Statistical Applications in Genetics and Molecular Biology

Volume 4, *Issue* 1 2005 *Article* 28

Reverse Engineering Galactose Regulation in Yeast through Model Selection

Vesteinn Thorsson* Michael Hörnquist†

Andrew F. Siegel‡ Leroy Hood**

*Institute for Systems Biology, first two authors contributed equally, thorsson@systemsbiology.org

†Linköping University, first two authors contributed equally, micho@itn.liu.se

‡University of Washington, asiegel@u.washington.edu

**Institute for Systems Biology, lhood@systemsbiology.org

Reverse Engineering Galactose Regulation in Yeast through Model Selection*

Vesteinn Thorsson, Michael Hörnquist, Andrew F. Siegel, and Leroy Hood

Abstract

We examine the application of statistical model selection methods to reverse-engineering the control of galactose utilization in yeast from DNA microarray experiment data. In these experiments, relationships among gene expression values are revealed through modifications of galactose sugar level and genetic perturbations through knockouts. For each gene variable, we select predictors using a variety of methods, taking into account the variance in each measurement. These methods include maximization of log-likelihood with Cp, AIC, and BIC penalties, bootstrap and cross-validation error estimation, and coefficient shrinkage via the Lasso.

KEYWORDS: Model selection, regression, AIC, BIC, MDL, Cp, Bootstrap, Lasso

---

*M.H. acknowledges financial support from The Swedish Foundation for International Cooperation in Research and Higher Education (STINT), the center for industrial IT (CENIIT), the Swedish research council (VR), and The Carl Trygger Foundation (CTS). V.T. was supported by NIH Grant P20 GM64361. A.F.S. holds the Grant I. Butterbaugh Professorship at the University of Washington. We thank J. Hertz and R. Bonneau for helpful discussions. The authors Thorsson and Hörnquist contributed equally to this work.

1 Introduction

DNA microarrays are an increasingly important technology for assaying gene expression on a genome-wide scale. The ability to measure the expression levels of tens of thousands of genes simultaneously has underscored the need for appropriate statistical analysis tools. Considerable effort has been devoted to developing methods to assess which genes are differentially expressed in comparative microarray measurements (Ideker *et al.*, 2000b). Other work has focused on clustering, or expression profiling, of such data from several different experimental conditions. The underlying idea has been that if genes behave similarly under different experimental conditions, they are probably co-regulated and may be involved in similar biological processes (Arkin *et al.*, 1997). This type of analysis is based on *pairwise* comparisons of expression profiles. Both of these ways of handling transcript data have also been used for forming networks among the genes. For example, the former approach has been utilized in order to form a "disruption network" among the genes in yeast by Rung *et al.* (2002), while the later has been used to obtain couplings among 287 transcripts from single gene deletion, also in yeast (Farkas et al., 2003).

However, in order to take into account possible *combined* effects of multiple genes, methods are needed to identify relations beyond those implied by pairwise similarity. Dependencies among gene levels or activities may be modeled as a network, which is a representation of the complex system of biological reactions underlying regulatory control. Networks representing dependencies among gene expression levels have been studied in a variety of different formalisms, including Boolean (Wuensche, 1998, Ideker *et al.*, 2000a), Bayesian (Friedman *et al.*, 2000, Yoo *et al.*, 2002), and a host of other descriptions, as surveyed in Jong (2002). In one class of models, gene dependencies are described by weighted linear combinations of other gene levels (Weaver *et al.*, 1999, Wahde and Hertz, 2001). These models have been used to describe dynamical evolution of expression timecourses, and involve a range of strategies for fitting weight coefficients, ranging from direct least squares (D'haeseleer *et al.*, 1999, van Someren *et al.*, 2000) to imposing constraints such as sparseness (Yeung *et al.*, 2002). With the growth in gene expression measurement technologies, interest in method development for the extraction of networks from gene expression data continues to increase (Pournara and Wernisch, 2004, Yu *et al.*, 2004, Bickel, 2005, Schäfer and Strimmer, 2005).

Here, we study the inference, or reverse engineering, of networks within the context of the model selection problem, and evaluate these methods using the well-characterized galactose utilization network in yeast. Yeast metabolizes

galactose with the aid of enzymes encoded by the genes GAL1, GAL5, GAL7, and GAL10. The production of these enzymes is initiated by a "switch", which is "turned on" by addition of galactose sugar. In the absence of galactose, the protein Gal80p (encoded by GAL80) inhibits the transcriptional activator Gal4p (GAL4). In the presence of galactose, this inhibition is lost, a process facilitated by the protein Gal3p and by increased transport of galactose into the cell by Gal2p (GAL2). The expression of the genes GAL1, GAL2, GAL3, GAL4, GAL5, GAL6, GAL7, GAL10, and GAL80 was assayed with DNA microarrays under 20 conditions. These conditions represent the wild type cell, and cells where each of nine genes in the galactose pathway has been knocked out, measured in both galactose rich and galactose poor media (Ideker *et al.*, 2001). A gene knockout is a genetic manipulation which forces the expression level of a gene to be zero, independent of other variables. Our network prediction methods take account of the measured experimental error in gene expression levels for each gene and condition.

The organization of the paper is as follows. Section 2 covers data considerations, including error assessment and data transformation. Section 3 is devoted to a description of the model. Section 4 forms the core of the paper, where we present different strategies for choosing a best model and present results of our calculations. Finally, in Section 5 we draw conclusions and discuss what can be learned from studies of this kind.

2 Data set and preprocessing

Experimental and data processing protocols are described in Ideker et al. (2001). Here, our initial data consists of mean *absolute* (*i.e.* not relative) expression level for each gene and each condition, as assayed in these experiments (Ideker *et al.* 2000b). In addition, from the replicated microarrays, we estimated the standard error of the mean for each gene and each condition, by second-order expansion of the maximum-likelihood function. The absolute values are logarithmically transformed, log (absolute value), and the standard errors accordingly as $\log[1+(\text{standard error})/(\text{absolute value})]$. (Absolute values less than 0.001 are set to 0.001 prior to transformation.) We henceforth refer to these transformed values as expression levels and standard errors, respectively.

3 Model

The nine genes in our gene network, GAL1, GAL2, GAL3, GAL4, GAL5, GAL6, GAL7, GAL10, and GAL80 will each in turn be considered as a response (dependent) variable. In each case, the 8 remaining genes are considered possible predictors. To simplify the notation and omit some indices, we will denote the response gene under consideration as y and number the remaining genes from 1 through 8, retaining the gene order above.

