA WITH COMPLEX MEAN STRUCTURE AND RANDOM EFFECTS

JESSICA KASZA*, GARY GLONEK AND PATTY SOLOMON

School of Mathematical Sciences, University of Adelaide

Summary

The estimation of Bayesian networks given high-dimensional data, in particular gene expression data, has been the focus of much recent research. Whilst there are several methods available for the estimation of such networks, these typically assume that the data consist of independent and identically distributed samples. It is often the case, however, that the available data have a more complex mean structure, plus additional components of variance, which must then be accounted for in the estimation of a Bayesian network. In this paper, score metrics that take account of such complexities are proposed for use in conjunction with score-based methods for the estimation of Bayesian networks. We propose first, a fully Bayesian score metric, and second, a metric inspired by the notion of restricted maximum likelihood. We demonstrate the performance of these new metrics for the estimation of Bayesian networks using simulated data with known complex mean structures. We then present the analysis of expression levels of grape-berry genes adjusting for exogenous variables believed to affect the expression levels of the genes. Demonstrable biological effects can be inferred from the estimated conditional independence relationships and correlations amongst the grape-berry genes.

Key words: Bayesian network; exogenous variable; grape-berry gene expression; regulatory network; score-based metric; variance components.

1. Introduction

The inner workings of a cell are very complex, with many interacting components. Determining how the genes within a cell interact with each other is an important, but difficult, field of research, often requiring the application of advanced statistical methods. Systems of these gene interactions are known as genetic regulatory networks, and the extent to which such networks may be inferred from observational gene expression data remains largely undetermined. To explore this question carefully and quantitatively, high-dimensional multivariate models, including Bayesian networks, need to be considered. The use of Bayesian networks for the modelling of genetic regulatory networks has been discussed by several authors: see, for example, de Jong (2002), Friedman (2004), Friedman et al. (2000), Markowetz & Spang (2007). Their popularity lies in the provision of a flexible framework for the estimation of conditional dependence relationships, thereby providing a means to estimate a covariance matrix given a high-dimensional sample when maximum likelihood methods are unavailable (Dykstra 1970). Estimation of such structures allows insight into how the expression levels

*Author to whom correspondence should be addressed.

School of Mathematical Sciences, University of Adelaide, North Terrace Campus, SA 5005 Australia. 
e-mail: jessica.kasza@adelaide.edu.au; gary.glonek@adelaide.edu.au; patty.solomon@adelaide.edu.au

Acknowledgments. The authors thank an editor and two anonymous reviewers for constructive comments and suggestions. They also thank Terry Speed for helpful advice on an earlier version of the paper.
of large groups of genes are related to one another, which, in turn, should help shed light on genetic regulatory networks involving the genes.

For the most part, it has been assumed that the data used to estimate the networks are independent and identically distributed (Koller & Friedman 2009, chap. 16). In the present paper, we consider the important case where the assumption of independent and identically distributed samples is not satisfied and propose new methods to allow for the estimation of effects of interest given such complexity. Our theoretical development has been motivated by an observational time course microarray study, involving expression levels of grape-berry genes observed over time and known to be associated with changes in temperature. The grapes were sampled from three vineyards in different regions of southern Australia and data on the ambient temperatures during the times leading up to the picking of each sample of grapes was also measured. We want to investigate the conditional dependence structure of the genes, adjusting for the exogenous effects of temperature and vineyards, and we aim to do this through the estimation of a Bayesian network. If the effect of temperature is unaccounted for in the estimation of a Bayesian network for these genes, because of their common relationship with temperature, many pairs of genes will exhibit strong correlations. Unless the gross effects of vineyard and temperature are removed, one cannot hope to detect more subtle associations between genes.

There are many methods available for the estimation of Bayesian networks given microarray and other high-dimensional datasets, and these may be divided into two broad categories, namely, score-based and constraint-based methods, (Spirtes et al. 1993). Score-based methods attempt to maximize some score metric associated with the estimated Bayesian network, whilst constraint-based methods estimate conditional independence relationships directly from the data, and combine these to form a Bayesian network. Constraint-based methods test for conditional independence relationships, so the networks obtained through their application can be quite sensitive to Type I and Type II errors, particularly when the sample sizes are small. Score-based methods, on the other hand, are not as sensitive to small sample sizes and, instead of finding the best local structure for each node, find the best global structure given the data, often resulting in more parsimonious models. Given that gene expression data sets tend to be high-dimensional with the attendant ‘small n, large p’ problem, we approach the problem of Bayesian network estimation from a score-based perspective, and extend it to include exogenous variables and dependent data.

The outline of the paper is as follows. In Section 2, Bayesian networks and score metrics are briefly reviewed, and our two new score metrics for datasets with complex mean structure and random effects are presented. The new score metrics are used to estimate Bayesian networks for simulated datasets with a known complex mean structure in Section 3, and then applied to the analysis of the grape-berry gene expression data in Section 4. In Section 5, we present a brief summary of our overall findings.

2. Bayesian networks and score metrics

2.1. BGe, the basic Bayesian score metric

A Bayesian network $B$ for a random vector $X = (X_1, \ldots, X_p)^\top$ consists of two components: a directed acyclic graph $G = (V, E)$ with $V = \{X_1, \ldots, X_p\}$, often written as $V = \{1, \ldots, p\}$, and assumed conditional distributions $f(x_i | x_P, \theta_i)$, $i = 1, \ldots, p$. 

© 2012 Australian Statistical Publishing Association Inc.
The set $P_i$ is the set of parents of $X_i$ in $G$ and $\Theta = \{\theta_1, \ldots, \theta_p\}$ is the set of parameters associated with the conditional distributions. The graph and conditional distributions then specify a joint distribution for $X$:

$$f(x \mid G, \Theta) = \prod_{i=1}^{p} f(x_i \mid x_{P_i}, \theta_i).$$

(1)

Bayesian networks encode information about the conditional independence relationships between the variables in $X$. The directed Markov properties, as described in Lauritzen (2004), for example, allow conditional independence statements about $X$ to be read from the graph $G$. Additionally, when the available data set is high-dimensional and the maximum likelihood estimate of the covariance matrix of $X$ is unavailable, estimation of a Bayesian network allows the estimation of the covariance matrix. As first shown by Wermuth (1980), any Bayesian network may be written as a system of linear recursive equations:

$$X = \Gamma X + \epsilon,$$

where $\Gamma$ is an upper triangular matrix, $[\Gamma]_{ij} = 0$ if and only if there is no edge from $j$ to $i$ in the corresponding graph, and $\epsilon \sim N(0, \Psi)$, with $\Psi$ a diagonal matrix. Re-arranging this equation then gives an expression for the covariance matrix of $X$ in terms of the regression parameters $\Gamma$ and $\Psi$. Hence, the covariance matrix of $X$ can be estimated given the estimated Bayesian network parameters.

