ABSTRACT

DUARTE, CHRISTINE W. A New Method for Genetic Network Reconstruction in Expression QTL Data Sets. (Under the direction of Professor Zhao-Bang Zeng).

Expression QTL (or eQTL) studies involve the collection of microarray gene expression data and genetic marker data from segregating individuals in a population to search for genetic determinants of differential gene expression. Previous studies have found large numbers of trans-regulated genes that link to a single locus or eQTL "hotspot". It would be of great interest to discover the mechanism of co-regulation for these groups of genes. However, many difficulties exist with current network reconstruction algorithms such as low power and high computational cost. A common observation for biological networks is that they have a scale-free or power-law architecture. In such an architecture, there exist highly influential nodes that have many connections to other nodes, but most nodes in the network have very few connections. If we assume that this type of architecture applies to genetic networks, then we can simplify the problem of genetic network reconstruction by focusing on discovery of the key regulatory genes at the top of the network. We introduce the concept of "shielding" in which a gene is conditionally independent of the QTL given the shielder gene, and we iteratively build networks from the QTL down using tests of conditional independence. We evaluate the confidence level of shielders using a two-part strategy of requiring a threshold number of genes to be shielded and requiring a high level of bootstrap support for shielders. We have performed a set of simulations to test the sensitivity and specificity of our method as a function of method parameters. We have found that our method has good performance using a significance level of 0.05 for testing the hypothesis that a gene is a shielder, with little gained by decreasing α further. The shielder boostrap confidence level depends on the desired balance between false positives and false negatives, but our recommendation is to use 80% bootstrap support for high confidence of discovered network features. With a small sample size (100) and a large number of network genes (as many as 622), our algorithm succeeds in finding a high percentage of the key network regulators (47% on average) with high confidence (95% specificity on average).

We have applied our network reconstruction algorithm to a yeast expression QTL data set in which microarray and marker data were collected from the progeny of a backcross of two species of *Saccharomyces cerevisiae* [8]. Networks have been reconstructed for 6 of the 11 largest eQTL hotspots in this data set. The regulation of shielder gene expression has been found to be primarily in trans. Bioinformatic analysis of three networks generated different hypotheses for mechanisms of regulation of the shielded genes by the primary shielders. One common theme was that the shielders modulated the effect of transcription factors of which they were themselves targets. Overall our method has created a list of potentially important regulatory genes in various yeast biological processes, and further bioinformatic analysis or laboratory experiments could lead to the generation and testing of many important hypthotheses.

A New Method for Genetic Network Reconstruction in Expression QTL Data Sets

by

Christine W. Duarte

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fullfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics

Raleigh, North Carolina

2009

APPROVED BY:

Dr. Ronald Sederoff	Dr. Russell Wolfinger
Dr. Zhao-Bang Zeng Chair of Advisory Committee	Dr. Jung-Ying Tzeng

DEDICATION

To Schuyler, Anya, and Summer

BIOGRAPHY

Christine Duarte, formerly Christine Woods, was born in Crofton, MD in 1974. She grew up enjoying studies of many different subjects including playing the piano and cello. She attended Cornell University and received a B.S. in Engineering Physics in 1996. Although she no longer uses physics very often, she appreciates the rigorous math training she received there. She went on to earn a M.S. in Biomedical Engineering from Duke University in 1999 and then work as a software engineer/analyst for IBM from 1999 through 2002 where she learned valuable skills in computer programming and database management.

She started her studies in Bioinformatics at NC State in 2003 under the direction of Zhao-Bang Zeng. Here she studied statistical methods applied to quantitative genetic data, and in particular became interested in network discovery algorithms in genomics. Christine started working for a small company called Nature Source Genetics in Ithaca, NY in 2006 and has been working there until the present. At NSG she has worked on developing optimization algorithms for the design and analysis of QTL experiments for large seed companies.

Christine married Schuyler Duarte in 2003 and has had two children: Anya (2005) and Summer (2008).

ACKNOWLEDGMENTS

I would like to acknowledge the financial support I received from an internship at SAS from 2003 to 2004 and by a VIGRE fellowship from the National Science Foundation through the Department of Statistics from 2004 to 2006. I would also like to thank my advisor Zhao-Bang Zeng for all of his help, and I would like to thank my other committee members Russell Wolfinger, Ronald Sederoff, and Jung-Ying Tzeng for their help as well. I would also like to thank my former committee member Bruce Weir for his help in the first couple of years of my studies.

I would also like to thank former members of the Zeng lab including Wei Zou, David Aylor, and Jessica Maia for their help in analyzing the yeast eQTL data set as well as general help in discussions of concepts involved in my dissertation. Finally I would like to thank the Kruglyak lab from the University of Washington for use of the yeast eQTL data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Literature Review

1.1 Introduction

The reverse engineering of genetic regulatory networks has been the focus of many researchers in recent years. This heightened interest has been spurred by the availability of large genomic data sets such as microarray expression data for various organisms under a variety of experimental conditions. Detailed observations about the change in expression of thousands of genes in response to various conditions can potentially be analyzed to learn regulatory relationships among genes by using appropriate data mining techniques.

Another source of information about genetic regulatory networks is the recent studies of genetical genomics or expression QTL analysis. In these studies, marker data and microarray data is collected from the offspring of an experimental cross of two parental lines, and QTL analysis techniques are used to find genetic loci that explain the variation in gene expression observed in the progeny. It has been observed in such studies that many gene expression variables that link to a common genetic locus are often functionally related and/or co-regulated and may represent modules in a gene regulatory network. To exploit the information contained in these data sets, some investigators have applied data mining techniques such as Bayesian Networks to analyze this data.

In this review, a summary of how data mining techniques have been used to discover genetic regulatory networks from microarray data will be given along with a summary

of recent Expression QTL studies. Because Bayesian Networks has been a popular technique for genetic network discovery, and also because it is used in this current work, a summary of the theory behind Bayesian Networks will be given in Section 1.2. In Section 1.3, a summary will be given of how data mining techniques such as Bayesian Networks have been used to analyze microarray data to discover genetic networks. In Section 1.4, a summary of eQTL studies will be given. Analysis of existing methods along with general conclusions and suggestions for future research will be given at the end of Section 1.3 and in Section 1.4.

1.2 Introduction to Bayesian Networks

This summary is drawn primarily from [52]. A Bayesian Network can be described as $G = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ represents a set of vertices or nodes in the graph, and $\mathbf{E}$ represents the edges between those nodes. If the Causal Markov Condition is satisfied, then the probability distribution for the vertices $\mathbf{V}$ can be decomposed as follows,

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V|\mathbf{Parents}(V)) \tag{1.1}$$

where $\mathbf{Parents}(V)$ represent all nodes with edges directed into the vertex V. The Causal Markov Condition basically states that in the probability distribution P over $\mathbf{V}$ generated by causal graph G, each variable or vertex is independent of its non-descendents given its parents. Two examples of Bayesian Networks along with their corresponding probability distributions are shown if Figure 1.2.

Thus a Bayesian Network describes the set of conditional independence relationships encoded in the probability distribution for a set of variables. Additional assumptions required for Bayesian Networks include the Causal Minimality assumption, the Faithfulness assumption, and the Causal Sufficiency assumption. Causal Minimality requires that no proper subgraph of G satisifes the Causal Markov assmption. The Faithfulness assumption requires that every conditional independence relationship true in the probability distribution P over the vertex set $\mathbf{V}$ is entailed by the Causal Markov Condition applied to G. Finally, the causal sufficiency condition requires that any common cause of two or more

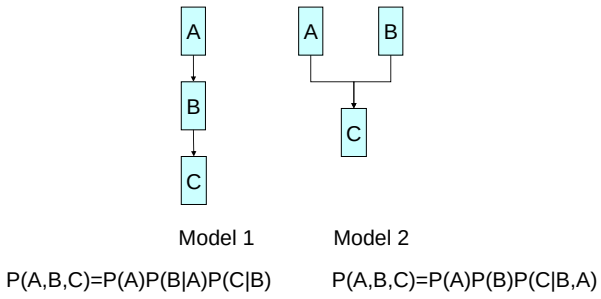

Figure 1.1: Examples of two Bayesian Networks along with their corresponding probability distributions.

variables in $\mathbf{V}$ be in $\mathbf{V}$, or in other words, that there are no latent variables that are not included in the vertex set.

The main problem in Bayesian Networks is finding the causal graph and corresponding probability distribution associated with a set of observational variables found in an experimental data set. There are two main classes of Bayesian Network learning algorithms that accomplish this task: the score-based approach and the conditional independence approach. In the score-based approach, an initial network is proposed, and a score is evaluated for that network. Then a number of changes to the network structure are proposed, and the score of the resulting network is evaluated. The goal is to find the highest-scoring network associated with a given data set. This problem is known to be NP-hard [15] and thus is computationally intensive as the number of variables grows. Usually heuristic methods such as greedy hill-climbing, simulated annealing, or other optimization techniques are used to search for the optimum solution in frequentist approaches and Markov Chain Monte Carlo (MCMC) sampling is used in Bayesian approaches.

The score can be any of a number of statistical measures of how well the network fits the data, and can include the maximum likelihood of the data given the network, the posterior probability of the network given the data, AIC, BIC, etc. Either frequentist or

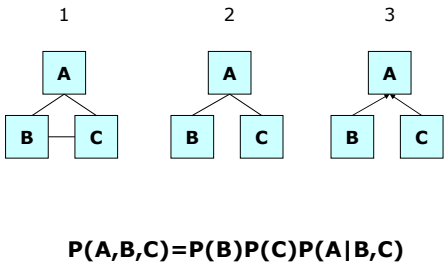

Figure 1.2: The three steps in the operation of the PC algorithm.

Bayesian approaches can be used in score-based methods. In Bayesian approaches, prior knowledge of the system can be incorporated by specifying prior probabilities for each of the possible network structures.

In conditional independence approaches, the conditional independence relations found in the data are directly used to select the appropriate network. One example of this approach is the PC or Peter Clark algorithm [52]. In the PC algorithm, first a fully-connected undirected network is constructed. Then edges are removed based on conditional independence relations of increasing order found in the data. Finally, the edges are directed if possible based on the types of conditional independence relationships found in the previous step. An example of the operation of the PC algorithm on a simple three node network is shown in Figure 1.2.

In step one the fully connected graph is constructed. In step two, it is found that B and C are marginally independent. Alternatively, B and C could have been found to be marginally dependent, but independent given their common neighbor A. Higher level conditional independence relations involve the conditional independence of B and C given two or more variables. Finally, in step three, the remaining edges are directed as because this is the only network consistent with the fact that B and C are marginally independent

but each are dependent on A. Alternatively, if it were found that B and C are conditionally independent given A, then any of the other three orientation combinations of the two edges found in Figure 1.2 would be valid ($B \rightarrow A$ $A \rightarrow C$, $C \rightarrow A$ $A \rightarrow B$, or $A \rightarrow C$ $A \rightarrow B$). Thus it can been seen that often the edges in a graph cannot be directed, or, one unique Bayesian Network cannot be specified. In such cases a family of networks that satisfy the conditional independence relations found in the data can be specified. Sometimes, prior knowledge can be used to direct edges that remain undirected after the Bayesian Network learning algorithm has been completed.

1.3 Reverse Engineering of Genetic Networks in Microarray Data

As mentioned previously, many data mining techniques have been applied to the analysis of microarray data for the purpose of reconstructing genetic regulatory networks. Studies that have used these various techniques will be reviewed here. First we will discuss the use of Bayesian Networks in genetic network reconstruction. Then we will discuss the use of Dynamic Bayesian Networks, as well as other techiques for modeling time-dependent microarray data. Then we will discuss the used of Boolean Networks, and then finally list other methods that have been used including maximum entropy, graphical gaussian models, structural equation models, biological model-based methods, and several custom-built algorithms for genetic network reconstruction. Finally we will discuss the limitiations of some of these approaches and suggest future areas of research.

1.3.1 Bayesian Networks for Reverse Engineering of Genetic Networks

In the first application of Bayesian Networks to microarray data, Friedman et al. [20] developed the "Sparse Candidate Algorithm" for Bayesian Network learning when the number of variables is in the thousands as is typically found in a microarray data set. The Sparse Candidate Algorithm reduces the size of the network search space by only allowing relatively few candidate parents for each variable based on a simple statistic such as correlation. Even with the simplified algorithm, many networks with roughly the same

score are found, and thus it is impossible to discern one correct network for a given data set. Friedman et al. decided instead to look for high confidence network features that tend to be found in a large majority of the high confidence networks. The network features they search for are Markov relations and order relations. Markov relations entail that one variable is a parent of another variable, and order relations imply that one variable is an ancestor of the other variable. The authors applied their method to a yeast cell cycle data set [51] and found a number of biologically interesting relationships among genes.

In [35], a Bayesian Networks approach is also used to resconstruct genetic networks from the analysis of a yeast cell cycle data set. They modeled the rate of transcription of each gene as a nonlinear function of its parents. They employed optimization algorithms to find the optimal solution and in applying their method to a yeast cell cycle data set containing 41 genes, found 70% (31/43) of true interactions. The authors mentioned the high CPU requirements of their method as well as the limitation due to noise in the data. They did not specifically address the issue of scaling their method to a larger number of genes.

Gamberoni et al. [21] used two different Bayesian Network learning algorithms to learn genetic networks from microarray data sets collected from experiments relating to Acute Myeloid Leukemia (AML). In [16], a new quantization method based on a model of the experimental error as well as a compromise between false negative and false positive interactions was used as a preliminary step in discrete Bayesian Network learning of genetic networks and was found to improve over existing methods.

In [12], the authors proposed an algorithm for genetic network discovery that improves over existing Bayesian Network hill-climbing techniques. In their approach, first an undirected network is constructed and split into a set of substructures. Then the edges are directed by optimizing a scoring function over each substructure. They showed that their method outperforms hill-climbing methods in terms of computational time and accuracy.

1.3.2 Dynamic Bayesian Networks and other Time-Dependent Methods for Reverse Engineering of Genetic Networks

An extension of Bayesian Neworks to time series data is called Dynamic Bayesian Networks (DBNs). In a DBN, any node at time t can be found to be causal for any other node (including itself) at time $t+1$. Thus DBNs remove the acyclic restriction that standard Bayesian Networks have, and they can better model the dynamic nature of the relationships among a set of variables. Given the importance of feedback loops and other cyclic network motifs in biological networks, DBNs are a popular choice.

The sensitivity and specificity of Bayesian Networks in inferring the correct genetic network structure is tested in a simulation study [24] in which a known genetic network is simulated and its structure is learned using a Dynamic Bayesian Network (DBN) approach using MCMC sampling. Receiver Operator Characteristic (ROC) curves are provided, and Husmeier found that the quality of the prior network probability distribution in addition to the time after experimental perturbaton at which the data is collected had large effects on the quality of the results. Husmeier also concluded that the number of false positive edges tends to grow as the square of the number of nodes, thus finding the true network structure for large networks is practically impossible. He agreed with the approach of Friedman et. al. [20] of looking for high confidence local network features rather than trying to find one correct global network structure.

Another simulation study was performed [55] for using Dynamic Bayesian Networks to reverse engineer genetic networks. In that study the dependence of algorithm performance on sample size, network connectivity, computational effort, and number of restarts was determined. It was found that the lower performance that occurs with large networks can be in part compensated by increased computational effort, and that the performance of the algorithm benefits more from a larger number of restarts rather than from a more sophisticated search strategy.

In [3], a new statistical method for using Dynamic Bayesian Networks to model gene expression data was proposed. In this technique, P-spline regression is used to solve some of the problems in the network inference from microarray data. Some of the problems dealt with include making the discrete time series data continuous, accounting for missing

data and noisy data, and accounting for the time dependent nature of the data. The authors found improved performance of their approach over the traditional linear model approach, although they admitted that the network search aspect of their algorithm still needed to be optimized.

In [17], the authors apply the Dynamic Bayesian Network (DBN) model to analyze data from perturbation experiments. They use an algorithm for finding the exact solution rather than a heuristic approach and found that that approach of incorporating data from perturbation experiments improves on the quality of discovered networks from time series data alone.

In [5], the authors used state-space models (SSMs) rather than Dynamic Bayesian Networks to model time-dependent microarray data. Specifically, the authors have expanded on previous work to model time series data recorded during T cell activation. In their previous work [44] they used a classical approach to determine confidence intervals for parameters representing gene-gene interaction over time, and in their current work they used a Bayesian approach for model selection. They have found that certain interactions are selected using either approach.

1.3.3 Boolean Networks for Reverse Engineering of Genetic Networks

Another popular network discovery technique is Boolean Networks. In Boolean Networks, all variables are forced to be binary, but uncertainty in the network relationships are incorporated in a manner similar to Bayesian Networks. Boolean networks are easier to implement, but are not as flexible or comprehensive as Bayesian Networks.