To set the notation more definitely, let N be the total number of possible predictors, K be the number of experiments, $x_i^{(k)}$ be the expression level of the predictor gene i observed in the kth experiment, and $y^{(k)}$ be the expression level of the response gene observed in the same experiment. Here, we have $N = 8$ and $K = 18$, since the two experiments in which the response gene itself was manipulated cannot contribute and hence these experiments are discarded. We posit an exact linear relationship between the true, unobservable expression levels $y_{\text{true}}^{(k)}$ and $x_{\text{true},i}^{(k)}$. As a consequence, the relationship between observed expression levels $y^{(k)} = y_{\text{true}}^{(k)} + \epsilon_y^{(k)}$ and $x_i^{(k)} = x_{i,\text{true}}^{(k)} + \epsilon_{x_i}^{(k)}$ is

$$y^{(k)} = \sum_{i=1}^{N} b_i (x_i^{(k)} - \epsilon_{x_i}^{(k)}) + a_{\theta(k)} + \epsilon_y^{(k)} \tag{1}$$

where $\epsilon_y^{(k)} \sim \mathcal{N}(0, \sigma_y)$ and $\epsilon_{x_i}^{(k)} \sim \mathcal{N}(0, \sigma_{x_i})$ are independently distributed. The intercept term $a_{\theta(k)}$ is estimated separately for experiments k in which galactose is present in the medium, $\theta(k) = 1$, and separately for experiments in which galactose is absent from the medium, $\theta(k) = 2$.

Since each measured entity has a unique standard error, parameter estimation, even without selection of the most important predictors, is not of the elementary textbook variety. We may nevertheless obtain maximum likelihood estimates of these parameters. By propagation of errors, we obtain the following total standard error of $y^{(k)}$ for the kth experiment:

$$s^{(k)}(b) = \sqrt{(s_y^{(k)})^2 + \sum_{i=1}^{N} b_i^2 (s_{x_i}^{(k)})^2} \tag{2}$$

where we explicitly notate that the total error depends on the coefficients b_i. For notational convenience, we define $b = (b_1, b_2, \cdots, b_N; a_1, a_2)$, and take $s^{(k)}(b)$ to be a constant function with respect to a_1 and a_2. The corresponding log-likelihood function is

$$\ell(b) = -\frac{K}{2}\log(2\pi) - \sum_{k=1}^{K} \log(s^{(k)}(b)) - \frac{1}{2}\sum_{k=1}^{K} \frac{(y^{(k)} - a_{\theta(k)} - \sum_{i=1}^{N} b_i x_i^{(k)})^2}{(s^{(k)}(b))^2}. \tag{3}$$

Consider first the special case $s_{x_i}^{(k)} = 0$, *i.e.*, all inputs are known exactly. In this case, both the maximum likelihood value of b and the variance-covariance matrix $\mathbf{C}$ (the negative inverse Hessian of Eq.(3)), may be expressed analytically:

$$b = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}y \quad (4)$$

$$\mathbf{C} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \quad (5)$$

Here, $\mathbf{V} = \mathrm{diag}([s_y^{(k)}]^2)$. To unify the estimation of intercept and the linear coefficients, we have defined the extended data matrix $\mathbf{X}$ such that $\mathbf{X}_{ki} = x_i^{(k)}$ for $i = 1, ..., N$. Additionally, we define $\mathbf{X}_{k,N+1}$ to be 1 if experiment k uses galactose and to be 0 otherwise, and similarly for $\mathbf{X}_{k,N+2}$ and the absence of galactose. The above estimates simplify further in the familiar case where $s_y^{(k)} = s$, a constant, for all k. In this case, $b = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$ and $\mathbf{C} = s^2(\mathbf{X}^T\mathbf{X})^{-1}$.

In the general case, it is difficult to solve for the maximum or Hessian of Eq.(3) analytically, due to the dependencies of $s^{(k)}(b)$ on b. The numerical solution presents itself as a $N+2$ dimensional nonlinear optimization problem. Computational time is greatly reduced by solving for $\nabla\ell = 0$ in a two-step iterative procedure, in which one step is linear. To do so, we express the gradient in terms of a diagonal, b-dependent matrix $\mathbf{M} = \mathrm{diag}(d)$

$$\nabla\ell = \mathbf{M}b + \mathbf{X}^T\mathbf{V}^{-1}(y - \mathbf{X}b) \quad (6)$$

$$d_i = \sum_{k=1}^{K} \frac{(s_{x_i}^{(k)})^2}{(s^{(k)}(b))^4}\left((y^{(k)} - a_{\theta(k)} - \sum_{i=1}^{N} b_i x_i^{(k)})^2 - (s^{(k)}(b))^2\right), \quad (7)$$

where $\mathbf{V} = \mathrm{diag}([s^{(k)}(b)]^2)$. In the first step, we keep both $\mathbf{M}$ and $\mathbf{V}$ constant and solve for

$$b = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} - \mathbf{M})^{-1}\mathbf{X}^T\mathbf{V}^{-1}y. \quad (8)$$

In the second step, we re-evaluate $\mathbf{M}$ and $\mathbf{V}$ with this value of b. We iterate these two steps until convergence is obtained. No non-linear optimization code is required. We verified that this procedure gives the correct result in two ways. First, we derived the Hessian $H_{ik} = \partial_i\partial_k l$ analytically, and found that the maximum by the iterative scheme above agrees with that obtained by Newton's method. Second, we used a genetic algorithm (GA) to find the global maximum.[1]

[1]The GA uses the likelihood as its fitness function, and tries to find the parameters

The derivation of the Hessian allows us to evaluate the covariance matrix $\mathbf{C} = -\mathbf{H}^{-1}$. We also obtained the covariance by simulation. Each simulated case is defined by adding to the measured y and $x_i^{(k)}$ a normally distributed error according to the distributions following Eq.(1), using the experimental standard errors. For each sample, we maximized Eq.(3), and computed the variance-covariance matrix for all samples. The resulting covariance matrix agrees very well with the analytically derived $\mathbf{C}$. In the simulation, we also observed that Eq.(3) is distributed as $\chi^2_{K-(N+2)}$, as expected.

4 Model selection

Optimizing by including all variables as predictors gives a maximal value of Eq.(3). However, the resulting complete network graph will have little value as an interpretation of the biological network structure. Somehow, we should pick a subset of the most significant predictors, generating hypotheses that these are the ones affecting the response gene. Strictly, the formalism we employ gives statistical associations and not causations (as, *e.g.*, is found by Yoo *et al.*, 2002). However, the causal relations are reflected in the found weight coefficients, and therefore these coefficients encapsulate information beyond more simple statements such as "these two genes have associated expression values".