Here we consider a vector of pre-processed, normalized expression levels for $p$ genes, and suppose that $X \sim N_p(0, \Sigma)$, where $\Sigma$ is unknown. We consider a data set $d = \{x_1, \ldots, x_p\}$, where $x_i = (x_{i1}, \ldots, x_{in})^T$ is the vector containing the $n$ samples of the expression levels of gene $i$. The estimation of a Bayesian network for $X$ given this data set $d$ consists of learning the structure of a directed acyclic graph encoding the conditional independence relationships between the variables, and the estimation of the parameters $\Theta$.

Taking a score-based approach to learning structure, we want to find a graph that maximises some score metric. The Bayesian score metric avoids the problems with overfitting associated with the maximum likelihood score, (see Koller & Friedman 2009, chap. 18.3), and following Geiger & Heckerman (1994), among others, the Bayesian score metric for a directed acyclic graph $G$ given $d$ is proportional to the posterior probability of that graph:

$$S(G \mid d) = Pr(G) f(d \mid G) = Pr(G) \int_{\mathbb{R}^{np}} f(d \mid G, \Theta) f(\Theta \mid G) d\Theta.$$  

(2)

Here the focus is on the second component of this score, the marginal model likelihood of the data given the graph $G$, where the density of $d$ given $G$ and $\Theta$ is assumed to be an $np$-dimensional normal density, with mean vector $0$ and covariance matrix $\Sigma \otimes I_n$. As per (1), this joint density may be decomposed into a product of $p$ conditional densities. When the data set $d$ consists of independent and identically distributed normal samples, $\theta_i = \{y_i, \psi_i\}$, and $x_i \mid x_{P_i}, y_i, \psi_i \sim N_n(x_{P_i}y_i, \psi_i I)$.

Given normal-inverse gamma priors for each $\theta_i$, or an equivalent inverse Wishart prior on $\Theta$, the score metric of (2) can be written as the product of the prior density on the space of directed acyclic graphs, $Pr(G)$, and $p$ multivariate $t$ densities. This score metric, only appropriate in the case of independent and identically distributed samples, is known as the BGe metric: ‘Bayesian metric for Gaussian networks with score equivalence’.

© 2012 Australian Statistical Publishing Association Inc.
2.2. BGeCM, the score metric for data sets with complex mean structure

We now consider the case of a more complex data set $d$, that does not consist of independent and identically distributed samples, such as the grape-berry gene data set described in Section 1. As explained there, contained within that data set is information about exogenous variables thought to affect the expression levels of the genes under study. Given such a data set, we now express the model for the vector of expression levels of gene $i$ as

$$x_i | x_P, y_i, \psi_i, b_i, \phi_i \sim N_n(x_Py_i + Qb_i, \psi_iI),$$

where $b_i$ is the $m$-vector of the effects of the $m$ exogenous variables on gene $i$, $\phi_i$ are the parameters associated with the (as yet to be selected) prior distribution for $b_i$, and $Q$ is the $n \times m$ matrix containing the data associated with the $m$ exogenous variables. It can be seen that in this specification we retain linear dependence upon expression levels of parent genes, but now, in addition to that dependence, more complex sampling schemes and the influence of exogenous variables are accounted for through the linear dependence of expression levels upon $b_i$.

Including exogenous variables in the estimation of a Bayesian network is important in order to obtain an unbiased estimate of the conditional dependence relationships between the genes of interest. For example, the expression levels of two genes may both be dependent upon changes in an exogenous variable, but conditionally independent of each other. If dependence upon exogenous variables is not accounted for, an edge between these two genes is likely to be present in an estimated graph. By accounting for the effects of exogenous variables, we can have more confidence that the conditional dependence relationships obtained represent actual dependence relationships, and are not due to common relationships with exogenous variables.

As can be seen by the definition of the Bayesian score metric given by (2), a joint prior distribution for $y_i, \psi_i, b_i$ and $\phi_i$ is required for the calculation of a score metric. Care is required in the specification of this prior distribution since, if priors are not properly selected, a score metric will be introduced that gives different scores to directed acyclic graphs encoding equivalent conditional independence restrictions. A score metric that does not discriminate between equivalent directed acyclic graphs is called an equivalent score metric. Discrimination between equivalent graphs is tantamount to assigning causal meaning to the directed edges of $G$, and since the emphasis here is on the estimation of graphs given observational data, the assignment of causal meaning to the estimated relationships is not appropriate.

Extension of the results of Geiger & Heckerman (2002) to our model indicates that the joint prior distribution for $y_i, \psi_i, b_i$ given $\phi_i$ must have a normal-inverse gamma form, and that the effects of the exogenous variables on one gene must be a priori independent of the effects upon another gene, for the induced score metric to satisfy equivalence. The following system of priors is used:

$$y_i | \psi_i \sim N_{|P_i|}(0, \psi_I), \quad \psi_i^{-1} \sim \text{Ga} \left(\frac{\delta + |P_i|}{2}, \frac{\tau}{2}\right), \quad b_i | \phi_i \sim N_m(0, \phi_iI).$$

There are several possible assumptions about the form of the prior distribution of the variance of the random effects for gene $i$, $f(\phi_i)$. Among other choices, the variance of the random effects could be assumed known, a uniform prior could be placed on $\phi_i^{1/2}$, or an inverse gamma prior could be placed on $\phi_i$. However, by an extension of the results in Geiger &

© 2012 Australian Statistical Publishing Association Inc.
Heckerman (2002), when $\phi_i \neq \psi_i^{-1}$, any choice of prior distribution on $\phi_i$ will result in a marginal model likelihood without a closed form, requiring numerical integration to compute and slowing down computations.

A simulation study in Kasza (2009) showed that the learnt network structure is quite robust to the misspecification of the prior density for $\phi_i$, provided the magnitude of $\phi_i$ is correctly specified. Hence, for computational simplicity, the following prior for the variance of the effects of exogenous variables is used:

$$b_i \mid \psi_i \sim N_m(0, \psi_i^{-1} I),$$

where $\psi$ is some positive parameter that is constant from gene to gene. If $\psi$ is taken to be very large, this is equivalent to assuming that the effects of exogenous variables do not contribute much to the overall variability of the expression levels of genes.