Directed acyclic Boolean (DAB) Networks were used to model the relationships among microarray gene expression variables in [34]. The DAB is characterized by a set of pairwise relations (prerequisite and similarity) between binary variables. The authors introduced a mechanism for generating random error and proposed a search strategy that minimizes false negatives and false positives, and applied their approach to simulated data and yeast pheromone response data.

Martin et al. [38] discovered Boolean activation-inhibition networks from clustered and discretized gene expression data. They tested their method on two immunology

microarray data sets and found that the discovered networks agreed with observed data.

1.3.4 Other Methods for Reverse Engineering of Genetic Networks

In [48], the authors circumvent the difficulty of Bayesian Network inference with the large number of variables and small sample size present in most microarray experiments by proposing the use of graphical Gaussian models (GGMs). Their method involves a small-sample calculation of partial correlation and an Empirical Bayes (EB) approach of finding the correct network topology. The GGM networks found are undirected and thus do not encode as much information as a Bayesian Network, but GGMs do appear to be more robust than Bayesian Networks for a small sample size.

In [57], a constraint-based approach was used in conjunction with graphical Gaussian modeling to create a method for causal structure learning of microarray data that handles the large number of variables better than in a Bayesian Network approach. In [40] the authors used a method that integrates estimating gene regulatory networks with estimating protein-protein interaction networks. In [22], the authors used quantitative association rules to find relationships among gene expression variables.

In [37], the authors proposed a method for reconstructing genetic networks in mammalian B cells. Their method involves an information theoretic approach (mutual information) for estimating gene coexpression. Their method achieves better results than Bayesian Networks. In [56], the authors proposed a constraint- based algorithm for discovering local causal structure in microarray data. Their approach improves on the computational complexity of Bayesian Networks.

The principle of entropy maximization has been used in [32] to infer gene interaction networks or higher-order interactions in the analysis of Saccharomyces cerevisiae chemostat cultures exhibiting energy metabolic oscillations. The authors found that their discovered gene interaction network reflects the intracellular communication pathways that adjust cellular metabolic activity and cell division to the limiting nutrient conditions that trigger metabolic oscillations.

A combination of orthogonal least squares, second order derivative for network pruning, and Bayesian model comparison was used in a method proposed in [28]. In this

study, the entire network was decomposed into a set of small networks that were defined as unit networks. A biological model-based strategy was proposed in [54]. The authors used an S- system based model for the transcription and translation process, and then applied an optimization-based regulatory network inference approach that uses time- varying data from DNA microarray analysis.

In order to infer transcriptional compensation interactions in yeast, a stepwise structural equation modeling algorithm (SSEM) was developed in [50]. An advantage of the SSEM approach is that it incorporates hidden variables to capture interactions or regulations from latent factors. A rank-based algorithm is used to specifically model networks with scale-free architecture in [11].

In [58], the authors discussed a system that recommends experiments for finding gene regulatory relationships. Their system calculates the posterior probability of a genetic regulatory relationship by analyzing gene knockout microarray data as well as observational microarray data, it recommends which experiments to perform and the sample size for those experiments, and then it analyzes the results of those experiments with previous results. The authors tested their system in a randomized study of ten biologists and found that those that used their system reached the correct causal assessment more often than those that did not use their system.

1.3.5 Analysis of Existing Methods for the Reverse Engineering of Genetic Networks

Many different methods for the reverse engineering of genetic networks in microarray data have been proposed, and many interesting biological results have been obtained. However, with the inability to experimentally verify the vast majority of predictions made using these approaches, the questions remains, how accurate are these approaches and with what level of confidence can the discovered networks be interpreted.

Simulation studies using known networks offer one way to evaluate existing methods. Examples of these include [55] and [24]. However, it is impossible to simulate all possible biological scenarios that may arise. Other solutions offered include bootstrap and other resampling-based methods for assessing confidence of observed network features ([20],

for example). These methods do give needed measures of confidence levels, but they often involve large computational costs. A common approach is to check the agreement of predicted networks with current biological knowledge or data (see [32] and [38], for example). This analysis is certainly helpful, but can only account for the validation of a fraction of the predictions made using most genetic network discovery techniques.

Another key problem with existing techniques as noted in [1] is the exclusion of many biologically-relevant data sources. Most genetic network reconstruction algorithms use only gene transcription levels as measured by microarray data sources, when in reality, gene regulation occurs at many different levels including translation and post-translational modification. By excluding "latent" variables such as protein levels from the analysis, it is possible for false inferences to be made. In particular, Margolin and Califano [1] discuss the effect of latent variables on the theory behind reverse engineering algorithms derived from three separate disciplines: system control theory, graphical models, and information theory.

Thus potential areas for improvement in genetic network discovery include more precise statistical methods for approximating confidence levels of predicted networks and network features, as well as the incorporation or integration of multiple sources of data. In addition, the field of genetic network discovery would benefit from some sort of automated experimental system for validating network discovery predictions such as that described in [58].

1.4 Expression QTL Analysis Studies

1.4.1 Introduction

Recently QTL analysis techniques have been applied to mRNA transcript abundances from microarray profiles of genetically distinct organisms to find the genetic basis for natural variation in gene expression. Because the effect of genetic variation on gene expression has been found to be significant in a number of studies, the importance of extending QTL analysis tools to gene expression data has been recognized, and expression QTL studies have been undertaken in a number of different organisms.

The first study in which QTL mapping techniques were applied to gene expression

data is presented in [8]. In this study the natural variation in gene expression among progeny from a cross between a laboratory strain and a wild strain of *Saccharomyces cerevisiae* was studied. General characteristics of the genetics of gene expression observed in this system include the high heritability of gene expression and the high complexity of gene expression with many loci contributing to transcript variation on average. This study also separated expression QTL or eQTL into two categories. The first category is called cis-acting which means that the eQTL is co-localized with the target gene, and the second category is called trans-acting which means that the eQTL is not co-localized with the target gene. This distinction is important to separate self-regulated or cis-regulated genes from trans-regulated genes which are regulated by other genetic factors. It was found that trans eQTL for many different genes often co-localize, possibly indicating co-regulation of those genes by a common locus.

Since the study by Brem et al. [8], investigators have applied expression QTL analysis to a wide range of organisms including yeast, mice, rats, humans, maize, and trees. eQTL analysis has been used to dissect complex molecular mechanisms underlying clinical traits of interest in many systems. Methodological advances in eQTL analysis include the modeling of multiple loci with epistatic interactions, the integration of clinical data with genotype and gene expression data, the reduction of dimensionality, the incorporation of prior information, and the development of software tools that make eQTL analysis accessible to more researchers. Applications of eQTL analysis include the recovery of genetic regulatory networks, disease gene discovery, and the dissection of the genetic basis of complex traits.

Issues in eQTL analysis that require further attention include the high dimensionality of the analysis, the acceptance of a convention for measuring the significance of discovered eQTL, the development of databases and software tools for integrating various sources of data relevant to eQTL analysis in specific organisms or crosses, and the refinement of existing statistical techniques and computational algorithms.

Existing eQTL studies will be summarized by organism including yeast, mice and rats, humans, and other organisms, and the specific issue of cis- and trans-acting eQTL will be discussed. Then studies of multiple loci and epistatis in eQTL data will be summarized.

Practical applications of the analysis of eQTL data will be discussed including genetic network reconstruction algorithms applied to eQTL data, and remaining methodological issues in analyzing eQTL data will be summarized in addition.

1.4.2 Studies in Yeast

In the eQTL study of yeast by Brem and her co-workers [8], the investigators found that differences in gene expression were highly heritable, with a median heritability of 84%. In performing simulations with their experimental results they found that most parental gene expression differences were caused by multiple loci, with most loci contributing less than one third to the overall genetic variation, and with greater than five loci contributing to each transcript level difference on average. They found that roughly 36% of genes with marker linkages were cis-regulated, with the remainder being trans-regulated. Furthermore, they found that in many cases just a few trans-regulators accounted for the genetic variation in a large number of genes. A more recent analysis of the yeast data set using multiple interval mapping has estimated the proportion of cis-regulated genes to be 14.2% [61]. In [61], overlap of the gene by the 1.5 LOD dropoff interval (expected 95% confidence interval) was used as a criteria for cis-regulation.

In [7], the authors performed an eQTL study in a cross of two strains of yeast with a larger number of segregants (112) and a larger number of genes (5,700) as compared with [8]. The authors wished to study general characteristics of the genetics of gene expression in yeast including the number of eQTL, the effects of eQTL, and interactions among eQTL. Consistent with their previous work, they found that most genes exhibit complex genetic patterns with only 3% of highly heritable loci found to be consistent with single-locus inheritance, 17-18% consistent with control by one to two loci, and half requiring over five loci under additive models. In addition, it was found that 40% of heritable genes had no detected QTL, underlying the small and thus undetectable effect of most QTL. In addition it was found that transgressive segregation, which is defined as transcripts with an excess of segregant values outside of the parental range, was the most prevalent form of segregation. This result was thought to be due to genetic drift or indicative of a mechanism for generating diversity. Epistasis was also found to be an important phenomenon because evidence was

found for it in 16% of highly heritable transcripts.

Other aspects of gene expression in yeast have been studied such as general characteristics of trans eQTL regulation [59], the importance of epistasis in the genetics of gene expression [7], and techniques for jointly estimating the effects of multiple loci on the expression of a gene [53].

1.4.3 Studies in Mice and Rats

Schadt et al. [47] used QTL mapping techniques to study the genetic basis of gene expression in a cross of two inbred strains of mice. Schadt and his colleagues also found the genetic basis of gene expression to be complex. They found that 40% of transcripts having at least one eQTL were found to have two eQTL, and 4% of such transcripts have more than three eQTL. One advantage of expression QTL analysis cited by these authors is that although positions of individual eQTL may not be considered significant due to the multiplicity of the tests being performed, the coincidence of many eQTL in one location imparts more confidence and can aid in the detection of loci that influence the expression of many traits. In this study the authors demonstrate how to combine eQTL results with clinical and microarray data to dissect the heterogeneous genetic basic of a clinical obesity phenotype.

Expression QTL analysis was used in a cross of two recombinant inbred lines of rats in [23]. In this study, the importance of tissue specificity in the genetics of gene expression was studied because microarray data was taken from kidney and fat tissue in rats. It was pointed out by the authors that performing expression QTL analysis with recombinant inbred lines has two major advantages. First, the heritability is increased because replicates of genetically identical animals are obtained, and second, since these lines are maintained over time, results from different experiments in different labs can be combined into a shared resource that can be jointly analyzed by many different investigators. Hubner and co-workers separated the discovered eQTL into cis-acting and trans-acting as in other studies, and in particular they found that cis eQTL can serve as positional candidates for clinical QTL when the QTL positions coincide. They demonstrated this phenomenon by investigating some of the most significant cis eQTL from their study and

finding functional and molecular data that supported the possible roles of the target genes in metabolic syndrome and hypertension, two diseases investigated in their study.

Expression QTL analysis was also performed using hematopoietic stem cells and brain tissue from recombinant inbred mice [30]. Tissue specificity of the discovered eQTLs was analyzed by comparing results from brain cells and stem cells. Also, the identities of previously-discovered clinical trait QTLs were hypothesized by looking for cis-regulated genes co-located with the clinical QTLs. The authors mentioned the potential to search for genetic networks by looking for cis-regulated genes that overlap with the eQTL of trans-regulated genes, which could represent potential downstream targets of the cis-regulated genes. Finally, the details of a new resource called WebQTL were described. WebQTL contains data from the current study as well as links to historical data on mice from the same inbred lines which can be used for custom analysis by other investigators. The construction of this database was described in [13], along with additional eQTL studies performed on brain tissue of mice from the same cross of recombinant inbred lines as in [30]. The ability to search for associative networks and cliques in WebQTL is described. In associative networks, correlated gene groups can be found as potential candidates for co-regulation, and cliques are defined as groups of highly interconnected genes thought to be prevalent in the network architecture of biological systems.

In [2], the use of high density SNP maps to find genes linked to expression differences in mice was shown to complement existing QTL techniques in unraveling the genetic architecture of gene expression.

1.4.4 Studies in Humans

The genetics of gene expression has been studied in humans in a few studies [14], [45], [39], [47]. In [14], F-ratios computed as variability among individuals divided by variability within an individual were calculated for gene expression profiles to determine the set of genes that were the most differentially expressed among individuals from the Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees in lymphoblastoid cells. They also determined that these differentially expressed genes had larger expression differences between unrelated individuals than between siblings, and larger differences between siblings

than between monozygotic twins.

In [39], microarray analysis was performed on B cells from unrelated individuals in the CEPH pedigrees. For genes differentially expressed between individuals, they performed linkage analysis using publicly available SNP data and found eQTL for several genes. The authors found evidence of both cis and trans regulation, and they found two eQTL hotspots in which many genes had eQTL that mapped to a specific region in the human genome. They confirmed many of the cis-acting loci using follow-up association and differential allelic expression studies.

Lymphoblastoid cells from CEPH families were also studied for the genetics of gene expression variation in [47] and [45]. Of the genes that were differentially expressed among individuals, 31% were heritable. In [45], multipoint-based identity-by-descent sharing was used to perform a linkage analysis across autosomal chromosomes, and eQTL were detected for several genes in both trans and cis. QTL hotspots, in which many eQTL clustered in small genome regions, were not observed as in other studies ([47], [8], [59]. One possible reason is that the study was underpowered. Another possible reason is that hotspots were found in studies of experimental populations derived from crosses of inbred lines likely differing in a few key QTL. These few QTL are of great importance in determining the genetic variation between lines and were therefore easier to detect. Another analysis performed in this study was the calculation of genetic correlation between genes once the QTL analysis of each gene had been accomplished. It was found that clustering of genes based on genetic correlation resulted in gene clusters consistent with known biological pathways. The authors proposed that this form of clustering offers a complementary or even superior approach compared with clustering based on correlation of microarray profiles alone.

1.4.5 Studies in Other Organisms

QTL analysis was applied to gene expression traits in a cross of two lines of maize in [47], and many eQTL were discovered both in cis-acting and trans-acting modes. In [29], expression QTL analysis was applied in a cross of two Eucaplyptus tree species to find the genetic determinants of growth differences between these species. Two major eQTL were found for a set of genes related to lignin biosynthesis whose expression was correlated

with tree growth traits. These eQTL also co-located with QTL for growth traits, providing further evidence that variation in lignin biosynthesis gene expression provides a molecular mechanism for the generation of growth differences in the two species of trees surveyed. To further dissect the interactions of the lignin biosynthesis genes, it was found that most of the gene expression differences were controlled in trans except for one, S-adenosylmethionine synthase, which may represent an important hub in the molecular pathway for growth determination.

1.4.6 Investigation of Cis-acting and Trans-acting eQTL

In [59], QTL analysis of gene expression in a cross of two strains of yeast was used to examine general characteristics of trans-acting eQTL. The authors noted that natural genetic variation serves as a "treatment" to use for cluster analysis of microarray profiles. In fact, it can be considered superior to other clustering mechanisms such as engineered genetic mutations or chronological phases in cell cycle processes because of the complexity of different combinations of parental polymorphisms occurring in the progeny. The authors performed QTL analysis on these clusters and also on individual genes and found that 75% of genes tested and 80% of clusters tested for genetic linkage did not show self-linkage (were not cis-regulatory), thus showing the importance of trans regulation. By clustering genes and then performing linkage analysis, the authors could more confidently declare the significance of eQTL regulators due to the lower number of tests being performed. They demonstrated two examples in which the eQTL responsible for the trans regulation of a cluster was mapped to a gene and functionally characterized. The authors found that clusters of genes were often highly enriched for specific transcription factor binding sites. However, in looking at whether clusters of genes co-regulated by a common transcription factor linked to the region of the genome containing that transcription factor, the authors found very few examples of this. They concluded that genetic variation in transcription factors does not account for most of the trans-genetic variation found in gene expression.

Doss et al. [18] have concentrated on the analysis of cis eQTL because of their importance in the identification of clinical QTL in mice and other organisms. They derived a calculation of the False Discover Rate (FDR) for cis eQTL, and they also found that

greater than 96% of cis QTL were in genetically distinct regions as characterized by SNP haplotype. The authors illustrated methods for analyzing the correlation among genotypes, gene expression data, and clinical phenotype data to further characterize cis-acting eQTL.