One way to pick a proper subset of independent variables is to start with one predictor—the one which correlates best with the response—and then add new predictors if an F-test yields that they contribute significantly. This is not a recommended approach, though, since there is no guarantee that, *e.g.*, the single predictor that best describes the data is one of the two predictors that best describe the same data. Other approaches taken in the literature with respect to transcript data have been to fix the number of predictors (Gardner *et al.*, 2003) and to apply the Akaike Information Criterion (AIC) (De Hoon

b that maximize it. Basically, the genetic algorithm consists of the following procedure. First, an initial set of candidate solutions, encoded as strings of digits and called the first generation, is randomly formed Each string is called a chromosome and the position of each digit within are referred to as a gene. The candidate solutions, *i.e.*, the chromosomes, in the first generation are evaluated by calculating their fitness, and a second generation is formed by selecting parent strings such that chromosomes with high fitness are more likely to be chosen. These parents are combined to new candidate solutions, whereafter a small degree of mutation is allowed, *i.e.*, a small number of the genes in each new chromosome is changed randomly. From the second generation the third generation is formed and the process goes on until a satisfactory solution has been obtained. A more thorough description of genetic algorithms can be found in many textbooks, *e.g.*, Mitchell (1998).

et al., 2003). Below, we describe various approaches to model selection as applied to the identification of gene regulatory networks. Throughout, we let ζ denote the set of indices in the current selection. Standard assumptions of multivariate regressions are employed: independence of observations, independent and normally distributed errors. For additional details and background please see the original publications for each method or one of the many reviews containing an overview and detailed comparison of the various model selection methods (Hastie *et al.*, 2001, Miller, 2002, George, 2000, Shao, 1997).

4.1 Selection by penalty terms

A popular and straightforward way of selecting the most important predictors in a regression model is to incorporate into the objective function a penalty term, which increases with the number of predictors (George, 2000). In our case, we can summarize all these methods as the minimization of an expression of the form

$$-\ell_\zeta + C(J+2), \tag{9}$$

where ℓ_ζ is the log-likelihood of the subset given the input parameters ζ from (3), $J+2$ is the number of predictors, including the two different values of the intercept term, and C is a constant following from the choice of penalty method.

A common choice is Mallows' C_p (Mallows, 1973), which corresponds to the choice $C = 2$. Mallows himself warns against using the C_p as a selection criterion, the reason being that the choice of ζ must not depend on the data for the estimate of the model error to be valid. Of course, if we have a huge resource of data, we can first use a part of it to pick predictors, and then another part to estimate the model error. This is almost never the case, and the use of the C_p is highly questionable when the same data is used for variable selection and error estimation. A similar, but more general, measure is the Akaike Information Criterion, AIC, which is based on a log-likelihood loss function. For a Gaussian distribution, however, it becomes identical to the C_p. We will henceforth refer to both these methods as "AIC".

Two other penalty methods, the Bayesian Information Criterion, BIC, and Minimum Description Length, MDL, are formally equivalent, although they are motivated very differently. The BIC can be motivated in a Bayesian approach to model selection see Hastie *et al.* (2001), while MDL stems from a description of data with a minimum of complexity, see Hansen and Yu (2001). They are obtained by choosing $C = \log K$ in (9). We note that these approaches suffer from the same drawback as the AIC, *i.e.*, they are doubtful

Table 1: Weight coefficients b_i for the predictors of GAL1 and GAL3, respectively, calculated with the penalties provided by the AIC and BIC. Here, these penalties resulted in identical model selections. The predictors are listed vertically.

Predictors	GAL1	GAL3
GAL1	—	0.06
GAL2	0.06	0.08
GAL3	0.26	—
GAL4	0.30	0.10
GAL5	0.12	0.06
GAL6	0.14	0.10
GAL7	0.18	0.08
GAL10	0.20	0.16
GAL80	0.00	0.04

for both selecting predictors *and* estimate errors from the same data. We will henceforth refer to these two measures collectively as BIC.

In Table 1 we show the results using GAL1 and GAL3, respectively, as the dependent variable. Although the penalty terms differ for AIC and BIC, they result in the same model selection, *i.e.* all the variables were selected as regulators. (Self-regulation is not considered (dashes in Table 1), as allowing self-regulation returns the uninformative solution that steady-state expression of GAL1 is determined solely by the level of GAL1, and so on.) This is not totally unexpected, since the AIC methodology is not specifically designed to make predictors zero. However, the prediction that all proteins are transcriptional regulators (particularly in cases such as regulation by Gal6p) is unlikely to be true. Furthermore, some predictions disagree with the known biological regulation: Gal80p is predicted to be a positive regulator of GAL3, and essentially irrelevant for GAL1 expression. To reduce the number of predictors, one might consider increasing the value of C in (9). However, without sufficient theoretical justification for such an action, we refrain from taking this step.

4.2 Selection by Cross-Validation

Although methods of the type described in 4.1 are widely used, they have the drawback that the same data is used both for estimation of model parameters and for estimation of prediction error. In cross-validation (see *e.g.* Hastie

Table 2: Weight coefficients b_i for the predictors of GAL1 and GAL3, using a cross-validation estimate of prediction error. The predictors are listed vertically. The dash(−) denotes an excluded or non-selected predictor.

Predictors	GAL1	GAL3
GAL1	—	—
GAL2	—	—
GAL3	0.21	—
GAL4	0.24	—
GAL5	—	—
GAL6	—	—
GAL7	0.10	—
GAL10	0.14	0.12
GAL80	–0.14	–0.09

et al., 2001) a subset of the data is used for parameter estimation, and the prediction error then evaluated on the complement set.

Here, we used leave-one-out cross-validation, repeated for each relevant experiment. The iterative procedure described above was seeded from (4), and to ensure convergence, a small random variation was added when convergence did not occur after a few steps. The subset of predictors having the least cross-validation error were selected. In Table 2 we give the parameters of the selected models. The prediction of Gal3p, Gal4p, and Gal80p as regulators of GAL1 recapitulates the known regulation, including the sign of the effect. In addition, there is evidence of positive regulation by GAL7 and and GAL10. For regulation of GAL1 and GAL3, the model predicts positive and negative regulation by Gal10p and Gal80p respectively. Gal10p thus seems to have a regulatory effect, a role which has hitherto not been established. The possible feedback inhibition of Gal10p on GAL1 expression was discussed in Ideker *et al.* (2001).