Although it will be application dependent, it may be that the assumption that the variance of the exogenous variables is related to the variance of the regression parameters in the same way for each gene is not valid. In this situation, a separate $\psi_i$ could be specified for each gene, but such specification would require information that is most probably unavailable. Alternatively, a hyperprior distribution could be placed upon the $\psi_i$. However, any choice of such a distribution would lead to a score metric without an exact form.

When $b_i \sim N_m(0, \psi_i^{-1} I)$, the marginal model likelihood for a particular random variable given its parents in the graph $G$ is shown in Appendix A to be

\[
x_i \mid x_{P_i} \sim t_{\delta+|P_i|}(0, \Sigma_{x_i|x_{P_i}}),
\]

\[
\Sigma_{x_i|x_{P_i}} = \frac{\tau}{\delta + |P_i|} \left\{ J - J x_{P_i} \left( \tau I + x_{P_i}^\top J x_{P_i} \right)^{-1} x_{P_i}^\top J \right\}^{-1},
\]

\[
J = I - Q(\psi I + Q^\top Q)^{-1} Q^\top.
\]

We call the resultant score metric the BGeCM metric the ‘Bayesian metric for Gaussian networks having score equivalence for data sets with a complex mean structure’. This score metric extends the BGe metric to situations where the dataset available for estimation does not consist of independent and identically distributed samples.

Instead of using a score metric as developed above to allow for the inclusion of exogenous variables in the model, an extended directed acyclic graph could be learnt, where exogenous variables are included as vertices in the graph. However, there are a couple of difficulties presented by such an approach. The first is that if the exogenous variables are discrete, methods for Bayesian networks on both continuous and discrete variables are required. Although such methods are available, they are not entirely appropriate for this setting, as we believe that all of the genes under consideration are affected by each exogenous variable. In other words, the dependence structure of the genes on the exogenous variables is assumed to be known. Additionally, many algorithms for learning directed acyclic graphs incorporate sparsity constraints, and if it is believed that many of the genes are affected by these exogenous variables, these sparsity constraints will require modification. Although there are methods that do not constrain the sparsity of learnt graphs, the dependence structure of the genes is still assumed to be sparse. Given that we assume each gene has each exogenous variable as a parent in the graph, methods that attempt to estimate a combined Bayesian network on genes and exogenous variables are not appropriate for our application.
2.3. Removal of random effects through analysis of residuals

In the derivation of the BGeCM score metric, it was assumed that the effects of exogenous variables on gene expression were of intrinsic interest. However, in many situations, the effects of exogenous variables can be thought of as nuisance variables, complicating the estimation of Bayesian networks for the given gene expression levels. It may be desirable to ignore the possible influences of such effects upon gene expression levels, and on the relationships between genes. Of course, simply ignoring such effects is not recommended. Instead, we develop an approach that is non-parametric in the effects of exogenous variables: adjusting for the effects of exogenous variables, without making assumptions about the form of their distributions. This approach, instead of directly using the gene expression data, is based upon the use of linear combinations of residuals left over after the data is regressed upon the effects of the exogenous variables. The vectors of residuals

\[ [I - Q(Q^\top Q)^{-1} Q^\top]x_i \]

obtained after regressing the gene expressions on the effects of exogenous variables have distributions which do not depend on the \( b_i \)s. However, the matrix \( I - Q(Q^\top Q)^{-1} Q^\top \) is not of full rank, so it is better to use linearly independent combinations of residuals.

We call this the ‘residual approach’ and it is inspired by the restricted maximum likelihood procedure used in inference for mixed linear models; see, for example, Davison (2003, chap. 12.2), or Speed (1997), which provide a good overview of REML.

The utility of the residual approach is that it makes no assumptions about the distributional form of the random effects of interest. Since no such assumptions are made, the approach may be used irrespective of the true distribution of the random effects.

We consider an \((n - m) \times 1\) random variable \( y_i = P^\top x_i \), where \( P \) is an \( n \times (n - m) \) matrix such that

\[ P^\top Q = 0, \quad P^\top P = I, \quad PP^\top = I - Q(Q^\top Q)^{-1} Q^\top. \tag{4} \]

Hence,

\[ y_i | \gamma_i, \psi_i, y_{\cdot i} \sim N_{n-m}(y_{P_i}, \psi_i I), \tag{5} \]

and the score metric associated with this set of marginal model likelihoods is invariant to the choice of \( P \), as shown in Appendix C. Implementation of the residual approach to the estimation of Bayesian networks is simple: after selection of an appropriate matrix \( P \) and computation of \( y_i = P^\top x_i \) for \( i = 1, \ldots, p \), the BGe score metric may be applied to this reduced data set in conjunction with the score-based method of choice.

A drawback of the residual approach is that posterior estimates of the random effects \( b_i \) are not admitted. However, any potential loss of information about the underlying covariance matrix when the residual approach is used, compared to the ‘full’ BGeCM score metric, has been investigated in Kasza & Solomon (2011), and found to be typically small.

3. Numerical study of BGeCM and the residual score metrics

In this section, the necessity of score metrics that take account of complex mean structure are demonstrated through the application of the residual approach and the BGeCM score metric to simulated data sets.
First, a note on implementation. The BGeCM score metric and the residual approach may be incorporated into any score-based algorithm for the estimation of Bayesian networks, without the need for any additional programming. In the case of the residual approach, the input to the algorithm is $P^T d$, where $P$ is as in (4). Similarly, when the BGeCM score metric is used, the input to the algorithm is $L^T d$, where $L$ is a matrix such that $J = LL^T$, where $J$ is as given in (3).

The residual approach and BGeCM score metrics, in conjunction with the high-dimensional Bayesian covariance selection algorithm of Dobra et al. (2004), were applied to data sets generated according to our examples 3.1 and 3.2 described below. This algorithm works by constructing and combining regression models for each $X_i$. Given that the problem of finding the highest-scoring directed acyclic graph given a decomposable score metric is NP-hard (Chickering et al. 2003), this algorithm is not guaranteed to find the highest-scoring graph. All that we can do is explore a small part of the space of graphs, and note that there may well be artefacts in the highest-scoring graph found.