1.4.7 Joint Analysis of Multiple Loci and Epistasis

To characterize the importance of epistasis between naturally-occurring polymorphisms in the genetics of gene expression, epistasis was studied using QTL analysis of gene expression in a cross of two strains of yeast [7]. Epistatic interactions were first searched for using an exhaustive search of all possible pairs of markers. Few epistatic interactions were found, probably due to the low power of this approach. In a second two-stage approach, a primary QTL was found for each transcript abundance, and a secondary QTL was then found by partitioning the segregants based on the genotype of the primary QTL and then testing the subgroups for secondary loci. Using this method in conjunction with a technique for estimating the fraction of transcripts for which both QTL are involved in the genetic basis of transcript variation, it was estimated that 57% of transcripts were jointly controlled by two loci. Among the transcripts predicted to have a QTL pair, 65% of the pairs were predicted to interact epistatically. Among the detected QTL pairs, 67% of the secondary loci were predicted to have effects too small to be detected in a genome-wide scan. Epistatic pairs that affected multiple transcripts were found, including a pair that affected the gene expression in 14 transcripts. Thus the authors concluded that epistatic interactions were highly prevalent in the genetics of gene expression.

In [53], a technique was developed for jointly estimating the effects of multiple loci on a gene expression trait. It was noted that the genetic basis of gene expression is thought to be complex with multiple loci required to explain the expression variation in most genes. However, existing methods for predicting the effects of multiple loci are not thought to be adequate. Existing methods include an exhaustive two dimensional scan for interacting loci, which has low power due to the large number of tests being performed, and model selection approaches like Multiple Interval Mapping (MIM, [26]) that involve the potentially problematic search over the large space of possible models. Their solution was to use a stepwise selection approach in which a secondary locus is selected conditional on

a primary locus being present. This approach could in theory be extended to more than two loci. To assess the significance of the joint participation of both loci, an empirical distribution for the test statistic that represents the probability that both loci contribute to the trait is plotted, and the null distribution of this test statistic is plotted after performing permutation tests. The empirical distribution is modeled as a mixture of the null and alternative distributions, and the mixing proportions of these two distributions is estimated as a function of test statistic value. In this way, a significance value can be computed for each test statistic. Using this procedure, the authors estimated that at least 37% of all transcript abundance traits in the yeast data set analyzed showed joint linkage to two loci. Furthermore, pairs of loci contributing to gene expression variation were estimated for 170 genes with 153 expected to be true positives, and epistatic interactions between loci were expected to contribute to variation in 14% of all expression traits.

1.4.8 Applications of Expression QTL Analysis

Schadt et al. [46] have demonstrated an application of expression QTL analysis that allows for prioritization of potential gene candidates underlying certain diseases. The authors outline a multi-step process; the first steps includes QTL analysis and model selection for a clinical trait associated with the disease, identification of gene expression profiles correlated with the disease phenotype, and filtering of this list by including only genes with eQTLs that overlap the disease QTLs. In the next step the authors fit the QTL, gene expression trait, and disease phenotype to one of three models using AIC (the Akaike Information Criterion) to determine the best-fitting model. The first model (causal model) assumes that the QTL causes the gene expression variation, which in turn causes the disease phenotype. The second model assumes that the QTL causes the disease phenotype through some other mechanism which in turns causes the gene expression variation (reactive model), and in the third model the gene expression and disease phenotype are effected independently by the QTL (independent model). Only those genes for which the causal model was the best fit were included in further analysis. In the final step, the remaining genes were rank-ordered according to the percentage of genetic variation in the disease phenotype that was causally explained by variation in the gene expression level.

Applying this technique to the analysis of gene determinants for obesity in mice, the authors identified one known gene and three previously unknown genes to be causally related to obesity. These gene candidates were validated using gene knockout experiments. This is a potentially powerful technique to narrow down the search for disease gene candidates. Traditional QTL analysis of clinical traits allows for a large number of potential genes, and typically a labor-intensive process is needed for narrowing down the list. Thus this new method that leverages the QTL analysis of gene expression, represents a big improvement over the current process.

Li et al. [33] discuss using QTL analysis of gene expression data to search for plausible networks of genetic interaction. Their approach involves narrowing the list of potential networks by looking for potential regulatory genes only in the region of a trans eQTL for a target gene. The list of potential modulators of the target gene is further reduced by only considering genes near SNPs found to be polymorphic between the parental strains crossed to form the progeny. Finally, Bayesian Network techniques are used to evaluate the probability of the remaining networks. The use of Bayesian Networks in conjunction with expression QTL studies to predict potential genetic networks is also presented in [25].

Bing and Hoeschele [6] also outline a technique for finding potential regulatory networks through QTL analysis of gene expression in yeast. In their technique, QTL analysis of expression profiles is performed, and the positions of discovered eQTL are refined using fine mapping. Then candidate regulatory genes are identified in each eQTL region, and the correlation of expression profiles of target and candidate genes are measured to filter out unlikely eQTL candidates. The retained candidate and target genes are connected by directed edges, edges are joined to form networks, and the networks are then validated and refined. It was found that specific biological processes were overrepresented within network structures discovered using this approach.

Two recent studies on the discovery of genetic networks from the analysis of eQTL data are described in [27] and [41]. In [27], the authors perform genome-wide expression variation analysis in a RIL population of Arabidopsis thaliana. Their apporach for regulatory network construction combines eQTL mapping and regulatory candidate gene selection. They evaluated their approach in a study of genes associated with flowering time, which

is a well-studied regulatory network in Arabidopsis. The authors found that clusters of coregulated genes and their regulators were in agreement with published data. In [41], the authors use previously-developed Bayesian Network algorithms for learning the structure of genetic networks from microarray data and then use the eQTL results to direct edges in the structure.

1.4.9 Methodological Issues in the QTL Analysis of Gene Expression

Though QTL analysis of gene expression holds much promise in dissecting the genetic and molecular basis of complex traits, many unresolved issues remain in how this technique is used. Some of these issues are outlined in [9]. For instance, one application of QTL analysis of gene expression is to help uncover the molecular basis for disease. However, Broman points out that this can only be successful when assaying the correct tissue at the correct time point for determining disease etiology. Another important issue in QTL analysis of gene expression is differentiating between gene expression differences that cause or are caused by disease, although this issue is subsequently addressed in [46]. Broman also points to the lack of inclusion of important factors such as proteins, metabolites, and the compartmentalization of molecular factors in most eQTL studies. Other important issues include the resolution of the genetic map in determining whether two QTL are shared or only closely linked, the multiple testing issue, and the need for better software tools for analyzing this complex data.

Another issue in eQTL studies is the reliability of gene expression data and the assessment of significance of discovered eQTL. In [10], the authors recommend using weighted least squares to model gene expression as a function of eQTL in order to weigh based on the repeatability of the trait and the number of replicates. They also recommend using prior genetic information to help assess the degree of confidence in eQTL discovery. Finally they recommend measuring the significance of an eQTL by calculating the False Discovery Rate (FDR) based on a single trait using prior information on the repeatability of the trait and the location of the gene transcript, with gene transcripts co-located with their eQTL (cis QTL) considered with more confidence. The use of weighting is also mentioned in [36] in which heritability weighting is used to adjust the definition of the transcript abundance for

each gene. They found that they were able to detect more eQTLs using this method.

The issue of high dimensionality in eQTL studies is addressed in [31] in which two techniques are used to account for the correlation structure among gene expression traits. These techniques include principle components and hierarchal clustering seeded by disease relevant traits. Their approach was used to find significant linkages for a cluster of lipogenic and gluconeogenic genes in a cross of mice. It was found that these linkages coincided with loci for type 2 diabetes in the same cross of mice.

To study the methodological issues of eQTL analysis in detail, Perez-Enciso et al. [43] undertook a simulation study in which genotypes and phenotypes were simulated based on publicly-available microarray data. The relationship between genotype and phenotype was modeled using partial least squares logistic regression with a continuous latent variable liability related to a set of measured cDNA expression levels. It was found that the power to detect linkage was increased when the number of cDNA levels decreased in the liability, and that the number of significant cDNA levels is increased if a large number of cDNAs are co-expressed. In a related study [42] the power to detect eQTL in outbred populations is tested by simulating linkage between available human microarray data and SNP data, and it was found that the estimation of QTL positions was unstable, and that a small decrease in sample size lead to a large decrease in power. Perez-Enciso et al. anticipated conflicting results in future whole-genome association studies for gene expression.

To aid in the visualization of expression QTL results, the tool Expressionview was developed within Ensembl as described in [19]. Both the location of genes and QTL are displayed across chromosomes with links to databases of additional information available by moving the mouse over specific QTL or genes. A tool for extrapolating from QTL linkage value to physical location is also provided. WebQTL as described in section 1.4.3 [30] is another tool for visualizing expression QTL results. It also provides access to large databases of information that can aid in the interpretation of eQTL results. A tool called eQTL Viewer [60] has been developed to view the position of discovered eQTL on the x axis and the physical position of the gene with the eQTL on the y axis. In this way cis-regulated genes can be visualized as a diagonal line, and trans-regulated genes mapping to eQTL "hotspots" can be seen as vertical lines. Other capabilities of this application include links

to gene annotation, ability to highlight eQTL for genes in a specified pathway, and other tools for exploring the results of an eQTL analysis.

1.4.10 Discussion

QTL analysis of gene expression data holds great promise for uncovering the complex molecular mechanisms underlying complex traits and disease phenotypes. One reason that eQTL analysis is such a powerful approach for understanding the genetic control of complex traits is that gene expression values serve as a molecular intermediate between clinical traits and their QTL. Traditional QTL analysis allows for the identification of loci associated with complex traits. However, the mapping of genes responsible for QTL is a time-consuming process, and the molecular mechanisms by which the mapped genes influence the clinical trait may still be unknown. QTL analysis of gene expression data of thousands of genes performed in conjunction with traditional QTL analysis of clinical traits can allow for the discovery of more detailed mechanisms of molecular action. For instance, gene expression traits that are correlated with a clinical trait may all map to a trans- acting eQTL that represents a key hub in the molecular pathway leading to variation in the clinical trait. Furthermore, the gene expression traits that map to that eQTL may be members of a regulatory network which ultimately influences the clinical trait, and methods for network discovery using expression QTL analysis have been described in section 1.4.8.

Another application of eQTL analysis is the mapping of clinical trait QTLs by looking for co-localized cis-acting QTL for gene expression traits [46]. The integration of many sources of data with expression QTL analysis has also been described as a powerful way to dissect molecular mechanisms involved in the variation of clinical traits. The discovery of disease genes and regulatory networks in many different organisms points to the utility of expression QTL Analysis.

Key methodological issues in eQTL analysis still need to be resolved as detailed in section 1.4.9. Included among these are the high dimensionality of the analysis, the difficulty in assigning significance levels to discovered eQTL, and the complexity thought to underlie most gene expression traits combined with the low power to resolve the influence of many QTL with small effects. One issue is to what extent results from eQTL studies

in experimental populations can be generalized to natural populations. In one instance, the lack of evidence for QTL hotspots in human data [45] illustrates possible differences in the analysis of experimental and natural populations. Thus, the use of expression QTL analysis in natural populations needs further study. In summary, though expression QTL analysis involves the use of genomic data which brings with it a host of statistical and computational issues, the potential benefits and the benefits already realized for using this data to help dissect the genetic architecture of complex traits definitely warrants further study into expression QTL analysis.

1.5 Discovery of Gene Regulatory Networks in Expression QTL Data Sets

We have outlined previous research in methods for discovery of genetic networks from microarray data (Section 1.3) as well as previous research in the genetics of gene expression (Section 1.4). In this work we build on these two areas of research by applying genetic network discovery techniques to expression QTL data sets. Expression QTL data sets are a natural choice for genetic network discovery because they allow for finding associations both among gene expression variables and also between genetic loci and genetic expression variables. However, the large number of variables in a typical microarray data set make the use of existing network discovery algorithms like Bayesian Networks impractical.

Some researchers have already applied modified network discovery algorithms to expression QTL data sets ([33], [6], [27], and [41]). In most of these studies the causal gene for the set of transcripts linked to an eQTL is searched for amongst the genes co-localized with the eQTL ([33], [6], and [27]). In [41], a more general algorithm is proposed that uses previously-developed Bayesian Network algorithms for learning the structure and then uses QTL results to direct edges in the structure. In this work we propose a method that is completely general and uses QTL and expression data in all parts of the network discovery process. Our approach differs from previous approaches using expression QTL data for network reconstruction because our method is completely general, it does not make assumptions about what genes will be causal in the network, and it takes into account the

unique structure of expression QTL data in proposing and evaluating potential network structures.

In this work the development of a new method for genetic network discovery is motivated by our desire to understand the mechanism by which eQTL regulate gene expression in existing and new data sets. To make sure that our method is relevant to existing data sets, we have chosen the yeast expression QTL data set originally presented in [8] as a "model" data set to base our analysis on. We look at basic properties of the eQTL found in that study to help us design our simulation study, and we also test our resulting method on that data set. Thus understanding the regulation of gene expression in yeast through analysis of the data set in [8] is a common theme that recurs throughout this work.

The rest of this thesis is organized as follows. In the next chapter (Chapter 2) we will describe the motivation behind our method, describe the steps in our method, and provide details for assessing the confidence of discovered network features. Then in Chapter 3 we will describe a simulation study we performed to assess the performance of our method. In Chapter 4 we apply our network discovery algorithm to a yeast expression QTL data set. We describe the networks discovered and also perform bioinformatic analyses to interpret the results. Finally, in Chapter 5, we summarize the findings in this work, we discuss potential areas for improvement, and we present ideas for future research in the discovery of genetic networks from expression QTL data sets.

Chapter 2

Methodology

2.1 Analysis of an expression QTL Data set

In this work we will be concerned specifically with expression QTL or eQTL data sets, although our method can be readily extended to any data set containing genotypic/genomic data and phenotypic data. In an eQTL data set, microarray expression data and marker data at a set of markers spanning the genome are collected for a set of individuals from an experimental cross or a sample population. Typically, QTL analysis is performed on each gene expression profile from the microarray data set to find QTL underlying differential gene expression in segregating individuals from a cross. It has been found in many studies that several gene expression variables will link to the same QTL, and oftentimes these genes are found to be functionally related. A pertinent follow-up question in these studies is how does the QTL jointly modulate the expression of all the genes that link to it, and can a network of gene interaction be proposed.

We will propose a method for reconstructing regulary networks from eQTL data sets using as nodes the gene expression profiles (or genes for short) and the QTL found to influence a set of genes. This type of data is unique in that the direction of edges between QTL and genes is known to be toward the genes (it is not logical for gene expression to cause a change in the DNA sequence of an individual, at least not in the short term). Thus networks constructed from eQTL data sets can be rooted at the QTL or QTLs. This is a

desirable property which greatly simplifies network construction as will be shown.

To test our method we will use a yeast eQTL data set in which microarray and marker data was collected from the progeny of a backcross of two species of *Saccharomyces cerevisiae* [8]. This data set is a good validation data set for our method for a couple of reasons. Firstly, it is the first expression QTL (eQTL) data set made available, and has been analyzed by many different researchers using many different methods. Thus we can compare the results of our method with other published methods. Secondly, the yeast genome is well-annotated, and networks predicted using our method can be interpreted using this wealth of information. We can check if genes found to be associated with the same QTL share functional, cellular, or molecular roles as has been done previously. Also we can query the annotations of predicted regulator and regulated genes and use these to propose possible biological mechanisms for the predicted regulatory relationships.

2.2 Introduction

Our method involves a modification of the PC (Peter Clark) algorithm described in [52]. We will first briefly describe the operation of the PC algorithm, and then we will describe our algorithm, which derives from the PC algorithm.

In the PC algorithm, first the complete undirected graph is constructed over all of the network nodes, which means that every node is connected to every other node with an undirected edge. Then edges are removed between each pair of adjacent nodes X and Y if X and Y are found to be independent either unconditionally or condional on one or more nodes adjacent to either X or Y (not including X or Y). Conditional independence tests of increasing order are performed, where order is the size of the set of nodes used to condition the correlation of X and Y. These tests are repeated until the order of the test performed equals the number of nodes adjacent to X or Y, and all possible conditioning subsets have been tested. This completes the construction of the undirected graph. Then the remaining edges are directed if possible based on the types of conditional independence relationships found in the previous steps (we will skip the details of this step for brevity).

