DISCOVERY OF COMPLEX REGULATORY MODULES FROM EXPRESSION GENETICS DATA

A Dissertation Presented

by

MANJUNATHA N. JAGALUR

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2010

Computer Science

DISCOVERY OF COMPLEX REGULATORY MODULES FROM EXPRESSION GENETICS DATA

A Dissertation Presented

by

MANJUNATHA N. JAGALUR

Approved as to style and content by:

David C. Kulp, Chair

Gary A. Churchill, Member

Erik G. Learned-Miller, Member

David D. Jensen, Member

Ramgopal R. Mettu, Member

Andrew G. Barto, Department Chair
Computer Science

To my parents, Sumangala and Nagendrappa, and my brother, Rajesh.

ACKNOWLEDGMENTS

First and foremost, I am profoundly grateful to David Kulp who has patiently advised me through years. His guidance and support were key to conceptualizing and completing this thesis.

I extend my deepest gratitude to Gary Churchill for being a great source of inspiration. I consider myself fortunate to have had the opportunity to work with him. I found my stay at his lab intellectually stimulating and motivating.

I am very thankful to my committee members: Erik Learned-Miller, David Jensen and Ramgopal Mettu for their support. Their timely and useful feedback were immensely helpful in creating this document. I extend special thanks to Rob Williams for being great mentor. I am also grateful to Olver Brock and Chris Pal for interesting research collaborations.

My friends and collegues at the Computational Biology Lab and Center for Genome Dynamics were always helpful and made my graduate student life interesting. I thank Tj Brunnette, Vidit Jain, Yimin Wu, Sharon Tsaih, Rachael Hageman and Hyuna Yang for their support.

ABSTRACT

DISCOVERY OF COMPLEX REGULATORY MODULES FROM EXPRESSION GENETICS DATA

MAY 2010

MANJUNATHA N. JAGALUR

B.E., UNIVERSITY OF MYSORE

M.E., INDIAN INSTITUTE OF SCIENCE

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor David C. Kulp

Mapping of strongly inherited classical traits have been immensely helpful in understanding many important traits including diseases, yield and immunity. But some of these traits are too complex and are difficult to map. Taking into consideration gene expression, which mediates the genetic effects, can be helpful in understanding such traits. Together with genetic variation data such data-set is collectively known as *expression genetics data*. Presence of discrete and continuous variables, observed and latent variables, availability of partial causal information, and under-specified nature of the data make expression genetics data computationally challenging, but potentially of great biological importance.

In this dissertation the underlying regulatory processes are modeled as Bayesian networks consisting of gene expression and genetic variation nodes. Due to the under-

specified nature of the data, inferring the complete regulatory network is impractical. Instead, the following techniques are proposed to extract interesting subnetworks with high confidence.

The *network motif searching* technique is used to recover instances of a known regulatory mechanism. The *local network inference* technique is used to identify immediate neighbors of a given transcript. Application of these two techniques often results in identification of hundreds of individual networks. The *network aggregation* technique extracts the most common subnetwork from those networks, and identifies its immediate neighbors by collapsing them into a common network.

In all the above tasks, simulation studies were carried out to estimate the robustness of the proposed methods and the results suggest that these techniques are capable of recovering the correct substructure with high precision and moderate recall. Moreover, manual biological review shows that the recovered regulatory network substructures are typically biologically sensible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

Genetical inheritance is influential in determining many important traits of an organism including physical attributes, behavior and disease immunity. Some of these traits such as sickle cell anaemia are determined by a single polymorphism (genetic variation) and, are called Mendelian traits. Other traits such as eye color are polygenic (determined by polymorphisms in multiple genes). However, in both of these cases the trait is solely determined by the polymorphisms. Quantitative traits such as adeposity are more complex and the trait is determined by the complex interactions between the polymorphisms, gene expression and the environment. In this dissertation methods are presented to infer such complex interactions using gene expression data and genetic variation data which are collectively known as expression genetics data.

1.1 Gene Regulation and Expression Genetics Data

A quintessential regulatory mechanism consists of a gene whose protein product influences the expression of a given trait (Figure 1.1). The variation in the trait is typically caused by two factors: the amount of protein that is transcribed, and the three dimensional structure of the protein. The main causes of variation of these two features are: environment, regulatory actions of other genes, and genetic inheritance.

An expression genetics experimental cross data set consists of genome-wide gene expression profiles and genetic variation data collected from a set of specially bred

strains. As these organisms are raised under identical conditions, the environmental source of variation is theoretically eliminated.

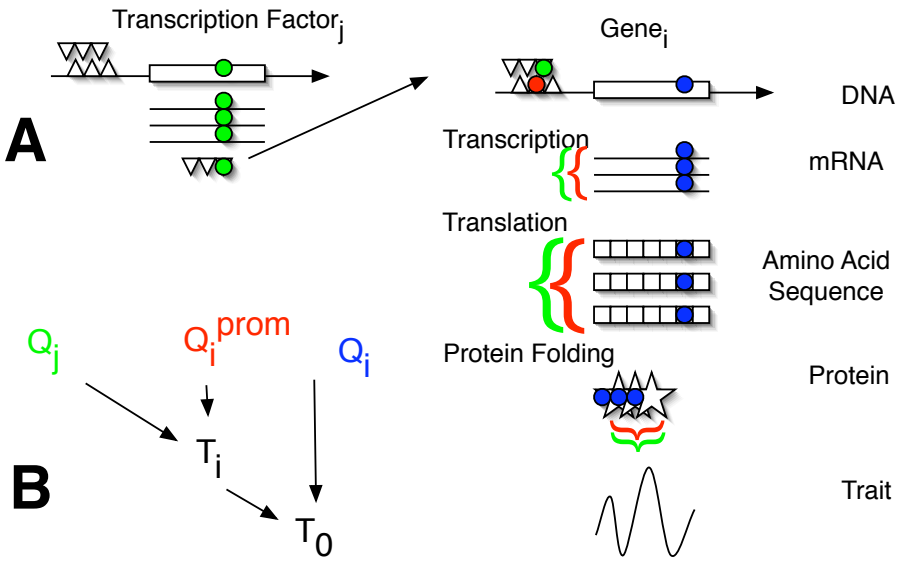


Figure 1.1. Example of a Regulatory Mechanism A. Regulatory mechanism modulating the trait: In this network gene i is responsible for modulating the trait. A variation in the coding region of gene i can change the amino acid sequence of the protein and, thereby, change its tertiary structure. In this example gene i is being regulated by transcription factor j. A variation in either promoter region of i or the coding region of j can impact the binding affinity of this transcription factor which changes the amount of ith protein. B. Using the data available from an expression genetics experiment this mechanism can be modeled as a Bayesian network. In this scenario the genotype of transcription factor j (Q_j) and the genotype of Gene i's promoter influence the variation of the trait (T_0) through modulation of i's transcript (T_i). The genotype of i (Q_i) directly affects the variation of the trait. The genotypes (Q_j, Q_i and Q_i^{prom}) have discrete values and the traits (T_i and T_0) are continuous quantities.

1.2 Modeling of Regulatory Modules and Bayesian Networks

One of the most useful tools used in analysis of correlation structures is Bayesian networks. It is a graphical model representing the conditional independencies betr-

ween a set of random variables. This model is attractive in our case because: direct and indirect influences can be easily modeled; in some cases it allows the inference of the causality, and the domain knowledge can be easily incorporated.

The challenges in applying this method to expression genetics data are related to the under-specified nature of the problem; the presence of hybrid data types and the fact that many important variables are unobserved.

To overcome these challenges I introduce a set of three novel techniques that reliably extract parts of the underlying network. The *network motif searching* technique allows modeling known regulatory mechanisms as Bayesian networks and allows recovering regulatory instances with similar mechanism. The *local network inference* technique finds the elements of the regulatory neighborhood of a given transcript and constructs its local network. The multiple networks inferred from application of these two methods are then assembled into a common network using the *network aggregation* technique.

1.3 Outline

The rest of this dissertation is organized as follows. Chapter 2 contains the key genetics concepts. Various breeding schemes to create the special strains are explained here. The microarray technique for extracting expression and genetic variation is described. And, the data-sets used in this dissertation are detailed.

In Chapter 3, previous work related to analyzing expression genetics data are discussed. In this chapter the notations used in this dissertation are introduced. The interval mapping technique, which is used in classical genetic mapping, is explained. Bayesian networks are introduced along with examples of application to expression data. Some of the existing Bayesian network based methods that have been applied on expression genetics data are also reviewed.

Chapter 4 deals with presenting an overview of methods that I adapted for the analysis of expression genetics data using techniques from information theory and Bayesian networks. In this chapter I explain some of the challenges with these approaches and detail how the methods presented in this dissertation overcome some of these challenges.

In Chapter 5, the application of the motif searching method to infer instances of *Quantitative Trait Gene* (QTG) model is discussed. The biological motivations for proposing this model are explained. The methodology for recovering this model is detailed. The results of the simulation studies to tests the robustness of this model are presented. The application of this model on a yeast cross resulted in recovery of thousands of QTG instances. Some of these instances are presented in this chapter along with analysis of these networks.

In Chapter 6, use of local network inferencing for recovery of local regulatory modules is discussed. The Markov blanket technique used for inferring elements of a regulatory neighborhood is explained. The method for constructing the local network is presented. The details of testing this method on simulated data are presented, and the results on the application of this method on mice are analyzed.

In Chapter 7, application of network aggregation in analyzing genetic hotspots is discussed. The biological significance of hotspots is explained. The method to infer a set of potentially important *primary transcripts* is presented. This method is then examined through simulation. Later, some of the aggregated networks inferred from the analysis of a mice cross are shown.

In Chapter 8, the contributions of this dissertation are summarized. The effect of dimensionality on the effectiveness of these methods are discussed. Some of the experimental challenges are mentioned.

CHAPTER 2

KEY EXPRESSION GENETICS TECHNIQUES AND DATA

Expression genetics data consists of mRNA transcript (gene expression) data and the genotype data from a specially created genetically diverse population. In this chapter various breeding schemes for creating such a population are discussed. Later, methods used to collect genotype and expression data are detailed. And finally, datasets used in this dissertation are introduced.

2.1 Inbred strains and breeding schemes

Inbred strains are plants and animals in which the copies of the chromosomes are identical and offspring resulting from intra-strain breeding also belong to the same strain. Such strains are created by sibling-mating for multiple generations. After 20 generations of mating, 99% of the dissimilarity between the haplotypes is lost and are technically deemed as inbred. Often selection is used to create strains that show a particular trait.

Inbred strains are valuable in a genetic cross experiment for multiple reasons:

- As the offspring of inbreeding are genetically identical to their parents, the experiments are reproducible with the same genetic background.
- Breeding two inbred strains for two or more generations results in offspring whose genome sequence is made up of long subsequences of the parental strains. The composition of each of the chromosomes can be reconstructed by genotyping only a few markers.

- The strains derived from breeding two different inbred strains have uniform distribution of alleles across the genome. Therefore the power of linkage studies is uniform across the genome.

Some of the popular breeding schemes used in genetic cross experiments are shown in Figure 2.1.

Mating of the inbred parental strains (F0, genotype AA and BB) results in offsprings (F1) containing a copy of each chromosome from its parents (AB). Sibling mating of F1 strains results in F2 offsprings whose allele at each position can be homozygous of any of the parental strains or heterozygous (AA, BB, or AB). The advantage of using F2 strains is that all the possible genotypes can be observed, which is helpful in recovering effects such as dominance. But the power of linkage studies can be lower for the following two reasons: there are three allele at each location (rather than two in other breeding techniques), and the allele frequencies are unequal (P(AA)=P(BB)=25%,P(AB)=50%).

On the other hand back-crosses, created by breeding F1 strain (AB) with one of the parent strain (for example BB), have two alleles (AB,BB). It is more powerful than an F2 cross but lacks complex inheritance patterns (e.g. dominance).

Both of these crosses, F2 and back-cross, result in non-inbred samples which makes the experiment non-reproducible. However F2 samples can be inbred for more than 20 generations to create *recombinant inbred strains* which contain only homozygous alleles(AA,BB). Recombinant inbred strains are very useful because they represent a reproducible stock for a set of traits derived from the crossing of the original F0 parental strains.

2.2 Genotyping and Phenotyping using Microarrays

The expression microarray is a technology used to measure the abundance of thousands of mRNA transcripts. In this technology a panel is created containing

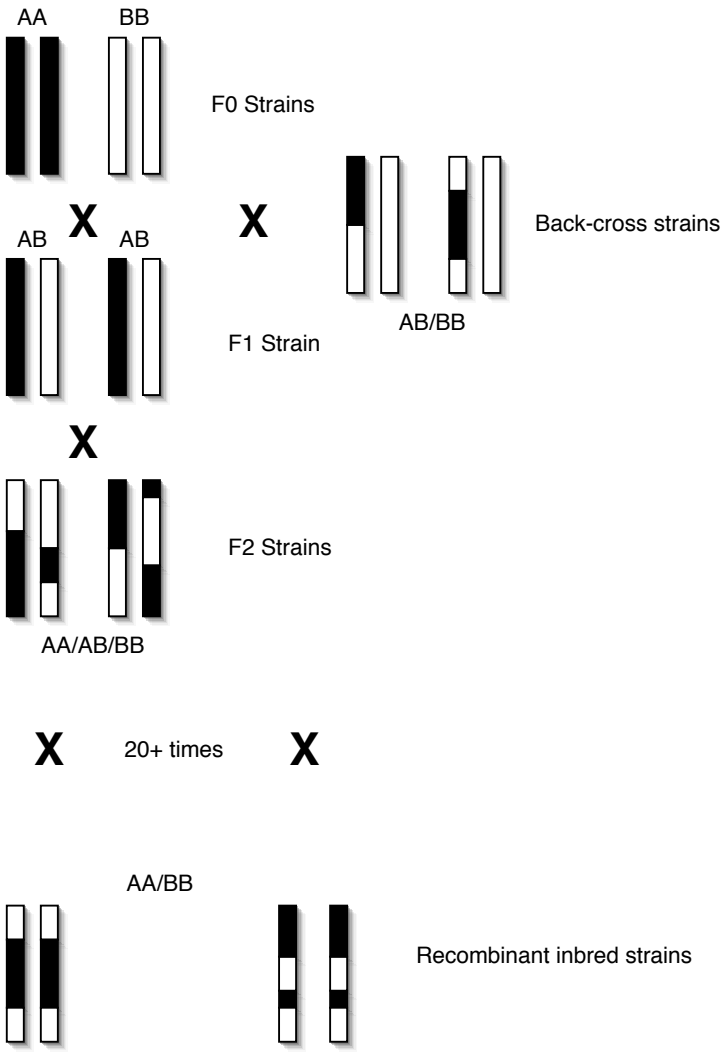


Figure 2.1. Cartoon view of the breeding schemes The composition of one of the chromosomes is shown. The exact number of crossovers depends on the frequency of recombination.

oligonucleotide probes specifically complementary to the sequence of each mRNA transcript that needs to be measured. When treated with a sample containing cDNA copies of mRNA, the free DNA molecules bind to their complementary sequences and the amount of hybridization provides information about the presence and quantity of each mRNA sequence. In this way, DNA expression microarrays quantify the expression level of thousands of genes.

The DNA genotyping microarray is a similar technology in which the sample is derived from an individual's genomic DNA and the probe-DNA hybridization indicates the presence or absence of specific genomic sequence. This allows for genotyping markers of interest, i.e. determining the un-phased diploid sequence at specific chromosomal locations. In other words, hundreds or thousands of markers can be genotyped along the chromosomes as either AA, BB, or AB.

In a genetic cross experiment, microarrays can be used to both genotype markers and measure transcript abundance.

2.3 Expression Genetics Data-set

A typical expression genetics data-set consists of genotypes of various markers and genomewide expression data along with other traits. The number of markers depends on the genotyping technology and breeding scheme and a typical data-set has anywhere between $10^2 - 10^5$ markers. The number of genes (which determine number of transcripts) ranges from 6000 in yeast to 20,000+ in mammals. Current available data-sets have anywhere between 30 and 300 samples. A schematic representation of expression genetics data is shown in Figure 2.2.

2.4 Data-sets used in this dissertation

The methods proposed in this dissertation were applied on one yeast data-set and two mouse data-sets.

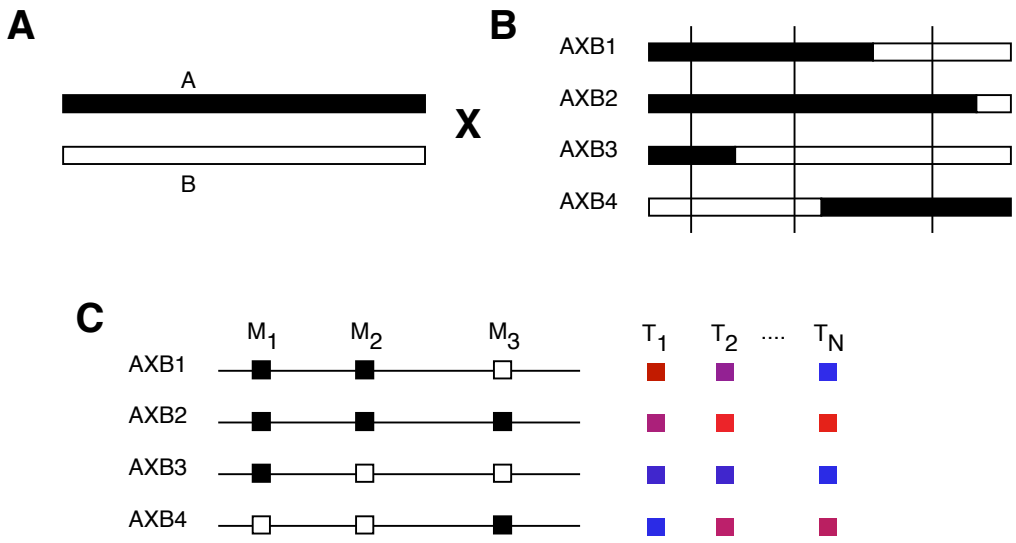


Figure 2.2. Schematic View of the Expression Genetics Data The strains A and B are bred to create samples $AXB1$, $AXB2$, $AXB3$ and $AXB4$. The genomes of these samples are mozaics of the parental strains. These samples are genotyped at many predetermined loci (also known as markers) M_1, M_2 and M_3. Also, expression of multiple genes are measured for each of these samples (T_1, T_1, $\cdots$, T_N). Collectively, the marker data and the expression are known as expression genetics data-set.

Yeast (RMXBY) In this study [6] yeast strains BY4716 and RM11-1a were used to create 113 samples. Each of the 113 strains were genotyped at 2957 SNP markers, and 6164 transcripts were measured.

Mice (BXA) In this study [22] 120 F2 crosses of mouse inbred strains C57BL/6J (B) and A/J (A) were created. The mice were genotyped using 173 SNP markers. The liver samples of these mice were used to measure 16,463 transcripts.

Mice (BXD) In this study [43] mouse inbred strains C57BL/6J (B) and DBA/2J were crossed to create 111 F2 samples. The resulting mice were genotyped using 134 micro-satellite markers. The liver samples of these mice were profiled for 23,574 transcripts.

CHAPTER 3

PREVIOUS WORK

Classical quantitative traits such as seed size [41] and hypertension [53] have been mapped to the genome since the beginning of 20th century. In the earlier years, as there was no genetic map, these traits were mapped in relation to other strongly inherited discrete traits. (For example, the correlation between seed color and seed size indicates that the genetic variations responsible for these traits are in the proximity of each other.) After the discovery of genetic markers, whose locations were known on the genome, more sophisticated mapping techniques were developed. *Interval mapping* [30] is one such technique where the relative locations of the consecutive markers are used to infer the putative loci more accurately.

The mRNA transcript abundances were measured using microarrays under different conditions such as stress, life cycle, tissue type and so on [48, 44]. Correlation between conditions and the variation of transcripts were used to implicate the role of specific genes with conditions. For example, using this approach genes involved in the yeast life cycle were identified [17]. Further, using the correlation structures between transcripts putative regulatory networks were constructed [19, 46].

When the first expression genetics experiments were conducted, the resulting data were analyzed using a mixture of the above techniques [7, 43, 42]. The rest of the chapter is organized as follows. In section 3.2, the interval mapping technique is explained. In section 3.3 use of Bayesian networks to analyze expression data is summarized. Later (section 3.4), methods that are currently being employed to analyze expression genetics data are presented.