**Example 1.** For this example, 10 data sets were generated according to the following system of linear recursive equations:

$$X_{ijk} = b_{ij} + \epsilon_{ijk}, \quad \epsilon_i \sim N(0, \psi_i) \quad (i = 1, \ldots, 100; \; j = 1, 2; \; k = 1, \ldots, 50).$$

The values of $\psi_i$ were obtained by sampling from an inverse gamma$(1, 1/2)$ distribution, and were constant for each of the samples generated. Similarly, $b_i = (b_{i1}, b_{i2})^T$, $i = 1, \ldots, 100$, were fixed across data sets, obtained by sampling from $b_{ij} \sim N(0, \psi_i)$, corresponding to $\nu = 1$. The non-zero mean structure of this example corresponds to two groups, and the true underlying directed acyclic graph is the empty graph.

**Example 2.** Data sets were simulated from the following densities:

$$x_i | \psi_i, b_i \sim N_{10}(Qb_i, \psi_i I) \quad (i = 1, \ldots, 18),$$

$$x_k | \gamma_k, \psi_k, b_k \sim N_{10}(x_{P_k} \gamma_k + Qb_k, \psi_k I), \quad (k = 19, 20)$$

where $x_i = (x_{i1}, \ldots, x_{i10})^T$, $x_{P_{19}} = (x_1, x_2)$, $x_{P_{20}} = x_{19}$. The random effects $b_i = (b_{i1}, b_{i2}, b_{i3})^T$ were constant across the 10 data sets generated, obtained by sampling from $b_{ij} \sim N(0, \psi_i)$, again corresponding to $\nu = 1$. The $10 \times 3$ matrix $Q$ was also constant across data sets, consisting of random samples from the standard normal distribution.

The values of the parameters $\psi_i$, $\gamma_{19} = (\gamma_{19.1}, \gamma_{19.2})^T$ and $\gamma_{20,19}$ were simulated from the following distributions:

$$\psi_i \sim \text{Inv Gamma}\left(\frac{2 + |P_i|}{2}, \frac{1}{2}\right), \quad |P_i| = 0 \quad (i = 1, \ldots, 18),$$

$$|P_{19}| = 2, \quad |P_{20}| = 1,$$

$$\gamma_{19} \sim N_2(0, \psi_{19} I), \gamma_{20,19} \sim N(0, \psi_{20}).$$

The true graph underlying this example has edges from $x_1$ and $x_2$ to $x_{19}$, and from $x_{19}$ to $x_{20}$.

For each of the simulated data sets, the number of spurious and correct edges in the highest-scoring network found by the algorithm are summarized in Tables 1 and 2.
Table 1 gives the mean and standard deviation of the number of spurious and correct edges in the highest-scoring Bayesian networks found when BGe, BGeCM and the residual approach, with $\nu = 1$, were used to estimate the true graph encoding the conditional independence relationships. Table 2 gives the numbers of correct and spurious edges when BGeCM was used, given a range of values of $\nu$. The results in Tables 1 and 2 show that by taking account of the exogenous variables, a better estimate of structure was obtained. In addition, Table 2 shows that the results obtained from the BGeCM score metric were quite robust to the misspecification of $\nu$, producing accurate results when the value of $\nu$ selected departed as much as one or two orders of magnitude from its true value. As $\nu$ became larger, the highest scoring graphs obtained became more and more similar to those obtained when the BGe metric was used. This is a result of the fact that as $\nu$ approaches $\infty$, the limit of the BGeCM metric is the BGe metric, a proof of which is given in Appendix B.
TABLE 3

The grape heat shock genes. The first column gives the gene reference numbers used in this study, the second column gives the Affymetrix reference numbers, and the third column gives the National Center for Biotechnology Information reference numbers. The fourth column provides a short description of the function of the genes. HSP stands for heat shock protein.

| Ref # | Affymetrix # | NCBI # | Protein Identity |
|-------|--------------|--------|------------------|
| 1     | 1616246.at   | Vvi.9142 | Heat shock protein 70, ATP binding |
| 2     | 1607002.at   | Vvi.4801 | Heat shock protein 70, ATP binding, luminal binding protein, glucose regulated chloroplast HSP 70-1, ATP binding |
| 3     | 1610684.at   | Vvi.2869 | Heat shock protein 70, ATP binding, luminal binding protein, glucose regulated chloroplast HSP 70-1, ATP binding |
| 4     | 1611740.at   | Vvi.295 | HSP70, luminal binding protein, ATP binding |
| 5     | 1620985.at   | Vvi.4530 | HSP21 chloroplast |
| 6     | 1616995.at   | Vvi.23518 | Ubiquitin conjugating enzyme 4e |
| 7     | 1614132.at   | Vvi.863 | Ubiquitin conjugating enzyme 4e |
| 8     | 1618265.at   | Vvi.15427 | Ubiquitin conjugating enzyme 4e |
| 9     | 1608052_s.at | Vvi.9085 | HSP81(early response to dehydration) |
| 10    | 1618009.at   | Vvi.9085 | HSP81(early response to dehydration) |
| 11    | 1619931_s.at | Vvi.7394 | HSP81(early response to dehydration) |
| 12    | 1608701.at   | Vvi.2083 | 10 kDa chaperonin |
| 13    | 1608164.at   | Vvi.6787 | Cytosolic class II 17.6 HSP |
| 14    | 1611052.at   | Vvi.6787 | Cytosolic class II 17.6 HSP |
| 15    | 1611192.at   | Vvi.6787 | Cytosolic class II 17.6 HSP |
| 16    | 1610032.at   | Vvi.6787 | Cytosolic class II 17.6 HSP |
| 17    | 1614330.at   | Vvi.6787 | Cytosolic class II 17.6 HSP |
| 18    | 1620956.at   | Vvi.3921 | 17.6 kDa class I small HSP |
| 19    | 1616538.at   | Vvi.7869 | 17.6 kDa class I small HSP |
| 20    | 1609554.at   | Vvi.7044 | 17.6 kDa class I small HSP |
| 21    | 1620960_a.at | Vvi.7044 | 17.6 kDa class I small HSP |
| 22    | 1621652.at   | Vvi.4464 | 17.6 kDa class I small HSP |
| 23    | 1622165.at   | Vvi.6156 | 17.6 kDa class I small HSP |
| 24    | 1612385.at   | Vvi.4422 | 17.6 kDa class I small HSP |
| 25    | 1622628.at   | Vvi.5040 | 17.6 kDa class I small HSP |
| 26    | 1610700_at   | Vvi.2537 | 17.6 kDa class I small HSP |

4. Analysis of the grape-berry microarray data

The data analysed here consist of 50 samples of gene expression levels for 26 grape genes, measured over a five-week period. The gene expression levels were derived from grape-berry tissue samples grown in three different vineyards in three different wine-growing regions of southern Australia. Twenty samples were taken from a vineyard in Clare, 20 from the Wingara Vineyard in Mildura and 10 from a vineyard in Willunga. Table 3 provides the reference numbers for the 26 grape genes, together with a brief summary of their functions. All the genes in Table 3 are known to code for heat shock proteins (HSPs, Wang et al. 2004), which are responsible for protecting the grapes against heat-induced stress. In addition to data on the gene expression levels, temperature in degrees Celsius was recorded during the time leading up to the picking of the grapes.