Conditional independence is measured using calculation of the partial correlation

coefficient of X and Y conditional on $\mathbf{C}$, where X and Y are nodes, and $\mathbf{C}$ is any set of nodes (including the empty set) adjacent to either X or Y. $\rho_{XY|\mathbf{C}}$ is calculated as using the following two equations,

$$\rho_{XY|C} = \frac{\rho_{XY} - \rho_{XC}\rho_{CY}}{\sqrt{1-\rho_{XC}^2}\sqrt{1-\rho_{CY}^2}} \tag{2.1}$$

where C is any member of $\mathbf{C}$. Then the additional members of $\mathbf{C}$ are added to the conditioning in a stepwise manner using the following equation,

$$\rho_{XY|\mathbf{Z}\cup R} = \frac{\rho_{XY|\mathbf{Z}} - \rho_{XR|\mathbf{Z}}\rho_{YR|\mathbf{Z}}}{\sqrt{1-\rho_{XR|\mathbf{Z}}^2}\sqrt{1-\rho_{YR|\mathbf{Z}}^2}} \tag{2.2}$$

where $R = C \in \mathbf{C}$ and members C are added until $\mathbf{Z} = \mathbf{C}$. $\rho_{X\hat{Y}|\mathbf{C}}$ is obtained by substituting sample estimates of the correlation parameters into Equations 2.1 and 2.2. $\rho_{X\hat{Y}|\mathbf{C}}$ is tested for significance using the Fisher's z transformation,

$$z(\rho_{X\hat{Y}|\mathbf{C}}) = \frac{1}{2}\sqrt{N-3-|\mathbf{C}|}\ln\left(\frac{|1+\rho_{X\hat{Y}|\mathbf{C}}|}{|1-\rho_{X\hat{Y}|\mathbf{C}}|}\right) \tag{2.3}$$

where $|\mathbf{C}|$ is the number of nodes in $\mathbf{C}$.

As an example of the operation of the PC algorithm, start with the undirected network shown in Figure 2.1. Assume that the edges shown are those that remain after testing all pairs of nodes for (unconditional) correlation. Then the first order tests will be performed to try to remove additional edges. First start with the 1-2 edge. The significance of this edge given node 3 will be tested, then the significance given node 4, and then given node 5. If all of these correlations are significant, then the second order tests will be performed: the correlation of nodes 1 and 2 conditional on each pair of nodes adjacent to node 1 will be tested (3 and 4, 3 and 5, and 4 and 5). If these tests are all significant, then the correlation of 1 and 2 on the triplet of nodes 3, 4, and 5 will be tested. If this is significant, then all possible combinations of nodes adjacent to 1 (not including 2) have been tested and found to be signficant, and thus the 1-2 edge will be retained. Then this process is repeated for all other edges in the graph (1-3, 1-4, and 1-5).

The PC algorithm is a flexible method for the construction of networks. However, it suffers some disadvantages as the number of nodes becomes large. The large number

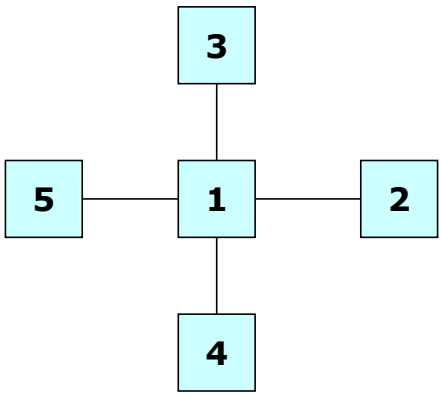

Figure 2.1: Illustration of the PC algorithm: undirected network in which the significance of the edge between nodes 1 and 2 is tested using conditional independence tests of increasing order for combinations of nodes adjacent to node 1 not including 2 (3, 4, and 5).

of tests performed leads to a multiple testing problem, thus calculation of the family-wise false positive rate needs to be performed. Also, as the number of tests increases, the power decreases, because true edges are tested multiple times, each with a chance for a Type II error (false negative).

In our current application we wish to determine the mechanism by which an eQTL regulates a large number of genes linked to that eQTL. With the number of genes linked to an eQTL numbering 500 or more in some cases, it is impractical to run the PC algorithm (or other standard Bayesian Network discovery algorithms). However, we can simplify the network search when we recognize that the network must be rooted at the QTL.

Another simplification to the network search can be realized if an assumption is made that the network follows a scale-free or power-law architecture. In a scale-free network, the distribution of the degree (number of edges) for each node is given by the power-law relationship,

$$P(k) \sim k^{-\gamma} \tag{2.4}$$

where k is the number of edges (or degree) and γ is a parameter typically between 2 and 3. See [4] for a more detailed description of scale-free networks as well some common examples. In this type of network, a few highly connected "super" nodes are responsible for regulating most of the nodes in the network. By combining the rooting of our networks at the QTL with an expected scale-free architecture, we can propose a new network discovery algorithm based on the PC algorithm that gives reasonable power even for large networks.

2.3 General Approach

The goal in Bayesian Networks is to decompose the joint likelihood of the vertices into the product of the conditional likelihood of each vertex given its parents. Thus the probability distribution for the vertices $\mathbf{V}$ is given by,

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V|\mathbf{Parents}(V)) \tag{2.5}$$

where $\mathbf{Parents}(V)$ represent all nodes with edges directed into the vertex V.

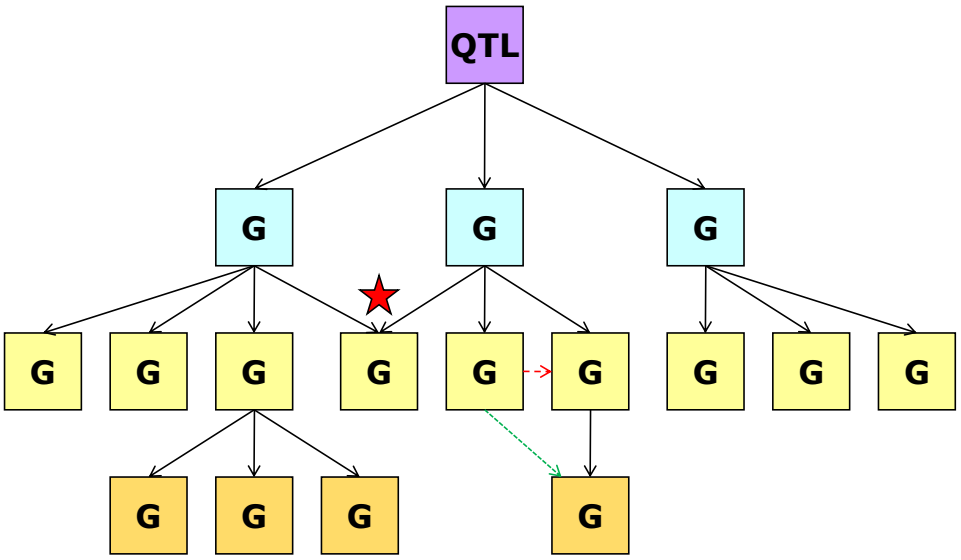

Figure 2.2: Sample network used to show how our algorithm decomposes the network into a series of hierarchal layers.

In an expression QTL data set, we have a set of genes linked to an eQTL, and we are interested in finding the mechanism by which the eQTL regulates the genes. Finding the exact network is difficult, but we propose a method for finding a "framework network" that decomposes the network into a series of hierarchal layers. To illustrate this concept, see the network in Figure 2.2. We start by recognizing that all genes are linked to the QTL, so all genes must have some connection to the QTL in the resulting network, either direct or indirect.

We first find the genes with direct connections to the QTL (the genes in green in Figure 2.2). Once these genes are discovered, then all the other genes in the network must

be connected to these "primary genes", either directly or indirectly. We designate these "primary genes", the genes in green with direct connections to the QTL, as "shielders". Specifically we call them "primary shielders" because they directly shield the QTL. Here "shield" means that the other genes in the network are independent of the QTL conditional on these genes, or in other words, the primary shielders "shield" the other genes from the influence of the QTL. We can thus decompose the joint likelihood of the genes as follows,

$$P(\mathbf{G}|T) = \prod_{S \in \mathbf{S}} P(S|T) \prod_{G \in \mathbf{G} \backslash \mathbf{S}} P(G|\mathbf{S}) \tag{2.6}$$

where T represents the target (the QTL), $\mathbf{G}$ represents the set of genes linked to the target, $\mathbf{S}$ represents the set of genes found to have direct connections to the target, and $\mathbf{G}\backslash\mathbf{S}$ represents the set of genes in $\mathbf{G}$ that are *not* in $\mathbf{S}$.

For the remaining genes in the network that are not primary shielders, $\mathbf{G}\backslash\mathbf{S}$, they may or may not have direct connections to the primary shielders. The genes with direct connections to the primary shielders are called secondary shielders, because they shield other genes in the network from the primary shielders. The secondary shielders are in yellow in Figure 2.2. The likelihood in Equation 2.6 can the be further decomposed once secondary shielders are found. Specifically, for each $G \in \mathbf{G}\backslash\mathbf{S}$ and each $S \in \mathbf{S}$, $P(G|S)$ is decomposed by setting $T = S$, $\mathbf{G} = \mathbf{G}\backslash\mathbf{S}$, and searching for a new set $\mathbf{S}$ for Equation 2.6. This process of finding shielders repeats until all genes in the network are shielders or have direct connections to shielders.

In this process we label as the target what is being shielded by a given set of shielders. For primary shielders, the target is the QTL. For secondary shielders, the target is one of the primary shielders. Each set of shielders of a given target form a hierarchal layer, and in general, edges can only be directed from a node in a higher layer to a node in a lower layer. One or more genes may be required to shield a gene in the lower layer. For instance, in Figure 2.2, the red star indicates a gene that is shielded by two genes in the green layer.

In general, it is possible for edges to exist within a set of shielded genes, like the red arrow in Figure 2.2. However, our algorithm will not detect these edges. Instead our algorithm will add the green arrow shown since both yellow genes with edges directed

into the orange gene are required to make the gene independent of the green gene in the preceding layer. It is for this reason that we call the network discovered with our algorithm a framework network rather than the full network.

2.4 Specific Algorithm

Now that we have introduced our algorithm in general terms, we will go into the specific operation here. Our goal is to find a series of hierarchal layers that contain a set of shielders, which renders the genes in the layers below independent of the QTL and genes in the layers above. We accomplish this by a recursive call to an algorithm called "FindShielders". The input to the algorithm is a target (T) and a group of genes with a signficant correlation to the target ($\mathbf{G}$). We initialize the algorithm with the target being the QTL and the gene group being the set of genes linked to the QTL. Here is the algorithm:

1. Discovery of shielders: test each gene in the gene group for a significant target-gene edge by performing conditional independence tests of increasing order as in the PC algorithm (see beginning of Section 2.2) using as conditioning variables all genes in the gene group. Mathematically, find the set of genes $\mathbf{S}$ such that for all $S \in \mathbf{S}$, $|\rho_{TS|\mathbf{G}\backslash S}| > 0$. This is also the set $\mathbf{S}$ that causes Equation 2.6 to be satisfied.

2. Discovery of shielded genes: For the remaining genes in the gene group that are not shielders, and for each shielder, test for a significant shielder-gene edge by performing conditional independence tests of increasing order using as conditioning variables the other shielders of the target. Mathematically, for each shielder $S \in \mathbf{S}$, find a set of genes $\mathbf{G}'$ where $\mathbf{G}' \subset \mathbf{G}\backslash\mathbf{S}$ such that for all $G \in \mathbf{G}'$, $|\rho_{SG|\mathbf{S}\backslash S}| > 0$. Then for each shielder S with a non-empty $\mathbf{G}'$, call "FindShielders" with the target being the shielder S and the gene group being the set of genes with edges remaining to the shielder, $\mathbf{G}'$.

We will illustrate this algorithm with an example (see Figure 2.3). The true network is shown in A. For this example we will look for direct connections simply by looking at the true network. In real data sets we use conditional independence tests. In the

first step shown in frame B, we call FindShielders with the QTL as the target (purple) and all of genes linked to the QTL (G1 through G8, yellow) as the gene group. In the next step (frame C), we search for shielders. We find that only G1, G2, and G3 (blue) have direct connections to the target when looking at the true network in A. Thus we declare G1, G2, and G3 to be shielders of the QTL target. For each shielder, we need to find which of the remaining genes (in yellow) have direct connections to the shielder conditional on the other shielders. In frame D, we show that for the shielder G1, only G4 has a direct connection given the other shielders, G2 and G3. Then we call FindShielders with G1 as the target and G4 as the gene group, but because there is only one gene in the gene group, there is no need to search for shielders, and we simply create an edge from the target, G1, to the single gene, G4.

Next, in frame E, we search for genes with direct connections to the next shielder, G2. We find that all the non-shielder genes, G4 through G8, have direct connections to G2 given the other shielders, G1 and G3. So we call FindShielders with G2 as the target and G4 through G8 as the gene group. In F, we discover that only G5 and G6 have direct connections to the target, G2; we thus label these as shielders. We search for which of the remaing genes in the gene group (G4, G7 and G8) have direct connections to each shielder given the other shielder. We find that all three genes (G4, G7, and G8) have direct connections to G5 given the other shielder, G6. In G we call FindShielders with G5 as the target and G4, G7, and G8 as the gene group. In H, we discover that all three genes in the gene group have direct connections to the target given the other genes, so we label each as shielders and create an edge from G5 into each of them.

In I we look for genes with direct connections to the last shielder from the previous recursion, G6. We find that no genes in the gene group have direct connections to G6, so we do not need to call FindShielders. Finally, in J we go back to the first recursion and look for genes with direct connections to the shielder G3. We find that only G6 has a direct connection, and thus we call FindShielders with G3 as the target and G6 as the gene group. Then we simply create an edge from G3 to G6 because there is only one gene in the gene group and no need to search for shielders. In K we show the final discovered network. Note that the edge between G7 and G8 was not detected, because it is within the set of shielded

genes of the shielder G5.

Although our algorithm in its current form is a simplification of the PC algorithm, it will still have low power and high compuational cost for a large number of variables because for higher order tests the number of tests increases exponentially with the number of variables. We can mitigate this problem by limiting the order of the conditional indpendence tests that are performed with a "maximum order" parameter. We will investigate the effect of this relaxation in the Simulations Chapter (Chapter 3).

2.5 Statistical Significance

To examine the confidence level for discovered network features using our algorithm, we must first examine what statistical tests are performed at each step. Each step of our algorithm involves a target, T, a gene group, $\mathbf{G}$, a set of discovered shielders, $\mathbf{S}$, and a set of shielded genes per shielder, $\mathbf{G}'$. When looking for shielders of a target among the members of the gene group, there are two requirements for each shielder $S \in \mathbf{S}$:

1. A significant correlation between the shielder (S) and the target (T) conditional on all combinations of group genes ($\mathbf{G} \backslash S$) with set size up to a specified maximum order, i.e., $|\rho_{TS|\mathbf{G}\backslash S}| > 0$.

2. A significant correlation between the shielder (S) and each shielded gene ($G \in \mathbf{G}'$) conditional on all combinations of other shielders ($\mathbf{S} \backslash S$) with set size up to a specified maximum order, i.e., for all $G \in \mathbf{G}'$, $|\rho_{SG|\mathbf{S}\backslash S}| > 0$.