3.1 Notation

For the rest of this dissertation T is used for measurable traits (both classical and transcript) and Q is used for genetic variation. Ts are always continuous and Qs are always discrete variables. Any feature corresponding to a gene is denoted by the corresponding subscript (T_i corresponds to transcript of gene i and Q_i its genotype). Subscript can also be a set of genes (eg. $S = \{1, 2, 3\}, T_S = \{T_1, T_2, T_3\}$). In some cases the subscript is dropped when there is no ambiguity. The bold type of these letters indicates set of all such variables ($\mathbf{T} = \{T_1, T_2, \cdots, T_N\}$ where N is the number of traits).

I borrow some concepts from information theory. The sign "$\perp$" is used for *data independence*. Eg. $X \perp Y|Z$ indicates X is independent of Y conditioned on Z. $I(X;Y|Z)$ corresponds to *mutual information* between X and Y conditioned on Z.

3.2 Interval Mapping

Quantitative trait loci (QTL) of a trait consists of locations on the genome that are responsible for variation of that trait. QTL mapping is done by finding the locations across a genome whose genotypes are correlated with the variation of the trait. Due to technological and cost constraints, the sample organism's complete genome is not sequenced, but rather a small set of markers spread across the chromosomes are queried. The actual genotype at each chromosomal location is estimated using one or multiple markers flanking the location.

For the simplest mapping case, where only the data at the markers are considered, ANOVA techniques can be applied to identify loci correlated with a trait. In this method the difference between means of traits in the genotype groups defined by the marker is used to calculate the significance of the association. For example, in a two allele (0,1) case,

$$s = \frac{\hat{\mu_0} - \hat{\mu_1}}{\hat{\sigma}\sqrt{n}}$$

where $\hat{\mu_0}$ and $\hat{\mu_1}$ are the empirical means of the trait when marker allele was 0 and 1 respectively, $\hat{\sigma}$ is the empirical standard deviation within the groups, and n is the number of samples. The cumulative distribution of s in a normal distribution is used to calculate the p-value of the linkage.

Many approaches have been used to extend this idea to any location. The means of the segregates at the QTL can be modeled to be a sum of means at the marker weighted according to recombination probability [51]. In the two allele case:

$$\mu_0 = (1-r)\mu_0^M + r\mu_1^M \text{ and } \mu_1 = (1-r)\mu_1^M + r\mu_0^M$$

are the trait means at a location which is at a recombination distance r from the marker M.

Alternatively the likelihood of the trait (T) can be modeled as a mixture [49]:

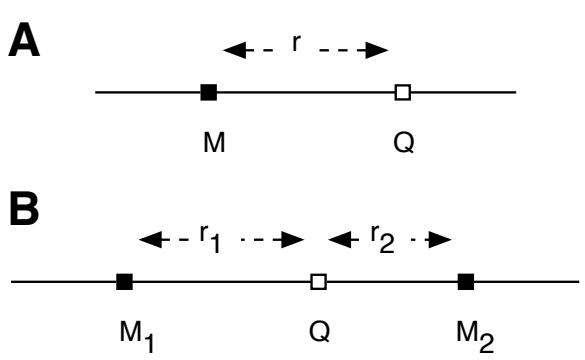
$$f(T) = \sum_Q b(r, Q, M)\phi(\frac{T-\mu_Q}{\sigma})$$

where function $b(r, Q, M)$ returns the probability of observing genotype (Figure 3.1 A) Q at a location which is at a distance r from a single nearby marker whose genotype is m. The parameters in this formulation, $\mu_Q, \forall Q$ and σ, can be estimated by using the expectation maximization (EM) strategy [30].

The relative positions of the consecutive markers can be used to provide a better estimate of the linkage [30]. The above mentioned formulation can be be expanded as:

$$f(T) = \sum_Q b(r_1, r_2, Q, M_1, M_2)\phi(\frac{T-\mu_Q}{\sigma})$$

where r_1 and r_2 are the distances from the flanking markers M_1 and M_2. The calculation of probability function $b(r_1, r_2, Q, M_1, M_2)$ is shown in Figure 3.1 B.



M	Q=0	Q=1
0	1-r	r
1	r	1-r

M_1	M_2	Q=0	Q=1
0	0	$(1-r_1r_2)$	$r_1\ r_2$
0	1	$r_2/(r_1+r_2)$	$r_1/(r_1+r_2)$
1	0	$r_1/(r_1+r_2)$	$r_2/(r_1+r_2)$
1	1	$r_1\ r_2$	$(1-r_1r_2)$

Figure 3.1. Probability observing a QTL genotype given genotypes of the flanking marker(s) A. This figure shows the calculation of observing genotype Q which is at a recombination distance r from $m(b(r, Q, M))$. B. This figure shows the calculation of observing genotype Q which is at a recombination distance r_1 and r_2 from M_1 and M_2 respectively $(b(r_1, r_2, Q, M_1, M_2))$. The figure on the left shows the schematic representation of the locations and the table on the right gives the probability distribution.

In some cases multiple QTLs synergistically influence the trait (example: logical AND where both QTLs must be in a particular state to have an effect). Such a phenomenon, known as epistasis [11], cannot be recovered using the single QTL methods. To address this problem multiple QTL mapping methods have been proposed [47].

Compared to the number of samples, the complexity of this model can be very high. For example, in the 2 allele situation, 4 mean parameters and the variance parameter have to be estimated from only 100 samples. Therefore, sample size effectively constrains the number of interacting QTLs that can be modeled. This problem also affects much of our work where model complexity must be sacrificed to achieve significant model fits.

A more detailed review of the QTL mapping methods can be found in [9] and [8].

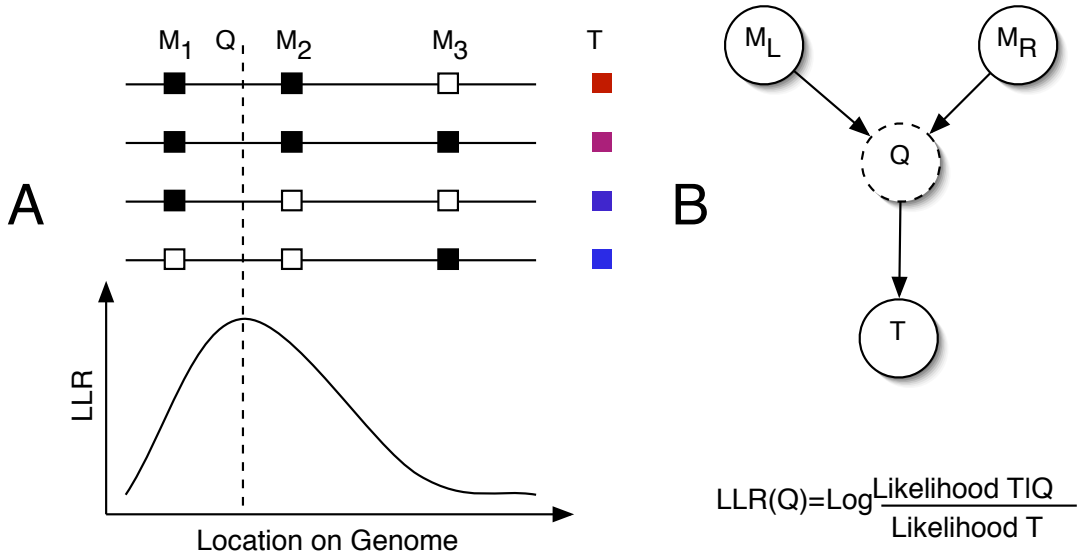


Figure 3.2. Schematic View of Interval Mapping A. The correlation between the marker genotype and the trait along with the recombination frequency is used to calculate the likelihood of a locus regulating the trait. B. This relation can be expressed as a Bayesian network.

3.3 Graphical Models and Expression Networks

Probabilistic graphical models provide an efficient way to represent the joint probability distribution of a set of random variables. In particular Bayesian networks are directed acyclic graphs representing joint distributions as products of conditional probabilities [20].

Formally, if a set of variables can be arranged as $T_1, ..., T_n$ such that

$$P(T_1, ..., T_n) = \prod_i P(T_i|Pa(T_i))$$

where $Pa(T_i) \subseteq \{T_1, ..., T_{i-1}\}$ is the parental set, then the joint probability can be represented as a Bayesian network. The corresponding graph of the network can be constructed with vertices $\{T_1, ..., T_n\}$ and drawing directed edges from $Pa(T_i)$ to T_i.

A Bayesian network representation consists of two parts: the graph (G) which represents the dependencies among the variables and is made up of a set of edges, and the probability distribution (P) representing the nature of dependency. For example the graphical models showing single QTL and multiple QTL models are shown in 3.3.

Learning an acyclic graph G from relational data $T_1, ..., T_n$ is called *structure learning*, and learning the conditional probability distribution P from the data for a given structure is known as *parameter learning*.

Multiple heuristics have been proposed for structural learning and they fall into two broad categories. The most used approach is to use a scoring function that determines the fitness of a graph in representing the given relational data [20]. Often posterior probability is used as the score function:

$$\begin{aligned} S(G:T) &= \log(P(G|T)) && (3.1) \\ &= \log P(T|G) + \log P(G) - \log P(T) && (3.2) \end{aligned}$$

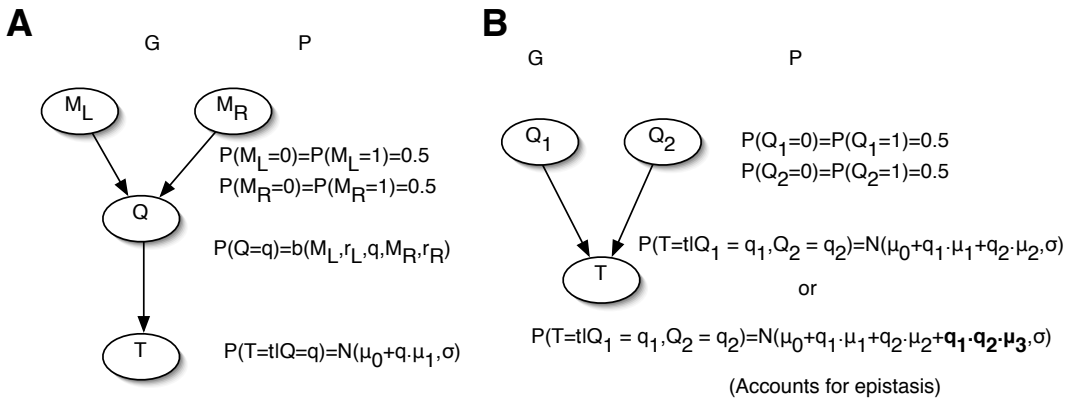


Figure 3.3. Bayesian networks showing one QTL and two QTL models In each instance the graph G showing the dependencies is shown on the left, and the probability distributions describing the nature of dependencies are shown on the right. In the single QTL case the relation between markers and the QTL is explicitly described whereas, for the sake of simplicity, such details are not shown in two-QTL case.

$$= \sum_i \log P(T_i|Pa(T_i:G)) + \log P(G) - \log P(T) \tag{3.3}$$

The individual terms in the first part can be calculated by learning parameters of the underlying distribution. The second part, $\log P(G)$, can be used to control network complexity (example: the Akaike information criterion [1] (AIC) $P(G) = k$, and the Bayesian information criterion [45] (BIC) $P(G) = k\log(n)/2$, where k is the number of free parameters and n is the number of samples) and incorporate regulations imposed by domain knowledge [21]. The third part is constant over all the structures and is ignored when choosing a G that maximizes the posterior probability.

The second approach in structure learning is using conditional independence statements. A set of coherent independence statements can be aggregated into a Bayesian network [57]. For example, given the statements $T_i \not\perp T_k$,$T_j \not\perp T_k$ and $T_i \perp T_j$, the network $T_i \rightarrow T_k \leftarrow T_j$ is constructed. These independence statements can be learned from analyzing relational data [50, 52].

Discovering the optimal Bayesian network from the relational data does not imply that all the causal relations can be inferred from the network structure. Multiple graphs can have the same score, the same undirected structure and denote the same set of conditional independence relationship. Such a set of graphs are said to be in the same *equivalence class* [20]. However, using domain knowledge about causation, some of the remaining undirected edges can be directed. For example, in expression genetics data only the state of the genotype determines the trait and not the opposite and, therefore, an edge from a QTL to a trait is causally correct.

3.3.1 Using Bayesian networks to analyze expression data

Bayesian network analysis has been applied to expression data (versus expression genetics data studied here) to identify potentially related genes from transcript abundance alone. The dimensions of a typical expression data set, with 100s of samples and 10,000s of genes, makes the task of structure prediction hard. The following two approaches heavily influence our work.

The first approach to the problem of structure prediction given small sample size is offered by Friedman et al. [19]. The authors propose a set of heuristics to efficiently predict features, network proximity and causal order, of the underlying network. In this method for a bootstrap instance (sampling with replacement) of the data, a graph is built incrementally by adding or deleting the best edge without violating the acyclic property of Bayesian networks. The Bayesian information criterion (BIC) is used as a score and the distribution among variables is assumed to be either discrete or linear Gaussian. This experiment is repeated over multiple bootstraps. Network proximity is measured as the fraction of instances where two variables were found to be in the Markov blanket (Markov blanket of a variable consists of its parents, children and spouse nodes in the Bayesian network) of each other, and causality is predicted as the fraction of instances where a variable is an ancestor of another.

Although efficient for data containing 100s of genes, this algorithm is computationally inefficient for the large data sets containing 1000s of genes. Furthermore as the parental set of any gene grows larger, the confidence on estimated parameters decreases.

The second approach is to construct local modules around the transcript of interest [37]. In this approach the elements of local regulatory network are inferred using the Markov blanket inferencing algorithm. As the number of nodes in the Markov blanket is typically very small, the optimal Bayesian network describing these nodes can be recovered using an exhaustive search over all the possible network structures.

3.4 Analysis of expression genetics data

An *Expression genetics data-set* consists of both genetic variation data and gene expression data. Given that the gene expression is a measure of an intermediate molecule in determining the trait, this data set can be used to construct more deliberate regulatory networks. Figure 3.4 describes the biological relation between these data and the summary of techniques that can be used to infer regulatory relations.

Most of the earlier expression genetics studies use interval mapping to analyze the data. In Brem et al. [7] the authors applied QTL mapping of transcripts in a yeast cross (described in 2.4) and discovered that 570 transcripts (out of 6164) were linked to at least one loci ($p < 5 \times 10^{-5}$). When these linkages were binned according to the genomic locations, eight unusually large groups were identified. Using the domain knowledge they were able to identify the putative regulator for six of these groups. In Schadt et al. [43] three sets of expression genetics data were analyzed, one each in mice, maize and human. The interval mapping of the mice transcripts revealed that 3,701 transcripts (out of 23,574) were linked to at least one QTL.

A few methods have been proposed to construct the gene regulatory modules using Bayesian networks. In Zhu et al. [62] $P(G)$ is composed of the product of individual

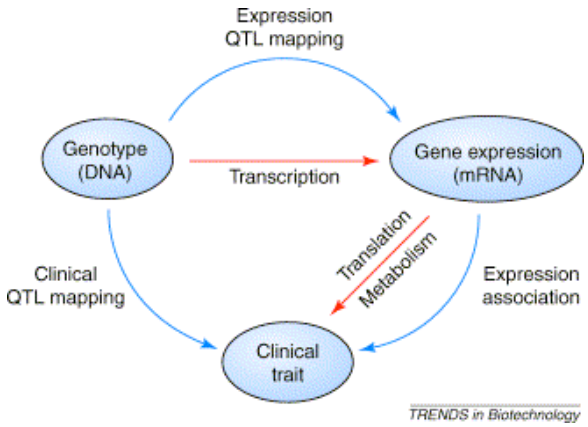


Figure 3.4. Strategies for using genotype, gene expression and trait data to study complex disease (From [27]) Red arrows represent the central dogma: the biological path from genotype to trait. Blue arrows represent statistical approaches. Quantitative trait locus (QTL) mapping uses genetic markers (genotypes) and is based on meiosis. Expression association uses clustering, classification, and gene filtering methods.

edge probabilities

$$P(G) = \prod_{E \in G} P(E).$$

Probability of an edge E_{ij} from T_i to T_j is calculated as:

$$P(E_{ij}) = r(T_i, T_j) \frac{N(T_j)}{N(T_i) + N(T_j)}$$

where r is the correlation coefficient calculated using overlap of QTL maps of T_i and T_j, and N is a functions that returns number of QTLs. The intuition behind such formulation is if T_i is upstream of T_j, then T_j will inherit variation from T_i.

Alternatively Li et al. [31] calculate probability of an edge using the QTL map of the target and checking if there is a QTL around the physical location of the regulator.

$$P(E_{ij}) = \begin{cases} 1 & \text{if there is a QTL for } T_j \text{ around the physical location of } T_i; \\ 0 & \text{otherwise.} \end{cases}$$

In both of these studies the putative regulatory network was constructed using the Bayesian network reconstruction algorithm [19] and the prior probability of the graphs ($P(G)$ in Equation 3.3) were calculated using QTL mapping.

3.5 Research Opportunity

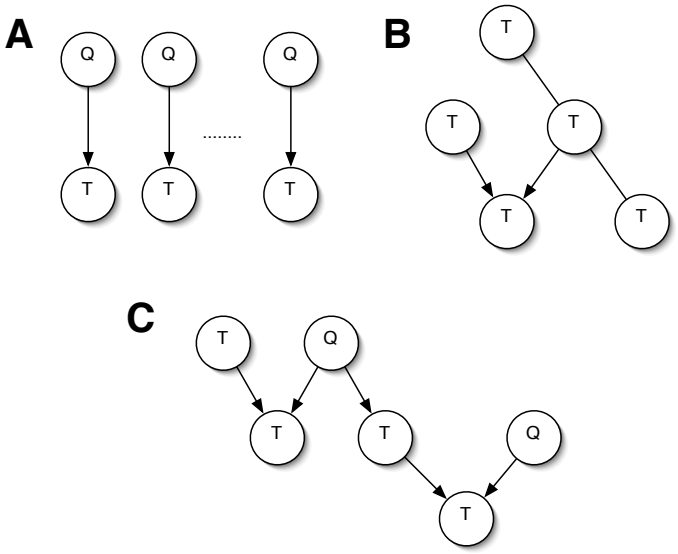


Figure 3.5. Examples of the networks recovered by each method A. Interval mapping recovers QTL and transcript pairs which are correlated. The correlation among multiple transcripts is ignored. B. The Bayesian networks of transcripts constructed using interval mapping as prior contain only transcript nodes. The direct modulation of the transcript by the genotype is ignored. C. We propose a method to recover complex networks that include genotype nodes. The explicit use of genotype nodes makes it useful in recovering more instances instances of modulation.

Although the above methods provide ways to analyze expression genetics data, there are several opportunities for improvement that we pursue in this work. The interval mapping technique is suited for traditional cases where only a few traits are collected. In an expression genetics data-set, thousands of transcripts are measured and each of them can be considered as a trait. The correlation structure among these

transcripts, which is ignored during interval mapping (Figure 3.5A), can also provide additional clues regarding the regulatory process.

The suggested Bayesian network construction methods, instead of relying on the direct correlation between genotype and expression, use latent observations such as overlap in their interval mappings to discover instances of genetic modulation (Figure 3.5B).

In this dissertation we present a unified method that considers both transcripts and genotypes simultaneously. Through Bayesian network modeling of these variables we show that complex and causally informative regulatory structures can be recovered (Figure 3.5C).

CHAPTER 4

SYSTEMATIC MINING AND ANALYSIS OF STATISTICALLY SIGNIFICANT SUB-NETWORKS

The objective of this dissertation is to use directed graphical models to infer regulatory relationships among the genotypes, transcripts and traits. Such a model would consider an arbitrary number of variables, incorporate complex interactions, and infer precise modulatory mechanisms. Due to the availability of very small number of samples (usually in 10^2s) compared to the number of variables (in 10^4s), we propose a method to recover sub-networks of interest rather than the complete network itself. We propose a combination of three basic techniques to achieve this: *network motif searching*, *local network inference*, and *network aggregation*.

This chapter begins with reasons for using Bayesian networks, some of the challenges in using it and how those problems are alleviated with use of the presented methods. Later each of the methods are briefly introduced and are linked to some of the actual implementations.

4.1 Why Bayesian Networks?

Bayesian networks are the natural choice for modeling gene regulation for the following reasons:

- In some cases Bayesian networks allow inference of the direction of causality. For example given three variables X, Y and Z, and if their correlation structure suggests that Y is correlated with both X and Z, but X and Z are not correlate, the only Bayesian network satisfying these conditions is $X \rightarrow Y \leftarrow Z$ which

implies the causal explanation. The regulatory networks are, by nature, causal and it would be immensely useful to be able to infer causal structure among genes.