4.3 Breiman's Little Bootstrap

Boostrap methods (see Hastie *et al.*, 2001) can also be adopted to reuse data in a more acceptable way than described in Section 4.1. Here, we describe the "Little" bootstrap of Breiman, essentially following the derivation in Breiman (1992).

The Little bootstrap is an x-fixed method, so we assume for the moment

that there are no uncertainties in the predictors, *i.e.* all $s_{x_i}^{(k)} = 0$. We also assume that $\epsilon_y^{(k)}$ are identically distributed. Let RSS_ζ be the residual sum of squares for the predictors ζ. With $\hat{y}_\zeta^{(k)} = \hat{a} + \sum_{i\in\zeta}[\hat{b}_i x_i^{(k)}]$ where $\hat{a}$ and $\hat{b}$ are point estimates of the intercept and coefficients, respectively, and $y_{\text{true}}^{(k)}$ as the "true" value of $y^{(k)}$, we have

$$RSS_\zeta = \sum_{k=1}^{K} |y^{(k)} - \hat{y}_\zeta^{(k)}|^2 = ME_\zeta + \sum_{k=1}^{K} |\epsilon_y^{(k)}|^2 + 2\sum_{k=1}^{K} \epsilon_y^{(k)} \left(y_{\text{true}}^{(k)} - \hat{y}_\zeta^{(k)} \right), \quad (10)$$

where the model error is given by $ME_\zeta = \sum_{k=1}^{K} |y_{\text{true}}^{(k)} - \hat{y}_\zeta^{(k)}|^2$. By introducing $\hat{y}_{\text{FULL}}^{(k)}$ as the estimated value from the full model (with all possible predictors included), we can expand the expression for the model error to yield

$$ME_\zeta = RSS_\zeta - RSS_{\text{FULL}} + \sum_{k=1}^{K} \left(y_{\text{true}}^{(k)} - \hat{y}_{\text{FULL}}^{(k)} \right)^2 - 2\sum_{k=1}^{K} \epsilon_y^{(k)} \left(y_{\text{true}}^{(k)} - \hat{y}_\zeta^{(k)} \right). \quad (11)$$

The expectation of the penultimate sum can be estimated as $N\hat{\sigma}_{\text{FULL}}^2$, but the last sum is more difficult to evaluate. Following the notation in the original paper, we introduce a function B_ζ, which has the same expectation as the term we want to estimate

$$E[B_\zeta] = E\left[\sum_{k=1}^{K} \epsilon_y^{(k)} \left(y_{\text{FULL}}^{(k)} - y_\zeta^{(k)} \right) \right]. \quad (12)$$

The estimate of the Little bootstrap for the model error is given by

$$\widehat{ME}_\zeta = RSS_\zeta - RSS_{\text{FULL}} + N\hat{\sigma}_{\text{FULL}}^2 - 2B_\zeta, \quad (13)$$

where the estimate of B_ζ is obtained by the resampling method described in Breiman. For comparison, the AIC (Mallows' C_p) sets $B_\zeta = \hat{\sigma}_{\text{FULL}}^2(K - J)$.

To implement the simulated resampling, standard errors of the response variables were used to generate "new" measured values. The Little bootstrap uses an arbitrary parameter t, which represents the tradeoff between bias and variance. Breiman recommends a value of $t = 0.6$, which we use here.

Results are summarized in Table 3. For most response variables, the coefficients for genes GAL3, GAL4, and GAL80 are non-zero, consistent with the known regulatory role of the corresponding proteins. Here, and elsewhere, we refer to a protein as a regulator if it determines the level of transcriptional activation of a gene, either by a direct molecular mechanism, or by an indirect

Table 3: Weight coefficients for the regulation of GAL1 through GAL80, calculated using the Little Bootstrap with the value $t = 0.6$. A dash denotes that the corresponding predictor candidate was not selected. The genes listed horizontally are regulated by the ones listed vertically, *e.g.*, GAL1 is here regulated by GAL3, GAL4, GAL10 and GAL80.

Predictors	GAL1	GAL2	GAL3	GAL4	GAL5	GAL6	GAL7	GAL10	GAL80
GAL1	–	–	–	–	–	-0.05	–	–	–
GAL2	–	–	–	–	-0.03	–	–	–	–
GAL3	0.16	0.05	–	–	–	–	0.15	0.14	0.05
GAL4	0.18	0.05	0.05	–	-0.05	–	0.18	0.15	–
GAL5	–	–	–	–	–	–	–	–	–
GAL6	–	–	–	–	–	–	–	–	–
GAL7	–	0.04	–	–	–	–	–	–	–
GAL10	0.18	0.04	0.11	–	–	–	0.24	–	0.04
GAL80	-0.19	-0.04	-0.08	–	–	-0.10	–	-0.22	–

effect. GAL4 is independent of all other variables, as may be expected given the knowledge that GAL4 expression is not thought to be regulated by any of the other proteins. We see no effects of GAL6 as a regulator, as was put forward as a possibility in Zheng *et al.* (1997). Worth noting is also that GAL80 always has a negative contribution in cases where it is non-zero, in clear concordance with the role of Gal80p as a negative regulator of transcription. As in 4.2, there is evidence for regulatory role by Gal10p.

In order to test the robustness of our results with respect to the value of t, we recalculated network predictions using the values $t \in \{0.4, 0.5, 0.7, 0.8\}$. Results are insensitive to variation of this parameter in the given range. Of the predicted regulatory interactions, only a single one changes: for some t, we do not pick up the the dependence of GAL3 on Gal4p, and coefficients for Gal10p and Gal80p regulation of GAL3 change slightly ($< 5 \cdot 10^{-3}$) as a result.

In summary, without introducing any prior knowledge of the Galactose utilization system, the calculation is able to derive from the observed expression levels a network describing regulatory influences which are consistent with those found from earlier experiments. In addition, the sign and relative strength of novel modes of influence is predicted.

4.4 Tibshirani's Lasso

The final method we will discuss involves the least absolute shrinkage and selection operator, or Lasso, introduced by Tibshirani (1996). The Lasso is a close relative of ridge regression (Draper and Smith, 1998), but the constraint

on the predictors is in the form of an L^1-norm, instead of an L^2-norm.

Explicitly, the problem can be formulated as a minimization of the residual sum of squares (RSS) under the constraint[2]

$$\sum_{i=1}^{N} |b_i| \leq t. \tag{14}$$

An alternative, but mathematically equivalent, formulation is to minimize the objective function

$$\sum_{k=1}^{K} \left| \frac{a_{\theta(k)} + \sum_{i=1}^{N} [b_i x_i^{(k)}] - y^{(k)}}{s_y^{(k)}} \right|^2 + \lambda \sum_{i=1}^{N} |b_i|. \tag{15}$$

Following the recommendation in Tibshirani, we normalize all predictors in order to make them potentially equally important.