These data are part of a larger dataset on grape berry tissue samples measured between 2003 and 2005. At each vineyard, grapes were sampled roughly weekly over the period of development of the berries, i.e., from the time buds formed on the vines, to the time when the grapes were ripe. In general, grape berries follow a double sigmoidal pattern of growth that consists of two distinct growth phases, with a lag period between these phases (see Coombe...
The second stage of grape-berry growth commences upon the occurrence of veraison, when the grape berries start to change colour. Robinson & Davies (2000) suggest that at veraison and during ripening, there are many changes in the expression levels of many different genes in grape berries. The observed developmental time period from bud formation to grape ripeness differed between vineyards in the present study. The shorter five-week time period we analysed occurred after fruit set, but well before veraison for all three vineyards. We restricted attention to these samples, corresponding to the third to seventh sampling weeks at each vineyard, because the relationships between genes are thought to be more stable during this period, and the modelling assumption of identically distributed samples is therefore more likely to be valid.

mRNA expression levels for each of the grape tissue samples was measured using Affymetrix Vitis vinifera oligonucleotide arrays. Background subtraction and normalisation was carried out using robust microarray analysis (RMA), as described in Irizarry et al. (2003). All samples were processed at the same laboratory.

Understanding the stress tolerance mechanisms of plants is important, and the heat shock protein network, as discussed by Kotak et al. (2007) and Wang et al. (2004), is very complex. The heat shock protein network of plants is believed to consist of interactions between small HSPs, HSP60, HSP70, HSP90 and HSP100 (Wang et al. 2004). Precisely how HSPs interact with one another and how they protect against heat stress is not yet completely understood, and here we seek to gain some insight into the heat shock protein network by examining the conditional dependence structure of the genes given in Table 3.

Given the known functions of the genes considered in this study, and the climatic and geographic disparities between the regions where the grape berries were sampled, it would be incorrect to ignore the effects of vineyard and temperature in the estimation of a Bayesian network for the grape genes. The essential point is that if the expression levels of these genes are strongly influenced by these exogenous variables, then accounting for variation due to such variables in the estimation of a Bayesian network should result in a network that more accurately encodes the dependence structure of the genes. A further important point is that causal interpretations should not be applied to the directed edges present in any network estimated from the data.

To begin, the initial (null) model omitted the effects of vineyard and temperature on the expression levels of the genes. That is, if \( x_i \) is the 50-vector of the expression levels for grape gene \( i \), it is assumed that

\[
x_i | x_P, y_i, \psi_i \sim N_{50}(x_P, y_i, \psi_i I),
\]

\[
y_i | \psi_i \sim N_{|P_i|}(0, \tau^{-1}\psi_i I), \quad \psi_i^{-1} \sim Ga\left(\frac{\delta + |P_i|}{2}, \frac{\tau}{2}\right),
\]

where \( x_P \) is a 50 \( \times \) \( |P_i| \) matrix. The columns of this matrix consist of the expression levels of the grape genes in the dataset that the expression level of gene \( i \) is dependent upon, \( y_i = (y_{ij})_{j \in P_i} \), and \( y_{ij} \) is the effect of the expression level of gene \( j \) on the expression level of gene \( i \). Following the analysis of Affymetrix gene expression data in Dobra et al. (2004), \( \tau = 1 \) and \( \delta = 2 \).

The highest scoring Bayesian network found through the application of the high-dimensional Bayesian covariance selection algorithm to the full dataset (ignoring the exogenous variables of vineyard and temperature) has 43 edges, and is shown in Figure 1(a).
Figure 1. The highest-scoring directed acyclic graphs obtained for the grape genes when (a) the effects of temperature and vineyard are ignored, and when the residual approach is used to include (b) vineyard effects, and (c) vineyard and temperature effects.

Next, graphs were estimated separately for each of the three vineyards. The highest-scoring directed acyclic graphs obtained for the Clare, Wingara and Willunga vineyards had, respectively, 22, 23 and 17 edges. These graphs were quite different from one another, with the three graphs having only two edges in common, the Wingara and Clare graphs sharing eight edges, and the Willunga graph having three edges in common with the Wingara and Clare graphs. Given the paucity of the data and the complexity of the models, this lack of concordance between the graphs obtained separately for each vineyard is not surprising.

In order to make more efficient use of the data, models incorporating data from all three vineyards simultaneously were then considered.

The question of how best to include temperature and vineyard effects in the model for gene expression was investigated using linear regression models with forward and backward selection. The largest model fitted for each gene contains separate intercepts for the data from
each vineyard, and terms for each of the temperatures recorded 30, 90, 150, 210, 270 and 330 minutes before the grapes were picked. We also considered the model including vineyard and temperature main effects and two-way temperature interactions. For the full model with interactions, it was observed that the adjusted $R^2$ of many of the regressions was above 0.99, indicating that some over-fitting was taking place. We therefore exclude two-way temperature interaction effects in what follows.

Results of the stepAIC function in R (R Development Core Team 2007), indicate that each of the vineyard or temperature variables is significant in at least one of the 26 regression models estimated. In any case, use of separate regression models for each gene is beyond the scope of the present score metrics. As such, the largest model considered is as follows:

$$x_i | x_P, y_i, \psi_i, b_i \sim N_{50}(x_P y_i + Q b_i, \psi_i I),$$

$$y_i | \psi_i \sim N_{|P_i|}(0, \tau^{-1}\psi_i I),$$

$$\psi_i^{-1} \sim Ga\left(\frac{\delta + |P_i|}{2}, \frac{\tau}{2}\right),$$

where $b_i = (b_{i1}, \ldots, b_{i9})^\top$ and $b_{ij}$ ($j = 1, 2, 3$) is the effect of vineyard $j$ on the expression level of gene $i$, and $b_{i4}, \ldots, b_{i9}$ are the temperature effects.