- Direct and the indirect influences can be modeled easily through Bayesian networks. In most cases the modulation of a trait takes place through a cascade of regulatory actions. When modeled as a Bayesian network the presence of an edge between two variables provides a strong indication of a direct regulatory relation between them. And the path between two variables provides the information about the regulatory cascade.

- Prior knowledge about the possible regulatory structure can be easily integrated in Bayesian networks. As shown in Equation 3.3 the prior probability term $P(G)$ can be set using the available domain information. For example, in a genetic cross cross experiment the genotype of a locus in an organism is determined by chance during meiosis, which in turn determines its traits. Therefore in expression genetics data, the genetic variation at a locus can influence the variation of a trait, but not otherwise. This information can be used in restricting the possible network space to only to Gs that do not have an edge from a trait node to to a genotype node.

4.2 Challenges in adapting Bayesian Networks

The most important challenge in adapting Bayesian Networks in analyzing expression genetics data is data insufficiency. A typical expression genetics data consists of thousands of variables and only a hundred samples. In this dissertation I present methods that handle this problem using the following two core ideas:

Controlling Size of the Network: Instead of attempting to infer the complete underlying network I focus on recovering small networks of interest with higher

reliability. In *network motif searching* the network is decomposed into smaller sub-networks with at most one dependent variable. The significant instances of these subnetworks are later glued together to recover the original network. In *local network inference* I restrict myself to a small subset of regulatory neighbors in the *Markov blanket* of the variable of interest. In *network aggregation* only a set of triplet of variables is considered at a time.

Clever Modeling of the Relation Between Genes: In this dissertation the recovery of complex modules is aided by the use of succinct, yet expressive modeling. Instead of using a standard linear model (with respect to regulator's transcript and genotype) an interacting component is included. This action enables recovering cases where the regulatory nature of a regulator changes because of a polymorphism.

4.3 Network Motif Searching

In many cases there is some information about a common regulation mechanism. That mechanism can be represented as a network motif i.e. a graph with known structure, but unknown labels. Different combination of variables can be substituted as labels, and for each combination the likelihood of the graph is calculated. The combination of variables with significant likelihoods represent instances of the regulation mechanism and I call this technique *network motif searching*. In fact, the process of interval mapping itself can be represented as a network motif: $Q_i \rightarrow T_j$ with T_j being the trait of interest and Q_i the unknown locus [29]. The set of Q_is for which the likelihood of this graph is significant are the QTLs of this trait.

In this dissertation a regulatory mechanism named *Quantitative Trait Gene* is introduced that represents the process of *Transcription regulation*. Here the relation between a regulator and a target was modeled to include both transcript and the genotype of the regulator ($Q_i \rightarrow T_j \leftarrow T_i$). This model is presented in Chapter 5.

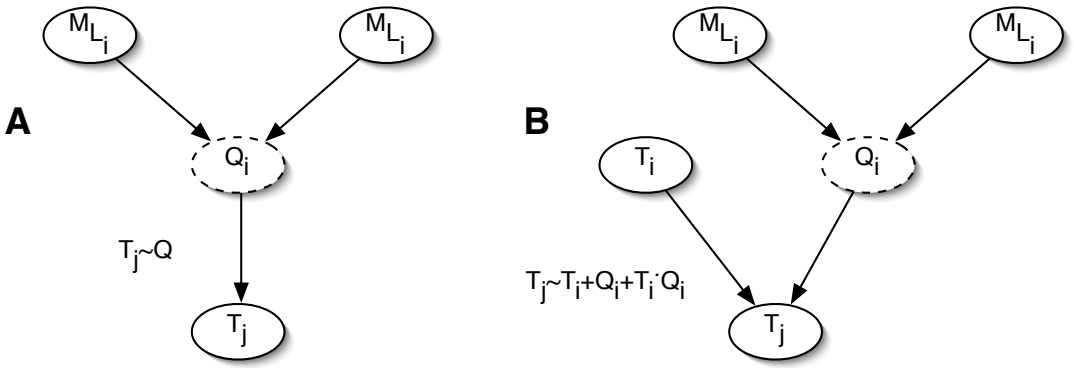


Figure 4.1. Network Motif Searching: A. Graphical model of the process behind Interval mapping: Q_I is the genotype of a location i flanked by markers M_{L_i} and M_{R_i}, and T_j is the trait of interest. The value of Q is determined by the flanking markers and T_j is modulated by Q. B. Quantitative Trait Gene model: In this case both transcript T_i and genotype Q_i of gene i influence j. The term $T_i \cdot Q_i$ is added to model changing rate of modulation due to change in genotype.

4.4 Local Network Inference

Instead of searching for data that fits the network model of a specific regulatory process, we may need to infer the regulatory processes associated with a single trait of interest. In such cases we find the variables in the regulatory neighborhood and reconstruct the best possible local network. Our definition of a regulatory neighborhood of a variable consists of its modulators, targets, and co-modulators (Figure 4.2). In directed graphical models, such a neighborhood definition corresponds to the Markov blanket of the variable. From the perspective of information theory the Markov blanket corresponds to the minimal set of variables that provide maximal knowledge about a particular variable. Knowledge of additional variables do not increase the knowledge about this variable.

In Chapter 6, a method to infer elements of a Markov blanket and a method for reconstruction of the network using information theoretic techniques are explained[23].

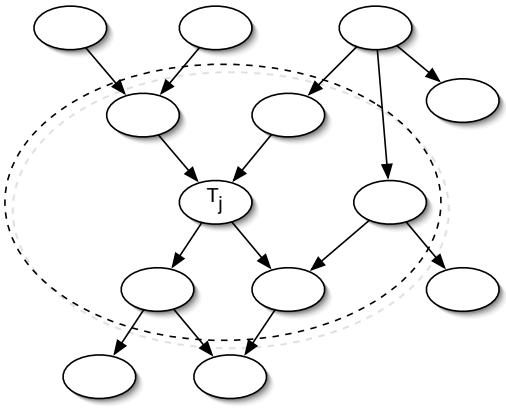


Figure 4.2. Local Network Inference The goal of local network inferencing is to (a) identify the variables in the regulatory proximity, (b) reconstructing the local network.

4.5 Network Aggregation

The *network motif searching* and *local network inference* steps result in detection of thousands of networks. Often a common subnetwork is found repeating in many of those network. The fact that it repeats frequently suggests that it is likely to be biologically important. The *network aggregation* algorithm finds such a *kernel subnetwork*, and infers its most likely neighborhood by aggregating over all the networks.

Analysis of interval mapping of transcripts often reveals the presence of pleiotropic genetic hotspots, i.e. loci that affect the expression of a large number of transcripts. In Chapter 7, I present an application of this method for analyzing such hotspots. In this chapter the relation among transcripts linked to the same hotspot is analyzed using conditional independence tests and a set of *primary transcripts*, transcripts that are directly modulated by the hotspot, is inferred [22] (Figure 4.3).

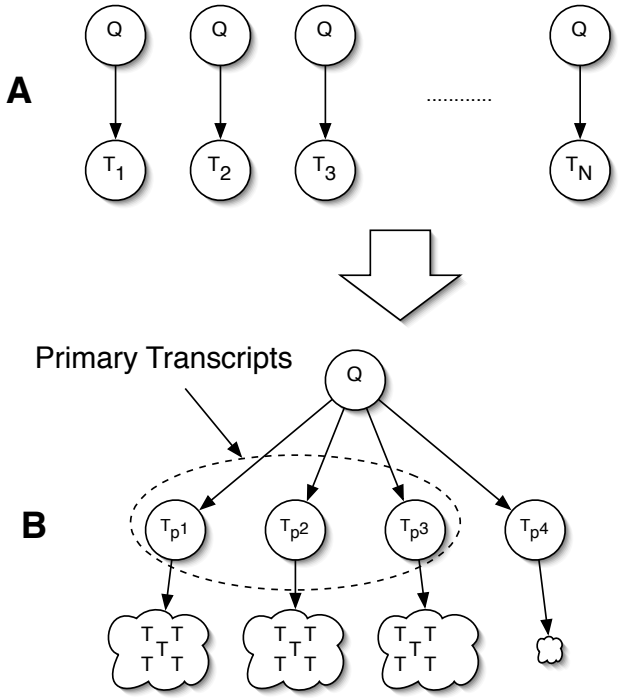


Figure 4.3. Example of a genetic hotspot analysis A. Transcripts $T_i, T_2, ...T_N$ linked to Q are inferred using interval mapping. B. Using conditional independence tests a set of primary transcripts $T_{P^1}, T_{P^2}, T_{P^3}$ are inferred. These primary transcripts are modulated directly by the hotspot Q, and they, in turn, modulate a large number of transcripts. T_{P^4} is not identified as a primary transcript as it does not modulate large number of transcripts.

CHAPTER 5

CAUSAL INFERENCE OF REGULATOR-TARGET PAIRS BY GENE MAPPING OF EXPRESSION PHENOTYPES

Correlations between polymorphic markers and observed phenotypes has been widely used to implicate regions on the genome responsible for modulation of different traits [41, 53]. When the phenotype is gene expression, then loci involved in regulatory control can theoretically be implicated. Recent efforts to construct gene regulatory networks from genotype and gene expression data have shown that biologically relevant networks can be achieved from an integrative approach[27].

Inspired by epistatic models of multi-locus quantiative trait (QTL) mapping, I propose a unified model of expression and genotype representing *cis*- and *trans*-acting regulation to identify quantitative trait genes (QTG). In this approach Bayesian networks are used to model the relation between the putative regulator and the target. In addition, the conventional linear model is extended to include both genotype and expression of putative regulator genes and their interactions. The model provides a high-resolution mapping of specific genes in contrast to standard linkage approaches that implicate large QTL intervals typically containing tens of genes.

5.1 Quantitative Trait Gene Model

The regulatory relationship between two genes is shown in Figure 5.1 A. This relation is represented as a Bayesian network (5.1 B). The genotypes and the expression measures as numeric random variables are represented in this graphical model. With respect to expression genetics data, where we do not have the genotypes of

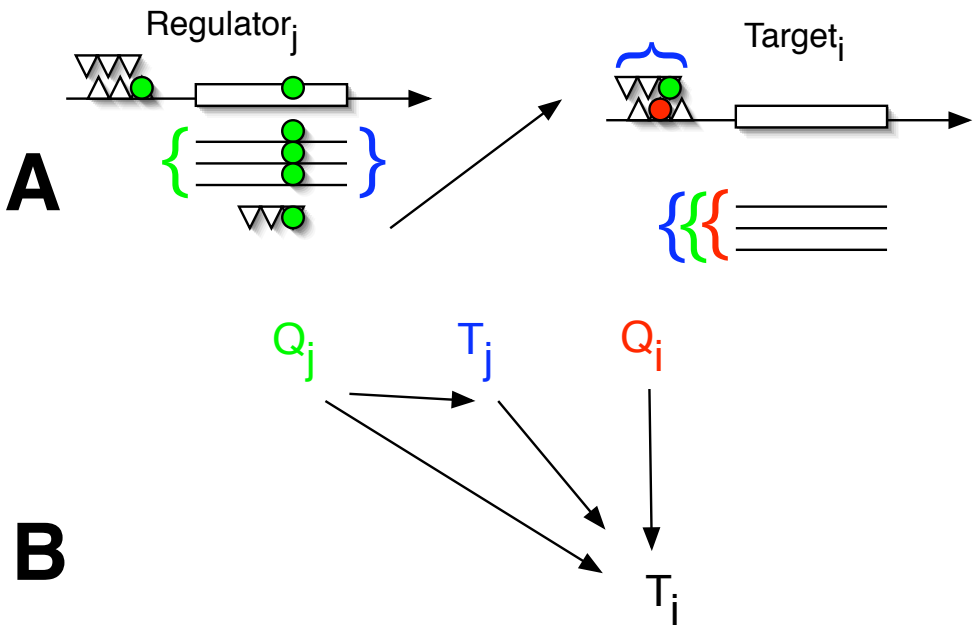


Figure 5.1. Transcription Regulation A. The transcription regulation process is depicted. The transcript of the regulator (T_j) is modulated by the variation in its promoter region (Q_j) (in expression genetics data, due to low resolution of crossing, the promoter and gene-coding regions are going to have same genotype). The variation in target's transcript (T_i) is modulated by the genetic variations in its promoter region (Q_i), coding region of the regulator (Q_j) and transcript variation of the regulator (T_j). B. This regulatory relation is modeled as a Bayesian network.

every location, a more refined version, the genotypes of the genes are further modeled as the function of their flanking markers (as shown in Figure 5.2b). In the general case of QTL interval mapping using sparse marker data, the genotype at a site of interest is an unknown random variable, Q_j, dependent on the observed genotypes of the nearest upstream and downstream flanking markers, $M_{j,L}, M_{j,R}$. The conditional probability of the unobserved genotype is a well-known function of the recombination distances among Q_j, and $M_{j,L}, M_{j,R}$[32]. Assuming that some observed phenotype (here gene expression, T_i, where i ranges over the number of genes) is dependent on Q_j, then the graphical model is shown in figure 5.2a. QTL interval mapping is then the likelihood of a mixture of each Q_j and the selection of those Q_j where the log likelihood exceeds some threshold.

In this work I am concerned with the class of *trans*-acting regulators in which the expression of the target is dependent on the expression of the regulating gene. I consider three sub-classes of genotypic effect: *cis*-, *trans*-, and *cis-trans*-acting sites. The *cis*- case corresponds to regulation by a variation around the physical location of the gene and the *trans*- case corresponds to regulation by non-proximal locus. The *cis-trans*- case corresponds to *trans*-regulation by a *cis*-regulated gene. For example, a variation in the promoter region or 3' end of the target gene may have a *cis*-acting effect on the expression level of the target; a variation in the coding region of the regulator may have a *trans*-acting effect, either directly or indirectly, on the expression of a target gene, such as through the modification of a DNA-binding motif in a transcription factor; and variation in or around the regulator gene may have a *cis*-acting effect on the regulator's expression which indirectly affords a *trans*-acting effect on the target, i.e. *cis-trans*. No specific assumptions is made in this model regarding the precise mechanism of the allelic effect even though it is convenient to imagine examples of transcription factor binding. Variation can have direct or indirect effects

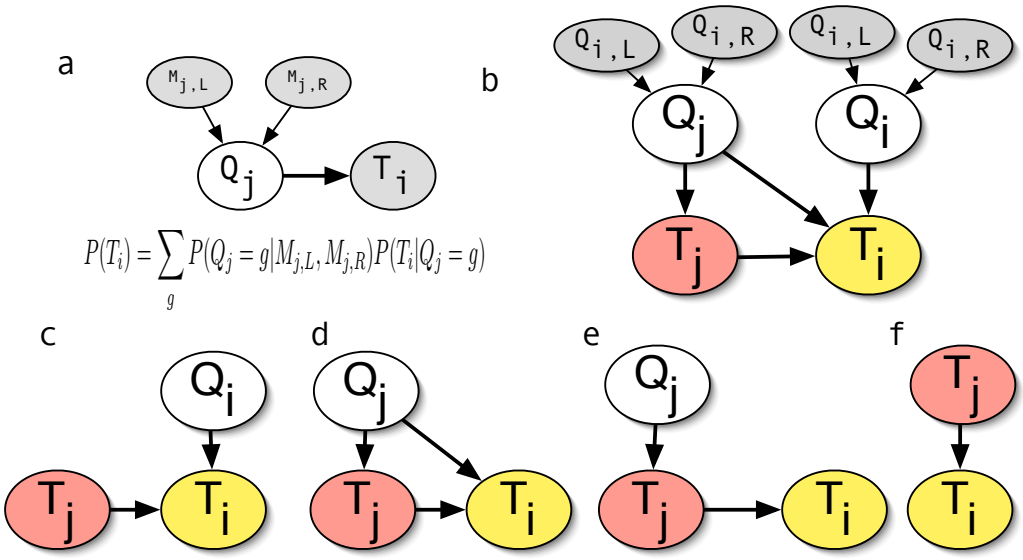


Figure 5.2. Graphical model representation of QTG model (a) The representation of conventional interval mapping as a graphical model. For an observed i, all candidate genotype sites j are considered. (b) The QTG model of a single regulator-target pair of genes (regulator is gene j and target is gene i). Subgraphs of (b) represent (c) *cis*-, (d) *trans*-, (e) *cis-trans*- cases, and (f) no genotypic effect, corresponding to the conventional BN. Colored and shaded nodes are observed.

on transcript abundance through a variety of mechanisms such as protein levels, RNA degradation rates, splicing, and so on.

If only the genotype sites at the locations of the protein-coding genes in a fully annotated genome are considered, then we can conveniently reference both genotypes and genes with a common index, i.e. Q_i represents the genotype for the gene i with expression T_i. Figure 5.2b naturally follows. This model is referred to as the full **QTG** model for a single *quantitative trait gene* and the process of estimating regulatory genes for a given target as "QTG mapping". The three genotype subclasses are subgraphs of the full model shown in figure 5.2c-f.

5.1.1 Inferring *trans*-acting Regulator

Here I address only the *trans*-acting regulator sub-class of figure 5.2d where the target is dependent on both the genotype and expression of the regulator. It is im-

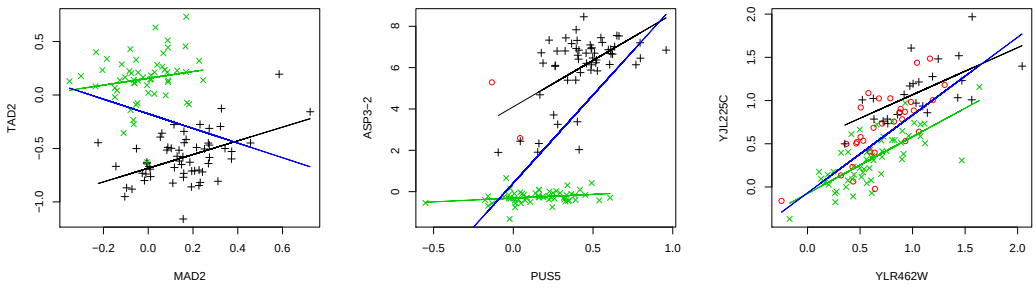


Figure 5.3. Examples of QTG model Three examples of the combined and interactive effects of the genotype and expression of a regulator gene on target gene expression in yeast. X and Y axes are expression of regulator and target, respectively. + and × are the two genotypes. Open circles are ambiguous genotypes when flanking markers differ. Regression lines are drawn for expression alone (blue) and by genotype (black and green). For the first example, the regulation and target gene expression appear anti-correlated, but are correlated with respect to genotype. The second example shows the importance of an interacting term to capture the change in the slope. The third example shows significant overlap in the range of target expression for the two alleles, but a clear separation with respect to regulator expression and genotype.

portant to recognize that this is a biologically reasonable scenario with many relevant examples in the data. For example, the scatter plots in figure 5.3 show the relationships among the expression of a target gene and the expression and genotype of putative regulators. In these cases only the combination, and sometimes interaction, of the regulator's genotype and expression can adequately model the target expression.

Therefore, to consider the possible interactions among genotype and expression, the full model is

$$P(T_i|Q_j, T_j, \theta) = \mathcal{N}(\beta_0 + \beta_1 T_j + \beta_2 Q_j + \beta_3 T_j Q_j, \sigma) \tag{5.1}$$

where θ is the β and σ model parameters.

As with standard interval mapping, Maximum likelihood estimation can be achieved using an expectation maximization (EM) approach in which the genotype, Q_j, and

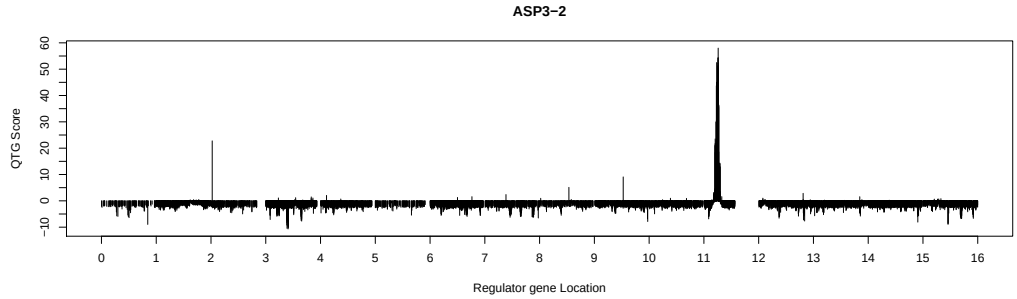


Figure 5.4. Sample QTG linkage map A sample QTG linkage map on the yeast genome using the *trans-* model for the target gene ASP3-2. A separate permutation test was performed per site and the corresponding threshold was subtracted from the full model score. Thus, positive values represent locally significant results.

the variables, θ, are alternatively estimated until convergence. But the advantage of this model over the standard mapping and multi-step approaches previously proposed is that individual loci are automatically mapped in a single step by simultaneously considering all available evidence.