In Figures 1 and 2 we show the result for GAL1 and GAL3 as response variable, respectively. The values of the eight possible predictors are shown for different values of the constraint parameter $t/t_{\max}$. Here, $t_{\max}$ is defined as the value of t where the constraint no longer is active, and hence the weight coefficients are the same as for ordinary least squares.

For the Lasso, the covariance can be obtained for all predictors, including those which are identically zero. Here, we use the formalism of Osborne *et al.* (2000) to estimate standard errors for the predictor coefficients. Table 4 lists coefficients and standard errors for GAL1, and GAL3 as dependent variables for t-values obtained by cross-validation, leaving out one experiment at a time. The values of $t/t_{\max}$ thus obtained are 0.65 and 0.22 for the regulation of GAL1 and GAL3, respectively.

For the regulation of GAL1, the genes GAL3, GAL4 and GAL80 play prominent roles, with values of the weight coefficients well above one standard error from zero, and the signs of the weight coefficients are as expected. The values of the GAL2 and GAL6 weight coefficients are not significant. Worth noting is that GAL7 and GAL10 both have values clearly above one standard error from zero. These may be false positives, or may indicate a novel regulatory role for the corresponding enzymes. For the regulation of GAL3, the only positive significant results are the positive regulation by GAL10 and the negative regulation by GAL80. The latter is well known, but the former is not. The absence of GAL4 as a regulator of GAL3 is most likely a false

[2]The Lasso parameter t should not be confused with Breiman's Little Boostrap parameter, also designated by t.

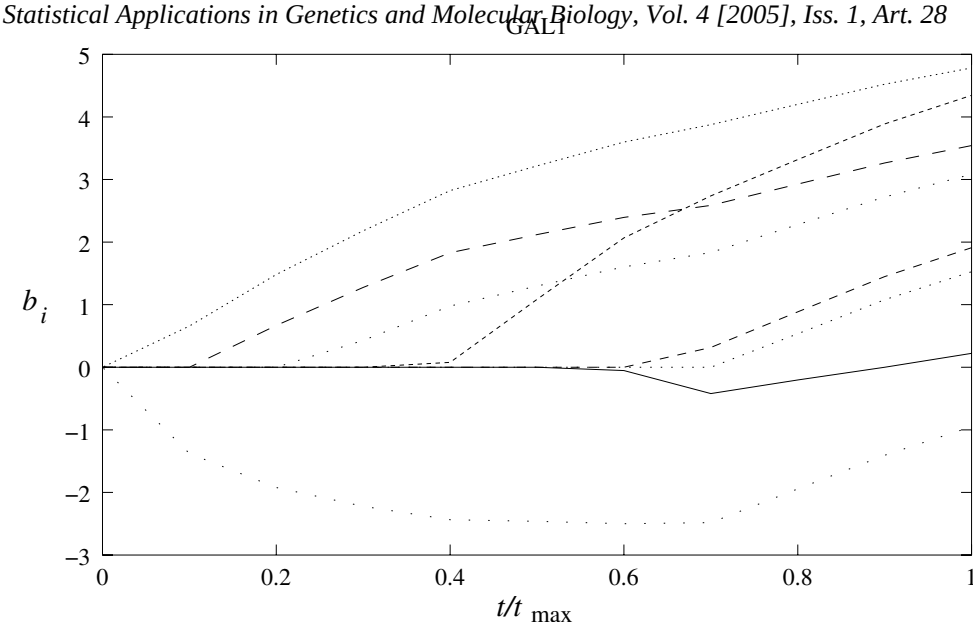

Figure 1: GAL1 predictor coefficients as function of normalized Lasso constraint. From top to bottom at $t/t_{\max} = 1$: GAL 3, 4, 10, 7, 6, 5, 2, 80.

negative. However, these results should be used with care, since, as is pointed out in Osborne *et al.*, the estimates will have a condensation of probability around zero. The coefficients may therefore be far from normally distributed, and hence the use of standard errors may not be appropriate.

Distribution of networks

At any given stage of the experimental process, researchers frequently consider multiple possible models. Models somewhat less likely than the most favored one may gain validity in subsequent rounds of experimentation. Here, we incorporate these considerations by examining probability distributions of networks of differing topology. Since DNA microarray data is often noisy, one expects that there will be a number of networks of similar likelihood, more or less consistent with the data. To find these networks, one can either consider the Bayesian approach of ranking solutions, or use brute force to simulate the measurement errors.

Here, we obtain a distribution of possible networks, by considering the standard error for each gene expression value for each condition. We simulate by disturbing each measured value with gaussian noise of zero mean and the

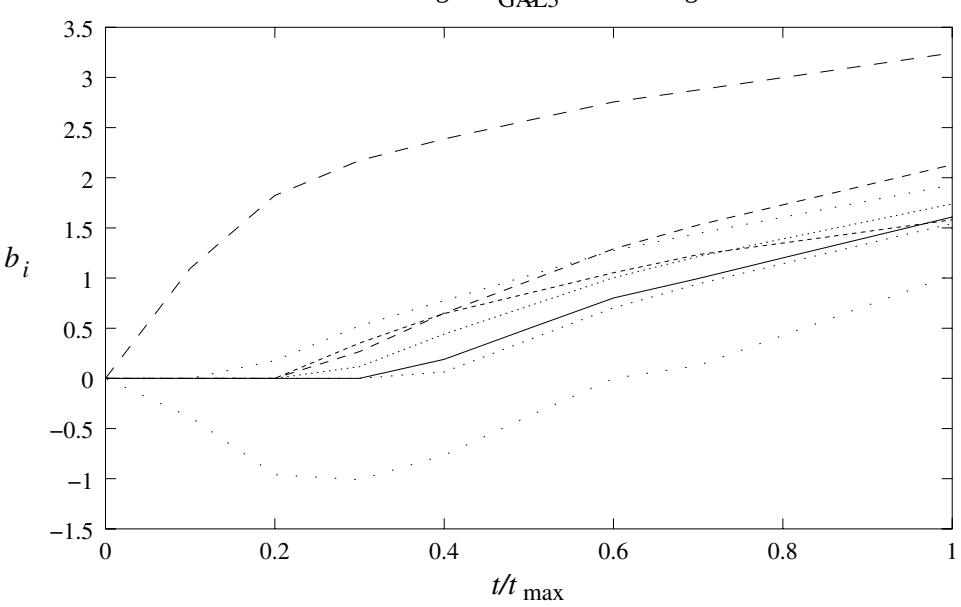

Figure 2: GAL3 predictor coefficients as function of normalized Lasso constraint. From top to bottom at $t/t_{\max} = 0.5$: GAL 10, 7, 6, 4, 2, 1, 5, 80

appropriate standard error, and obtain the network from these disturbed data.[3] Tallying the number of times any specific network wiring is optimal gives us an estimate about the distribution of networks. The value of the constraint parameter t is not straightforward to choose. Here, we use half the value yielded by the cross-validation procedure mentioned above—the reason being that we do not calculate which values are significant and hence we obtain too many non-zeroes with the full cross-validation value (cf. Table 4).