Histograms of the marginal standard deviations of the expression data for each gene, and the residual standard errors from the regressions containing just vineyard, and vineyard and temperature together, are shown in Figure 2. There are three genes with very small standard deviations. These plots show that vineyard variables account for only some of the variation in the gene expression levels. Changes in temperature and vineyard account for much more of the observed variation, but there remains some residual variation to be explained. On the basis of these histograms, we expect that the graph obtained when only vineyard is included as an exogenous variable will be somewhat similar to that obtained when the exogenous variables are ignored, whilst we would expect to see a reasonably different structure when both vineyard and temperature are accounted for.

In accounting for the effects of vineyard and temperature, we find high-scoring Bayesian networks using the BGeCM score metric, first fitting the model with vineyard effects only, then the model with vineyard and main temperature effects. The highest-scoring Bayesian networks found for $\nu = 0.5, 1$ and 10 were recorded and are summarized in Table 4. It can be seen that as more covariates are included in the model, more of the variation in the expression levels of the grapes is explained, and the highest-scoring graphs obtained have fewer edges. Edges that are removed as more exogenous variables are included in the model can be interpreted as being explained by common relationships of genes with these additional covariates.

The BGeCM score metric assumes that the effects of the exogenous variables are independent and identically distributed, an assumption that must be questioned. The effects of temperature and vineyard are almost certainly not i.i.d. There is, however, little information available to provide a useful estimate of the covariance structure of the effects of these exogenous variables. Therefore, the residual approach, which makes no assumptions about the covariance structure of the effects included in the model, is preferred here.
Figure 2. Histograms of (a) the marginal standard deviations of the grape gene expression levels, (b) the residual standard errors after regressing the expression levels on vineyard only, and (c) the residual standard errors after regressing expression levels on vineyard and temperature.
The number of edges in the highest-scoring networks obtained using the residual method are summarized in Table 4. When only the effects of vineyards are included in the model, the results obtained using the residual method are similar to those obtained when the BGeCM score is used, as expected on the basis of the histograms in Figure 2. When the effects of temperature are included in the model, the residual method produces high-scoring graphs with fewer edges than the BGeCM score metric. This indicates that whilst the BGeCM score metric may account correctly for the covariance structure of the effects of vineyard, the effects of temperature may have a more complicated variance structure, that is not adequately modelled by the iid assumption.

The directed acyclic graphs obtained from the residual method are displayed in Figure 1. These graphs, drawn using GraphViz (Ellson et al. 2004), show that as more of the variation in gene expression due to exogenous sources is accounted for in the model, the highest-scoring networks obtained have fewer edges. The graph obtained by including both temperature and vineyard as exogenous variables, Figure 1(c), is preferable to that obtained when only vineyard is included, Figure 1(b). For most genes, very little variation in the gene expression values is accounted for by the relationship with vineyard alone.

There are a number of interesting features to be observed in Figure 1(c), which is the graph obtained when both vineyard and temperature effects are accounted for in the model. We observe that seven nodes in this graph are completely disconnected from all other nodes, which implies that once relationships with temperature and vineyard have been accounted for, the expression levels of each of these genes are independent of the expression levels of all other genes. (Recall that absence of an edge between two genes in Figure 1(c) indicates that the expression levels of these nodes are independent, a relationship which can be refined through application of Markov properties.)

It is apparent that three of these seven disconnected nodes, corresponding to genes 14, 18 and 23, are already disconnected from the rest of the graph when only vineyard is included in the model; see Figure 1(b), where it is observed that these are the only three unconnected nodes. The expression levels of these three genes are observed to have the lowest standard deviations of all the genes, at 0.037, 0.034 and 0.068 respectively. When these three genes are regressed on vineyard, the residual standard deviations are even smaller (0.029, 0.023 and 0.047). In other words, there is no variation in the expression levels of these genes to be explained by relationships with other genes, so it is not surprising that they are unconnected in the graph. Very small gene standard deviations can be problematic in microarray data analysis, and methods have been proposed for adjusting the standard deviation estimates upwards by adding a constant term or by application of empirical Bayes methods when constructing

| Included Covariates              | BGeCM $\nu = 0.5$ | BGeCM $\nu = 1$ | BGeCM $\nu = 10$ | Residual |
|----------------------------------|-------------------|-----------------|------------------|----------|
| Vineyard                        | 42                | 43              | 48               | 37       |
| Vineyard and temperature         | 35                | 37              | 41               | 23       |
Genes 9, 10 and 11 are also disconnected in the final graph, Figure 1(c), and correspond to HSP81, which is an early response to dehydration. According to the KEGG data base (Kanehisa et al. 2010), they are predicted to be similar to HSP90. The role of these sets of genes in the heat shock network of grapes is not entirely understood. The role of HSP81 proteins in *Arabidopsis thaliana*, more commonly known as thale cress, has been discussed in Yabe et al. (1994), who suggest that an increase in the expression level of HSP81-1 is possibly caused by a regulatory pathway other than the heat shock pathway. Our analysis supports this finding for *Vitis Vinifera*, indicating that HSP81 may not be implicated in the heat shock protein network of grapes, at least over the four-week time period studied. We have established that variation in the HSP81 genes is accounted for directly by the effects of the exogenous variables, and that they are uncorrelated with the other HSPs in the final graph.

The seventh gene, number 26, which is a mitochondrial small HSP, is not implicated in the final network either. This gene is the only mitochondrial gene considered. That it is unconnected from the rest of the network indicates that variation in this gene is explained purely by exogenous temperature and vineyard effects, and is not dependent upon any of the other genes in the dataset. This suggests that the mitochondrial HSPs are not regulated in the same way as other cellular HSPs, over the time period considered.

On the whole, relatively little is known about the heat shock regulatory network for grapes. Typically in the representation of the heat response network for plants, relationships between classes of genes, such as small HSPs or HSP70s, are the focus of discussion (Wang et al. 2004). The final graph obtained here provides a good starting point for the development of a finer regulatory structure, which can then be further developed. In particular, the edges between the genes in Figure 1(c) may be interpreted as encoding conditional dependence relationships. Detailed investigation of the connected nodes in these graphs, whilst of interest, is beyond the scope of the present analysis.