Note that the strength of the genotypic effect is directly related to the ability to infer causality. That is, as the contribution of the β_2 and β_3 terms decreases, the confidence in the causal direction between genes i and j is reduced. We can be precise about this directionality by comparing this model with the simpler model of no genotypic effect (figure 5.2f). From equation(5.1), for each tested gene pair, i and j, we can determine the strength of a relationship (the *full model score*) as

$$\log_{10} \frac{P(T_i|Q_j, T_j, \theta)}{P(T_i|Q_j, T_j, \theta : \beta_1 = \beta_2 = \beta_3 = 0)} \tag{5.2}$$

and the directionality (*genotype reduced-model score*) according to

$$\log_{10} \frac{P(T_i|Q_j, T_j, \theta)}{P(T_i|Q_j, T_j, \theta : \beta_2 = \beta_3 = 0)} \tag{5.3}$$

Moreover, if the β_2 and β_3 terms are weak, then it indicates that the major effect is the QTL interval and so the confidence in the *specific* regulator gene is correspondingly

weak. Thus, confidence in the gene, T_j, as the actor in the relationship is found with the *expression reduced-model score*

$$\log_{10} \frac{P(T_i|Q_j, T_j, \theta)}{P(T_i|Q_j, T_j, \theta : \beta_1 = \beta_3 = 0)} \tag{5.4}$$

5.1.2 QTG Mapping

This model can be used to produce a QTG map (figure 5.4) for each target gene. These maps are similar to conventional QTL maps, but differ in that peaks are usually narrow (unless confounded by local linkage) and there is no genome-wide LOD significance threshold since the distribution of regulator transcript varies across the genome. Instead, the local significance threshold at each test site is subtracted from the LOD score such that positive values are significant.

5.2 Results

5.2.1 Function enrichment

As with networks derived from gene expression alone, the connectivity does not necessarily imply physical interactions between genes. Yvert et al. previously observed that genes within QTLs of gene expression traits were not enriched for transcription factors or any other function[60]. Nevertheless, it was wondered whether this lack of functional enrichment was due to the imprecise mapping of intervals that contain usually tens of candidate genes. It was hypothesized that the QTG mapping method, which identifies specific candidate genes, might show enrichment for transcription factors or other functional categories.

This hypothesis was tested by analyzing the yeast set consisting of 6164 gene expression measurements and 2957 genotype markers across 113 matings between two distinct isogenic strains[6]. I computed the pairwise dependency among all pairs of genes according to the full and reduced model scores, selecting those pairs with a

Table 5.1. Functional enrichment of regulators in GO: The set of GO terms from the "molecular function" (F) and "biological process" (P) categories showing significant enrichment among the candidate QTG. Total number of genes in yeast genome is 6164. Total number of regulators in filtered set is 823.

# in genome	# of regulators	P-value	GO Type	GO ID	GO Term
216	50	10^{-30}	F	GO:0003735	structural constituent of ribosome
264	57	10^{-29}	P	GO:0006412	protein biosynthesis
60	19	10^{-22}	P	GO:0006364	rRNA processing
62	19	10^{-21}	P	GO:0006365	35S primary transcript processing
46	14	10^{-16}	P	GO:0030490	processing of 20S pre-rRNA
39	12	10^{-14}	P	GO:0000027	ribosomal large subunit assembly and maintenance
91	19	10^{-12}	P	GO:0006468	protein amino acid phosphorylation
145	27	10^{-12}	F	GO:0003723	RNA binding
8	5	10^{-11}	P	GO:0006109	regulation of carbohydrate metabolism
25	8	10^{-11}	P	GO:0000074	regulation of cell cycle
14	6	10^{-10}	P	GO:0000183	chromatin silencing at ribosomal DNA
33	9	10^{-10}	F	GO:0003899	DNA-directed RNA polymerase activity
53	12	10^{-10}	F	GO:0004672	protein kinase activity

$p < 0.00001$ based on exhaustive permutation tests (required for each pair for the full model and expression reduced-model). This p-value was biologically reasonable in that we expect about 10 regulators per target gene — consistent with conventional wisdom and recent studies[6]. This resulted in 22,923 predicted interacting pairs yielding a modest false discovery rate of 1.7%. Finally, to avoid linkage disequalibrium effects, putative *cis-trans*-acting regulators (using a conventional p-value<0.05 cutoff) were excluded and regulator-target pairs residing on the same chromosome were removed. This filtering likely removed some true pairs, but a conservative selection was chosen in order to detect any group-wide trends that would be obscured by noise from false positives. After this filtering, the final set consisted of 4268 pairs.

I then considered the significance of each Gene Ontology (GO [3]) category in the "biological process" and "molecular function" ontologies with respect to the known GO assignments to the candidate regulators using the standard hypergeometric distribution test. Unlike previous reports, I found some highly significant classes shown in table 1. However, there was no enrichment among transcription factors or related activity, in agreement with Yvert et al.[60]. It is interesting, however, that there is enrichment in many different regulatory and control related activities, including cell cycle regulation, metabolism, and kinase activity. However, most enrichment is for functions and processes related to protein translation. Ribosomal proteins and related genes are well known to be highly co-expressed, but this analysis confirms the stronger claim that these genes are auto-regulated to a high degree[61].

Even though no functional enrichment in transcription factors was found, we still examined the predicted targets of transcription factors for evidence of physical interaction. Considering all the predicted targets of each transcription factor that met the selection criteria above, I searched 500bp upstream of the target for matches to known binding site motifs (TRANSFAC [34]). I found no significant enrichment for targets containing known binding regardless of sequence similarity thresholds. For example,

only 35 of 719 putative targets contained matches to known binding sites. And of those, only 8 were known targets of their respective transcription factor regulators.

In a final attempt to recover a bias for transcription factors, I hypothesized that QTGs associated with multiple target genes would be enriched for transcription factors. Those regulators were extracted from the set of 4268 pairs that had ten or more target genes. The set included well known transcription factors FKH1, FKH2, MSN1, KSP1, and ZAP1, but there was no significant enrichment in the total set for transcriptional regulators. All these observations further confirm that regulatory behavior captured in genotype/expression networks is not likely to be physical interactions, but more complex, indirect relationships as suggested by the functional enrichment found above.

5.2.2 Robustness

Next I wondered how well a causal relationship could be inferred when the regulator was part of a multifactor regulation. Using the yeast data set of $n = 6164$ genes, an $n + 1$ target gene was simulated according to an additive model of $k = 2 \ldots 5$ regulators, with only one regulator having genotypic effect. Specifically, I simulated

$$T_{n+1} = \beta_1 T_1 + \ldots + \beta_k T_k + \beta_{k'} T_k Q_k + \epsilon$$

where $\beta_{k'}$ was set at random values such that the genotypic effect between the two alleles, $(\mu_a - \mu_b)/\sigma$, was uniformly selected between 0.5 and 3.0. The other β's were selected from $\mathcal{N}(0, 1)$. Using the QTG *trans* model it was attempted to recover the *causal* regulator of the simulated target among the background of the other n genes. By modifying the full model threshold for equation 5.2, I obtained different trade-offs between recall and precision. It was found that the QTG model was successful in identifying the correct regulating gene, even for larger values of n (figure 5.5). Not surprisingly, conventional QTL mapping alone, being a function of only the flanking

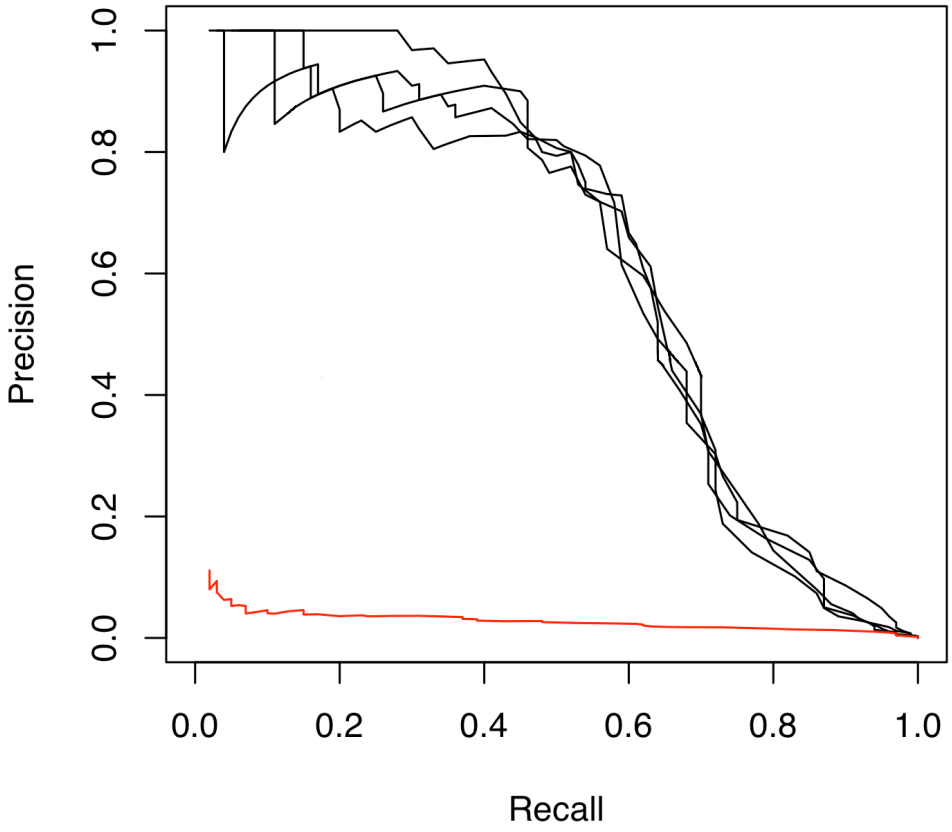


Figure 5.5. Performance of QTG model A plot of recall ($\frac{TP}{TP+FN}$) versus precision ($\frac{TP}{TP+FP}$) for varying full model scores. The red line corresponds to the performance of QTL mapping and each of the black lines correspond to performance of QTG mapping with varying number of regulators.

markers, failed to accurately predict the precise regulating gene, but the QTL interval was typically identified with reasonable success in the simulation cases.

5.2.3 Prediction of novel regulators

Six candidate QTL intervals analyzed in [7] representing highly pleiotropic loci were considered. The intervals were each predicted as containing a regulator gene associated with a large number of target genes, although the precise gene was unknown. In the paper, a putative gene within each interval was predicted manually by the authors according to published gene function annotations of regulator and target genes. For the most part, QTG analysis was disappointing in these cases; as it turned out, loci 1 through 5 were coincident with cis-acting QTLs. As a result, the gene expression of most of the putative regulators are highly linked and the manually predicted genes are no better fit than the neighboring genes.

However, a likely alternative regulator for the second of the six loci was predicted. The region on chromosome III represented a common QTL for 21 genes identified by Brem, who predicted that LEU2 was the putative regulator based on its similar function to these 21 target genes. But I identified ILV6, about 13kb from LEU2, as the more likely candidate. ILV6 is the best fit for the full and reduced models for 12 of the 21 genes with no other candidate gene showing significant fit for more than a few targets. Scanning the genome, I also found an additional five target genes not previously identified (table 2). This set of 17 putative targets of ILV6 are significantly enriched for genes associated with branched chain family amino acid biosynthesis (p-value 1.8×10^{-8}) and related amino acid metabolism GO terms. Moreover, ILV6 has been shown through direct assays to be part of the superpathway for leucine, isoleucine, and valine as the regulatory subunit of acetolactate synthase[15]. Thus, ILV6 and its targets are functionally related and its highly plausible that modulation of ILV6 directly affects the abundance of these other genes.

5.3 Discussion

In this chapter I presented the Quantitative Trait Gene model which improves upon interval mapping. This method infers instances of regulation where genotype and expression of the regulator interact while modulating the target.

Robustness of this method was tested through simulation and the results suggest that the regulatory relation could be recovered with a precision of 80% and recall of 60% (Figure 5.5). This performance remained fairly constant even with simulating larger networks.

The application of this model on a yeast cross returned numerous instances where both genotype and transcription variation of the regulator were explaining a large part of target gene's transcript variation 5.3 and the number of such instances was significantly higher than expected due to chance. The set of regulators found through my analysis was not enriched for transcription factors as I had expected 5.1 which indicates that the transcription regulation in yeast much more complex.

CHAPTER 6

RECOVERY OF LOCAL NETWORK

In this chapter I present an extension of the QTG model where, instead of restricting to a pre-determined graph, an arbitrarily complex regulatory network is constructed around a transcript of interest. This method is a two-step process: starting with a seed gene of interest, a Markov blanket over genotype and gene expression observations is inferred according to differential entropy estimation; a Bayesian network is then constructed from the resulting variables with important biological constraints yielding causally correct relationships.

This method was tested by simulating a five-node regulatory network within the background of of a real data set. It was found that 45% of the genes in a regulatory module can be identified and the relations among the genes can be recovered with moderately high accuracy ($> 70\%$). Since the sample size is a practical and economic limitation, I considered the impact of increasing the number of samples and found that recovery of true gene-gene relationships only doubled with an order of magnitude increase in samples, suggesting that useful networks can be achieved with current experimental designs, but that significant improvements are not expected without major increases in the number of samples. When this method was applied to an actual data set of 111 BXD mice cross, I was able to recover local gene regulatory networks supported by the biological literature.

The rest of the chapter is organized as follows. In section 6.1 the problem is concretely defined and methods for Markov blanket Inference and Bayesian Network construction are introduced. A new regulatory network inference method is presented

in section 6.2, and experiments are described in section 6.3.1. The results on application on a mice data-set is presented in section 6.3.2.

6.1 Local Networks from Expression genetics Data

In some experiments it is important to understand the regulatory neighborhood of a given trait. Here I present an improved Bayesian network reconstruction algorithm for facilitating this goal. In particular the contributions of my approach are:

- Regulatory modules, instead of global regulatory networks, are inferred, which mitigates some of the difficulties of BN structure inference when sample size is small relative to the number of variables;
- Genotype values and expression levels are modeled together in a single BN, which provides simultaneous integration of data types and the identification of different kinds of regulatory control;
- Multiple genes and genetic effects are considered together, rather than a single gene or a single QTL;
- Gene "self effects" are included, which incorporates the often significant effect of *cis*-acting polymorphisms;
- and the interacting effect between genotype and expression level is modeled (QTG model), which allows for complex regulatory behavior.

6.1.1 Markov blanket

The Markov blanket of a variable $X_s \in \mathbf{X}$ is defined as the minimal set of variables $MB \in \mathbf{X} - \{X_s\}$ that provide the maximum possible information about X_s. Knowing

the value of other variables outside of MB does not provide additional information. Formally,

$$\forall_{\bar{X} \subseteq \mathbf{X} - MB - \{X_s\}} (\bar{X} \perp X_s | MB).$$

In a Bayesian network, the Markov blanket is the union of parent, child and spouse (i.e. parents of children) nodes. In a gene regulatory network, the Markov blanket of a gene contains its regulators, targets and co-regulators. Thus, a Markov blanket of a gene of interest corresponds to the biological concept of a local gene regulatory module (figure 6.1).

Recovering the Markov blanket using raw data is well-studied in the context of feature selection[26, 56, 36]. Here I describe one particularly attractive approach.

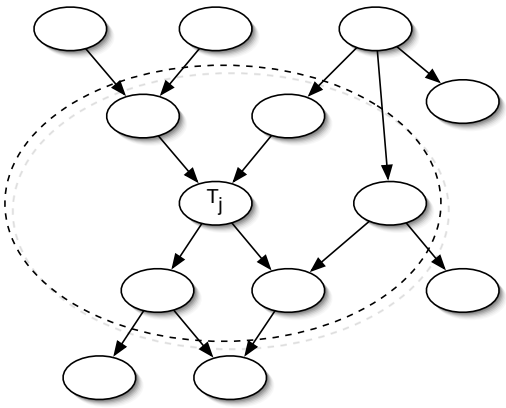


Figure 6.1. Local regulatory network The local regulatory network of a transcript T_j consists of its modulators (parent nodes), targets (children nodes) and co-modulators (spouse nodes). These nodes are inferred from the variable set containing other transcripts and genotypes.

6.1.1.1 Incremental Association Markov blanket

Incremental association Markov blanket (IAMB) is an information theoretical approach to infer a Markov blanket (MB) from data[56]. This is a two-step algorithm. In the first step, nodes are added to an interim MB^* based on a greedy search for variables that are not conditionally independent. Since it is a greedy algorithm some

Algorithm 1 IAMB algorithm

INPUT: Data: $X = \{X_1, X_2, \ldots, X_n\}$, Target: s, Threshold: θ **OUTPUT:** Markov blanket: MB

$MB = \emptyset$
repeat
  $i = \arg\max_{i \neq s} I(X_i; X_s | MB)$
  if $I(X_i; X_s | MB) > \theta$ **then**
    $MB = MB \cup \{X_i\}$
  end if
until MB does not change
repeat
  $i = \arg\min_{X_i \in = MB} I(X_i; X_s | MB - \{X_i\})$
  if $I(X_i; X_s | MB - \{X_i\}) < \theta$ **then**
    $MB = MB - \{X_i\}$
  end if
until MB does not change

nodes that should not be in the final MB might be present in MB^*. These nodes are removed in the second step through an exhaustive search of all subsets of MB^*. When the data set is faithful to the true distribution (i.e. empirical distribution is equal to the true distribution) and the measure of conditional independence is accurate, then this algorithm is guaranteed to give correct results. Usually conditional mutual information is used for measuring conditional independence [56, 36]. In practice the conditional independence test is deemed reliable only when the number of samples is at least five times the number of degrees of freedom. For discrete data this imposes a requirement of an exponential number of samples with respect to the number of variables in the conditioning set. However, when data is continuous and Gaussian distributed, as assumed here, then the number of required samples is only quadratic with respect to the number of variables in the conditioning set.

Conditional independence for continuous data can be computed using the differential entropies of the involved variables. Differential entropy is a relative measure that quantifies the amount of surprise (or information) of a continuous variable. It is equal to the expected log of the probability density.

$$\begin{aligned} h(x) &= E(log(f(x))) \\ &= \int_{-\infty}^{+\infty} f(x)\log(f(x))dx \end{aligned}$$

where f is the probability density function of x. For a multivariate Gaussian variable $X = \{X_1, X_2, \ldots, X_N\}$ differential entropy $h(X)$ is equal to

$$h(X) = \frac{1}{2} ln\{(2\pi e)^N det(\Sigma)\}$$

where Σ is the co-variance matrix of X. Conditional relative entropy is defined as the amount of surprise in one variable when the condition variable is known.

$$\begin{aligned} h(X|Y) &= E(log(f(X|Y))) \\ &= h(X,Y) - h(Y) \end{aligned}$$

Mutual information quantifies the amount of information that is contained in a random variable (X) about the other variable (Y). It is equal to the difference between the amount of information in one of the variables (which is entropy, $h(X)$) and the amount of information in it that is unexplained by the other variable (which is conditional entropy, $h(X|Y)$). Under condition Z it is equal to:

$$I(X;Y|Z) = h(X|Z) - h(X|Y,Z)$$

6.1.2 Bayesian Networks

Constructing Bayesian networks is a well-studied problem[20, 18, 12]. For a given network structure, the conditional probability distribution function of each variable can be calculated using maximum likelihood estimates. Using these functions, the posterior probability of the data can be calculated and a network can be scored. Let

$X = \{X_1, X_2, \ldots, X_N\}$ be the set of variables in the network. The posterior likelihood of an observation x is given by:

$$P(x) = \prod_{i=1}^{N} P(x_i|Pa(x_i), \Theta)$$

where $Pa(x_i|\Theta)$ is the set of parent nodes corresponding to node X_i and Θ is the hyper-parameter set determining the conditional probability distribution. For a data set $\mathcal{X} = \{x^1, x^2, \ldots, x^M\}$ the posterior likelihood is given by:

$$P(\mathcal{X}|\Theta) = \prod_{j=1}^{M} \prod_{i=1}^{N} P(x_i^j|Pa(x_i^j))$$

Log likelihood is used as the scoring function:

$$LL(\mathcal{X}, \Theta) = \sum_{j=1}^{M} \sum_{i=1}^{N} \log(P(x_i^j|Pa(x_i^j)))$$

Since the hyper parameter Θ is estimated using the finite number of samples, it is always possible to increase the log likelihood of a graph by increasing its connectivity. This over-fitting phenomenon can be be avoided by using a scoring scheme that takes connectivity into consideration. Bayesian information criterion (BIC, also known as Schwarz information criterion) is one such scheme.

$$Score_{BIC}(X, \Theta) = 2LL(\mathcal{X}, \Theta) - k \log(M)$$

where k is the number of free parameters in Θ. For linear Gaussian models k is equal to the total number of edges in the network.