In Fig. 3 we present results for GAL1 and GAL3 as the dependent variables. In Fig. 3 (a), we see that GAL80 turns up in every distribution. This is very reasonable, since it is a known regulator. GAL3 and GAL4 are two other known regulators, and they show up at least in the two most frequent wirings. It is interesting that GAL2 is identified as a potential regulator in wirings just below the most likely ones, a hypothesis we would miss if we considered *only* the model most likely under this round of experiments. In Fig. 3 (b) we see which genes best explain the expression levels of GAL3. As expected, GAL80 is a major regulator which is present in almost all the most likely wirings. A

[3]The disturbance is here obtained by adding noise to the original, non-logarithmic data, which thereafter is logarithmically transformed. The same threshold at 0.001 as before is used.

Table 4: Coefficient values, and standard errors (in brackets) for the predictors of GAL1 and GAL3 obtained using the Lasso method. Each row corresponds to a possible predictor.

Predictors	GAL1	GAL3
GAL1	—	0.0 (0.4)
GAL2	-0.27 (1.1)	0.0 (0.3)
GAL3	3.7 (0.4)	—
GAL4	2.4 (0.5)	0.0 (0.3)
GAL5	0.0 (0.6)	0.0 (0.4)
GAL6	0.1 (0.6)	0.0 (0.3)
GAL7	1.7 (0.5)	0.3 (0.3)
GAL10	2.5 (0.5)	1.9 (0.3)
GAL80	-2.5 (0.7)	-1.0 (0.5)

peculiarity for the regulation of GAL3 is that GAL4 turns up in our results only for the second most common wiring. GAL4 is a known transcription factor for the galactose regulatory system, and its absence is clearly a false negative.

The gene GAL10 is not known to have any regulatory effect on the galactose metabolic system, but nevertheless, it turns up as one of the important regulators of GAL1 and GAL3. This may be a false positive, or may highlight the need for further experimental characterization. With respect to the GAL6 gene, for which there was prior evidence of a regulatory role (see Zheng *et al.*, 1997), we see almost no evidence of such a role for it in our results. When considering possible regulators of GAL3, GAL6 turns up as a regulator in only 4% of the cases.

The computed distributions could be used to assist in the selection of subsequent experiments for network refinement. For example, a researcher may be considering developing strains with deletions of two different genes, or introducing a constitutive expression construct for a gene, and would like to select constructs that are most likely to be informative. Predictions of system response could be made for a perturbation under consideration, not only for the most likely model but also for models with slightly less likelihood. Repeating this simulation for other possible perturbations could aid in the prioritization of choices, for example by highlighting perturbations with the greatest potential power to discriminate among possible networks (Ideker *et al.*, 2000a). This approach may be of particular relevance for unraveling interdependencies in

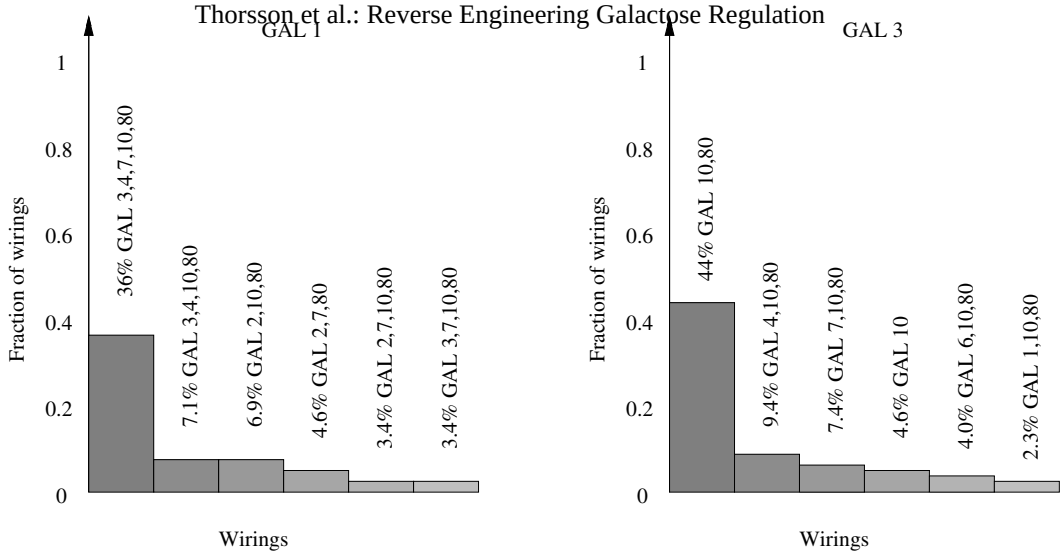

Figure 3: The five most frequent wirings when we let the dependent variable be (a) GAL1 and (b) GAL3.

complex biological networks, in which initial rounds of experiments may be consistent with a multitude of possible network structures.

5 Conclusions and discussion

We compared four methods for model selection, as applied to the study of galactose metabolic system in the yeast *S. cerevisiae*. The analysis was based on steady-state expression measurements obtained in the wild-type state and data obtained after deletion of single genes, both in galactose-rich and galactose-poor media. We concentrated on nine genes considered to be involved in this metabolic system. Our underlying assumption is that a specific expression level can be approximately described by a linear combination of the expression levels of other genes in the system, possibly following a global transformation of the levels. The expectation is that by exploring these linear dependencies one can capture the main features of the network. Understanding relationships between these genes in detail is likely to require more complex descriptions, such as non-linear relations. However, with limitations in the number of examples available, a linear description can often capture most of the variability in the data, and there is little to be gained by employing more complex descriptions (van Someren *et al.*, 2000).

In this study, we sought to incorporate the measured error in the microarray

gene expression measurements, specific to each gene and to each condition, into our model selection process. These errors are often quite large. In this regard, we derived a log-likelihood function incorporating these gene- and condition-specific errors. The maximization of this function is a non-linear problem, which we solved both by a computationally efficient iterative procedure, and by a genetic algorithm. A second order expansion of the log-likelihood function yielded analytic expressions for the standard errors of the estimated linear coefficients.

