AN INFORMATION THEORETIC METHOD FOR RECONSTRUCTING LOCAL REGULATORY NETWORK MODULES FROM POLYMORPHIC SAMPLES

Manjunatha Jagalur, David Kulp*

Computational Biology Lab, University of Massachusetts Amherst, Amherst, MA-01002, USA

**Email: {manju,dkulp}@cs.umass.edu*

Statistical relations between genome-wide mRNA transcript levels have been successfully used to infer regulatory relations among the genes, however the most successful methods have relied on additional data and focused on small sub-networks of genes. Along these lines, we recently demonstrated a model for simultaneously incorporating micro-array expression data with whole genome genotype marker data to identify causal pairwise relationships among genes. In this paper we extend this methodology to the principled construction of networks describing local regulatory modules. Our 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 Bayes Net is then constructed from the resulting variables with important biological constraints yielding causally correct relationships.

We tested our method by simulating a regulatory network within the background of of a real data set. We 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 sample size is a practical and economic limitation, we considered the impact of increasing the number of samples and found that recovery of true gene-gene relationships only doubled with ten times the number of 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 we applied this method to an actual data set of 111 back-crossed mice we were able to recover local gene regulatory networks supported by the biological literature.

1. INTRODUCTION

Understanding the function of every gene and its role in expression of a particular complex trait is one of the fundamental aims of genomics. Availability of genome-wide data has made it possible to tackle this problem from a systems biology perspective. Global putative gene regulatory networks have been constructed using mRNA abundance data collected through micro-array experiments. In some cases supplemental information like chip-CHIP binding data [13] and single or multiple gene perturbation data [5] have been used to construct more robust networks. Recently there has been growing interest in a quantitative genetics strategy wherein, along with gene expression data, genetic marker data is used for constructing such networks[20, 33]. In this strategy, crosses are made from inbred strains that differ in physical and genetic attributes. Resulting individuals can be considered the result of thousands of gene perturbations. Whole genome markers are genotyped and the abundance of transcripts are measured for each individual. For example, Brem and colleagues used a cross of a wild strain of yeast and baker's yeast to create one of the first such data sets[17]. Schadt and colleagues have collected such data for mouse and maize[1].

Figure 1(a) describes the data. Gene expression (T_j) represents transcript abundance. Discrete genotype values (M_k) for bi-allelic markers are measured at relatively uniform positions across the genome. For an F2 diploid cross, if the parent genotypes are AA and BB, then markers may take values AA, AB and BB. We assume alleles have additive effect and represent genotypes as integers (0, 1, 2). The genotype of a gene (Q_j) is not directly measured, but can be estimated by maximum likelihood using the flanking marker genotypes and genetic linkage distances to those markers (D_L and D_R in the figure)[35]. Our aim is to find genetic and genomic factors (i.e. some subset of $(\mathbf{T} \cup \mathbf{Q})$) that affect a particular complex trait and infer the relationships among these factors. We generally refer to $(\mathbf{T} \cup \mathbf{Q})$ as an *expression genetics* data set.

There has been several efforts to infer regulatory relationships among genes using expression genetics data sets. A key quantitative genetics concept in all these approaches is the quantitative trait locus (QTL), which refers to a region along a chromosome

*Corresponding author.

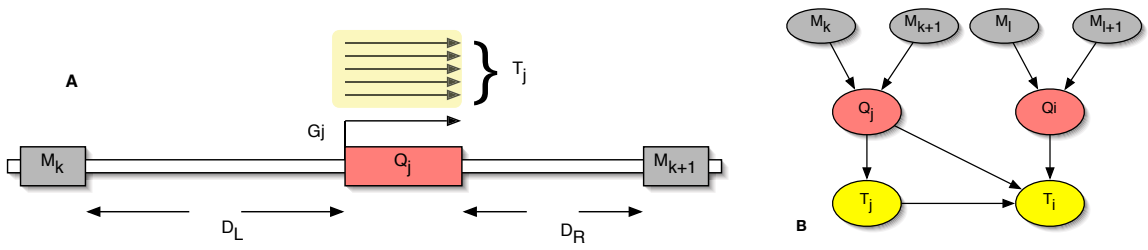

Fig. 1. Expression genetics data description. Gene G_j is located between markers M_k and M_{k+1}. Its genotype is Q_j and amount of mRNA translated is T_j. Note that Q_j is unobserved and must be estimated. (b) The QTG model of a single regulator-target pair of genes (regulator is gene j and target is gene i). M_l and M_{l+1} are flanking markers of gene i.

where the markers are significantly correlated with a measured trait. QTLs are determined using interval mapping[34] or other related methods[35]. In our case, the gene expression level is the trait of interest and QTLs of this kind are called eQTLs.