Given that the possible network structure space is super-exponential with respect to the number of nodes, an exhaustive search through all possible graphs is usually not feasible. Reasonable heuristics like node ordering[18] can be used when the number

of samples is high and the number of variables is low. But those algorithms are infeasible when the number of dimensions is high and inaccurate when the number of samples is low. Another class of algorithms use information theory to construct these networks. A polynomial time algorithm exists[12] when an oracle, which determines if two variables are dependent conditioned on a set of variables, is available and the data is DAG-faithful. Such an oracle can be constructed by calculating conditional mutual information for the set of variables. But calculation of mutual information can be problematic when the number of samples is low, just as with the Markov blanket algorithms, as mentioned above, and when the number of variables is high. The proposed method overcomes this limitation by restricting to building local networks around the gene of interest. As the number of genes in the regulatory neighborhood of a gene is usually low, the network searching problem remains tractable.

6.1.3 Extending the QTG Model

The conventional model for mapping linkage of loci to phenotypes is a linear model of the form

$$P(T_i|Q_j) = \mathcal{N}(\beta_0 + \beta_1 Q_j, \sigma)$$

where T_i is the phenotype of interest (expression of a target gene) and Q_j are inferred genotypes of genes G_j along a chromosome.

In chapter 5, I suggested an alternative model that explicitly incorporated the genotype and expression level at gene G_j as well as the potential interacting effect of genotype and expression level, yielding

$$P(T_i|Q_j, T_j, \theta) = \mathcal{N}(\beta_0 + \beta_1 T_j + \beta_2 Q_j + \beta_3 T_j Q_j, \sigma) \tag{6.1}$$

where θ is the β and σ model parameters. (Equation 5.1.)

A scanning method, like conventional QTL mapping, can be used in which pairwise relationships are found by computing the log posterior odds for all G_j in the

genome. Equation 6.1 has the advantage of capturing complex dependency relationships. However, the scanning method does not incorporate multi-locus regulatory control.

6.2 Methods

Now I present an algorithm that finds the loci that are in the regulatory neighborhood of a gene of interest and reconstructs the corresponding partial network. The main advantage of this new method over QTG scanning method[29] is here I construct networks involving multiple genes to specifically model the joint distribution, whereas the previous approach could only identify putative pairwise relationships akin to a relevance network[10].

6.2.1 Mixed Type Bayesian Network Under Biological Constraints

We model a gene regulatory network as a highly constrained Bayesian network subject to the biological conditions as graphically described in Figure 6.2. A "gene" is modeled as a meta-node, such that a node (G_a) consists of expression (T_a), genotype (Q_a) and interaction (T_aQ_a) variables (Figure 6.2a). Edges denote regulation between genes where edges are drawn from the regulator meta-node to a target meta-node. The kind of regulatory control between two genes depends on which terms in the meta-nodes were used (Figure 6.2b). Since genotypes are determined by chance during the meiosis, it is implausible that phenotypes are causally upstream of genotypes. Therefore, whenever a direct relation is found between a genotype and a phenotype node, the edge is always directed away from genotype.

6.2.2 Markov blanket Inference

Algorithm 2 for inferring a Markov blanket is very similar to the IAMB algorithm with several domain specific differences. The candidate variable set C consists of all *gene expression values* ($T_i, 1 \leq i \leq n$, where n is the number of genes), all *marker*

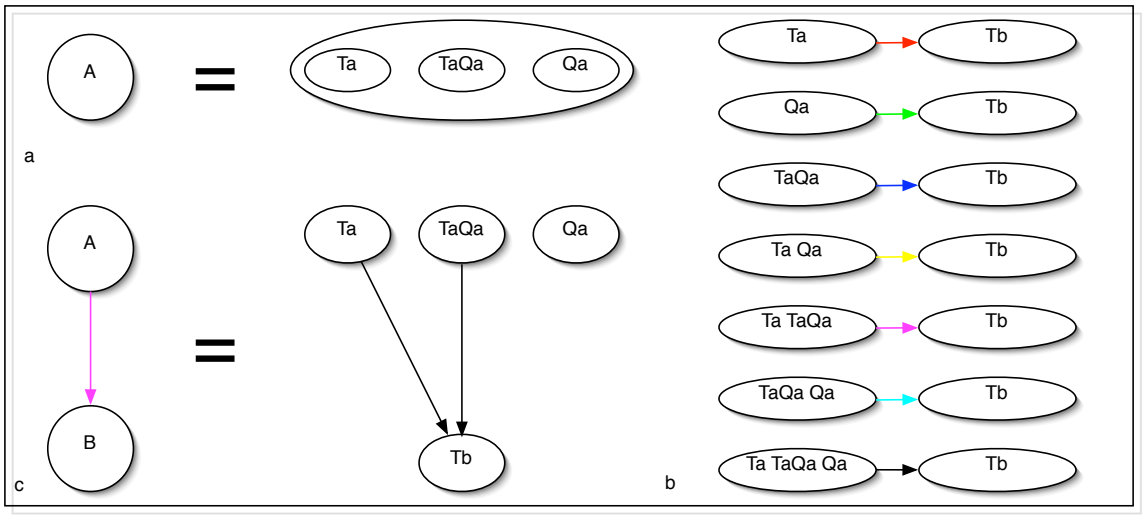


Figure 6.2. Modeling regulatory relation between genes (a) Elements of gene A. T_a is the expression, Q_a is the genotype and T_aQ_a is the interaction variable. (b) All edge types. Colors are used to visually code predicted networks (such as in figure 6.5. (c) Example of gene-gene relationship with two edge types involved.

genotypes ($M_j, 1 \leq j \leq k$, where k is the number of polymorphic markers) and *approximate interacting terms* estimated from the product of expression and flanking marker genotypes (where TQ_i^l is used to mean $T_iM_{L(i)}$, TQ_i^r to mean $T_iM_{R(i)}$, and $M_{L(i)}$ and $M_{R(i)}$ are the flanking left and right markers of gene G_i). In the forward step, based on conditional independence, variables from C are incrementally added to the Markov blanket MB and in the backward step false positives are removed. A continuous form of conditional mutual information (as explained in section 6.1.1.1) is used as the measure of conditional independence. Variables are assumed to follow a multinomial Gaussian distribution. With the reasonable biological assumption that any gene has no more than about ten genes in its local regulatory network[6], then only (≈ 100) samples are required to accurately calculate conditional mutual information.

Algorithm 2 Inferring Markov blanket of a gene. I calculates the conditional mutual information as described in section 6.1.1.1. Functions *max* and *min* return maximum/minimum element in the array and its index.

INPUT: Expression Levels: $T = \{T_1, T_2, \ldots, T_n\}$,
Marker Genotypes: $M = \{M_1, M_2, \ldots, M_k\}$,
Interaction terms: $I = \{TQ_1^l, TQ_1^r, \ldots, TQ_n^l, TQ_n^r\}$,
Seed Gene s, Threshold α
OUTPUT: Markov blanket $MB \in T \cup M \cup I$

1: MB=$\emptyset$
2: C=$(T \cup M \cup I) - \{T_s, TQ_s^l, TQ_s^r\}$
3: **repeat**
4: **for** $C_i \in C$ **do**
5: $score_i = I(C_i; T_s | MB)$
6: **end for**
7: $[maxI, maxi] = max(score)$
8: **if** $maxI \geq \alpha$ **then**
9: $MB = MB \cup \{C_{maxi}\}$
10: **end if**
11: **until** $maxI < \alpha$
12: **repeat**
13: **for** $C_i \in MB$ **do**
14: $score_i = I(C_i; T_s | MB - \{C_i\})$
15: **end for**
16: $[minI, mini] = min(score)$
17: **if** $minI < \alpha$ **then**
18: $MB = MB - \{C_{maxi}\}$
19: **end if**
20: **until** $maxI < \alpha$
21: **return** MB

6.2.3 Gene regulatory network reconstruction

Here an incremental algorithm similar to Cooper et al.[14] for constructing the local network for a seed gene, s (Algorithm 3) given its Markov blanket, MB_s is used. The novelty of this method is that the unobserved genotype values Q_i must be simultaneously be estimated while constructing the graph edges.

I begin with an MB_s that contains zero or more expression and genotype terms (e.g. T_i, TQ_i^r, etc.) for each gene G_i. The regulatory neighborhood of seed gene s the defined as $RN_s = MB_s \cup \{T_s\}$. For all genes with a flanking marker in the MB_s I introduce the unobserved genotype Q_i and estimate its maximum likelihood value according to the distances to the flanking markers. Similarly any TQ_i^l and TQ_i^r terms are replaced with TQ_i.

Next, the variables in RN_s are consolidated into gene meta-nodes, such that all variables associated with gene G_j are grouped. Then, beginning with an empty graph, edges are added, removed, or reversed between variables in separate meta-nodes based on an increase in the network score. Unlike a conventional Bayes Net construction, I explicitly consider combined genotype and expression effects including interacting effects. These different kinds of regulatory effects are represented as different types of edges (figure 6.2b). The score is computed as the log of the joint probability with a Bayesian Information Criterion (BIC) penalty term to control for complexity of the network.

Finally, the Q_i terms are re-estimated based on the new graph structure (connected genes and flanking markers). With the new values of Q_i, a new graph structure is generated. This EM-like iterative process is repeated until convergence, which happens quickly in practice.

Purely genetic hyper-nodes are an interesting special case. In some cases a marker variable M_i might not have a gene in MB_s that it can be grouped with. In those cases a dummy gene hyper node is created for this marker. These dummy genes are

Algorithm 3 Algorithm for constructing local regulatory network. *EstimateGenotype* function estimates the genotype of a locus by using the genotypes of the flanking markers and the distance to those markers. *Score* calculates the optimal score of a network using EM strategy. In expectation step all the Qs and TQs are estimated using the current value of hyper parameter set (Σ) and their priors. Later in maximization step the Σ is re-calculated using the re-estimated values of Qs and TQs. *AddScore* is the score of the new network when a edge is added, reversed or removed. This function also checks for DAG consistency of the network and if that is violated returns $-\infty$. *getPossibleEdges* returns the set of possible edges ($edge = \{from, to, kind\}$) in the network depending on the contents of hyper-nodes. *from* can be any node, *to* node needs to have expression term in it and *kind* can be any kind of edge shown in 6.2 or of kind *no edge* (used when an edge needs to be deleted).

INPUT: Markov blanket MB_s,
Expression profiles: $T = \{T_1, T_2, \ldots, T_n\}$,
Marker Genotypes: $M = \{M_1, M_2, \ldots, M_k\}$,
Interaction terms: $I = \{TQ_1^l, TQ_1^r, \ldots, TQ_n^l, TQ_n^r\}$
Seed Gene s, Threshold β
OUTPUT: Local Network BN_s

$RN_s = MB_s \cup T_s$
for each gene i **do**
  $Q_i = EstimateGenotype(M_{Leftmarker(i)}, M_{Rightmarker(i)}, Location(i))$
end for
for each gene i **do**
  $G_i = \{T_i, T_iQ_i, Q_i\}$
end for
$CG = \{G_i | T_i \in RN_s \vee T_iQ_i^l \in RN_s \vee T_iQ_i^r \in RN_s\}$
$BN_s = \emptyset$
$curMaxScore = Score(BN_s, CG)$
while forever **do**
  $\{from, to, kind\} = argmax_{\{from,to,kind\} \in getPossibleEdges(RN_s)}$
  $Addscore(BN_s, \{from, to, kind\}, CG)$
  if $AddScore(BN_s, \{from, to, kind\}, CG) - curMaxScore > \beta$ **then**
    if $\exists \overline{kind}$ $s.t$ $\{from, to, \overline{kind}\} \in BN_s$ **then**
      $BN_s = BN_s - \{\{from, to, \overline{kind}\}\}$
    end if
    if $\exists \overline{kind}$ $s.t$ $\{to, from, \overline{kind}\} \in BN_s$ **then**
      $BN_s = BN_s - \{\{to, from, \overline{kind}\}\}$
    end if
    $BN_s = BN_s \cup \{\{from, to, kind\}\}$
  else
    return BN_s
  end if
end while

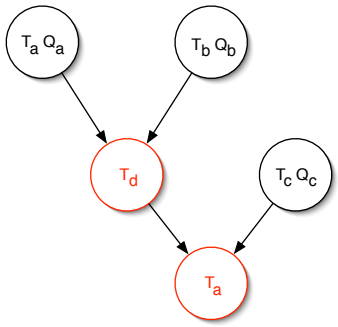


Figure 6.3. Simulation strategy for local network inferencing Black nodes were selected from the existing data and the red nodes were simulated using a linear Gaussian model.

assigned a range of locations (determined using the location of markers M_{i-1} and M_{i+1} that flank M_i) instead of having one exact location as with regular gene hyper nodes. During the network optimization the exact location of this dummy gene is re-calibrated to maximize the score. This strategy allows us to detect genetic elements that are either not associated with any of the known genes. Such effects include, for example, *cis*-acting QTLs and non-coding genes.

6.3 Experiments and Results

Simulations were performed to test the fidelity of the model, to set appropriate threshold parameters, and to calculate the sample size needed to achieve good accuracy and recovery.

6.3.1 Simulations

Synthetic data was generated to test the viability of this approach. To keep the simulation as realistic as possible and to preserve the distribution of the real data, only a small set of simulated data was added to the existing data. Networks of various size were simulated. Importantly, parent and spouse genes were not simulated, but selected from existing genes. Target genes and their children were simulated using

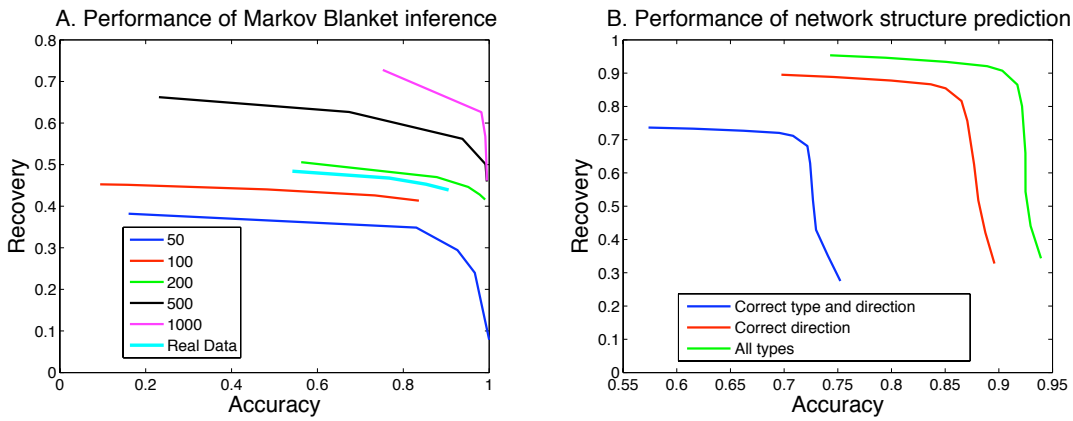


Figure 6.4. Performance of local network inferencing A. Accuracy vs recovery plot for classification of variables in Markov blanket of candidate seed gene. Different lines show results for different sample sizes. B. Accuracy vs recovery plot for graph reconstruction using different graph evaluation criteria.

a linear model with Gaussian noise. A example of such simulated network is shown in Figure 6.3. The coefficients of this linear model were selected from a Gaussian distribution. To test the data requirement for sample sizes greater than the available 111 samples, I simulated additional expression values as Gaussian and genotypes from linkage probabilities.

Results of these simulations are presented in Figure 6.4 for a 5 node network. (For network sizes greater than 5, accuracy did not decrease substantially and the number of recovered genes remained almost the same; data not shown.) Figure 6.4a describes the performance of the Markov blanket recovery. Each line in the figure corresponds to a sample size. Results suggest that this algorithm can recover parts of the network with high accuracy at useful recovery rates. For example, greater than 45% of genes in the true Markov blankets were recovered at an accuracy of about 75%. Reducing the threshold did not result in increased recovery but caused accuracy to drop substantially. When I increased sample size to one thousand (ten times the current available data) there was a marked improvement in recovery($> 75\%$) and accuracy($> 85\%$).

Figure 6.4b describes the performance of network inference, i.e. edge prediction, over the Markov blanket variables. Considering only gene meta-node connectivity, the algorithm exceeded 90% accuracy and 90% recovery for the correct placement of edges. When the correct direction is also taken into account, accuracy of 85% could be achieved with recovery of about 85%. Edges of correct direction and correct edge type could be recovered with 70% accuracy and 70% recovery. Thus, a quite reasonable reconstruction of a network could be achieved with a large majority of edges properly labeled and oriented.

6.3.2 Biological Significance

For practical experimental results I used data collected by Schadt et al.[43], consisting of gene expression profiles for 111 F_2 mice derived from crossing C57BL/6J and DBA/2J. The data-set contains expression for 23,574 genes and genotypes for 134 markers spread over 19 chromosomes.

When this algorithm was applied to construct local networks seeded by 400 highly cited mouse genes in PubMed database, under the assumption that well-annotated seeds are more useful when performing a manual, qualitative review of predicted regulatory networks. Many networks found which were enriched for shared functions. Several of these networks are shown in Figure 6.5 with the biological interpretations and analysis. The inferred local regulatory network of *Dlx2* is shown in figure 6.5a. Three of the genes in the network, *Dlx2*, *Aebp1* and *Dnmt3a*, are known transcription factors. This indicates that these genes might be involved in a transcriptional cascade. The local regulatory network of *Rela* (figure 6.5b) contains *Mapk1* and both of these are involved in organ morphogenesis. *Rela* seems to be regulating *Usmg5*, which is involved in skeletal muscle growth, which suggests that *Rela*'s role is skeletal muscle growth. The inferred local regulatory network of *Pcna* (figure 6.5c) suggests that *Pcna* and *Dmap1* might be co-regulating *Prim1*. This is interesting as these two

genes are known to interact with similar domains[33]. The local network of *Fgfr2* (figure 6.5d) is interesting in many ways. Biologically this network makes sense as there is reasonable functional overlap among the genes n the network. *Fgfr2* and *Ptk2* are involved in regulation of actin cytoskeleton. *Fgfr2*, *Ptk2* and *Gnaq* are all nucleotide binding proteins. This network is also interesting computationally as we can predict the causality of this network though there are no genetic variables. In this network all the genes are well correlated with the seed gene, but *Ptk2* and *Ppt* are uncorrelated. This is the only network that is able to capture these informational dependencies accurately[57].

6.4 Discussion

In this chapter I presented an extension of the QTG model for analyzing regulation involving multiple genes as a directed acyclic graph. In this study I investigated the use of an information theoretic method for accurately constructing local gene regulatory network from a seed gene. This model allows use of both expression and genotype in the same network thereby exploiting the natural dependencies. The method combines conventional quantitative genetic mapping and model-based network inference in one unified algorithm compared to approaches where genetic analysis is done first and results are used to refine genomic study results.

The simulation results suggest that reasonably accurate small networks can be constructed using this approach. Importantly, it was also found that small sample size is the most important limitation on the utility of these data sets. The simulation study also suggests that an order of magnitude increase in number of samples is needed to identify reliable and complete gene regulatory networks, but such large experiments are impractical in the near term.