For the first requirement, a series of conditional independence tests of increasing order are performed at a level α, and only if *all* of these tests are positive will the first requirement be satisfied. Thus the Type I error rate for simultaneously rejecting all of these tests must be $\leq \alpha$. However, this Type I error rate applies to a single shielder test, and in our algorithm we will test all n genes linked to a QTL for having a direct edge to that QTL, so multiple testing is an issue. To give further stringency to our test of putative shielders, we go to the second requirement, which is that a shielder have a signficant correlation to other genes in the gene group conditional on the other shielders. If correlation with these

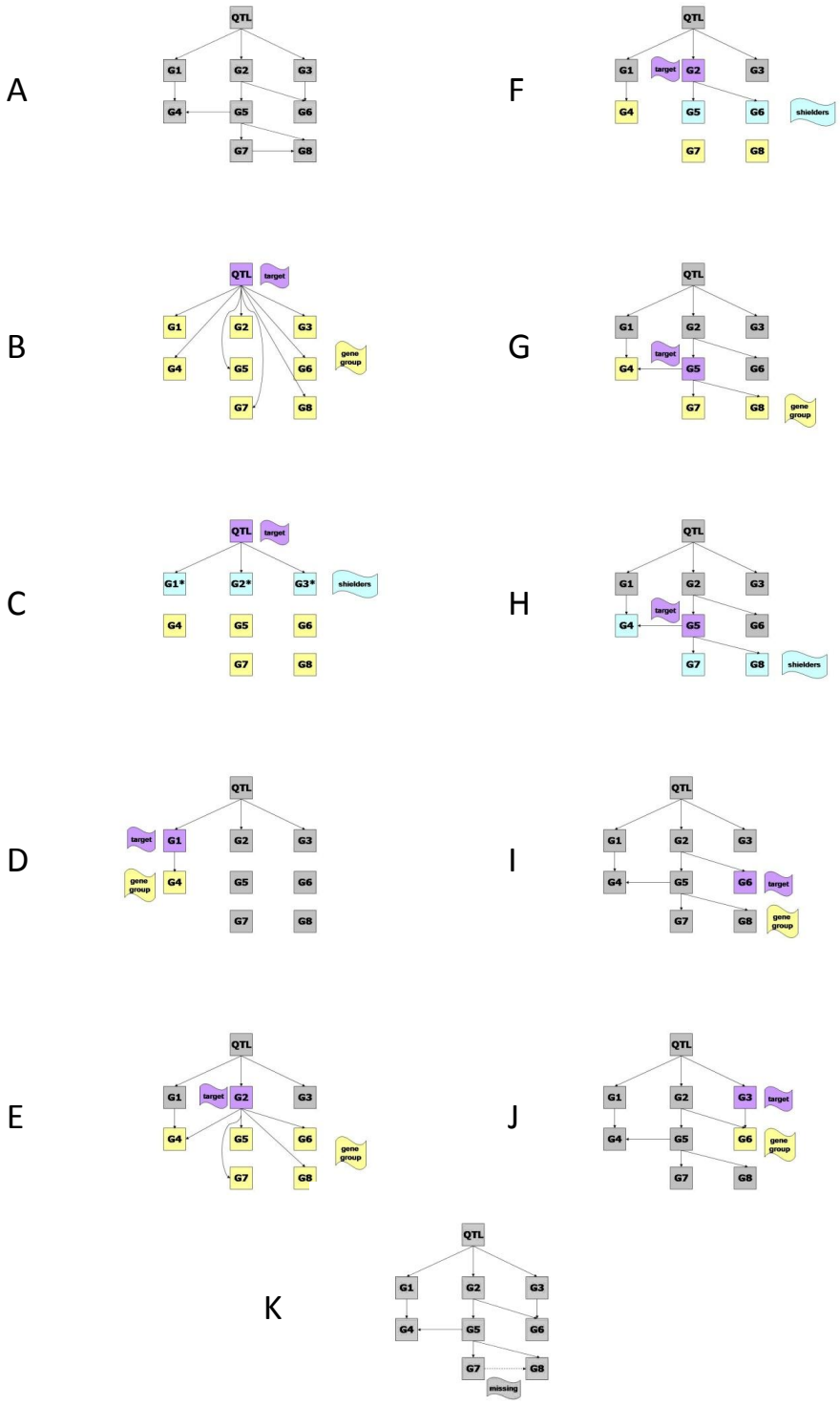

Figure 2.3: Illustration of the step by step operation of our algorithm on a sample network.

genes conditional on the other shielders exists, we say these genes are "shielded" by the shielder.

Intuitively, a putative shielder which shields more genes has stronger evidence for being a shielder. Thus, for all the genes (putative shielders) with direct connections to the QTL, we look for further evidence of shielder status by looking at the number of genes shielded by each shielder. To look at this question more quantitatively, we set up a hypothesis test for declaring a shielder. If a gene is not a shielder, than it does not shield any genes in the gene group. Specifically, if T is the target, $\mathbf{G}$ is the set of genes linked to the target, $\mathbf{S_T}$ is the set of *true* shielders of the target, and P is a putative shielder which is *not* a true shielder, then the correlation of the putative shielder with each gene in the gene group conditional on the true shielders should be zero, i.e.,

$$\rho_{PG|\mathbf{S_T}} = 0 \tag{2.7}$$

where $G \in \mathbf{G} \backslash \mathbf{S_T}$. If we assume that we have found all shielders of the target, then the Type I error rate for declaring a significant correlation between a shielded gene and a shielder conditional on the other shielders must also be $\leq \alpha$ because again, a series of tests each performed at a level α must *all* be positive for there to be a significant conditional correlation declared between the shielder and a shielded gene. Thus for each putative shielder, the probability that a gene is found to be correlated to it given the other shielders when that gene is *not* directly correlated to the shielder is $\leq \alpha$.

Thus we set up the hypothesis test as follows. The null hypothesis is that the putative shielder P is not a shielder and thus it shields no genes,

$$\mathrm{H}_0\text{: for all } G \in \mathbf{G}, \rho_{PG|\mathbf{S_T}} = 0 \tag{2.8}$$

where P is the putative shielder, $\mathbf{S_T}$ is the set of true shielders of the target T, and $\mathbf{G}$ is the set of genes that have significant unconditional correlation with T. The alternative hypothesis is that the putative shielder P is a shielder and thus shields one or more genes,

$$\mathrm{H}_1\text{: there exists } G \in \mathbf{G} \text{ such that } |\rho_{PG|\mathbf{S_T}}| > 0. \tag{2.9}$$

If we assume that the tests of correlation between each putative shielder and each gene given the other shielders are independent (a conservative assumption), then we can

model the number of falsely declared shielded genes per shielder as a binomial distribution with probability α. Thus if a putative shielder P is found to shield a set of genes $\mathbf{G'}$, and the null hypothesis that P is not a shielder is true, then the probability of obtaining $|\mathbf{G'}|$ or more positive tests is given by the binomial distribution,

$$P(x \geq |\mathbf{G'}|) = \text{pbinom}(p = \alpha, n_{\text{trials}} = |\mathbf{G} \backslash \mathbf{S}|) \tag{2.10}$$

where $|\mathbf{X}|$ designates the number of members of $\mathbf{X}$ and $\mathbf{S}$ is the set of discovered shielders. If this probability is $\leq \alpha_1$ where α_1 is the desired level of the test, then we can reject the null hypothesis that P is not a shielder.

The assumption of independence of tests, though probably not true, results in a conservative upper bound on the number of false positive results expected under the null hypothesis. The assumption that all of the true shielders of a target have been discovered, however, will not be true in general, and this issue will be investigated in the Simulations Chapter (Chapter 3).

In theory, α_1 can be decreased as much as necessary to achieve a suitably low shielder false positive rate. However, power will decrease as α_1 is decreased. In addition, errors due to using a maximum order parameter (i.e. not performing all orders of conditional tests) do not necessarily decrease with increasing α_1. Also, violation of the assumption of our test of finding all of the true shielders may contribute to a higher than predicted false positive rate. The effect of the size of α_1 on the shielder false positive rate and power will thus be investigated further in the Simulations Chapter (Chapter 3).

Another way to assess confidence in shielder status is to use bootstrapping. Bootstrapping attempts to recreate the sampling distribution of the data from a single data set, and thus confidence levels of network features given variablility in the data can be measured. We sample with replacement from the data set a number of times and apply our algorithm to see in what percentage of data sets a gene is declared to be a shielder. Requiring different levels of bootrap support for shielders gives an added level of confidence in the shielder status of genes. The level of bootstrapping support needed to ensure low shielder false positive rates will be investigated in the Simulations Chapter (Chapter 3).

Chapter 3

Simulations

3.1 Introduction

In order to evaluate the properties of our method for genetic network reconstruction in an expression QTL data set, we have performed a set of simulations. When generating a simulated data set, there are a number of parameters we have to specify. First we have to specify the structure of the network. To specify the structure we need to specify how many genes the QTL directly influences, and then how many genes are in turn influenced by each of those genes, etc. In these simultaions we will look at networks with scale-free architecture. Specifically, the parameters we have to specify include the number of network genes, the scale-free parameter γ, the number of network edges, the number of genes dierctly influenced by the QTL, the QTL effect or regression coefficient for the QTL on each network gene, and the regression coefficient for each gene-gene interaction.

In order to derive the relevant simulation parameters, we have analyzed some basic statistical properties of the yeast eQTL data set that we use throughout this work. We will use this data set as representative of other eQTL data sets that may be analyzed using our method. We cannot directly find all of the parameters from the data set without actually performing a network reconstruction, but we can look at some basic statistical properties of the yeast data set and make sure the parameters we choose in our simulations result in statistical properties that are consistent with those seen in the yeast data.

The basic statistical properties of the yeast data set we have analyzed include:

1. Size of the QTL as measured by the R^2 (genotypic variance of QTL as a percentage of the total trait variance) for each network gene.
2. Size of the network (number of genes typically linked to a given QTL).
3. Level of gene-gene correlation among network genes.

We then chose scale-free parameters, QTL and gene regression coefficients, and network sizes and edge densities that gave approximately the QTL sizes, level of gene-gene correlation, and gene group sizes observed in the yeast data set.

3.2 eQTL Analysis of Yeast Data Set

The yeast data set was analyzed using Multiple Interval Mapping (MIM) as described in [61]. The analysis consisted of two steps:

1. (Unconditional) single QTL analysis for each trait
2. For QTL found to be significant in the previous step, analysis for additional QTL conditional on the QTL found to be significant in the previous step.

LOD thresholds were determined for each step based on the expected FDR (false discovery rate, for details see [61]). Because our current network reconstruction method only allows for single QTL models, we used the results of the first step only. The position of the QTL for each gene is unique as determined by the MIM output, but our method requires a list of genes that link to a single QTL. There are several ways to group genes with nearby QTL. For simplicity we have used the method used orignally in [8] that involves dividing the genome into bins and finding eQTL "hotspots" as bins with eQTL for many different genes.

We used a sliding window approach with bin size of 20 cM and a bin increment size of 10 cM. Previous analysis [61] shows that the average 1.5 LOD dropoff interval was 25 cM in this data set, so the bin size we use roughly correponds to the expected 95%

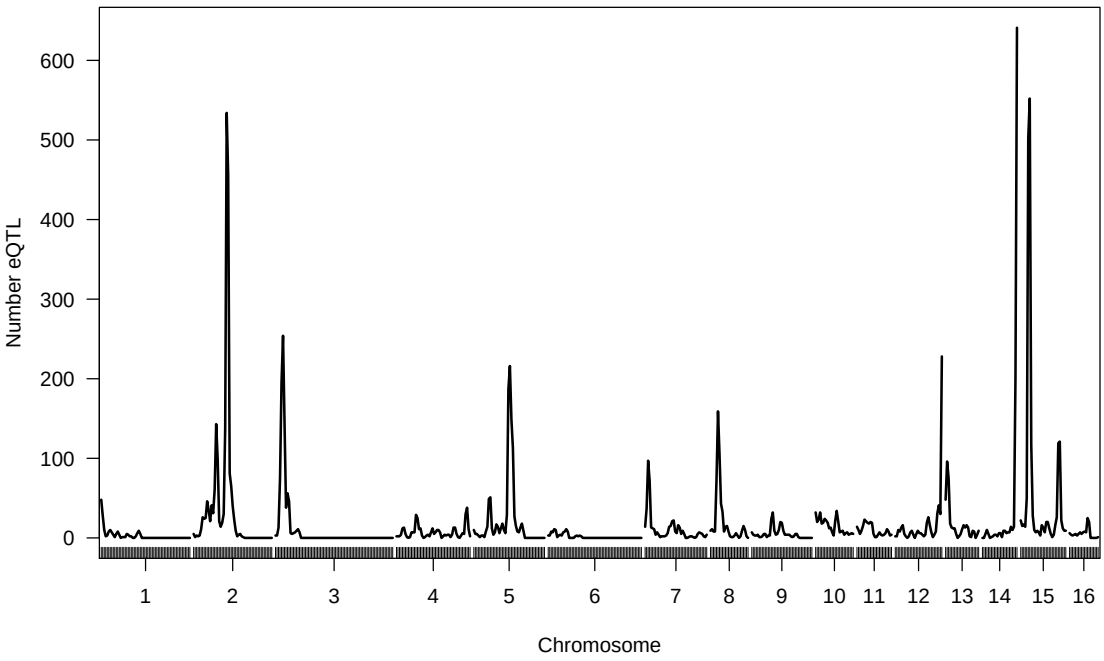

Figure 3.1: Number of eQTL per 10 cM bin across 16 yeast chromosomes

confidence interval for each eQTL. Figure 3.1 shows a plot of the number of genes with eQTL in each bin across the genome. Several eQTL "hotspots" are immediately apparent on chromosomes 2, 3, 5, 8, 12, 13, 14, and 15.

We selected eQTL hotspots by finding the bins with the largest number of linked genes, or the peaks of the curve shown in Figure 3.1. We chose to analyze eQTL hotspots with around 100 or more linked genes because QTL with many linked genes represent important hubs of transcriptional control.

Some relevant statistics about each gene group (eQTL hotspot) are summarized in Table 3.1. The following attributes are described for each eQTL: the position by chromosome and position in cM, the percent genetic variance explained by the eQTL (R^2), the percentage of other genes linked to each gene, and the average correlation among genes for positively- and negatively-correlated genes, respectively. The number of linked genes is

Table 3.1: Basic Statistics for the gene groups from several eQTL hotspots

QTL		Size	R^2		Correlated Genes			
chrom	pos	n genes	mean	sd	mean	sd	pos	neg
2	160	143	11%	5%	87%	15%	0.46	-0.42
2	230	517	16%	10%	81%	18%	0.42	-0.35
3	60	253	17%	12%	81%	16%	0.44	-0.31
5	250	213	12%	6%	75%	19%	0.39	-0.31
7	30	91	11%	3%	93%	11%	0.56	-0.44
8	60	155	14%	8%	66%	17%	0.30	-0.24
12	330	234	20%	14%	64%	15%	0.28	-0.25
13	20	93	14%	5%	63%	15%	0.28	-0.25
14	240	622	19%	9%	65%	14%	0.30	-0.24
15	70	531	16%	8%	79%	16%	0.38	-0.32
15	270	120	10%	6%	78%	19%	0.43	-0.28

seen to vary from around 100 to greater than 600. The average R^2 is seen to be large and varies from 10% to 20%. The average percentage of correlated genes is also large and is seen to vary from 63% to 93%. The average gene-gene correlation is also high and varies from 0.24 to 0.56. Gene-gene correlation is expected among genes linked to a common QTL due to the covariance induced by that common QTL. Additional correlation among genes due to direct gene-gene interactions can be deduced by applying our network reconstruction algorithm.

3.3 Selection of Simulation Parameters

In selecting parameters for simulating network structures, we hope to mimic the QTL sizes, number of genes, and degree of gene-gene correlation found in the yeast eQTL hotspot gene groups. The parameters that need to be specified include the scale-free parameter γ, the number of edges in the network, the size of the network (number of genes), the number of genes directly connected to the QTL, and the size of the QTL-gene and gene-gene regression coefficients.

Since γ values between 2 and 3 are commonly observed in biological networks [4], γ values of 2, 2.5, and 3 were used. A maximum network size of 500 genes was used since

this is near the maximum size of networks observed in the yeast data set (see Table 3.1). The number of edges in each network was between 0.65 and 2 times the number of genes in the network: specifically edge to gene ratios of 0.65, 1, 1.5, and 2 were used. Some combinations of the smaller γ values and the larger number of edges could not be generated due to constraints of the scale-free architecture.

Networks of various sizes were generated by varying the number of edges in the network as well as the scale-free parameter γ. By network size here we mean the number of network genes with significant assocation with the QTL. The total number of genes in the network remains constant at 500 (including genes linked and unlinked to the QTL). This distinction is made because only those genes with significant association with the QTL will be observed in the eQTL data set, and these are the ones used in the network reconstruction. Networks with more edges and a less scale-free topology (smaller γ) resulted in larger networks (near the maximum size of 500). This is because a highly connected network results in many genes with some connection to the QTL, either direct or indirect. More sparse networks resulting from less edges or a more scale-free topology (higher γ) resulted in smaller networks since less genes had connections to the QTL.

The number of genes directly connected to the QTL was chosen to be between 1 and 5. The QTL-gene and gene-gene regression coefficients were chosen to give QTL R^2 values and gene-gene correlation values similar to those observed in the yeast data set. Different ranges of values were experimented with, and finally it was decided to draw the QTL-gene regression coefficients from a uniform distribution between 1.5 and 3, and to draw the gene-gene regression coefficients from a uniform distribution between 0.5 and 1.5. It will be shown in the next section that this range of regression coefficients resulted in a distribution of QTL R^2 values and gene-gene correlation values similar to that of the yeast data set.

Table 3.2: Basic Statistics of the simulated gene networks.