Finding pairwise regulatory relations between a regulator gene and a target gene is a simpler problem than that of constructing more elaborate networks. Bing and Hoeschele chose a regulator gene that is maximally correlated to the target among those found in a target's eQTL[24]. In our previous work, we generalized this idea by mapping eQTLs using a modified interval mapping model that simultaneously fit the joint contributions of both genotype and expression level of each candidate regulator along a chromosome. We called this model QTG[2]. Its important new feature was the ability to capture the varying nature of regulation with respect to a regulator's genotype. Through this approach we could discover regulators that act as enhancers or repressors depending on their genotypes. (Described in more detail in section 1.3.)

Network structure prediction has also been attempted by [25] and [26]. In these works the network is represented by a Bayesian Network (BN) where the nodes are observed gene expression levels and the edges represent conditional dependencies, which are assumed to correspond to causal relationships. In both of these works, the key idea is to place strong priors over possible network structures according to the eQTLs associated with each gene. Li et al[25] selected candidate regulators with non-synonymous polymorphisms that are positioned within eQTLs. Later an exhaustive search over BN structures was performed to reconstruct a global regulatory network. Zhu et al [26] used a set of heuristics based on the characteristics of eQTLs to determine probable edge direction and connectivity.

In this paper, we present an improved BN reconstruction algorithm with the following major contributions:

- 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.

The rest of the paper is organized as follows. Subsections 1.1-1.2 present important concepts regarding Markov blankets and Bayesian networks. More details of the QTG model are described in 1.3. We present our new regulatory network inference method in section 2, describe our experiments in section 3.1 and end with a discussion and conclusion in section 4.

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 2).

Recovering the Markov blanket using raw data is well-studied in the context of feature selection[29, 3, 32]. Here we describe one particularly attractive approach.

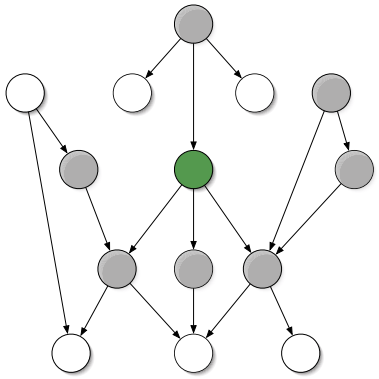

Fig. 2. Example of Markov Blanket. Nodes marked gray belong to Markov blanket of node marked in green

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[3]. 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 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 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 such as in [3, 32]. 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.

Algorithm 1.1 IAMB algorithm

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

1: $MB = \emptyset$
2: **repeat**
3: $\quad i = \arg\max_{i \neq s} MI(X_i; X_s | MB)$
4: $\quad$ **if** $MI(X_i; X_s | MB) > \theta$ **then**
5: $\quad\quad MB = MB \cup \{X_i\}$
6: $\quad$ **end if**
7: **until** MB does not change
8: **repeat**
9: $\quad i = \arg\min_{X_i \in MB} MI(X_i; X_s | MB - \{X_i\})$
10: $\quad$ **if** $MI(X_i; X_s | MB - \{X_i\}) < \theta$ **then**
11: $\quad\quad MB = MB - \{X_i\}$
12: $\quad$ **end if**
13: **until** MB does not change

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:

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

1.2. Bayesian Networks

A Bayesian network (BN) is a minimal graphical representation of a joint probability distribution over a set of random variables[22]. Each variable in a BN corresponds to a node and each dependency corresponds to an edge. Nodes are connected by a directed edge and the resulting graph will be a directed acyclic graph. The distribution of a variable conditionally depends only on its parents.

Like Markov blanket selection, constructing Bayesian networks is also a well-studied problem[22, 30, 28]. 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[30] 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[28] 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. In our method we overcome this limitation by restricting ourselves to building local networks around our gene of interest. As the number of genes in the regulatory neighborhood of a gene is usually low, we can keep our network searching problem tractable.