A brief analysis of the local networks that are constructed around some well known genes suggest that the proposed method is capable of recovering biologically relevant

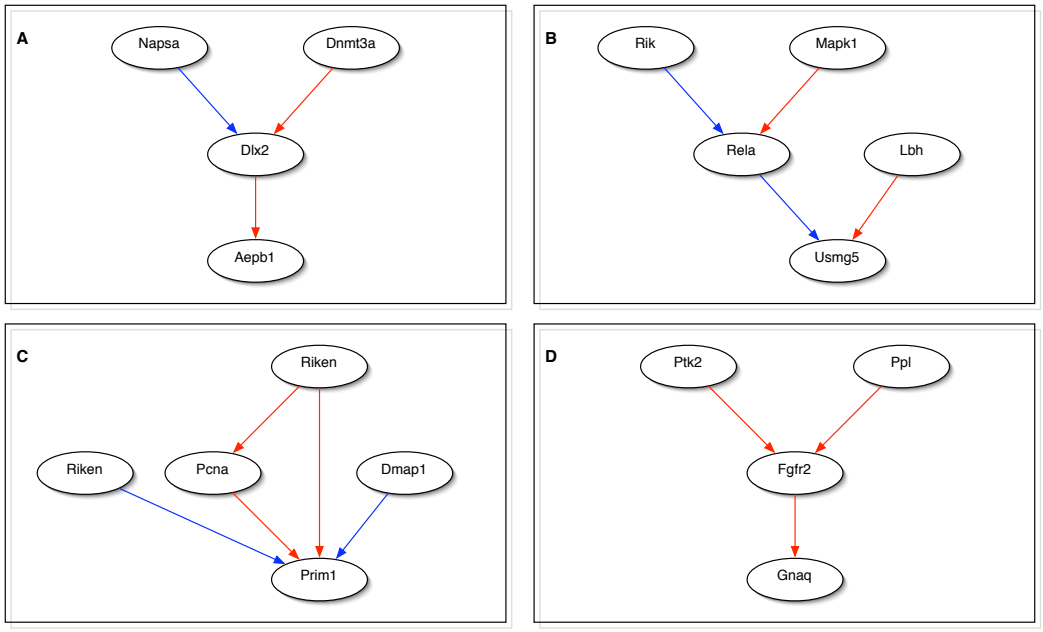


Figure 6.5. Sample local regulatory networks A. *Dlx2*. *Dlx2*, *Aebp1* and *Dnmt3a* are expressed inside nucleus and they are transcription factors. *Napsa* and Aebp1 are involved in proteolysis. B. *Rela*. *Mapk1*, *Rela* and *Lbh* are nuclear proteins. *Mapk1* and *Rela* are involved in organ morphogenesis and *Usmg5* is up-regulated during skeletal muscle growth. This suggests that *Rela* might be involved in a signaling cascade that controls muscle growth. C. *Pcna*. *Pcna* and *Dmap1* are known protein and DNA binding proteins. They are also known to interact with similar domains [33]. D. *Fgfr2*. In this network, though there is no genetic component we can still predict that this graph is causally correct. In this graph *Fgfr2* is well-correlated with other genes in the network: *Ptk2*, *Ppl* and *Gnaq* (with correlations of 0.53, 0.70 and 0.62). But *Ptk2* and *ppl* are uncorrelated (with correlation 0.16). These numbers suggest that this is the only plausible network that can be constructed with these genes [57].

networks from the expression genetics data. Most of the networks have edges between the genes that are known to be functionally similar and/or are active in the same cellular locations.

CHAPTER 7

ANALYSIS OF GENETIC HOTSPOTS

Interval mapping of genomewide expression data often reveals existence of highly localized regions that appear to regulate a number of transcripts. It is not always clear whether these aggregations reflect real biology or whether they may be due to experimental artifacts. In this chapter I present a method that analyzes such hotspots and presents plausible regulatory elements responsible for such phenomenon. In this method conditional independence analysis is applied to determine which transcripts are directly modulated by allelic variation and which transcripts can be explained by another transcript acting as a causal intermediary. Using simulated data it is shown that it is possible to reliably detect these primary transcripts even when it is difficult to elucidate the entire network of interactions. This method was applied to data from a mouse intercross population to characterize a number of prominent hotspots. In most cases it was found that a small set of local transcripts are primary and that each influences a non-overlapping set of downstream transcripts. However one case was identified in which a single distant transcript was acting as a causal intermediate for most of the transcripts in the hotspot. Functional analysis of groups of transcripts that are downstream from a common primary transcript revealed some biologically interesting enrichments.

7.1 Genetic Hotspots

Expression genetics studies are now being conducted in many organisms to identify genetic elements that influence genomewide transcriptional profiles and to infer their

causal role in determining phenotypes and diseases. Transcript abundance, measured using a microarray or similar platform, can be mapped as a quantitative trait in a genetically variable population. Mapped loci associated with variation in transcript abundance are referred to as expression quantitative trait loci (eQTL) [24]. In many of these studies it has been observed that highly localized regions of the genome seem to regulate an unexpectedly large number of transcripts [7, 43, 59]. These transcripts are often associated with functionally coherent sets of genes involved in common biological processes. There are several plausible mechanisms that would give rise to such hotspots such as presence of a pleiotropic gene [7, 59]. The simplest model for a hotspot (Figure 7.1 A) would consist of a *master locus* with a polymorphic variation that directly affects numerous downstream transcripts. The master locus could be a transcription factor but this is not necessary [7]. A more general *hierarchical model* (Figure 7.1B) assumes that polymorphism at the hotspot, influences the expression of one or more primary transcripts, which in turn, mediate the affects on multiple secondary transcripts. Given a location on the genome and a list of all transcripts linked to that locus, the objective is to identify the primary transcripts that are the immediate downstream targets of the local polymorphisms. We also aim to identify groups of secondary transcripts, which are modulated by a common primary transcript, and characterize common functional features of these groups.

Brem et al. [7] identified eight hotspots in a segregating population of yeast and for six of these they were able to identify putative regulator gene based on annotation of transcripts linked to these loci. In Schadt et al.[43] the authors discovered seven hotspots in a mouse intercross that each accounted for more than 1% of all eQTLs. In Wu et al. authors were able to find many hotspots in a diverse panel of inbred mice, that were linked to functionally coherent sets of transcripts, and they were able to relate these to the known transcriptional factors [59].

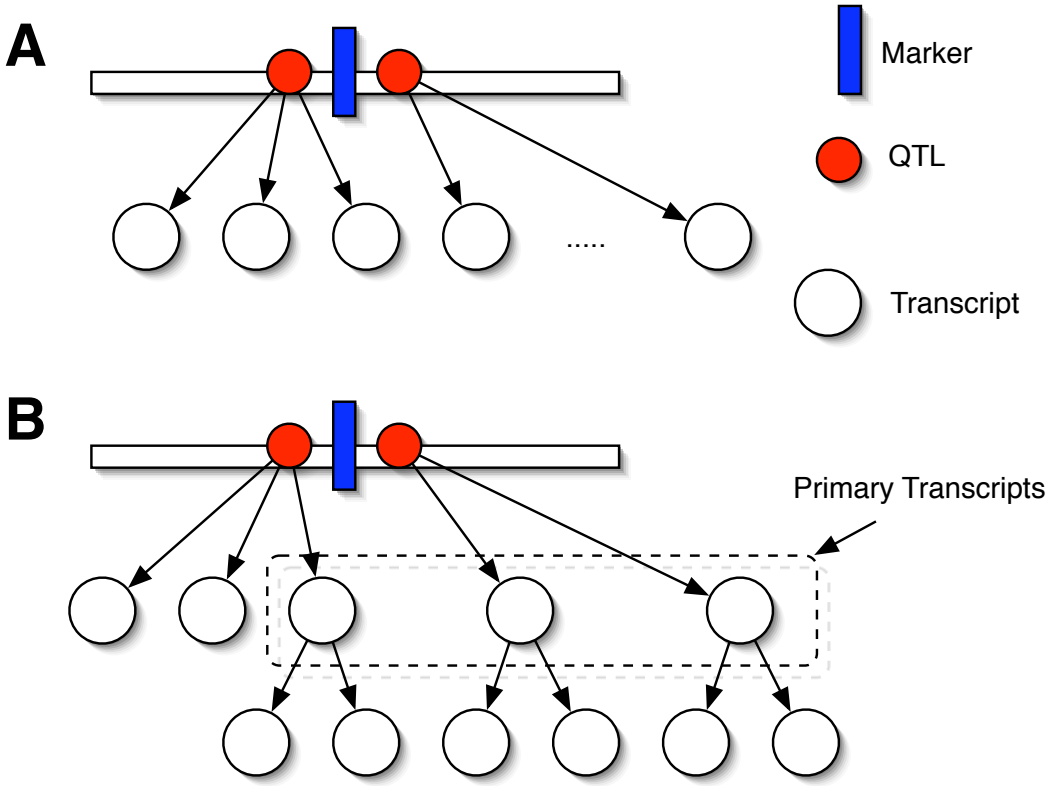


Figure 7.1. Genetic Hotspot models A. In the master locus model, polymorphisms linked to a single marker are directly regulating many transcripts. B. In the hierarchical model local polymorphisms are regulating a few primary transcripts and these in turn regulate additional secondary transcripts.

However subsequent work by Breitling et al. [5] suggest that some of the apparent hotspots may be artifacts due to confounding factors introduced during the microarray experiment. Perez et al. [39] used simulated genotypes in combination with real expression data to show that apparent hotspots can occur by chance. Wang et al. [58] support this result by showing that the expression of the transcripts in a hotspot can be highly correlated and that removing these correlated transcripts results in reduced numbers of hotspots. They suggest pursuing hotspots where the correlation among transcripts is low. Peng et al.[38] suggests calculating residuals for each transcript with transcripts of stronger linkage as covariates and consider only the transcripts whose residuals are also linked to the allele. These latter two approaches ignore the fact that two independently regulated transcripts can still be highly correlated due to shared genetic effects (correlation does not always imply causation). Furthermore a strongly linked transcript can act as surrogate for the allelic state of the linked locus and make the residual information less significant. Also it has been observed that strongly linked transcripts can correspond to artifacts such as polymorphism in the probes [2]. A more general algorithm to remove systemwide correlation was proposed by Kang and Eskin [25].

There is no definitive method to determine whether an apparent hotspot is due to a spurious unknown confounding factor or due to a pleiotropic polymorphism. In this paper we try to understand the correlation structure among the transcripts linked to the hotspot and predict a putative regulatory cascade that explains the data. In specific a set of *primary transcripts* which are (a) directly modulated by the hotspot, and (b) in turn, modulate multiple transcripts, are inferred.

7.2 Methods

Two random variables (X and Y) are said to be independent ($X \perp Y$) if knowing the state of one does not provide any information about the state of the other, other-

wise the variables are said to be dependent($X \not\perp Y$). If two variables are dependent, but knowledge of the state of a third variables (Z) makes this information redundant then the variables are said to be conditionally independent given the third variable ($X \perp Y|Z$).

In this setting we consider a genotype variable that defines the allelic states of a locus (Q) and two correlated transcript abundances T_i and T_j that are both linked to Q. When these conditions are met, there are four possible relationships among these variables:

(i). Q is modulating T_i and T_j independently in which case $T_i \perp T_j|Q$. In this case T_i and T_j are correlated due to the common influence of Q but additional variation in T_i and T_j that is not due to Q will be uncorrelated

(ii). Q is modulating T_i which is in turn modulating T_j. In this case $Q \perp T_j|T_i$ and the correlation of T_j with Q is entirely explained by variation in T_i

(iii). Q is modulating T_j which is in turn modulating T_i. In this case $Q \perp T_i|T_j$

(iv). the relationship is complex involving mutual dependence among all three variables or involves unknown latent variable.

The conditional independence relations described in (i)-(iii) each provide insight into causal relationships among the transcripts. If a transcript (T_i) is a causal intermediate between a genotype (Q) and another transcript (T_j) such that $Q \perp T_j|T_i$ then we say that T_i *shields* T_j. If a transcript shields one or more other transcripts that are linked to a common hotspot but itself is not shielded then we say it is a *primary transcript* and the transcripts that are shielded by *primary transcript* are referred to as *secondary transcripts*.

7.2.1 Testing Conditional Independence Relationships

Conditional independence implies that the joint distribution is the product of the marginal distributions, thus

$$X \perp Y|Q \Leftrightarrow P(X,Y|Q) = P(X|Q) \cdot P(Y|Q)$$

If the probability distributions of the variables are known then just checking for the above condition would be sufficient to infer conditional independence. However, when the probability distributions are estimated from the data the above conditions for independence will never be precisely met due to random variation. Therefore we consider the difference between the joint distribution and the product of the marginal distributions and if the difference is sufficiently small, conditional independence is inferred. One widely used measure of the difference between two probability distributions is Kullback-Leibler divergence [28] which is equivalent to the conditional mutual information.

$$I(X;Y|Q) \quad = \quad \sum_{q} \int_{x,y} P(x,y,q) \cdot \log_2 \frac{P(x,y|q)}{P(x|q) \cdot P(y|q)} dxdy$$

We will assume that the distribution of transcripts is Gaussian and that Gaussian mixtures define the relation between transcript levels and genotypes. With these distributional assumptions we can compute conditional mutual information from data. In this work, we will use base 2 logarithm, thus mutual information will be measured in *bits*.

For any set of three variables that are pairwise dependent ($X \not\perp Y$, $X \not\perp Q$ and $Y \not\perp Q$) at most one of the three possible conditional independence statements can be true. For example, $X \perp Y|Q$ and $X \perp Q|Y$ cannot be true simultaneously if $X \not\perp Y$, $X \not\perp Q$ and $Y \not\perp Q$. Thus we can use the criterion

$$X \perp_\delta Y|Q \Leftrightarrow I(X;Y|Q) < min(\delta, I(X;Q|Y), I(Y;Q|X))$$

to establish which, if any, of the conditional independence statements is true. The sensitivity of this criterion can be modulated through the choice of δ. A lower value of δ will result in fewer inferred conditional independencies. For our analysis we are going to consider a threshold equivalent to the Bayesian Information Criterion (δ_{BIC}).

7.2.2 Detecting Primary Transcripts

Given a marker (Q) we first identify all transcripts ($T_1, T_2, \cdots, T_k$) that are linked to that marker. We used standard genome scans [30] applied to each transcript to establish linkage. Significant linkages are determined by permutation tests [13].

To identify the primary transcripts that are being directly modulated by the allelic variation near the marker, we created a binary *shield matrix* (Figure 7.2A and Figure 7.5) whose values indicate when one transcript shields another transcript. The shield matrix elements are

$$S_{ij}^{\delta} = \begin{cases} 1 & if \qquad T_j \perp_\delta Q|T_i \\ 0 & otherwise \end{cases}$$

Transcripts (T_i) that are not shielded themselves, but shield a significant number of other transcripts are determined to be *primary transcripts*. We restricted our attention to primary transcripts that shield 10% or more of the total number of transcripts linked to a given hotspot. The shield matrices were computed for a range of values of δ and the results are summarized using a *variable threshold plot* (Figure 7.2B and Figure 7.6). Each line on the plot corresponds to an unshielded transcript. The X-axis of this plot is threshold δ and the Y-axis is the number of secondary transcripts shielded.

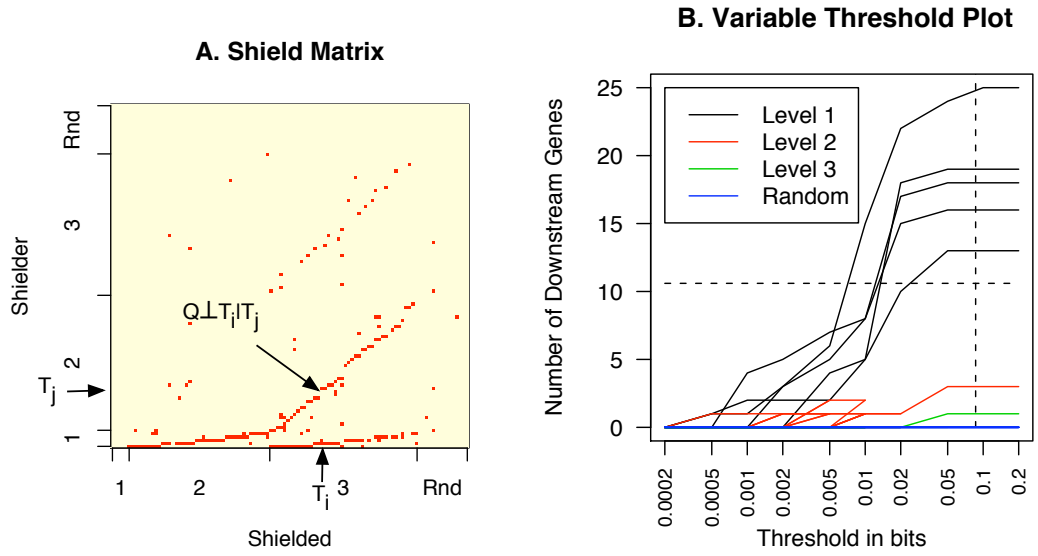


Figure 7.2. Analysis of a simulated network The shield matrix (A) was computed for the simulated data using threshold $\delta_{BIC} = 0.0863$ bits. Nodes in the simulated network are ordered on both axes according to their depth, where 1,2,3 indicates primary, secondary and indirect secondary nodes respectively. Random nodes are included at right and top of the axes. Red point indicates that X-axis node is being shielded by the Y-axis node ($Q \perp T_i | T_j$). In simulation the structure of the graph is known and we can arrange the nodes according to their level. With real data this is not possible and we will arrange transcripts along the axis according to their location on genome. The variable threshold plot (B) shows the number of *downstream* transcripts as a function of threshold δ, i.e. the number of transcripts a particular transcript shields for some δ. Each line on the plot corresponds to a transcript and the line is drawn for only the range of δ over which it is unshielded (i.e. the line is terminated at the δ at which this transcript is shielded by another transcript) . The dotted horizontal line corresponds to 10% of total number of nodes found to be linked and the vertical line corresponds to δ_{BIC}. The five black lines correspond to primary nodes in the simulated network.

7.2.3 Functional Analysis

The sets of secondary transcripts shielded by each primary transcript were analyzed for functional enrichment using DAVID [16] and the top functional groups are mentioned in the experimental results section (Section 7.3.2).

7.2.4 Simulations

To test this algorithm we carried out a simulation by generating data that are similar to real data in size and structure, but with a known network of conditional relationships embedded (Figure 7.3). Briefly, we selected one location from the genome to be a QTL that directly affects 5 transcripts (Level 1), each of which affect 10 transcripts (Level 2. Total number: 50) directly and in turn, each of these affect 2 transcripts(Level 3. Total number: 100) (Figure 7.3). *Level 1* transcripts are being directly modulated by the QTL and they are the set of *primary transcripts* we are aiming to recover. *Level 2 and 3* transcripts are essentially *secondary transcripts*. Although it in possible to further analyze *secondary transcripts* to classify them into *Level 2* and *Level 3* transcripts, we do not attempt to do so here. The remaining 16,844 transcripts are drawn independently from a Gaussian distribution.

A linear model with Gaussian noise was used to model the values of descendants.

$$T_i = c_i T_{parent(i)} + \epsilon\sqrt{1 - c_i^2}\epsilon$$

where ϵ is white noise sampled from $\mathcal{N}(0, 1)$, and *parent* of each transcript can be looked-up from the network structure shown in Figure 7.3. For example the primary transcripts were simulated as:

$$T_i = c_i T_1 + \epsilon\sqrt{1 - c_i^2}, \text{For } i \text{ in } 2, .., 6$$

and so on.

We sample the values of c from the distribution of correlations between primary and secondary transcripts we discovered from real hotspots.

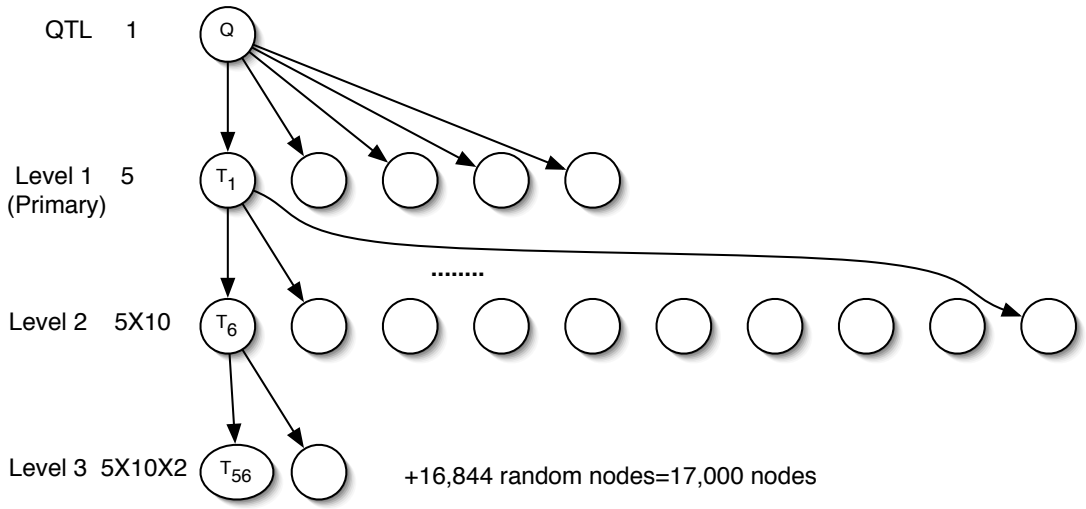


Figure 7.3. Network model used in genetic hotspot simulations A total of 156 transcripts were sampled from this tree-structured Bayesian network. The correlation for each linear model was sampled from the correlation between primary and secondary transcripts in real data.