The first model selection scheme we explored is based on the addition of a penalty term to the log-likelihood function. We examined both an AIC penalty, which here is equivalent to Mallows' C_p, and a BIC term, which is equivalent to MDL. For either choice, all but one predictor was selected to be non-zero. Hence, for this data set, these methods did not yield sufficient discrimination among possible predictors.

In the second scheme, we selected the model with the lowest prediction error, as estimated by cross-validation. GAL3 and GAL4 were selected as positive regulators, while GAL80 was selected as a negative regulator, consistent with expectations from earlier literature. The prediction showed evidence of positive regulation by GAL7 and GAL10.

The third scheme we considered is the "Little Bootstrap" of Breiman. This method was designed for the x-fixed case, and we hence took only the uncertainties in the responses into account. Many of the results found using this method are supported by the known biology. Both GAL3 and GAL4 were found to be positive regulators, and GAL80 a negative regulator. GAL6 was found to be irrelevant. Some results here were also outside the scope of the standard model of the control of galactose utilization, such as the positive regulatory role of GAL10.

The fourth scheme was based on Tibshirani's Lasso. This method introduced a tuning parameter, which we determined by cross-validation. By simulating measurement noise, we found probability distributions of possible wiring diagrams. The results obtained are similar to those found using the Little Bootstrap, including the possible regulatory role of Gal10p. There are differences, as well. For example, the Lasso method indicates that Gal7p may regulate GAL1 expression, whereas Breiman's method does not provide evidence for such regulation.

Numerous other model selection methods can be evaluated for the purpose of reverse engineering gene regulatory networks. Even within the class of methods involving multivariate regression, other variants are possible besides those described here (see *e.g.* Hastie *et al.*, 2001). One can, for example consider various methods for generating a class of models, such as exhaustive

search over all subsets (Sections 4.1, 4.2, and 4.3), the Lasso (Sec. 4.4), Ridge Regression, or Principal Components Regression. On the other hand, one can explore various ways to assess model error, including complexity penalties (Section 4.1), bootstrap methods (Sec. 4.3) or cross-validation (Secs. 4.2 and 4.4). Here, we sought to adopt and evaluate a limited number of established methods, maintaining a degree of connectedness and inter-comparability.

The four schemes considered here thus have contrasting computational requirements. Methods involving exhaustive search over all subsets have the potential drawback that they are not scalable to large systems, since one has to search through all possible combinations of non-zero predictor coefficients, which increases exponentially with the number of predictors. This increase may be lessened, but not eliminated, by setting an explicit limit on the possible number of predictors, as done by Gardner *et al.* Boostrap error estimation comes at a fairly high computational cost. Lasso parameter estimation is scalable, and can be applied to very large datasets, and is therefore an attractive model selection method to consider for the task of studying larger gene regulatory networks. The cross-validation step for determining the Lasso tuning parameter adds somewhat to the computational cost, and the extension to the calculation of distributions over possible networks requires considerably more. It is likely that under some circumstances an adaptive strategy is advantageous, in which a relatively large number of possible predictors is first determined from a from a fast method, and refinement of the network is subsequently done using the more computationally intensive steps of bootstrapping or examining distributions of models. It is also possible that other more integrated formulations may be derived, exploiting the advantages of each method.

For any gene under consideration, the methods described here provide predictions for which factors are likely to determine the expression level of that gene, and conversely, which factors unlikely to be determinants of that expression level. Among the determining factors, conclusive information on which of the corresponding proteins acts more directly (*e.g.*, fewer molecular reaction steps away from transcription) or less directly is not provided in that prediction, although those factors that are more strongly statistically related to gene expression level may tend to be those factors that act more directly. Somewhat separately, causal dependencies *among* the determining factors may be revealed from *their* expression levels. Identifying the molecular mechanism for any implied influence, either direct or indirect, can only be determined using assays directed at those data types. The methods discussed here are thus best regarded as a starting point for hypothesis generation and further experimentation.

In applying model selection methods to gene expression data, the underly-

ing dependence of one gene on another may be difficult to discern, as the two gene expression values may simply be correlated over the experimental conditions, and thus no directionality is implied. Recently, efficient algorithms have been developed for identifying partial correlations among gene network variables (Schäfer and Strimmer, 2005). Here, our data is such that each individual variable under consideration was perturbed to a fixed, known value (forced to zero by gene deletion, one gene at a time), thus allowing directional relations to be revealed. By only considering correlations, such relations cannot be inferred. Also, these relations would be more difficult to discern using wild-type data alone. In the general case, methods taking particular consideration of variable correlation (Schäfer and Strimmer, 2005) are likely to form particularly valuable components of the inference process.

The network model selection methods described here can be generalized to any data in which levels are measured for multiple genes, such as microarray data, proteomics and assays of protein activities. Low abundance transcripts, however, are typically below the limit of detection of microarray technology, and small changes in transcript level may not be detected. Regulatory influences involving low abundance proteins or subtle changes are therefore less likely to be revealed by the measurements and by the model selection methods described herein.

The methods described here require no prior knowledge or alternate lines of evidence for interactions. As is well known, integrating information from other sources is of great value in model generation and refinement. These methods are readily incorporated as components of a more general iterative refinement procedure. Prior knowledge on interactions can be incorporated as bounds on linear coefficients. Other promising directions to follow are the extension to generalized linear models, and methods for selecting additional experiments to discriminate among the more likely models implied by the probability distributions.

As discussed above, these methods provide information as to which proteins may play a role in determining the expression level of regulated genes, and whether such regulators play the role of activators or repressors. In addition, a number of predictions emerge which may not at all be apparent by qualitative or visual assessment of perturbation responses. For example, our results by all four methods indicate that no single factor is a *sole* determinant of GAL1 or GAL3 expression level, and knowledge of the levels of multiple factors is needed to predict their expression. Secondly, we obtain information on the relative strengths of regulatory influences. For example, both Breiman's Little Bootstrap and the Lasso method indicate that the transcriptional activator Gal4p and transcriptional repressor Gal80p are of equal influence in

determining the level of GAL1 expression, and that Gal10p plays a comparable or even somewhat greater role in determining expression levels of GAL1. We confirmed evidence from our earlier study implying that Gal10p or Gal7p levels may be of importance for determining transcript levels and found that the influence of the former is considerably greater than the latter. In addition, we demonstrated how taking account of the measured experimental error can yield a ranked list of regulatory networks differing in structure, or wiring.