5. Discussion

The BGeCM score metric and the residual approach presented in this paper enable Bayesian network structures to be learnt given datasets that do not consist of independent and identically distributed samples, and may be used in conjunction with any score-based method for the estimation of a Bayesian network. Furthermore, the residual approach allows the estimation of a Bayesian network for datasets with a complex mean structure without the need to specify the variance structure of the mean effects. This approach proved useful for the analysis of grape-berry gene data, where it could not reasonably be supposed that the effects of the exogenous variables were independent and identically distributed. Our analysis of the grape-berry gene microarray data has resulted in biologically plausible conclusions on the heat shock regulatory network of grape genes. These inferences could not have been drawn without the availability of suitable score metrics to account for the effects of exogenous variables.
Appendix A: Derivation of BGeCM

Assuming \( x_i | x_p, \psi, \phi_i \sim N_n(x_p, \psi b_i, \phi_i I) \), with prior densities given by

\[
\psi_i | \tau_i \sim \text{Ga} \left( \frac{\delta + |P_i|}{2}, \frac{\tau}{2} \right), \quad \phi_i | \psi_i \sim N_m(0, \psi^{-1} \psi_i I),
\]

we now derive the marginal model likelihood for \( x_i \), as given in (3), where \( a_i = |P_i| : \)

\[
f(x_i, \psi, b_i | x_p) = (2\pi)^{-\frac{(n+a_i+m)/2}{2}} \frac{\tau^{a_i/2} \psi^{m/2}}{\Gamma(a_i/2)} \exp\left\{ -\frac{\tau}{2} \psi_i \right\} \times \exp \left\{ -\frac{\tau}{2}\psi_i \frac{b_i^T b_i - x_i - x_p, \psi_i \phi_i - Q b_i}{2} \right\}
\]

The terms involving \( b_i \) may be rearranged to give

\[
-\frac{\nu}{2\psi_i} b_i^T b_i - \frac{\tau}{2\psi_i} \psi_i \phi_i - 1 \frac{1}{2\psi_i} (x_i - x_p, \psi_i \phi_i - Q b_i)^T
\]

Substituting this into \( f(x_i, \psi, b_i | x_p) \) gives a normal kernel in \( b_i \), so integrating over \( b_i \) gives

\[
f(x_i, \psi, | x_p) = (2\pi)^{-\frac{(n+a_i)/2}{2}} \frac{\tau^{a_i/2} \psi^{m/2}}{\Gamma(a_i/2)} \exp\left\{ -\frac{\tau}{2} \psi_i \right\} \times \exp \left\{ -\frac{\tau}{2}\psi_i \frac{I - Q (v I + Q^T Q)^{-1} Q^T (x_i - x_p, \psi_i)}{2} \right\}
\]

Letting \( J = I - Q (v I + Q^T Q)^{-1} Q^T \), the terms involving \( \psi_i \) may be dealt with similarly, and integrating over \( \psi_i \) then gives

\[
f(x_i, \psi_i | x_p) = (2\pi)^{-\frac{(n/2)}{2}} \frac{\tau^{a_i/2} \psi^{m/2}}{\Gamma(a_i/2)} \exp\left\{ -\frac{\tau}{2} \psi_i \right\} \times \exp \left\{ -\frac{\tau}{2}\psi_i \frac{J - J x_p, (v I + Q^T Q)^{-1} Q^T J x_p)}{2} \right\}
\]

© 2012 Australian Statistical Publishing Association Inc.
This is an inverse-gamma kernel in $\psi_i$, with

$$\alpha = \frac{n + a_i + \delta}{2}$$

$$\beta = \frac{\tau}{2} + \frac{1}{2} x_i^\top \{ J - J x_{p_i} (\tau I + x_{p_i}^\top J x_{p_i})^{-1} x_{p_i}^\top J \} x_i$$

so integrating over $\psi_i$ gives

$$f(x_i | x_{p_i}) = (2\pi)^{-n/2} \frac{\Gamma((\delta + a_i + n)/2)}{\Gamma(\delta + a_i)/2} \left( \frac{\tau}{2} \right)^{(\delta + a_i)/2}$$

$$\times \tau^{a_i/2} u^{m/2} |v I + Q^\top Q|^{-(1/2)} \left| \tau I + x_{p_i}^\top J x_{p_i} \right|^{-(1/2)}$$

$$\times \left[ \frac{\tau}{2} + \frac{1}{2} x_i^\top \{ J - J x_{p_i} (\tau I + x_{p_i}^\top J x_{p_i})^{-1} x_{p_i}^\top J \} x_i \right]^{-(\delta + a_i + n)/2}$$

$$= \pi^{-(n/2)} \tau^{(a_i-n)/2} \frac{\Gamma((\delta + a_i + n)/2)}{\Gamma(\delta + a_i)/2} u^{m/2} |v I + Q^\top Q|^{-(1/2)} \left| \tau I + x_{p_i}^\top J x_{p_i} \right|^{-(1/2)}$$

$$\times \left[ 1 + \frac{1}{\tau} x_i^\top \{ J - J x_{p_i} (\tau I + x_{p_i}^\top J x_{p_i})^{-1} x_{p_i}^\top J \} x_i \right]^{-(\delta + a_i + n)/2}.$$

Hence, when $b_i \sim N_n(0, \psi^{-1} \psi_i I)$, $x_i | x_{p_i} \sim t_{\delta+\alpha_i}(0, \Sigma x_i | x_{p_i})$, with

$$\Sigma x_i | x_{p_i} = \frac{\tau}{\delta + a_i} \{ J - J x_{p_i} (\tau I + x_{p_i}^\top J x_{p_i})^{-1} x_{p_i}^\top J \}^{-1}.$$

The case when $x_i$ has no parents also needs to be considered. In such a case, $a_i = 0$, and

$$x_i \sim t_\delta(0, \Sigma x_i),$$

with

$$\Sigma x_i = \frac{\tau}{\delta} J^{-1} = \frac{\tau}{\delta} \{ I - Q (v I + Q^\top Q)^{-1} Q^\top \}^{-1}.$$