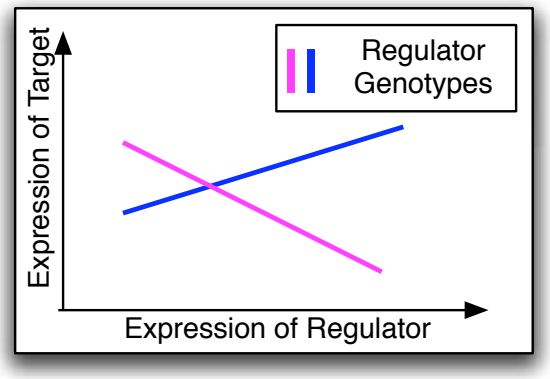

Fig. 3. *Trans*-acting effect as a function of regulator genotype allows for complex enhancer/suppressor relationships. Note that expression and genotype may be marginally independent of the target, but the regulatory relation can still be identified.

1.3. 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 [2], we 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

$$P(T_i|Q_j, T_j, \theta) = \mathcal{N}(\beta_0 + \beta_1 T_j + \beta_2 Q_j + \beta_3 T_j Q_j, \sigma) \quad (1)$$

where θ is the β and σ model parameters. (Figure 1b.)

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 1 has the advantage of capturing the types of dependency relationships shown in figure 3. However, the scanning method does not incorporate multi-locus regulatory control.

2. METHODS

Now we 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 our previous scanning method[2] is that we 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[37].

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 4. 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 4a). 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 4b). Since genotypes represent independently random recombination events, edges are always directed away from genotype variables.

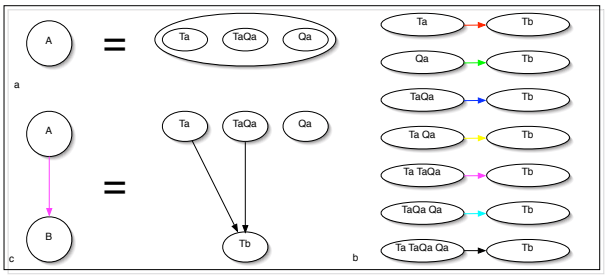

Fig. 4. (a) Elements of gene A. T_a is the expression, Q_a is the genotype and T_aQ_a is the interaction variable. (b) All edge types. Colors are used to visually code predicted networks (such as in figure 7. (c) Example of gene-gene relationship with two edge types involved.

2.2. Markov Blanket Inference

Algorithm 2.1 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 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 we write TQ_i^l to mean $T_i M_{L(i)}$, TQ_i^r to mean $T_i M_{R(i)}$, 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 1.1.1) is used as the measure of conditional independence. Variables are assumed to follow a multinomial Gaussian distribution. If we make the reasonable biological assumption that any gene has no more than about ten genes in its local regulatory network[38], then we require only (≈ 100) samples to accurately calculate conditional mutual information.

Algorithm 2.1 Inferring Markov Blanket of a gene. *MI* calculates the conditional mutual information as described in section 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 $MB \in T \cup M \cup I$

MB=$\emptyset$
C=$(T \cup M \cup I) - \{T_s, TQ_s^l, TQ_s^r\}$
repeat
  for $C_i \in C$ **do**
    $score_i = MI(C_i; T_s | MB)$
  end for
  $[maxMI, maxi] = max(score)$
  if $maxMI \geq \alpha$ **then**
    $MB = MB \cup \{C_{maxi}\}$
  end if
until $maxMI < \alpha$
repeat
  for $C_i \in MB$ **do**
    $score_i = MI(C_i; T_s | MB - \{C_i\})$
  end for
  $[minMI, mini] = min(score)$
  if $minMI < \alpha$ **then**
    $MB = MB - \{C_{maxi}\}$
  end if
until $maxMI < \alpha$
return MB

2.3. Gene regulatory network reconstruction

We use an incremental algorithm similar to [31] for constructing the local network for a seed gene, s (Algorithm 2.2) given its Markov blanket, MB_s. The novelty of our method is that we must simultaneously estimate the unobserved genotype values Q_i while constructing the graph edges.

We 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. We define the regulatory neighborhood of seed gene s as $RN_s = MB_s \cup \{T_s\}$. For all genes with a flanking marker in the MB_s we introduce the true but unobserved genotype Q_i and estimate its maximum likelihood value according to the distances to the flanking markers. Similarly we replace any TQ_i^l and TQ_i^r terms 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, we explicitly consider combined genotype and expression effects including interacting effects. These different kinds of regulatory effects are represented as different types of edges (figure 4b). 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.

Algorithm 2.2 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 $-\infty$. *from* can be any node, *to* node needs to have expression term in it and *kind* can be any kind of edge shown in 4 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**
  $Q_i = EstGenotype(M_{L(i)}, M_{R(i)}, Location(i))$
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**
  $\{from, to, kind\} = \text{argmax}_{fr,to,kd}\, Addscore(BN_s, \{fr, to, kd\}, CG)$
  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

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 to which it can be grouped with. In those cases a dummy gene hyper node is created for this marker. These dummy genes are 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.