7.3 Results

7.3.1 Simulations

From 100 instances of our simulation with $\delta_{BIC} = 0.0863$ we found that the primary transcripts were found with 84.8%($\pm$13.06) sensitivity and 100.0% specificity, which is to say that most of the 5 simulated primary transcripts were recovered and no secondary transcripts were misidentified as primary. Lowering the correlation, c, in our model reduced sensitivity but not specificity. More generally, the 155 dependent primary and secondary transcripts were found with 55.7%($\pm$6.22) sensitivity and 83.6%($\pm$4.04) specificity. This indicates that all primary and most secondary dependent transcripts can be identified, but some additional false associations with additional transcripts are found. But these false discoveries have unique character-

istics. As an example, figure 7.2A show the shield matrix for one instance of the simulation (delta=0.1), which shows that primary transcripts are not shielded by others but do shield downstream transcripts. The primary (level 1), direct secondary (level 2), and indirect secondary (level 3) are ordered by index according to figure 3 so that stepwise lines can be seen indicating the parent-descendent relationships from the simulated dependency tree. In addition, "random" transcripts are shown in the shield matrix, which are falsely discovered transcripts, i.e. those identified as being part of the network, but were not. Note, that these false positive transcripts are rarely being shielded and none shield others.

The threshold of $\delta_{BIC} = 0.0863$ clearly separates the primary transcripts from others without falsely introducing a large number of dependencies among the secondary transcripts and introducing no false parent relationships to the random transcripts that do not belong in the network to begin with.

7.3.2 Data Example

We identified a total of 10,784 significant (LOD score > 3.2) linkages involving 3,945 transcripts between the 173 markers and the 16,463 transcripts (Figure 7.4A). Of these 6,818 were *local linkages* (i.e transcript and marker are located on the same chromosome) and 3,968 *distant linkage* (i.e transcript and marker are located on different chromosomes). The number of linkages per marker was computed separately for distant and local linkages (Figure 7.4). The distribution of distant linkages is concentrated in a few hotspots whereas local linkages are more evenly distributed. In total, 21 out of 173 markers were found to have significant number of distant linkages ($p < 0.001$).

We selected four hotspots, on chromosomes 1, 5, 14 and 18, each of which had more than 100 distant linkages, for further analysis. Once we identified hotspots based on this strategy, we used permutation tests (described in 7.2.2) to refine the

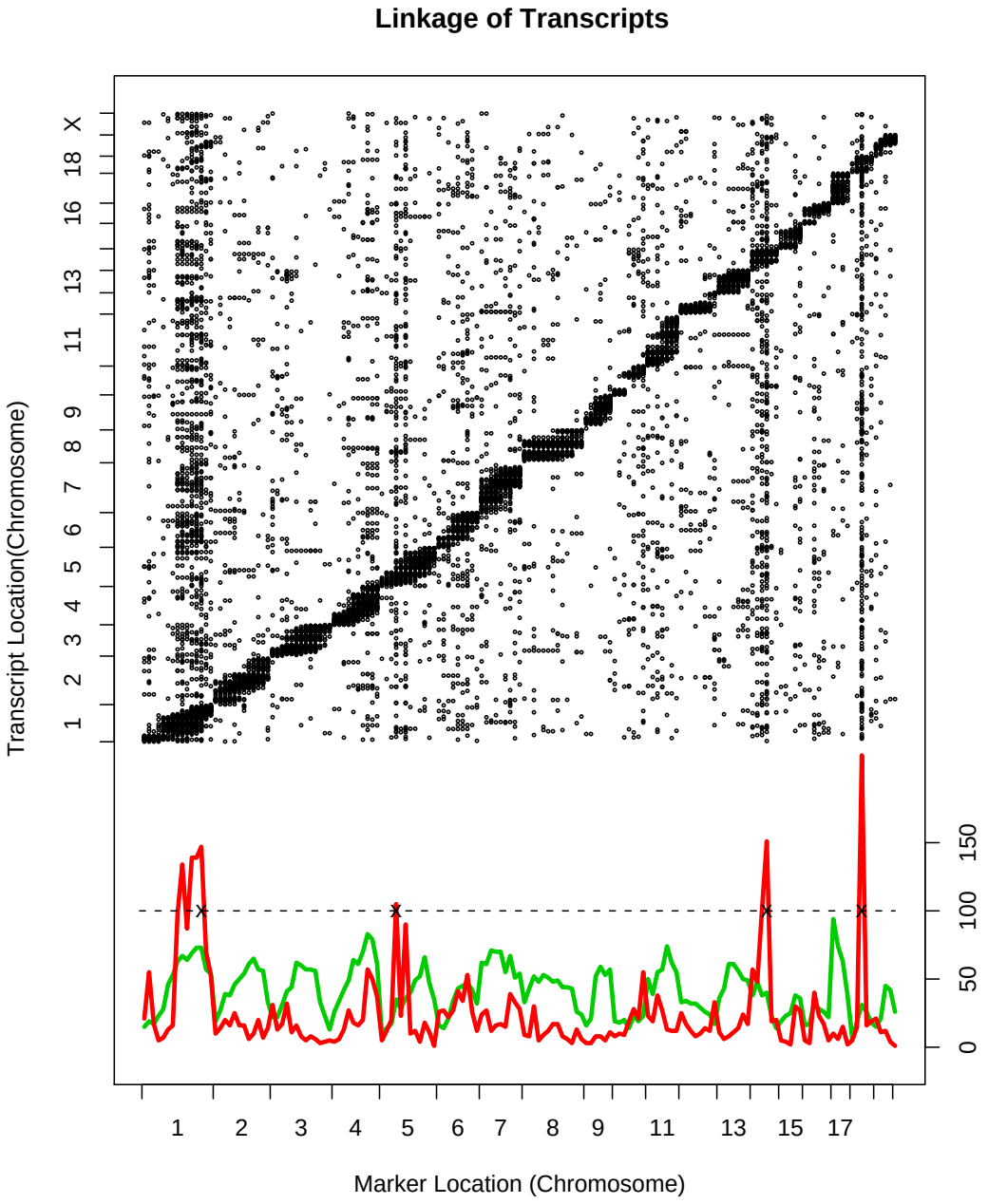


Figure 7.4. Significant marker-transcript linkages Linkages ($LOD > 3.2$) are indicated as points on the scatterplot with genomic location of transcript on Y-axis and QTL location on X-axis. A prominent diagonal band of local QTL is apparent. The lines at the bottom show the linkage counts for each of the markers (scale located on the lower-right edge): the red line show the number of distant linkages and the green line shows the number of local linkages. For our work we consider four markers that have at least 100 distant linkages and are located on different chromosomes (Chromosomes 1, 5, 14 and 18). The selected markers are indicated with crosses.

list of linked transcripts. With a p-value of 0.001 we could identify more than 150 transcripts for each of the markers (Table 7.1). When corrected multiple testing, false discovery rate (FDR) ranged from 5 to 10% which is less than FDR obtained with simulated data for this step.

Table 7.1. Hotspots with more than 100 distant linkages.

Marker	Location	Local QTL	Distant QTL	Total QTL	FDR
rs3724524	Chrom 1 @ 157.80 Mb	69	178	247	6.67%
rs3659933	Chrom 5 @ 46.36 Mb	36	125	161	10.23%
rs3676913	Chrom 14 @ 86.92 Mb	38	152	190	8.67%
rs3713429	Chrom 18 @ 47.80 Mb	33	254	287	5.74%

The chromosome 1 hotspot is linked to 247 transcripts. A visual inspection of the shield matrix indicates that the local transcripts are less shielded than the distant transcripts. The variable-threshold plot identified 7 primary transcripts, 5 of which are on chromosome 1. The gene Fkbp9 codes for an enzyme involved in *isomerase activity* and it shields a set of 78 transcripts. This set is enriched for *steroid biosynthesis* (3 out of 78, p=0.0035). he gene Ensa is a known regulator of insulin secretion. It shields 41 transcripts out of which 12 are glyco-proteins (12 out of 41, p=0.01). The gene F11r is involved in *protein binding* and is active in membrane. It shields a set of 40 transcripts which are enriched for *oxidoreductase activity* (8 out of 40, p=0.002) and many of them are active in *endoplasmic reticulum* (7 out of 40, p=0.003).

The chromosome 5 five out of seven identified primary transcripts are local. The gene Ppat, a local primary transcript, is annotated as having transferase activity and many of its descendants are involved in *regulation of cell proliferation* (6 out of 53, p=0.005) and many of them are *phospho-proteins* (19 out of 53, p=0.02).

On chromosome 14 has 3 primary transcripts. The gene Entpd4, a local primary transcript on chromosome 14, is an ion binding protein and many of its descendants

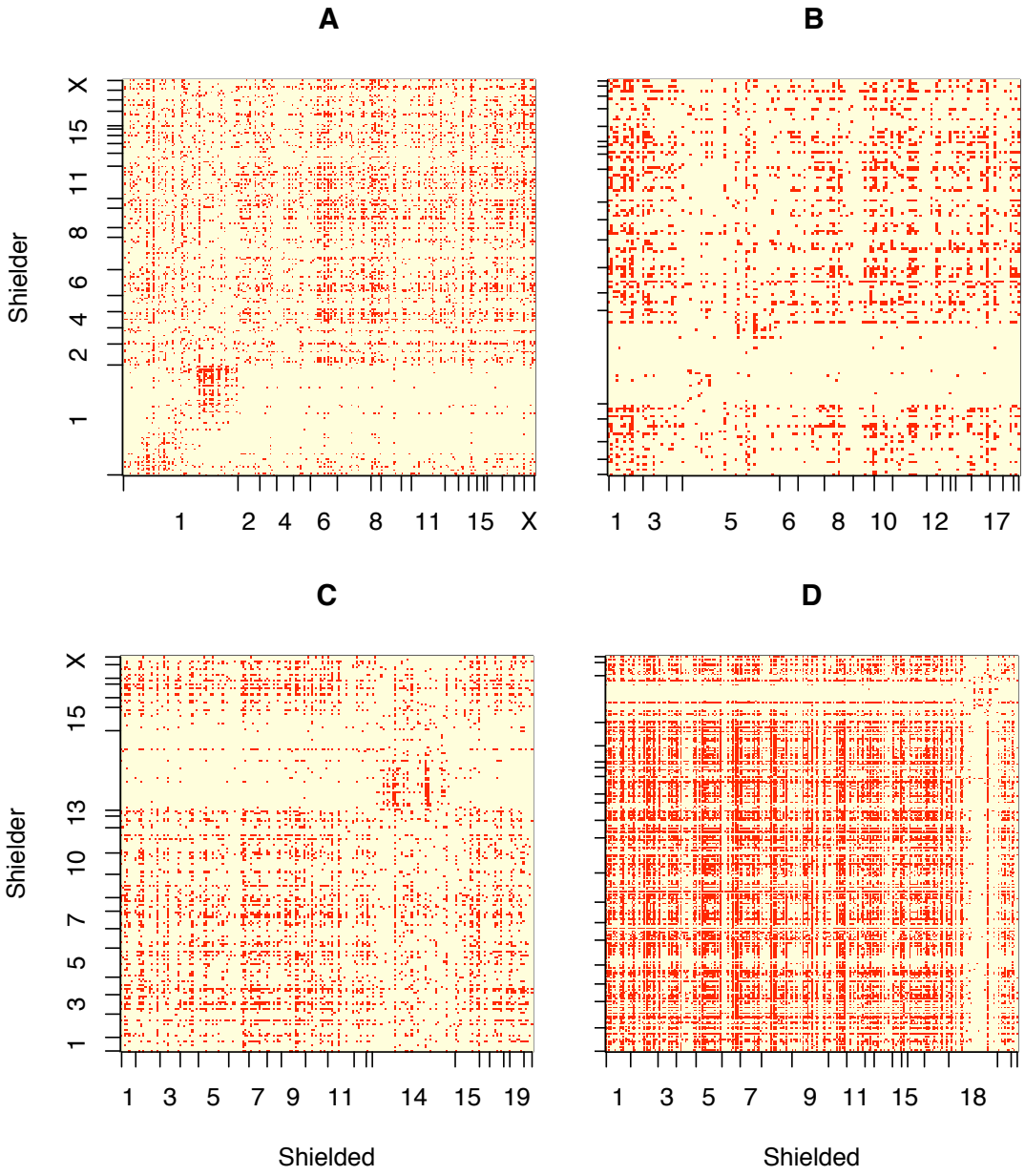


Figure 7.5. Shield matrices for hotspots on chromosomes 1,5,14 and 18 In every case transcripts are arranged according to their genomic location. A red dot indicates that the corresponding transcript on the X axis shields the one on the Y axis. A vertical run of points indicates a transcript that shields many other transcripts. In A-C we can see that local transcripts shield many distant transcripts, but themselves are not shielded as indicated by the white horizontal band. In the case of chromosome 18 (D), local transcripts shield very few distant transcripts. A threshold of 0.1 bits was used to construct these shield matrices.

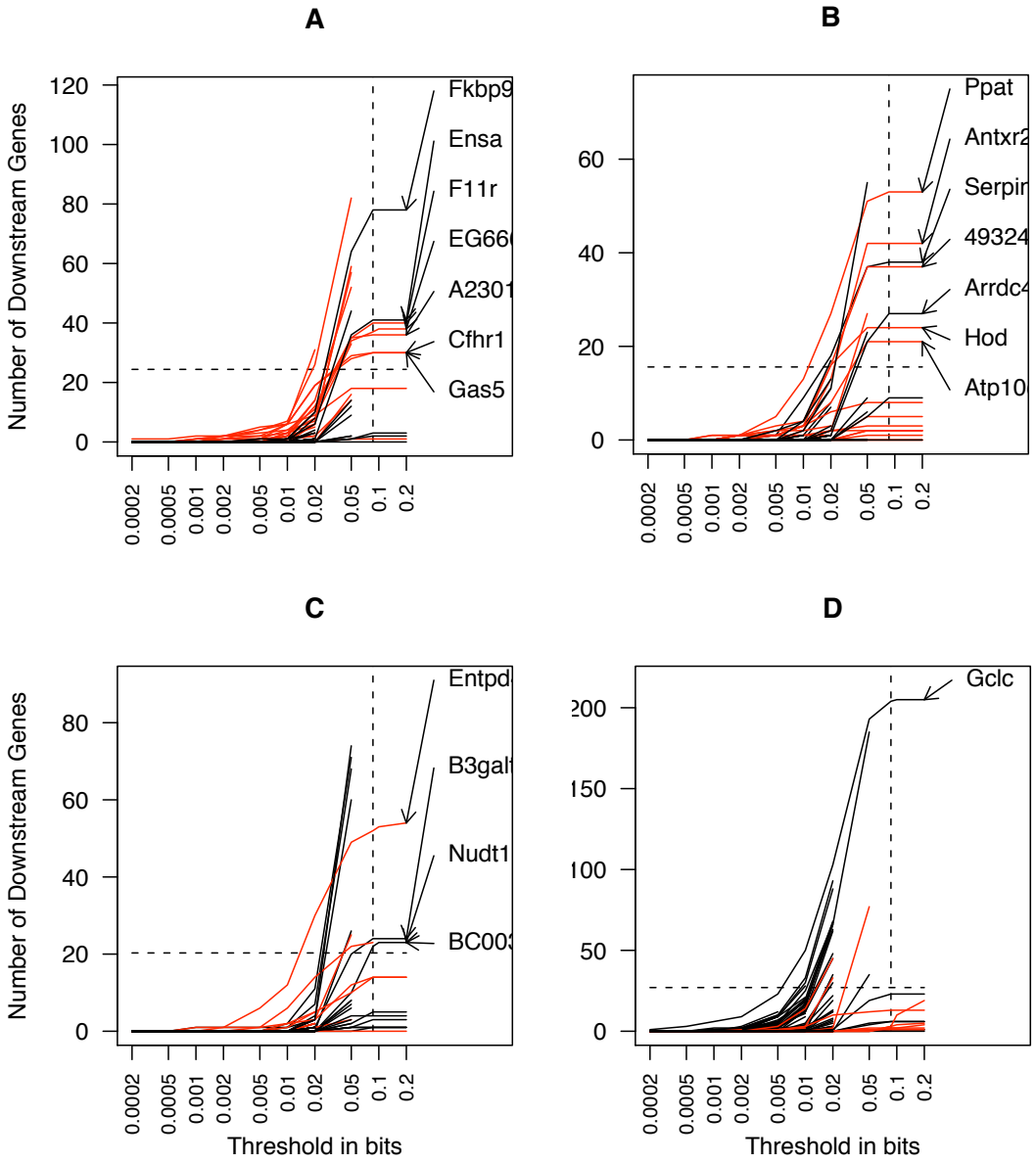


Figure 7.6. Variable threshold plots for four hotspots in the BXA cross (Chromosome 1 (A), Chromosome 5 (A), Chromosome 14 (A) and Chromosome 18 (A)). Each line on the plot corresponds to a transcript in the hotspot. The lines extend only through thresholds where the transcript is not shielded by any other transcript. Y-axis denotes the number transcripts it is shielding. Red lines indicate local transcripts and the black lines indicate the distant transcripts. The dotted horizontal line corresponds to 10% of total number of nodes found to be linked and the vertical line corresponds to δ_{BIC}. Transcripts appearing on top right corner of this plot are the most probable primary transcripts.

were involved in *electron transport* (6 out of 59, p=0.006) and *vesicle mediated transport* (6 out of 59, p=0.006).

The chromosome 18 hotspot has a different character from other hotspots. It is very narrow in extent with little linkage spillover to the adjacent markers. There are a few local eQTL that shield other transcripts. The shielding pattern is dominated by a single distant transcript Gclc, that shields 202 out of 287 transcripts in the hotspot. Gclc is involved in negative regulation of apoptosis and is under-expressed in heterozygote state of this marker. Many of its descendants are involved in *cellular lipid metabolic process* (15 out of 202, p=0.001).

7.4 Discussion

We have presented a method for analyzing the causal association structure of genetic hotspots in expression genetics data. Our method uses conditional independence tests to infer a set of primary transcripts that are candidates for direct regulation by polymorphic loci near the hotspot marker as well as secondary transcripts that can be grouped according to primary transcript that shields them. The secondary transcripts, in many cases, are enriched for functional annotations that are related to the known functions of the primary transcript.

My results provide insights into the complexity of the hotspots and it is possible frame specific hypothesis about the mechanisms that produce this prominent feature of eQTL data.

Simulation results shows that our method is capable of accurately identifying primary transcripts in realistic data. Application to four hotspots spread over different chromosomes suggest that these results maybe biologically sensible although additional experimental work may be needed to validate these relationships.

For each hotspot, we computed conditional independence tests for all pairs of transcripts and constructed shield matrices (Figure 7.5). For chromosomes 1, 5 and 14

local transcripts are rarely shielded by other transcripts, however and they often shield numerous distant transcripts. This indicates that possibly one or more polymorphism around the marker are regulating a small set of primary transcripts, mostly local, that are in turn regulating a large number of secondary transcripts. However, the pattern on chromosome 18 is distinct; local transcripts are rarely shielding other transcripts.

In our method we use a shield of size one to detect secondary transcripts that are not directly downstream of the hotspot. But it is unlikely that any given network is that simple. Our method eliminates candidate primary transcripts conservatively and some of the putative primary transcripts might be in fact modulated by multiple real primary transcripts. However, when we increase the shield size, our power to detect independence relationships decreases and our results are difficult to interpret.

In addition most of the primary transcripts correspond to phenotypic differences between the parent strains. A/J strain (http://jaxmice.jax.org/strain/000646.html) is known to be resistant to obesity and diabetes. Among the primary transcripts Lrpap1 is known to be related to obesity [35], and Ensa [54], Fmo4 [40], Ppat [55] and Gclc [4] are known to be related to diabetes.

Presence of artifacts like SNP overlapping probes can make it seem that the corresponding transcript is strongly linked to the nearest marker and is likely to appear on the top of the list if arranged according to the strength of linkage [2]. But these transcripts are less likely to shield other transcripts as the additional variation, which is not due to the SNP, will not be reflected in expression of other transcripts.

This approach is not guaranteed to find all the primary transcripts, however it appears to provide an effective explanatory tool for interpretation of expression genetics data. The power of this method is derived in part because we do not attempt to search all possible models or to construct a large scale graphical model. Although true dependence relationships are likely to be more complex than triplets, this simple strategy finds many of them with high power.

CHAPTER 8

CONCLUSION

In this thesis I presented a set of three methods to infer statistically significant components of the underlying regulatory network by analyzing expression genetics data. Specifically I modeled the regulatory network as a Bayesian network and presented techniques to recover instances of a known regulatory mechanism, infer local regulatory networks of a transcript, and consolidate multiple networks.

Modeling of expression genetics data as a Bayesian network presented many challenges. The data was severely under-specified, many of the relevant variables were not directly observed, and it consisted of multiple data types. These challenges were alleviated by reducing the problem into subproblems of manageable complexity and creating detailed models of the relationships between genes.