Appendix B: Proof that BGeCM $\rightarrow$ BGe as $v \rightarrow \infty$

We examine the behaviour of BGeCM as $v \rightarrow \infty$. Examination of the marginal model likelihood given in (3), shows that $v$ appears in the marginal model likelihood in the terms $v^{m/2} |v I + Q^\top Q|^{-(1/2)}$ and $I - Q (v I + Q^\top Q)^{-1} Q^\top$. These terms may be written as follows:

$$v^{m/2} |v I + Q^\top Q|^{-(1/2)} = |I + v^{-1} Q^\top Q|^{-(1/2)}$$

$$I - Q (v I + Q^\top Q)^{-1} Q^\top = I - v^{-1} Q (I + v^{-1} Q^\top Q)^{-1} Q^\top.$$
Appendix C: Residual approach invariant to choice of $P$

It will now be shown that the residual approach score metric is not dependent upon the (somewhat arbitrary) choice of $P$. Consider the marginal model likelihood of $y_i$, from (5):

$$y_i | y_{P_i} \sim t_{\delta+a_i} \left( 0, \frac{\tau}{\delta+a_i} \{I_{n-m} - y_{P_i} (\tau I + y_{P_i}^T y_{P_i})^{-1} y_{P_i}^T \}^{-1} \right).$$

This implies that

$$\bar{f}(y_i | y_{P_i}) = \frac{\Gamma(\delta + a_i + n - m)/2}{\Gamma(\delta + a_i)/2 (\pi \tau)^{(n-m)/2}} \left| I - y_{P_i} (\tau I + y_{P_i}^T y_{P_i})^{-1} y_{P_i}^T \right|^{-1/2} \times \left[ 1 + \frac{1}{\tau} y_i^T \{I - y_{P_i} (\tau I + y_{P_i}^T y_{P_i})^{-1} y_{P_i}^T \} y_i \right]^{-((\delta+a_i+n-m)/2)}$$

$$= \frac{\Gamma(\delta + a_i + n - m)/2}{\Gamma(\delta + a_i)/2 (\pi \tau)^{(n-m)/2}} \left| I - P^T x_{P_i} (\tau I + x_{P_i}^T PP^T x_{P_i})^{-1} x_{P_i}^T P \right|^{-1/2} \times \left[ 1 + \frac{1}{\tau} x_i^T \{PP^T - PP^T x_{P_i} (\tau I + x_{P_i}^T PP^T x_{P_i})^{-1} x_{P_i}^T PP^T \} x_i \right]^{-((\delta+a_i+n-m)/2)}.$$

Using the identity

$$\left| I - P^T x_{P_i} (\tau I + x_{P_i}^T PP^T x_{P_i})^{-1} x_{P_i}^T P \right|^{1/2} = \tau^{a_i/2} \left| (\tau I + x_{P_i}^T PP^T x_{P_i})^{-1} \right|^{1/2},$$

it can be seen that $P$ only appears in the marginal model likelihood as $PP^T = I - \mathbf{Q}(\mathbf{Q}^T \mathbf{Q})^{-1} \mathbf{Q}^T$.

References

CHICKERING, D. M., MEEK, C. & HECKERMAN, D. (2003). Large-sample learning of Bayesian networks is NP-hard. In Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence, 124–133.

COOMBE, B. G. (1973). The regulation of set and development of the grape berry. *Acta Horticulturae* **34**, 261–271.

DAVISON, A. C. (2003). *Statistical Models*. Cambridge, U.K.: Cambridge University Press.

DE JONG, H. (2002). Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.* **9**, 67–103.

DOBRA, A., HANS, C., JONES, B., NEVINS, J. R. & WEST, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90**, 196–212.

DYKSTRA, R. L. (1970). Establishing the positive definiteness of the sample covariance matrix. *Ann. Math. Stat.* **41**, 2153–2154.

ELLSON, J., GANSNER, E. R., KOUTSOFIOS, E., NORTH, S. C. & WOODHULL, G. (2004). Graph drawing software. In *Graphviz and dynagraph - static and dynamic graph drawing tools*, pp. 127–148. Springer-Verlag.

FRIEDMAN, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **303**, 799–805.

FRIEDMAN, N., LINIAL, M., NACHMAN, I. & PE’ER, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* **7**, 601–620.

HECKERMAN, D. & GEIGER, D. (1994). Learning Gaussian networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*.

© 2012 Australian Statistical Publishing Association Inc.
GEIGER, D. & HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30**, 1412–1440.

IRIZARRY, R. A., HOBB, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLI, K. J., SCHERF, U. & SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

KANEHISA, M., GOTO, S., FURUMICHI, M., TANABE, M. & HIRAKAWA, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research* **38**, D355–D360.

KASZA, J. & SOLOMON, P. (2011). A comparison of score-based methods for estimating Bayesian networks using the Kullback-Leibler divergence arXiv:1009.1463v2.

KASZA, J. E. (2009). Bayesian networks for high-dimensional data with complex mean structure. PhD thesis, The University of Adelaide.

KOLLER, D. & FRIEDMAN, N. (2009). *Probabilistic graphical models*. Cambridge, U.S.A.: MIT Press.

KOTAK, S., LARKINDALE, J., LEE, U., VON KOSKULL-DÖRING, P., VIERLING, E. & SCHARF, K. D. (2007). Complexity of the heat stress response in plants. *Current opinion in plant biology* **10**, 310–316.

LAURITZEN, S. L. (2004). *Graphical Models*. Oxford: Clarendon Press.

MARKOWETZ, F. & SPANG, R. (2007). Inferring cellular networks - a review. *BMC Bioinformatics* **8**(Suppl 6), S5.

R Development Core Team. (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

ROBINSON, S. P., & DAVIES, C. (2000). Molecular biology of grape berry ripening. *Australian Journal of Grape and Wine Research* **6**, 175–188.

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3, Article 3.

SPEED, T. P. (1997). Restricted maximum likelihood (REML). In *Encyclopedia of Statistical Sciences Update Volume 1*. New York: Wiley.

SPIRITES, P., GLYMOUR, C. & SCHEINES, R. (1993). *Causation, Prediction, and Search*. New York: Springer-Verlag.

WANG, W., VINOCUR, B., SHOSEYOV, O. & ALTMAN, A. (2004). Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends in Plant Science* **9**, 244–252.

WERMUTH, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.* **75**, 963–972.

YABE, N., TAKAHASHI, T. & KOMEDA, Y. (1994). Analysis of tissue-specific expression of *Arabidopsis thaliana* HSP90-family gene HSP81. *Plant cell physiology* **35**, 1207–1219.