Network Params				Size	R^2		Correlated Genes			
scenario	n primary	γ	ratio	n genes	mean	sd	mean	sd	pos	neg
1	1	2	0.65	29	12%	15%	15%	11%	0.14	-0.14
2	1	2	0.9	41	19%	17%	34%	20%	0.24	-0.24
3	1	2	1.5	181	24%	14%	78%	18%	0.39	-0.39
4	1	2	2	333	27%	12%	90%	11%	0.49	-0.49
5	1	2.5	0.65	31	14%	18%	19%	14%	0.15	-0.15
6	1	2.5	0.9	78	22%	13%	62%	25%	0.35	-0.35
7	1	3	0.65	34	14%	15%	22%	15%	0.17	-0.17
8	3	2	0.65	37	19%	20%	28%	19%	0.19	-0.19
9	3	2	0.9	67	27%	19%	54%	24%	0.29	-0.28
10	3	2	1.5	305	31%	16%	79%	16%	0.38	-0.38
11	3	2	2	430	30%	14%	84%	14%	0.44	-0.44
12	3	2.5	0.65	42	19%	18%	31%	20%	0.20	-0.20
13	3	2.5	0.9	108	27%	16%	63%	23%	0.31	-0.31
14	3	3	0.65	58	26%	19%	48%	25%	0.26	-0.27
15	5	2	0.65	47	21%	20%	35%	21%	0.21	-0.21
16	5	2	0.9	107	26%	16%	59%	23%	0.28	-0.29
17	5	2	1.5	386	33%	16%	82%	15%	0.38	-0.38
18	5	2	2	437	36%	17%	84%	14%	0.41	-0.41
19	5	2.5	0.65	57	25%	20%	45%	24%	0.24	-0.25
20	5	2.5	0.9	200	29%	14%	73%	20%	0.32	-0.32
21	5	3	0.65	70	29%	20%	53%	25%	0.28	-0.28

3.4 Simulation Results

3.4.1 Basic Statistics

Four networks and ten data sets per network were generated for each combination of simulation parameters. We used a maximum order parameter of 1 in all of these simulations. The basic statistics for each set of network parameters are shown as an average over the four networks and ten simulated data sets per network in Table 3.2.

The number of genes in each network family varied widely from 29 to 437. The average R^2 values varied from 12% to 36%. The average R^2 values for the yeast data set were typically between 10% and 20%, so R^2 values covering the range found in the yeast data set as well as above that range were simulated. The average percentage of correlated

genes ranged from 15% to 90%. The average percentage of correlated genes for the yeast data set was on the higher end of this range between 63% and 93%. The average positive and negative gene-gene correlation values ranged from 0.14 to 0.44 for the simulated data sets. The yeast data set was on the higher end of this range between 0.24 to 0.56. Thus the range of statistics seen in the simulated data sets was equal to or greater than the range seen in the yeast data set.

3.4.2 Sensitivity and Specificity

There are two paramaters we wish to study using simulation: the number of shielded genes per shielder constraint which is controlled by the parameter α, and the level of boostrap support for shielders. α is the level of the test used to test the null hypthesis that a shielder shields no genes versus the alternative hypothesis that it shields one or more genes (see Section 2.5). Note that the level of the test for ordinary conditional independence relations was held constant at 0.05.

We tested α values of 1 (i.e. no required minimum number of shielded genes), 0.2, 0.05, 0.01, and 0.002. The specificity for shielders in each of the 21 scenarios tested (averaged over the four networks and ten simulations per network) is shown in Figure 3.2. Shielder specificity is the percentage of discovered shielders that are true, or in other words, 1 - Specificity is the false discovery rate for shielders. Error bars are approximate 95% confidence intervals estimated by 1.96 times the standard deviation among network means (with network means obtained by averaging over simulated data sets). Missing bars indicate no shielders were detected.

It can be seen that a substantial increase in shielder specificity is obtained when going from an α of 1 to 0.2 (blue to green). A small increase in specificity for some scenarios is obtained when increasing α to 0.05 (green to yellow), but increases in α beyond 0.05 do not appear to increase the shielder specificity. This leveling off of shielder specificity may be due to the use of the maximum order parameter which allows for exclusion of some higher order conditional independence tests (see Section 2.4), or it may be due to the violation of some of the assumptions of the method (see Section 2.5). Since the maximal increase in shielder specificity is obtained with $\alpha = 0.05$, we have chosen to use this value in subsequent

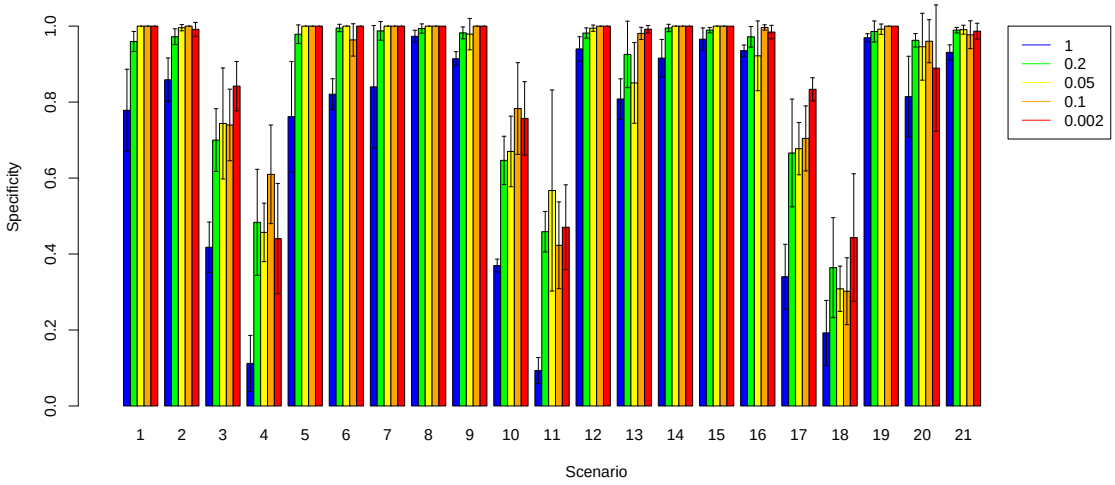

Figure 3.2: Shielder specificity for different required numbers of shielded genes per shielder given by the parameter α. We applied our network discovery algorithm to 21 different network scenarios characterized by different numbers of genes with direct connection to the QTL, different scale-free parameters, and different numbers of edges (see Table 3.2). 4 networks and 10 data sets per network were simulated. Results are averaged over all simulated networks; error bars are approximate 95% confidence intervals estimated from the standard deviation among network means.

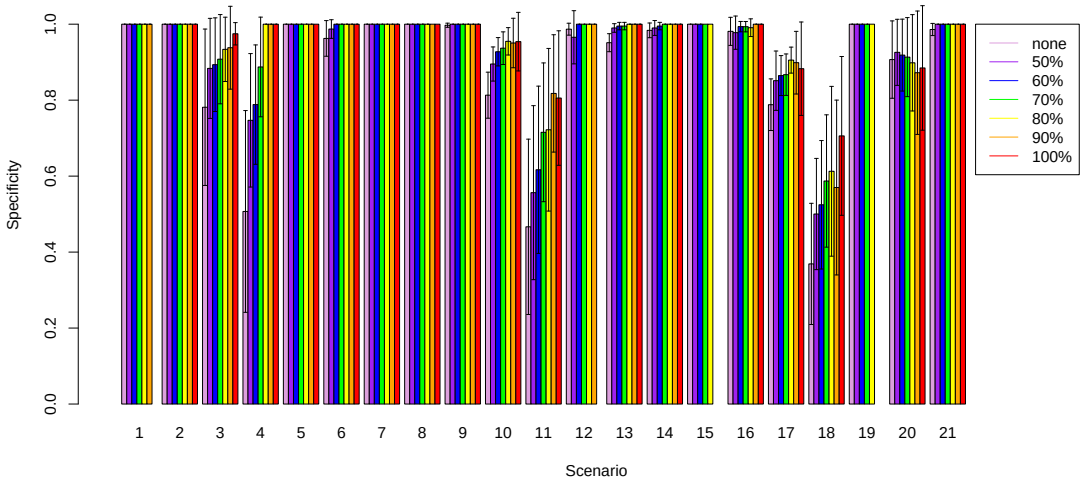

Figure 3.3: Shielder Specificity at different levels of bootstrap support in simulated data sets. We applied our network discovery algorithm to 21 different network scenarios characterized by different numbers of genes with direct connection to the QTL, different scale-free parameters, and different numbers of edges (see Table 3.2). 4 networks and 10 data sets per network were simulated. Results are averaged over all simulated networks; error bars are approximate 95% confidence intervals estimated from the standard deviation among network means.

simulations.

Next we wanted to investigate the level of bootstrap support for shielders needed to obtain good shielder specificity. We took ten bootstrap samples from each simulated data set and recorded discovered shielders occuring in 50%, 60%, 70%, 80%, 90%, or 100% of tested bootstrap samples. Figure 3.3 shows these results.

Shielder specificity is seen to be high for many scenarios even without bootstrap support. However, some scenarios, including 4, 11, and 18, have somewhat low specificity that is improved by increasing the bootstrap confidence level. Edge specificity is shown as a function of bootstrap confidence level in Figure 3.4.

Increased bootstrap confidence does not necessarily improve edge specificity for all scenarios; however, many scenarios had nearly 100% shielder specificity even without boos-

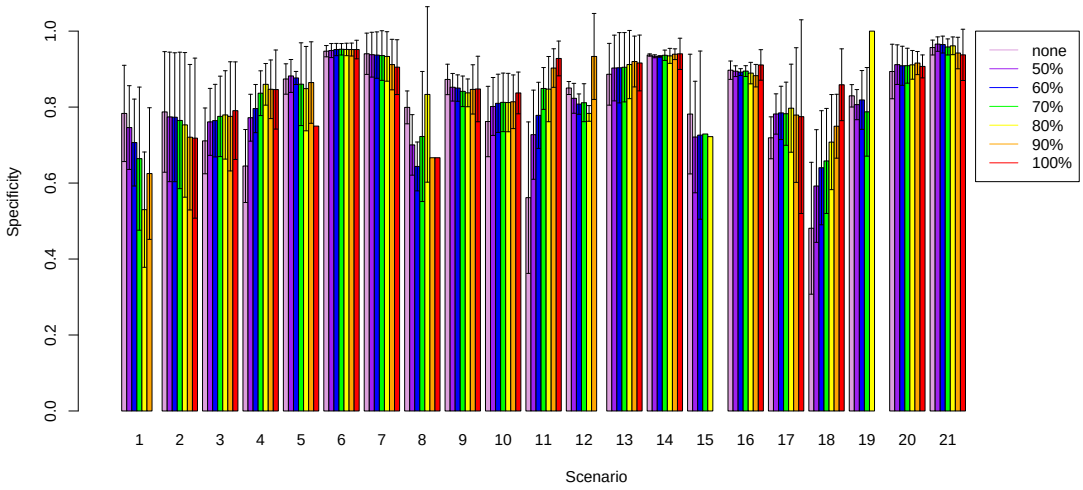

Figure 3.4: Edge Specificity at different levels of bootstrap support in simulated data sets. We applied our network discovery algorithm to 21 different network scenarios characterized by different numbers of genes with direct connection to the QTL, different scale-free parameters, and different numbers of edges (see Table 3.2). 4 networks and 10 data sets per network were simulated. Results are averaged over all simulated networks; error bars are approximate 95% confidence intervals estimated from the standard deviation among network means.

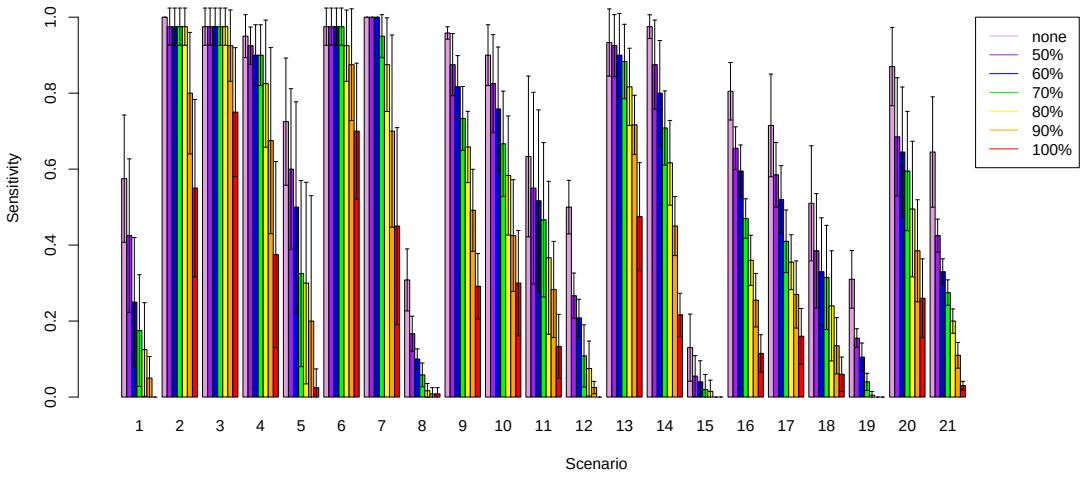

Figure 3.5: Primary Shielder Sensitivity at different levels of bootstrap support in simulated data sets. We applied our network discovery algorithm to 21 different network scenarios characterized by different numbers of genes with direct connection to the QTL, different scale-free parameters, and different numbers of edges (see Table 3.2). 4 networks and 10 data sets per network were simulated. Results are averaged over all simulated networks; error bars are approximate 95% confidence intervals estimated from the standard deviation among network means.

trap support. It is these scenarios in which edge specificity does not increase with increased bootstrap confidence. In the scenarios 4, 11, and 18, however, where increased shielder specificity is realized with increased boostrap confidence, an increase in edge specificity is obtained as well.

Consideration has to be given to the power of our analysis in addition to the false discovery rate. The sensitivity to detect primary shielders, where primary shielders are the shielders with direct connection to the QTL, is shown in Figure 3.5. Shielder sensitivity is defined as the percentage of true shielders that are detected.

Shielder sensitivity is seen to be quite high for most scenarios without boostrap support, but it decreases quite rapidly with increased boostrap confidence. Edge sensitivity (percentage of true edges that are detected) is also shown in Figure 3.6. Edge sensitivity is

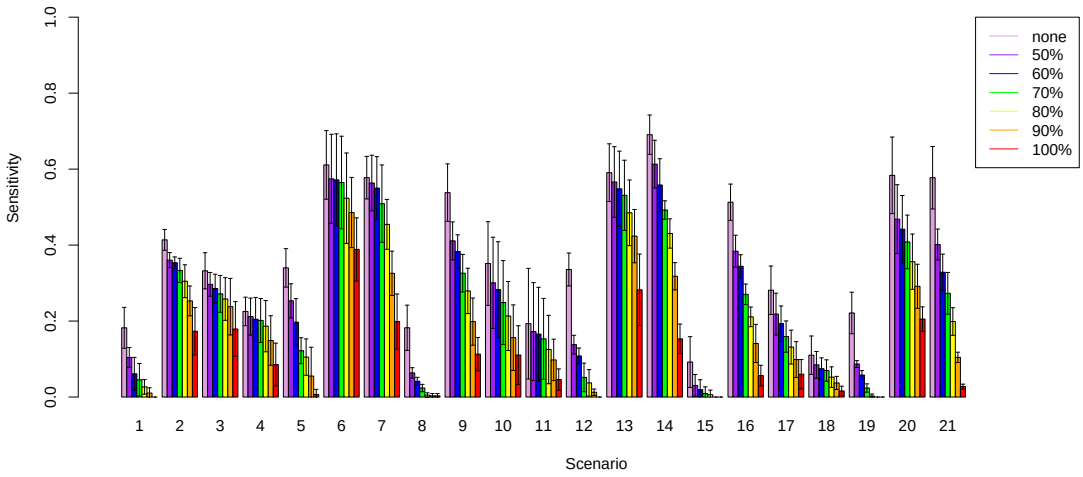

Figure 3.6: Edge Sensitivity at different levels of bootstrap support in simulated data sets. We applied our network discovery algorithm to 21 different network scenarios characterized by different numbers of genes with direct connection to the QTL, different scale-free parameters, and different numbers of edges (see Table 3.2). 4 networks and 10 data sets per network were simulated. Results are averaged over all simulated networks; error bars are approximate 95% confidence intervals estimated from the standard deviation among network means.

seen to be noticably lower than shielder sensitivity, which is expected given that our method is designed to detect shielders and not individual edges with high power. Edge sensitivity is also seen to decrease rapidly with increase in bootstrap confidence.

It is important to make an appropriate tradeoff between false positives and false negatives (power). In order to obtain a high degree of confidence in discovered shielders and edges as well as maintain moderate power, we have chosen to use 80% bootstrap support. The statistics for sensitivity and specificity for 80% boostrap confidence are shown in Table 3.3.

For 80% boostrap confidence, the average shielder and edge specificities are 95% and 84%, respectively. The minimum shielder specificity is 61%, and all but two scenarios have shielder specificities greater than 90%. The average shielder and edge sensitivities are

Table 3.3: Sensitivity and specificity for shielders and edges at a 80% bootstrap confidence level.