The Quantitative Trait Gene (QTG) model infers causal regulatory relations between genes. The interacting term in the QTG model helps it to recover more complex instances of regulation. Unlike the Quantitative Trait Loci model, it also provides a finer mapping of the causal elements. Also, this model provides an example of representing a regulatory mechanism as a directed graphs and inferring similar regulatory instances.

The local network inferencing method extends the QTG model to recover regulatory modules around a transcript of interest. As compared to the QTG model this method can infer modules of arbitrary size. Furthermore, as compared to other existing methods, it recovers both transcript and genetic variation nodes, which provides a better biological understanding of the data.

The hotspot analysis method presented in this thesis analyzes pleiotropic loci to infer plausible explanations for observing such phenomena. This method considers each linkage as a directed graph and aggregates multiple such graphs into a biologically meaningful network. The set of primary transcripts inferred from this analysis are of greater influence in the global regulation. This analysis also shows that many local linkages, even those with strong linkage, might not be involved in regulation.

The application of these methods on simulation data suggests that these techniques were largely successful in recovering a majority of the regulatory relations with modest error rates. Application on real world data-sets revealed biologically interesting instances of regulation.

8.1 Challenges and Limitations

8.1.1 Dimensionality

The most important challenge in analyzing expression genetics data is its dimensionality. The number of variables is very large as compared to the number of available samples. In this thesis I address this challenge by considering much smaller subnetworks in turn. The QTG model considers a network containing only three variables at a time. In the local regulatory network, inferencing the size of the network is determined by the number of elements in the Markov blanket – a much smaller number than all possible variables. And, the hotspot analysis involves calculating the shield matrix by considering just three variable at a time. This approach helped us to uncover many instances of biologically sensible networks, but it also misses many of the true but non-significant relations in favor of statistical significance.

Even with only three variables and knowing the right probability distribution, it is not always possible to empirically detect the right conditional independence (See Figure 8.1).

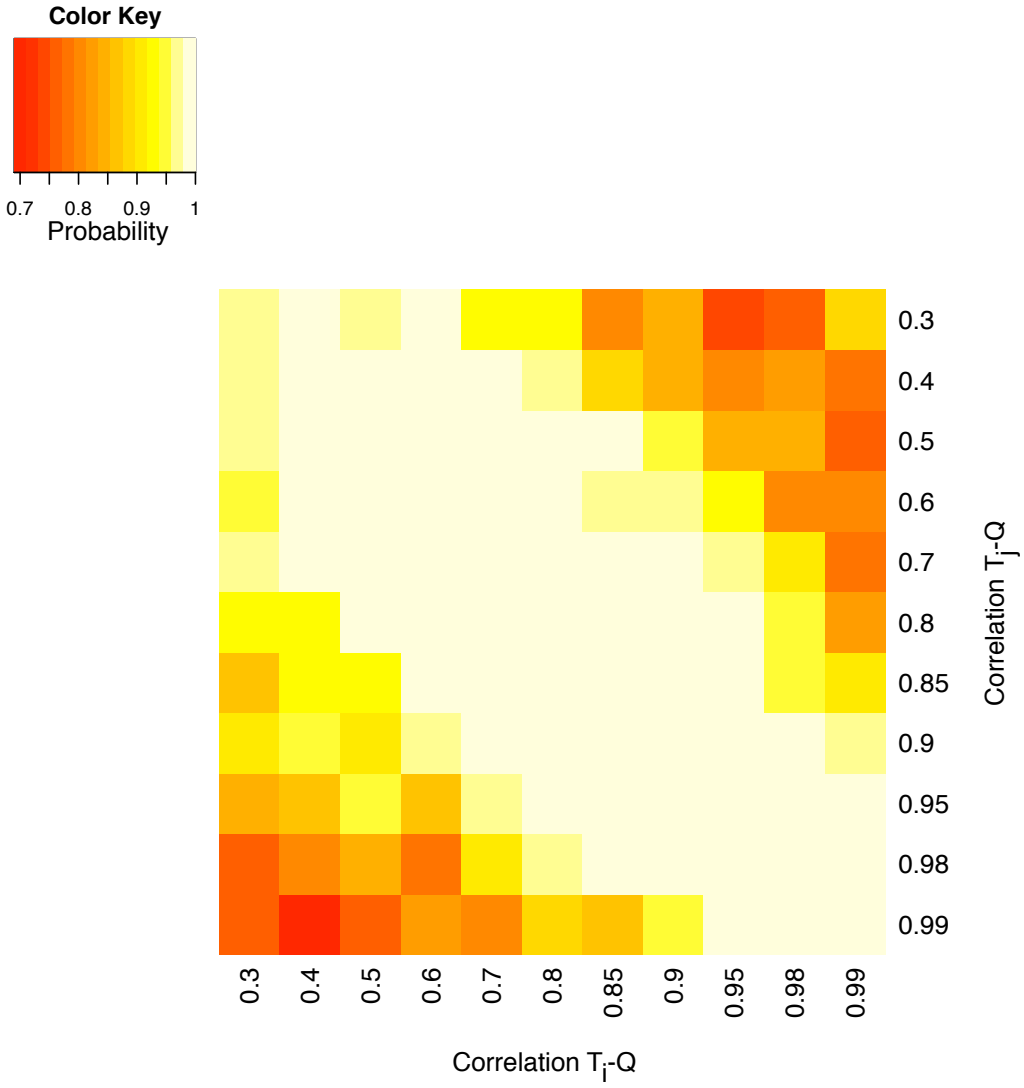


Figure 8.1. Accuracy of detecting conditional independence as a function of correlation between the variables using mutual information Data was simulated with known correlations between Q and T_i, and Q and T_j, satisfying $T_i \perp T_j|Q$. The correlations are indicated on the axes of the figure and color of each square denotes the probability of detecting the correct relationship. Simulated samples of 120 samples were used to generate this plot. Lopsided correlation decreases the probability of recovering the right conditional independence relation.

Moreover, even a perfect conditional independence testing algorithm for three variables does not always return the right conditional independence relations in a network of size four as shown in figure 8.2.

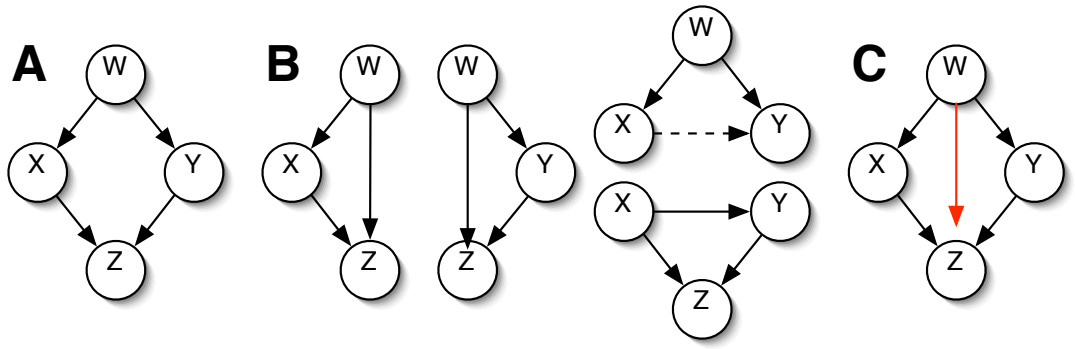


Figure 8.2. Detecting Conditional Independence in larger networks In this figure it is shown that applying a perfect conditional independence testing for three nodes fails when applied on a network of four nodes. A. A simple network with four nodes. B. The resulting networks when a perfect conditional independence testing method for three nodes is applied. The dotted line indicates conditional independence. C. The aggregated network constructed from the conditional independence testing. The red line indicates the erroneous dependency.

The use of just three variables at a time is just a starting point. If there are enough samples available, these methods should be extended as needed. Searching for smaller networks reduces the number of false positives, but also increases the number of false negatives. An accurate probability distribution in necessary to achieve a low false positive rate. But accurate distributions are achieved from larger samples and smaller networks at the expense of reduced model complexity. Thus, there remains a critical trade-off between accurate inference and model complexity.

A similar trade-off is observed in the QTG model. Compared to the conventional QTL model this model is more expressive, but at the cost of estimating an additional parameter. Also, being a parametric model this approach misses many of the relations that cannot be approximated by this model. Ideally a more expressive model would

capture more complex relations, but it is unlikely that such models can be inferred with high statistical significance.

8.1.2 Unobserved data and related problems

There are multiple sets of unobserved variables in this data-set. The samples are genotyped only at the marker locations and genotypes at other locations need to be estimated from these values. In this thesis the distribution of these variables are estimated from the flanking markers and the recombination distance. Many other important variables, such as protein abundance, are not directly measured and instead transcript abundance is used as a proxy measurement. However, the correlation between proteins and transcripts is known to be poor.

The data collecting methodology often induces artificial correlations among variables. If such errors are not corrected, applying these methods can result in the inference of spurious networks. If the samples are collected or processed in different batches, a large number of transcripts are going to be correlated to the grouping. Applying hotspot analysis, without correcting for such confounding factors, would reveal false regulatory cascades[5, 39, 38, 25]. In this thesis such corrections are not made and it is not possible to conclude whether the detected networks are real or spurious without the availability of additional data.

The probes of microarrays used to measure transcript abundance are assumed to be designed using the genome sequence of a reference strain. When used with a different strain, due to the genetic variations between the strains, these probes can fail. In expression genetics data, where the samples are mixtures of their parental genomes, the transcripts can be correlated to their nearest markers because of this strain-specific probe hybridization problem, not because of actual changes in expression level. If such probes are not filtered out before data analysis many spurious causal relation are inferred.

8.2 Conclusions

The expression genetics data is computationally challenging but is biologically important. In this thesis I presented a set of methods to model this data as a Bayesian network. This model consisted of both transcript nodes and genotype nodes, which provided a thorough description of the underlying regulatory process.

I proposed three different methods: *network motif searching*, *local network inferencing*, and *network aggregation*. The *network motif searching* method is effective when there is sufficient information about the common regulatory mechanisms, the *local network inferencing* is useful when there is a list of interesting variables, and the *network aggregation* is used to merge the networks obtained in the previous methods.

An instance of each of these methods were implemented and applied on real datasets. The process of pairwise transcription modulation was modeled as *Quantitative Trait Gene* and application of this method on a yeast cross revealed many instances where both genotype and transcript abundance of the regulator were interacting to modulate the target. This method was extended to find the *local regulatory network* of transcripts. The *genetic hotspots* were analyzed by aggregating the pairwise networks and inferring the set of primary transcripts for many hotspots in a mice cross data.

Through these implementations I show that the Bayesian networks are an effective modeling tool in analyzing the expression genetics data. This model allows detailed and flexible modeling of the underlying mechanism. Although it is not possible to decipher the complete underlying regulatory network, I present a case in favor of using a combination of the presented methods, *network motif searching*, *local network inferencing*, and *network aggregation*, to reliably recover its significant components. This argument is supported by systematic simulation studies and the biological reasonable networks recovered from application on real data.

BIBLIOGRAPHY

[1] Akaike, H. A new look at the statistical model identification. *IEEE transactions on automatic control 19*, 6 (1974), 716–723.

[2] Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J.P., and Jansen, R.C. Sequence Polymorphisms Cause Many False cis eQTLs. *PLoS ONE 2*, 7 (2007), e622.

[3] Ashburner, M., Ball, C.A., et al. Gene ontology: tool for the unification of biology. *Nat Genet 25*, 1 (2000), 25–9.

[4] Bekris, LM, Shephard, C., Janer, M., Graham, J., McNeney, B., Shin, J., Zarghami, M., Griffith, W., Farin, F., Kavanagh, TJ, et al. Glutamate cysteine ligase catalytic subunit promoter polymorphisms and associations with type 1 diabetes age-at-onset and GAD65 autoantibody levels. *Exp Clin Endocrinol Diabetes 115*, 4 (2007), 221–8.

[5] Breitling, R., Li, Y., Tesson, B.M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L.V., de Haan, G., Su, A.I., et al. Genetical genomics: spotlight on QTL hotspots. *PLoS Genetics 4*, 10 (2008).

[6] Brem, R.B., and Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences 102*, 5 (2005), 1572–1577.

[7] Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science 296*, 5568 (2002), 752.

[8] Broman, K.W., and Sen, S. *A Guide to QTL Mapping with R/qtl*. Springer Verlag, 2009.

[9] Broman, K.W., and Speed, TP. A review of methods for identifying QTLs in experimental crosses. *Lecture Notes-Monograph Series* (1999), 114–142.

[10] Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R., and Kohane, I.S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences 97*, 22 (2000), 12182.

[11] Carlborg, Ö., and Haley, C.S. Epistasis: too often neglected in complex trait studies?

[12] Cheng, J., Greiner, R., Kelly, J., Bell, D., and Liu, W. Learning Bayesian networks from data: an information-theory based approach. *Artificial Intelligence 137*, 1-2 (2002), 43–90.

[13] Churchill, GA, and Doerge, RW. Empirical Threshold Values for Quantitative Triat Mapping. *Genetics 138*, 3 (1994), 963–971.

[14] Cooper, G.F., and Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Machine learning 9*, 4 (1992), 309–347.

[15] Cullin, C., Baudin-Baillieu, A., Guillemet, E., and Ozier-Kalogeropoulos, O. Functional analysis of YCL09C: evidence for a role as the regulatory subunit of acetolactate synthase. *Yeast 12*, 15 (1996), 1511–1518.

[16] Dennis Jr, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol 4*, 3 (2003), 2003–4.

[17] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns, 1998.

[18] Friedman, N., and Koller, D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning 50*, 1 (2003), 95–125.

[19] Friedman, N., Linial, M., et al. Using Bayesian networks to analyze expression data. *J Comput Biol 7*, 3-4 (2000), 601–20.

[20] Heckerman, D. A tutorial on learning with Bayesian networks. *Learning in graphical models* (1998), 301–354.

[21] Heckerman, D., Geiger, D., and Chickering, D.M. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning 20*, 3 (1995), 197–243.

[22] Jagalur, M., and Churchill, GA. Deciphering genetic hotspots of transcriptional regulation. *Genetics* (Submitted).

[23] Jagalur, M., and Kulp, D. An information theoretic method for reconstructing local regulatory network modules from polymorphic samples. In *Computational Systems Bioinformatics: Proceedings of the CSB 2007 Conference* (2007), Imperial College Press, p. 133.

[24] Jansen, R.C., and Nap, J.P. Genetical genomics: the added value from segregation. *Trends Genet 17*, 7 (2001), 388–91.

[25] Kang, H.M., Ye, C., and Eskin, E. Accurate Discovery of Expression Quantitative Trait Loci Under Confounding From Spurious and Genuine Regulatory Hotspots. *Genetics 180*, 4 (2008), 1909.

[26] Koller, D., and Sahami, M. Toward Optimal Feature Selection. In *Machine learning: proceedings of the Thirteenth International Conference (ICML'96)* (1996), Morgan Kaufmann Pub, p. 284.

[27] Kraft, P., and Horvath, S. The genetics of gene expression and gene mapping. *Trends in Biotechnology 21*, 9 (2003), 377–378.

[28] Kullback, S., and Leibler, R. A. On information and sufficiency. *Annals of Mathematical Statistics 22* (1951), 79–86.

[29] Kulp, D.C., and Jagalur, M. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics 7*, 1 (2006), 125.

[30] Lander, ES, and Botstein, D. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics 121*, 1 (1989), 185–199.

[31] Li, H., Lu, L., Manly, K.F., Chesler, E.J., Bao, L., Wang, J., Zhou, M., Williams, R.W., and Cui, Y. Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Human Molecular Genetics 14*, 9 (2005), 1119–1125.

[32] Lynch, M., and Walsh, B. *Genetics and analysis of quantitative traits*. Sinauer, Sunderland, Mass., 1998. 97017666 Michael Lynch, Bruce Walsh. Includes bibliographical references (p. 891-[948]) and indexes.

[33] Margot, J.B., Ehrenhofer-Murray, A.E., and Leonhardt, H. Interactions within the mammalian DNA methyltransferase family. *BMC Molecular Biology 4*, 1 (2003), 7.

[34] Matys, V., Fricke, E., et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res 31*, 1 (2003), 374–8.

[35] McCarthy, J.J., Meyer, J., Moliterno, D.J., Newby, L.K., Rogers, W.J., and Topol, E.J. Evidence for substantial effect modification by gender in a large-scale genetic association study of the metabolic syndrome among coronary heart disease patients. *Human Genetics 114*, 1 (2003), 87–98.

[36] Pena, JM, Bjorkegren, J., and Tegner, J. Growing Bayesian network models of gene networks from seed genes. *Bioinformatics 21*, 90002 (2005).

[37] Pena, J.M., Nilsson, R., Björkegren, J., and Tegner, J. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning 45*, 2 (2007), 211–232.

[38] Peng, Jie, Wang, Pei, and Tang, Hua. Controlling for false positive findings of trans-hubs in expression quantitative trait loci mapping. *BMC Proceedings 1*, Suppl 1 (2007), S157.

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

[40] Rouer, E., Rouet, P., Delpech, M., and Leroux, JP. Purification and comparison of liver microsomal flavin-containing monooxygenase from normal and streptozotocin-diabetic rats. *Biochem Pharmacol 37*, 18 (1988), 3455–9.

[41] Sax, K. The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris. *Genetics 8*, 6 (1923), 552–560.

[42] Schadt, E.E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S.K., Monks, S., Reitman, M., Zhang, C., et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics 37* (2005), 710–717.

[43] Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature 422*, 6929 (2003), 297–302.

[44] Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes, 1996.

[45] Schwarz, G. Estimating the dimension of a model. *The annals of statistics* (1978), 461–464.

[46] Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics 34* (2003), 166–176.

[47] Sen, S., and Churchill, G.A. A statistical framework for quantitative trait mapping. *Genetics 159*, 1 (2001), 371–387.

[48] Shalon, D., Smith, SJ, and Brown, PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research 6*, 7 (1996), 639.

[49] Simpson, SP. Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *TAG Theoretical and Applied Genetics 77*, 6 (1989), 815–819.

[50] Singh, M., and Valtorta, M. An algorithm for the construction of Bayesian network structures from data. *International Journal of Approximate Reasoning* (1993).

[51] Soller, M., Brody, T., and Genizi, A. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *TAG Theoretical and Applied Genetics 47*, 1 (1976), 35–39.

[52] Spirtes, P., Glymour, C.N., and Scheines, R. *Causation, prediction, and search*. The MIT Press, 2001.

[53] Sugiyama, F., Churchill, G.A., Higgins, D.C., Johns, C., Makaritsis, K.P., Gavras, H., and Paigen, B. Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics 71*, 1 (2001), 70–77.

[54] Thameem, F., Farook, V.S., Yang, X., Lee, Y.H., Permana, P.A., Bogardus, C., and Prochazka, M. The transcribed endosulfine α gene is located within a type 2 diabetes-linked region on 1q: sequence and expression analysis in Pima Indians. *Molecular Genetics and Metabolism 81*, 1 (2004), 16–21.

[55] Thuresson, ER. Inhibition of glycerol-3-phosphate acyltransferase as a potential treatment for insulin resistance and type 2 diabetes. *Curr Opin Investig Drugs 5*, 4 (2004), 411–8.

[56] Tsamardinos, I., Aliferis, C.F., and Statnikov, A. Algorithms for large scale markov blanket discovery. In *The 16th International FLAIRS Conference* (2003), vol. 103.

[57] Verma, T., and Pearl, J. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the eighth conference on Uncertainty in Artificial Intelligence table of contents* (1992), Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp. 323–330.

[58] Wang, Shuang, Zheng, Tian, and Wang, Yuanjia. Transcription activity hot spot, is it real or an artifact? *BMC Proceedings 1*, Suppl 1 (2007), S94.

[59] Wu, C., Delano, D.L., Mitro, N., Su, S.V., Janes, J., McClurg, P., Batalov, S., Welch, G.L., Zhang, J., Orth, A.P., et al. Gene Set Enrichment in eQTL Data Identifies Novel Annotations and Pathway Regulators. *PLoS Genetics 4*, 5 (2008).

[60] Yvert, G., Brem, R.B., et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet 35*, 1 (2003), 57–64.

[61] Zhao, Y., Sohn, J.H., and Warner, J.R. Autoregulation in the biosynthesis of ribosomes. *Molecular and Cellular Biology 23*, 2 (2003), 699.

[62] Zhu, J, Wiener, M C, Zhang, C, Fridman, A, Minch, E, Lum, P Y, Sachs, J R, and Schadt, E E. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol 3*, 4 (Apr 2007).