In summary, we have studied the galactose regulatory system in yeast as a relatively well-known test case for different model selection procedures. We have demonstrated how to include known errors in the measurements into the selection procedure. Introducing the complexity penalty terms AIC and BIC did not yield a reduction in model complexity for these data. Methods involving cross-validation and bootstrap estimation of prediction error performed well, as did use of the Lasso shrinkage estimator, in the sense that most of the known regulatory interactions were reproduced, and the algorithms introduced only one or two potential regulators to account for the observations. The Lasso method has a number advantages in terms of computational requirements. Our results indicate that describing regulatory influences as a linear relationship can capture the main features of the regulatory network while maintaining computational tractability.

References

Arkin, A., Shen, P. & Ross, J. (1997), 'A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements', Science **277**, 1275–1279.

Bickel, D.R. (2005), 'Probabilities of spurious connections in gene networks: application to expression time series', Bioinformatics **21**, 1121–1128.

Breiman, L. (1992), 'The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error', J. Am. Stat. Ass. **87** (419), 738–754.

D'haeseleer, P., Liang, S., & Somogyi, R. (1999), 'Linear Modeling of mRNA expression levels during CNS development and injury', in: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E., and Lauderdaule, K. (Eds.), Pacific Symposium on Biocomputing **4**. World Scientific Publishing Co, Singapore, pp. 41–52.

Draper, N.R. & Smith, H. (1998), *Applied Regression Analysis, 3rd ed.*, Wiley, New York.

Farkas, I., Jeong, H., Vicsek, T., Barabsi, A.-L., & Oltvai, Z.N. (2003), 'The topology of the transcription regulatory network in *Saccharomyces cerevisiae*', Physica A **381**, 601–612.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000), 'Using Bayesian Networks to Analyze Expression Data', J. Comp. Biol. **7**, 601–620.

Gardner, T.S., Bernardo, D., Lorenz, D, & Collins, J.J. (2003), 'Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling', Science **301**, 102–105.

George, E.I. (2000), 'The Variable Selection Problem', J. Am. Stat. Ass. **95** (452), 1304–1308.

Hansen, M.H. & Yu, B. (2001), 'Model Selection and the Principle of Minimum Selection Length', J. Am. Stat. Ass. **96** (454), 746–774.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer, New York.

De Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N., & Miyano, S. (2003), 'Inferring Gene Regulatory Networks from Time-Ordered Gene Expression Data of Bacillus Subtilis Using Differential Equations', in: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E., and Lauderdaule, K. (Eds.), Pacific Symposium on Biocomputing **8**. World Scientific Publishing Co, Singapore, pp. 17–28.

Ideker, T.E., Thorsson, V., & Karp, R. (2000a), 'Discovery of Regulatory Interactions through Perturbation: Inference and Experimental Design, Pacific Symposium on Biocomputing 2000', in: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E., and Lauderdaule, K. (Eds.), Pacific Symposium on Biocomputing **5**. World Scientific Publishing Co, Singapore, pp. 302–313.

Ideker, T.E., Thorsson, V., Siegel, A.F., & Hood, L.E. (2000b), 'Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data', J. Comp. Biol. **7**, 805–817.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. & Hood, L. (2001), 'Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network', Science **292**, 929–934.

Jong, H.D. (2002), 'Modeling and Simulation of Genetic Regulatory Systems: A Literature Review', J Comp Biol **9**(1), 67–103.

Mallows, C.L. (1973), 'Some Comments on C_p', Technometrics **15** (4) 661–676.

Miller, A. (2002), *Subset Selection in Regression, 2nd ed.*, Chapman & Hall, London.

Mitchell, M. (1998), *An Introduction to Genetic Algorithms*, MIT Press, Cambridge MA.

Osborne, M.R., Presnell, B. & Turlach, B.A. (2000), 'On the LASSO and its Dual', J Comp and Graphical Stat **9** (2), 319–37.

Pournara J. & Wernisch, L. (2004), 'Reconstruction of gene networks using Bayesian learning and manipulation experiments', Bioinformatics **20**, 2934–2942.

Rung, J., Schlitt, T., Brazma, A., Freivalds, K. & Vilo, J. (2002), 'Building and analyzing genome-wide gene disruption networks', Bioinformatics **18**, S202–S210.

Schäfer J. & Strimmer, K. (2005), 'An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks', Bioinformatics **21**, 754–764.

Shao, J. (1997), 'An asymptotic theory for linear model selection', Statistica Sinica **7**, 221-264.

van Someren, E.P., Wessels, L.F.A. & Reinders, M.J.T. (2000), 'Linear Modeling of Genetic Networks from Experimental Data', in: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB00), AAAI, La Jolla, California, pp. 355–366.

Tibshirani, R. (1996), 'Regression Shrinkage and Selection via the Lasso', J. R. Stat. Soc. B **58** (1), 267–288.

Wahde, M. & Hertz, J. (2001), 'Modeling genetic regulatory dynamics in neural development', J. Comp. Biol. **8**(4), 429–442.

Weaver, D.C., Workman, C.T. & Stormo, G.D. (1999), 'Modeling regulatory networks with weight matrices', in: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E., and Lauderdaule, K. (Eds.), Pacific Symposium on Biocomputing **4**. World Scientific Publishing Co, Singapore, pp. 112–123.

Wuensche, A. (1998), 'Genomic regulation modeled as a network with basins of attraction', in: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E., and Lauderdaule, K. (Eds.), Pacific Symposium on Biocomputing **3**. World Scientific Publishing Co, Singapore, pp. 89–102.

Yeung, M.K.S., Tegnér, J. & Collins, J.J. (2002), 'Reverse engineering gene networks using singular value decomposition and robust regression', Proc. Natl. Acad. Sci. USA **99**, 6163–6168.

Yoo, C., Thorsson, V. & Cooper, G. (2002). 'Discovery of Causal Relationships in a Gene-regulation Pathway from a Mixture of Experimental and Observational DNA Microarray Data', in: Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E., and Lauderdaule, K. (Eds.), Pacific Symposium on Biocomputing **7**. World Scientific Publishing Co, Singapore, pp. 498–509.

Yu, J., Smith, A., Wang, A.P., Hartemink, A.J. & Jarvis, E.D. (2004), 'Advances to Bayesian network inference for generating causal networks from observational biological data', Bioinformatics **20**, 3594–3603.

Zheng, W., Xu, H.E. & Johnston, S.A. (1997), 'The cysteine-peptidase bleomycin hydrolase is a member of the galactose regulon in yeast' J. Biol. Chem. **272** (48) 30350–30355.