	Network Params			Specificity		Sensitivity	
scenario	n primary	gamma	ratio	p shielder	edge	p shielder	edge
1	1	2	0.65	100%	53%	13%	3%
2	1	2	0.9	100%	75%	98%	30%
3	1	2	1.5	93%	78%	98%	26%
4	1	2	2	100%	86%	83%	19%
5	1	2.5	0.65	100%	85%	30%	11%
6	1	2.5	0.9	100%	95%	93%	52%
7	1	3	0.65	100%	93%	88%	45%
8	3	2	0.65	100%	83%	2%	1%
9	3	2	0.9	100%	84%	66%	28%
10	3	2	1.5	95%	81%	58%	21%
11	3	2	2	72%	85%	37%	13%
12	3	2.5	0.65	100%	78%	8%	4%
13	3	2.5	0.9	100%	91%	82%	48%
14	3	3	0.65	100%	93%	62%	43%
15	5	2	0.65	100%	72%	2%	1%
16	5	2	0.9	99%	89%	36%	21%
17	5	2	1.5	91%	80%	36%	13%
18	5	2	2	61%	71%	24%	5%
19	5	2.5	0.65	100%	100%	1%	0%
20	5	2.5	0.9	90%	91%	50%	36%
21	5	3	0.65	100%	96%	20%	20%
average				95%	84%	47%	21%

Table 3.4: Dependence of network reconstruction performance on network parameters

stat	regressor	effect	p value	r squared
shielder spec	ratio	-0.26	$\leq$0.0001	0.81
edge spec	gamma	0.2	$\leq$0.0001	0.58
p shielder sens	n primary	-0.1	0.0072	0.32
edge sens	gamma	0.3	0.0048	0.35

47% and 21%, respectively.

To see how the network reconstruction performance depends on network parameters, a regression model was fit to each of the four statistics measured (shielder and edge specificity, and primary shielder and edge sensitivity) as a function of the simulation parameters including the scale-free parameter (γ), the number of primary shielders (n primary), and the ratio of network edges to network nodes (ratio). Backward selection was used to find the best model for each statistic, and each was found to significantly depend on only one parameter for 50% bootstrap confidence. The results are shown in Table 3.4 (the results for 80% boostrap confidence were similar).

Shielder specificity was seen to decrease with the ratio of edges to genes, a parameter that heavily influences network size ($\rho = 0.94$, $p \leq 0.0001$), where network size is defined as the number of genes significantly linked to the QTL. Edge specificity is seen to increase with γ, which means that networks that are more "scale-free" have higher edge specificity. Primary shielder sensitivity is seen to decrease with number of primary shielders, and edge sensitivity is seen to increase with γ. In summary, our algorithm performs best with few primary shielders, smaller networks, and networks that are more scale-free.

3.5 Conclusions

In conclusion, our network reconstruction algorithm is found to have good performance using a significance level of 0.05 for testing the null hypothesis of a gene not shielding any genes, will little gained by decreasing α further. The shielder boostrap confidence level depends on the desired balance between false positives and false negatives, but our recommendation is to use 80% bootstrap support for high confidence in discovered

shielders.

We maintain high average shielder specificity (95%) for 80% bootstrap confidence while achieving moderate shielder and edge sensitivity (47% and 21%, respectively). Our algorithm is found to perform best with fewer primary shielders, more scale-free networks, and smaller network sizes, although acceptable performance was obtained for all combinations of network parameters simulated. Considering the small sample size (100) and the large number of network genes (as many as 500), our algorithm succeeds in finding a high percentage of the key network regulators (47% on average) with high confidence (95% specificity on average).

Chapter 4

Genetic Network Reconstruction in a Yeast Expression QTL Data Set

4.1 Introduction

The algorithm described in Section 2.3 was applied to discover networks for eleven of the largest eQTL hotspots in the yeast data set (see Figure 3.2). We used an α of 0.05 both for the ordinary conditional independence tests and for the number of shielded genes test for declaring a shielder (see Section 2.5). We used a maximum order of one for conditional independence testing just as we did in the simulations to give maximum power and to reduce the multiple testing problem. We used a bootstrap confidence level of 80% because we found this to give a good tradeoff between shielder confidence level and power (see 3.4). Three of the discovered networks, listed in Tables 4.1, 4.2, and 4.3, will be analyzed in more detail. The remaining networks are listed at the end of this chapter in Tables 4.4, 4.5, and 4.6. The networks are described by a list of shielders or modulating genes and the set of genes each shielder modulates.

Table 4.1: Discovered Network 1: eQTL at Chromosome 3 at 79,091 bp.

shielder	shielded
LEU4	ARG2 ILV6 MET10 MET13 MRPL50 RTG3 SDT1 SER1 TEA1 YAT2 YDR531W
BAT1	AAT2 CTI6 IDP1 ILV3 LEU4 MAE1 SER33 TRP3 YBR012C YBT1 YJR111C YOR271C

Table 4.2: Discovered Network 3: eQTL at Chromosome 8 at 98,513 bp.

shielder	shielded
FUS1	AFR1 AGA1 ASG7 BUD14 CHS1 EST3 FUS3 GFA1 HAL1 HYM1 INP52 KAR5 KNH1 MCM3 NIS1 NTA1 PDS1 PRM4 PRM5 RGD2 SHU1 SMY1 SNL1 SPP1 SST2 STE14 TEC1 UME6 YBP2 YDR124W YDR249C YDR282C YFR026C YIL080W YJR039W YJR054W YOR343C

Table 4.3: Discovered Network 6: eQTL at Chromosome 15 at 572,410 bp.

shielder	shielded
GCR1	YDR170W-A VPS8 ATP17 QCR8 COX12 ENO1 PGK1 YLR224W MAD2 MIR1 IDH1 APQ12 COX9

4.2 Cis versus Trans Regulation

Much attention has been given to whether eQTL regulate gene expression in cis or in trans. Cis regulation occurs if the expression of a gene links to a QTL that is at the same position as that of the linked gene. This is cis regulation because presumably a polymorphism in the gene causes variation in that gene's expression. Trans regulation occurs when a gene is linked to a QTL that does not co-localize with the gene. Both cis- and trans-regulating eQTL have been discovered, but eQTL hotspots which we are concerned with here tend to involve large numbers of trans-regulated genes.

Even though we expect the eQTL regulation to be primarily in trans in the eQTL hotspots, it is possible that a single cis-regulated gene could be the causal gene for the other genes linked to a particular locus. This is the assumption of many authors with regard to searching for causal genes in eQTL data sets ([33], [6], and [27]). However, it was found that none of the shielders significant at the 80% bootstrap level co-localized with the eQTL. Thus, a different mechanism must account for the regulation of shielded genes by these shielders.

One important note is that in order for a cis-regulated shielder gene to cause variation in a set of shielded genes, its own expression must be regulated by the eQTL (or presumably the polymorphism in that shielder gene). The change in the shielder gene's expression in turn causes a change in the expression of the shielded genes. However, other mechanisms are possible. One example is that the polymorphism in the gene co-localized with the eQTL does not alter that gene's expression but rather changes its function. An example of this mechanism is given in [59] for a known polymorphism in GPA1 that causes it to be constitutively active.

If the function but not the expression of a gene co-localized with the eQTL is effected, then the co-localized gene may not even link to that eQTL. In this case, the shielder gene would be a gene that responds to the change in functionality of the causal gene and subsequently influences the expression of the other genes linked to the eQTL. In this case the shielder gene would not co-localize with the eQTL.

In the mechanism just described, a regulatory gene modulates the expression of other genes in response to a functionally-altered gene co-localized with the eQTL. This

mechanism may account for the regulatory action of the shielder genes discovered using our network reconstruction algorithm.

4.3 Bioinformatic Analysis of Networks

We have performed bioinformatic analysis on three networks: network 1 (eQTL on chromosome 3), network 3 (eQTL on chromosome 8), and network 6 (eQTL on chromosome 15) (see Tables 4.1, 4.2, and 4.3).

4.3.1 Network 1

In network 1 (eQTL on Chromosome 3 at 79,091 bp), the primary shielder was BAT1 which was found to shield 23 genes, and the other shielder was LEU4 which was found to shield 11 genes. There is a known loss-of-function mutation in LEU2 in one of the strains. LEU2 is an enzyme in the leucine biosynthesis pathway, as are LEU4 and BAT1, and they are all activated by the transcription factor LEU3. LEU3 has been shown to be activated by α-isopropylmalate, the product of the LEU4 enzymatic reaction. With loss of function of LEU2, there would be a build up of α-isopropylmalate in the susceptible strain which would cause activation of the LEU3 transcription factor.

Many of the genes linked to this locus are targets of the GCN4 transcription factor (39.3%) and the LEU3 transcription factor (21.4%). GCN4 is a transcriptional activator of amino acid biosynthetic genes, and LEU3 is a transcription factor that regulates genes involved in branched chain amino acid biosynthesis and ammonia assimilation. Even a larger percentage of genes in the discovered network are targets of these two transcription factors (79.2% for GCN4, 45.8% for LEU3).

One hypothesis is that a loss of function mutation in Leu2 causes activation of the GCN4 and LEU3 transcription factors in the susceptible strain, and that LEU4 and BAT1, key genes in the Leucine biosynthesis pathway, modulate the response to these transcription factors. One possible mechanism for the modulation by LEU4 and BAT1 is the following. Since the product of LEU4 is α-isopropylmalate which activates LEU3, LEU4 could influence the activity of LEU3 in a feedback loop. Specifically, if LEU4 is highly

expressed, it will generate lots of α-isopropylmalate, which will activate LEU3, which will cause up-regulation of LEU4 (and other LEU3 targets), which will in turn cause higher concentrations of α-isopropylmalate, and so on and so forth. BAT1 is one step upstream of LEU4 in this pathway and thus its expression level may modulate the levels of LEU4.

4.3.2 Network 3

In network 3 (eQTL on chromosome 8 at 98,513 bp), the primary shielder was found to be FUS1 which was found to shield 37 genes. There is a known mutation in the GPA1 gene which co-localizes with this eQTL (GPA1 is at 114,911 bp) in one of the yeast strains in the cross for this experiment. GPA1 is the alpha subunit of the G protein coupled to mating factor receptor and is involved in the mating pheromone signal transduction pathway. Previous authors [59] have hypothesisized that the mutation in GPA1 is a loss of function mutation that causes it to be constitutively active, which explains the differential expression of genes downstream of the pheromone signal transduction pathway in this data set.

In particular, 40.5% of linked genes and 70.3% of genes shielded by FUS1 are targets of the STE12 transcription factor. STE12 is a transcription factor that is activated by a MAP kinase signaling cascade which activates genes involved in mating or pseudohyphal/invasive growth. In Figure 4.1, the MAPK signaling pathway is shown (from KEGG). It is seen that FUS1 is the first downstream target of STE12. FUS1 has been proposed to coordinate signaling, fusion, and polarization events required for fusion. Itself a target of Ste12, it may modulate the transcriptional effects of Ste12 on other targets.

4.3.3 Network 6

In network 6 (eQTL on Chromosome 15 at 572,410 bp), the primary shielder was GCR1 which was found to shield 13 genes. GCR1 is described as a transcriptional activator of genes involved in glycolysis, and 57.1% of genes in this network are targets of GCR1 according to Yeastract, including GCR1 itself. In addition, the genes in this network show enrichment for the GO biological processes "generation of precursor metabolites and energy" ($p < 2.7^{-4}$) and "death" and "cell death" ($p < 2^{-3}$). It makes sense that a transcription

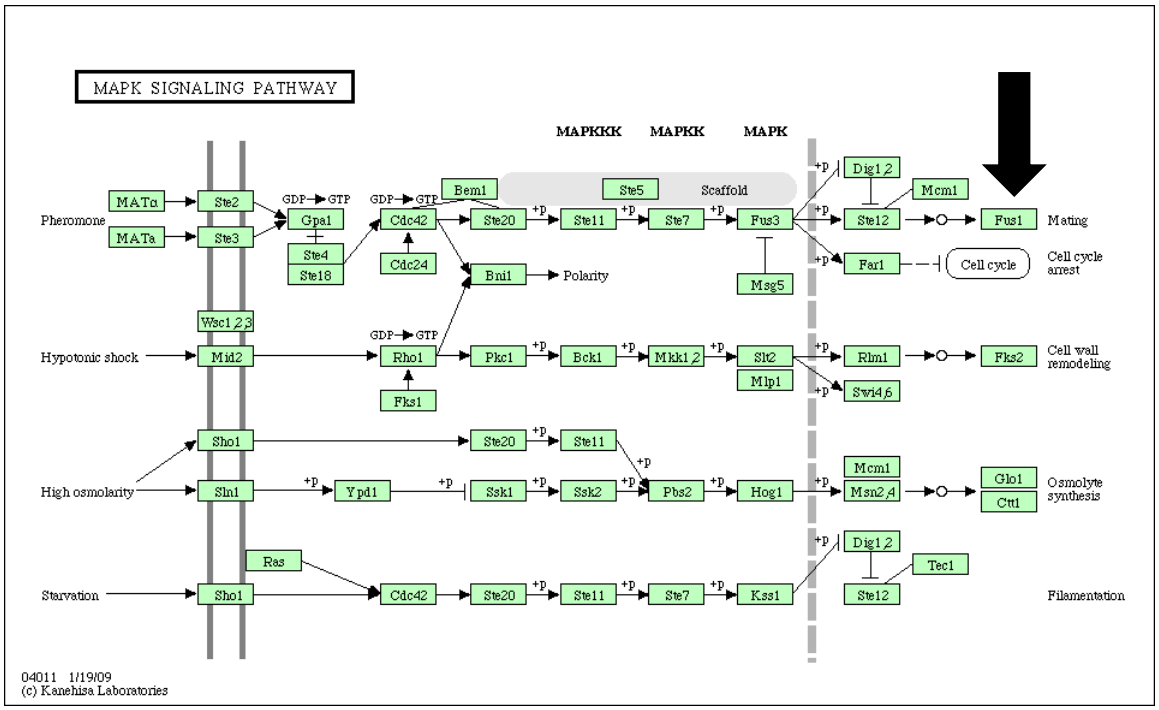

Figure 4.1: MAPK Signaling Pathway from the KEGG database.

Table 4.4: Discovered Network 2: eQTL at Chromosome 5 at 395,442 bp.

shielder	shielded
ECM34	CDC5 FUS3 GLE2 MNN10 MUC1
	NCB2 NMA1 PCK1 POL32 PSF1
	RIM101 SHO1 SPC42 SPT15 VPS65
	YER158C YLR108C YLR118C YNL217W
	YNR029C

factor would be a shielder since transcription factors modulate gene expression in general, and the annotation of the shielded genes as transcription factor targets of the discovered shielder is consistent with our findings.

4.4 Discussion

Networks have been reconstructed for 6 of the largest hotspots in a yeast eQTL data set. The regulation of shielder gene expression has been found to be in trans, which is in contrast to the assumption of cis-regulation used by other researchers who have created methods for discovering networks from eQTL data sets ([33], [6], [27], and [41]). Bioinformatic analysis of three networks generated different hypotheses for mechanisms of regulation of the shielded genes by the primary shielders. In networks 1 and 2 (for eQTL on chromosomes 3 and 8, respectively), it was hypothesized that the shielders modulated the effect of transcription factors of which they were themselves targets. In network 6 (for eQTL on chromosome 1), the shielder was iteself a transcription factor of which the shielded genes and itself were targets. Overall our method has created a list of potentially important regulatory genes in various yeast biological processes, and further bioinformatic analysis or laboratory experiments could lead to the generation and testing of many important hypthotheses.

Table 4.5: Discovered Network 4: eQTL at Chromosome 12 at 674,651 bp.

shielder	shielded
UBX6	CYT1 DLD1 GAS4 GLO3 GYP6
	HEM13 IZH2 LAC1 NDE1 PRE1
	PUT4 RPT5 UPC2 YGL160W YJL049W
	YLR280C YOR175C

Table 4.6: Discovered Network 5: eQTL at Chromosome 14 at 449,639 bp.

shielder	shielded
YFR044C	AIR2 BYE1 CLG1 DPP1 ENT5
	ERG4 ERP6 ESC8 FAS1 GAL11
	HMG2 HOM6 ISR1 LAS21 MCX1
	MRS2 PIS1 QRI7 RBK1 SEC11
	TDH3 TFP3 TSA1 VAC8 VPS73
	YCP4 YDL173W YFR006W YGR263C
	YHC3 YHL017W YHR048W YHR113W
	YIP5 YJL086C YNL086W YPL105C

Chapter 5

Discussion

5.1 Summary and Evaluation of Research

We have presented a new method for the discovery of genetic networks from expression QTL data sets. Collection of marker and microarray data for various organisms to perform expression QTL analysis has become quite popular in recent years. While searching for genetic loci that underly expression differences in specific genes is an important first step in understanding the genetics of gene expression, it is desirable to have analytical tools to allow researchers to go one step farther and actually predict the structure of genetic interactions entailed by eQTL data sets.

The discovery of genetic networks from large genomic data sets is fraught with challenges, many of which are listed in Chapter 1. The "large P small N" problem is compounded in genetic network discovery because in principle an edge can exist between any pair of variables, so P becomes P choose 2. Another important and frequently overlooked issue is the assessment of statistical confidence in discovered networks. While standard statistical methods exist for assessing selected models and estimates of parameters in regression and ANOVA models, predicting estimates of confidence in discovered networks is more complicated and has not been solved analytically to our knowledge.

In order to deal with the issues of power and computational complexity intrinsic in expression QTL data sets with their large number of variables and small sample sizes,

we have chosen to search for "framework" networks that consist of some of the major regulatory genes in the network along with some of their targets. Such a skeleton network can be discovered more easily and with a greater degree of confidence than the full detailed network, as we believe we have shown in Chapters 3 and 4. However, to apply our method in full generality involves assumptions that are not likely to be true in real data sets, and thus simulation and resampling techniques are used to assess the confidence levels of discovered network features. It would be of great benefit to find a more principled approach that could be generally used to assess confidence in discovered networks.

5.2 Specific Areas for Improvement

While our method represents an important first step in detecting high confidence genetic networks from expression QTL data, there are many possible areas of improvement that will make the method more powerful and more robust. First, the power of our method is not particularly high due to the conservative Bonferroni-like correction for false positive shielders. It may be possible to improve the power by adapting existing methods for estimating false discovery rate to our method. This can be a particularly difficult issue due to the existence of multiple correlated and nested hypothesis tests in our method. One possibility is to look into methods proposed by Westfall and Young that have closure under nested hypothesis testing.

Another way to improve power is to find a better balance between false positives and false negatives. In this work we have been particularly concerned with controlling the shielder false positive rate and thus have emphasized minimizing false positives, but perhaps a more balanced approach that looks jointly at false positives and false negatives would give higher power with not too much of an increase in shielder false positive rate. Using ROC (Receive Operator Characteristic) curves or the Youd index to assess that balance are two possibilities.

Another potential source of false negatives in our method exists if there are duplicate genes on the microarray, which has been known to occur. The two copies of a true shielder gene would "cancel each other out" in our method. That is, each would be found

to shield the other, and neither would be found to be a shielder. One way to correct this problem is to do a search for collinearity in the microarray genes, for instance using K-means clustering.

5.3 Future Research Directions

There are many potential research directions to take for improving our method and for better understanding genetic network discovery in general. One direction involves making our method more flexible and allowing it to include a broader array of data types. It would be desirable to include multiple QTL in a single network model to account for polygenic control of transcription. In addition, we would like to include other soures of data in the model besides genotypic and gene expression data. For example, by including clinical data in a genetic network model, we can potentially "close the loop" by discovering how genetic polymorphims influence gene expression and how gene expression in turn influences clinical phenotype. One example of building this type of three-level network model is given in [46]. It would also be of interest to extend this multi-level network modeling to include proteomic or metabolomic data; thus multiple levels of genetic regulation can be jointly estimated. In addition, network building directly from genotype to clinical trait may be possible as is shown in [49]. It would also be beneficial to make our method available for the analysis of natural populations. Possible applications include linkage or association studies in human populations.

Another direction to take in research on genetic network discovery is basic statistical research on the problem of network inference from a data set. Network discovery is essentially a model selection problem and as such can borrow from the literature on model selection. Issues of power and robustness are of particular interest. For instance, it would be advantageous to be able to determine sample size requirements for networks of different size, architecture, and levels of complexity. Ultimately a tool that could be developed to determine what size and types of networks can be confidently inferred given a specific data set would be very useful.

Another important area of improvement is to address the scalability of the method

to larger data sets. The size of genomic data sets is increasing quickly over time, and it is important that our method be able to handle larger data sets. Investigations into parallel processing or other ways to speed up the algorithm would be useful.

One additional future research direction in genetic network discovery is in the application of our method to additional real data sets. This analysis would benefit from collaborations with subject-matter experts in order to analyze predicted networks to see if biological mechanisms can be proposed, and possibly to design follow-up experiments to test generated hypotheses.

In summary, the area of genetic network discovery from expression QTL data sets and from genomic or other -omic data sets in general holds great promise for allowing researchers to learn more detailed genetic and molecular mechanisms for biological processes of interest. However, this research area is complicated with many methodological challenges. Basic research into statistical inference of network structure as well as generalization of existing network models are both key areas for advancement. In addition testing and exploring discovered networks with subject-matter experts is of great importance too.

Bibliography

[1] Margolin AA and Califano A. Theory and limitations of genetic network inference from microarray data. *Ann N Y Acad Sci.*, 2007.

[2] Cervino AC, Li G, Edwards S, Zhu J, Laurie C, Tokiwa G, Lum PY, Wang S, Castellani LW, Lusis AJ, Carlson S, Sachs AB, and Schadt EE. Integrating qtl and high-density snp analyses in mice to identify insig2 as a susceptibility gene for plasma cholesterol levels. *Genomics*, 86(5):505–517, 2005.

[3] Tomohiro Ando, Seiya Imoto1, and Satoru Miyano. Functional Data Analysis of the Dynamics of Gene Regulatory Networks, 2004.

[4] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[5] Matthew J. Beal, Francesco Falciani, Zoubin Ghahramani, Claudia Rangel, and David L. Wild. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3):349–356, 2005.

[6] N Bing and I Hoeschele. Genetical genomics analysis of a yeast segregant population for transcription network inference. *GENETICS*, 170(2):533–542, JUN 2005.

[7] Rachel B. Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1572–1577, 2005.

[8] RB Brem, G Yvert, R Clinton, and L Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *SCIENCE*, 296(5568):752–755, APR 26 2002.

[9] Karl W Broman. Mapping expression in randomized rodent genomes. *Nature Genetics*, 37:209–210, 2005.

[10] O. Carlborg, D. J. De Koning, K. F. Manly, E. Chesler, R. W. Williams, and C. S. Haley. Methodological aspects of the genetic dissection of gene expression. *Bioinformatics*, 21(10):2383–2393, 2005.

[11] Guanrao Chen, Peter Larsen, Eyad Almasri, and Yang Dai. Rank-based edge reconstruction for scale-free genetic regulatory networks. *BMC Bioinformatics*, 9(1):75, 2008.

[12] Xue-wen Chen, Gopalakrishna Anantha, and Xinkun Wang. An effective structure learning method for constructing gene networks. *Bioinformatics*, 22(11):1367–1374, 2006.

[13] Elissa J Chesler, Lu Lu, Siming Shou, Yanhua Qu, Jing Gu, Jintao Wang, Hui Chen Hsu, John D Mountz, Nicole E Baldwin, Michael A Langston, David W Threadgill, Kenneth F Manly, and Robert W Williams. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, 37:233–242, 2005.

[14] Vivian G. Cheung, Laura K. Conlin, Teresa M. Weber, Melissa Arcaro, Kuang-Yu Jen, Michael Morley, and Richard S. Spielman. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics*, 33:422–425, 2003.

[15] D Chickering. Learning Bayesian networks is NP-complete, 1996.

[16] Barbara Di Camillo, Fatima Sanchez-Cabo, Gianna Toffolo, Sreekumaran Nair, Zlatko Trajanoski, and Claudio Cobelli. A quantization method based on threshold optimization for microarray short time series. *BMC Bioinformatics*, 6(Suppl 4):S11, 2005.

[17] Norbert Dojer, Anna Gambin, Andrzej Mizera, Bartek Wilczynski, and Jerzy Tiuryn. Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics*, 7:249–259, 2006.

[18] Sudheer Doss, Eric E. Schadt, Thomas A. Drake, and Aldons J. Lusis. Cis-acting expression quantitative trait loci in mice. *Genome Research*, 15(5):681–691, 2005.

[19] Gertrud Fischer, Saleh Ibrahim, Gudrun Brockmann, Jens Pahnke, Ezio Bartocci, Hans-Jurgen Thiesen, Pablo Serrano-Fernandez, and Steffen Moller. Expressionview: visualization of quantitative trait loci and gene-expression data in ensembl. *Genome Biology*, 4(11):R77, 2003.

[20] N Friedman, M Linial, I Nachman, and D Pe'er. Using Bayesian networks to analyze expression data. *JOURNAL OF COMPUTATIONAL BIOLOGY*, 7(3-4):601–620, 2000.

[21] Giacomo Gamberoni, Evelina Lamma, Fabrizio Riguzzi1, Sergio Storari, and Stefano Volinia. Bayesian Networks Learning for Gene Expression Datasets, 2005.

[22] Elisabeth Georgii, Lothar Richter, Ulrich Ruckert, and Stefan Kramer. Analyzing microarray data using quantitative association rules. *Bioinformatics*, 21(suppl2):ii123–129, 2005.

[23] Norbert Hubner, Caroline A Wallace, Heike Zimdahl, Enrico Petretto, Herbert Schulz, Fiona Maciver, Michael Mueller, Oliver Hummel, Jan Monti, Vaclav Zidek, Alena Musilova, Vladimir Kren, Helen Causton, Laurence Game, Gabriele Born, Sabine Schmidt, Anita Mller, Stuart A Cook, Theodore W Kurtz, John Whittaker, Michal Pravenec, and Timothy J Aitman. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, 37:243–253, 2005.

[24] Dirk Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19(17):2271–2282, 2003.

[25] Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB, and Schadt EE. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res*, 105:363–374, 2004.

[26] Chen-Hung Kao, Zhao-Bang Zeng, and Robert D. Teasdale. Multiple Interval Mapping for Quantitative Trait Loci. *Genetics*, 152(3):1203–1216, 1999.

[27] Joost J. B. Keurentjes, Jingyuan Fu, Inez R. Terpstra, Juan M. Garcia, Guido van den Ackerveken, L. Basten Snoek, Anton J. M. Peeters, Dick Vreugdenhil, Maarten Koornneef, and Ritsert C. Jansen. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 104(5):1708–1713, JAN 30 2007.

[28] Chang Kim. Bayesian orthogonal least squares (bols) algorithm for reverse engineering of gene regulatory networks. *BMC Bioinformatics*, 8(1):251, 2007.

[29] Matias Kirst, Christopher J. Basten, Alexander A. Myburg, Zhao-Bang Zeng, and Ronald R. Sederoff. Genetic Architecture of Transcript-Level Variation in Differentiating Xylem of a Eucalyptus Hybrid. *Genetics*, 169(4):2295–2303, 2005.

[30] Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, and de Haan G. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nature Genetics*, 37(3):209–210, 2005.

[31] Hong Lan, Jonathan P. Stoehr, Samuel T. Nadler, Kathryn L. Schueler, Brian S. Yandell, and Alan D. Attie. Dimension Reduction for Mapping mRNA Abundance as Quantitative Traits. *Genetics*, 164(4):1607–1614, 2003.

[32] Timothy R. Lezon, Jayanth R. Banavar, Marek Cieplak, Amos Maritan, and Nina V. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc Natl Acad Sci U S A*, 103(50):19033–19038, 2006.

[33] HQ Li, L Lu, KF Manly, EJ Chesler, L Bao, JT Wang, M Zhou, RW Williams, and Y Cui. Inferring gene transcriptional modulatory relations: a genetical genomics approach. *HUMAN MOLECULAR GENETICS*, 14(9):1119–1125, MAY 1 2005.

[34] Lei M. Li and Henry Horng-Shing Lu. Explore biological pathways from noisy array data by directed acyclic Boolean networks. *Journal of Computational Biology*, 12(2):170–185, 2005.

[35] S.P. Li, J.J. Tseng, and S.C. Wang. Reconstructing gene regulatory networks from time-series microarray data. *Physica A*, 350(1):63–69, 2005.

[36] Kenneth Manly, Jintao Wang, and Robert Williams. Weighting by heritability for detection of quantitative trait loci with microarray estimates of gene expression. *Genome Biology*, 6(3):R27, 2005.

[37] A Margolin, I Nemenman, K Basso, U Klein, C Wiggins, G Stolovitzky, Riccardo D Favera, and A Califano. ARACNE: An algorithm for reconstruction of genetic networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl.1):S1–S7, 2006.

[38] Shawn Martin, Zhaoduo Zhang, Anthony Martino, and Jean-Loup Faulon. Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*, 23(7):866–874, 2007.

[39] Michael Morley, Cliona M. Molony, Teresa M. Weber, James L. Devlin, Kathryn G. Ewens, Richard S. Spielman, and Vivian G. Cheung. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430:743–747, 2004.

[40] Naoki Nariai, Yoshinori Tamada, Seiya Imoto, and Satoru Miyano. Estimating gene regulatory networks and protein-protein interactions of Saccharomyces cerevisiae from multiple genome-wide data. *Bioinformatics*, 21(suppl2):ii206–212, 2005.

[41] Elias Chaibub Neto, Christine T. Ferrara, Alan D. Attie, and Brian S. Yandell. Inferring causal phenotype networks from segregating populations. *GENETICS*, 179(2):1089–1100, JUN 2008.

[42] Miguel Perez-Enciso. In Silico Study of Transcriptome Genetic Variation in Outbred Populations. *Genetics*, 166(1):547–554, 2004.

[43] Miguel Perez-Enciso, Miguel A. Toro, Michel Tenenhaus, and Daniel Gianola. Combining Gene Expression and Molecular Marker Information for Mapping Complex Trait Genes: A Simulation Study. *Genetics*, 164(4):1597–1606, 2003.

[44] Claudia Rangel, John Angus, Zoubin Ghahramani, Maria Lioumi, Elizabeth Sotheran, Alessia Gaiba, David L. Wild, and Francesco Falciani. Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372, 2004.

[45] Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, and Schadt EE. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet.*, 75(6):1094–1105, 2004.

[46] EE Schadt, J Lamb, X Yang, J Zhu, S Edwards, D GuhaThakurta, SK Sieberts, S Monks, M Reitman, CS Zhang, PY Lum, A Leonardson, R Thieringer, JM Metzger, LM Yang, J Castle, HY Zhu, SF Kash, TA Drake, A Sachs, and AJ Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *NATURE GENETICS*, 37(7):710–717, JUL 2005.

[47] Eric E. Schadt, Stephanie A. Monks, Thomas A. Drake, Aldons J. Lusis, Nam Che, Veronica Colinayo, Thomas G. Ruff, Stephen B. Milligan, John R. Lamb, Guy Cavet, Peter S. Linsley, Mao Mao, Roland B. Stoughton, and Stephen H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, pages 297–302, 2003.

[48] Juliane Schafer and Korbinian Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[49] Paola Sebastiani, Marco F Ramoni, Vikki Nolan, Clinton T Baldwin, and Martin H Steinberg. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *NATURE GENETICS*, 37:435–440, 2005.

[50] Grace S Shieh, Chung-Ming Chen, Ching-Yun Yu, Juiling Huang, Woei-Fuh Wang, and Yi-Chen Lo. Inferring transcriptional compensation interactions in yeast via stepwise structure equation modeling. *BMC Bioinformatics*, 9:134, 2008.

[51] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Biol. Cell*, 9(12):3273–3297, 1998.

[52] Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, Prediction, and Search, 2000.

[53] John D Storey, Joshua M Akey, and Leonid Kruglyak. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol*, 3(8):e267, 07 2005.

[54] Reuben Thomas, Sanjay Mehrotra, Eleftherios T. Papoutsakis, and Vassily Hatzimanikatis. A model-based optimization framework for the inference on gene regulatory networks from DNA array data. *Bioinformatics*, 20(17):3221–3235, 2004.

[55] R. J. P. van Berlo, E. P. van Someren, and M. J. T. Reinders. Studying the Conditions for Learning Dynamic Bayesian Networks to Discover Genetic Regulatory Networks. *SIMULATION*, 79(12):689–702, 2003.

[56] Mingyi Wang, Zuozhou Chen, and Sylvie Cloutier. A hybrid bayesian network learning method for constructing gene networks. *Comput. Biol. Chem.*, 31(5-6):361–372, 2007.

[57] Xintao Wu and Yong Ye. Exploring gene causal interactions using an enhanced constraint-based method. *Pattern Recognition*, 39(12):2439–2449, 2006.

[58] Changwon Yoo, Gregory F. Cooper, and Martin Schmidt. A control study to evaluate a computer-based microarray experiment design recommendation system for gene-regulation pathways discovery. *J. of Biomedical Informatics*, 39(2):126–146, 2006.

[59] G Yvert, RB Brem, J Whittle, JM Akey, E Foss, EN Smith, R Mackelprang, and L Kruglyak. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *NATURE GENETICS*, 35(1):57–64, SEP 2003.

[60] Wei Zou, David L Aylor, and Zhao-Bang Zeng. eqtl viewer: visualizing how sequence variation affects genome-wide transcription. *BMC Bioinformatics*, 8(7), 2007.

[61] Wei Zou and Zhao-Bang Zeng. Multiple interval mapping for gene expression QTL analysis. *Genetica*, 2008.