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.

3.1. Simulations

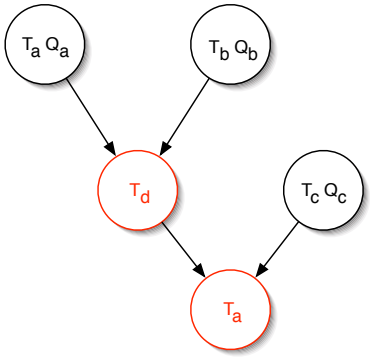

Fig. 5. Simulation Strategy. Black nodes were selected from the existing data and the red nodes were simulated using a linear Gaussian model.

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 a linear model with Gaussian noise. A example of such simulated network is shown in Figure 5. 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, we simulated additional expression values as Gaussian and genotypes from linkage probabilities.

Results of these simulations are presented in Figure 6 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 6a 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 we increased sample size to 1000 (ten times the current available data) there was a marked improvement in recovery($> 75\%$) and accuracy($> 85\%$).

Figure 6b 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.

We found that a threshold of $\alpha = 0.1$ on conditional mutual information and $\beta = 50.0$ for adding an edge in network reconstruction yielded the best results.

3.2. Biological Significance

For practical experimental results we used data collected by Schadt et al[1], consisting of gene expression profiles for 111 F_2 mice derived from crossing C57BL/6J and DBA/2J. The dataset contains expression for 23,574 genes and genotypes for 134 markers spread over 19 chromosomes.

We applied our algorithm 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. A simple analysis showed that 69% of these networks seed gene shared common Gene Ontology annotation with at least one other gene in the network. Further, in 31% of the cases seed gene shared annotation with two or more neighbors. Several of these networks are shown in Figure 7 with the biological interpretations and analysis. The inferred local regulatory network of *Dlx2* is shown in figure 7a. 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 7b) 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 7c) suggests that *Pcna* and *Dmap1* might be co-regulating *Prim1*. This is interesting as these two genes are known to interact with similar domains[36]. The local network of *Fgfr2* (figure 7d) 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[39].

4. DISCUSSION

Expression genetics data has helped scientists to understand the genetics behind expression of many simpler traits that are affected by very few genetic factors. Understanding genetics behind more complex traits needs careful modeling of the interaction between the quantitative (gene expression) and qualitative (genotype) traits.

We presented an extension of our QTG model

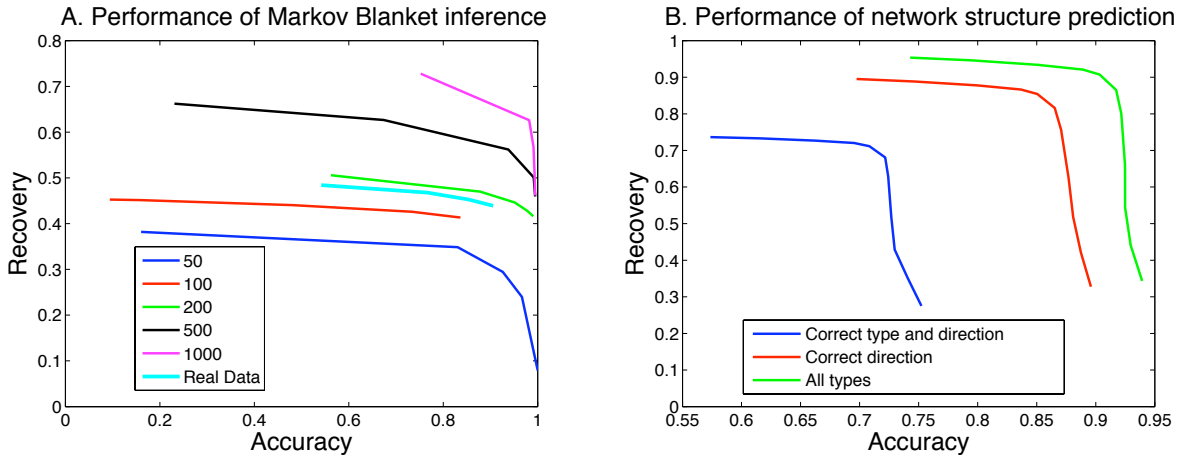

Fig. 6. A. Accuracy vs recovery plot for classification of variables in Markov Blanket of candidate seed gene. Different lines show results for different sample sizes. B. Accuracy vs recovery plot for graph reconstruction using different graph evaluation criteria.

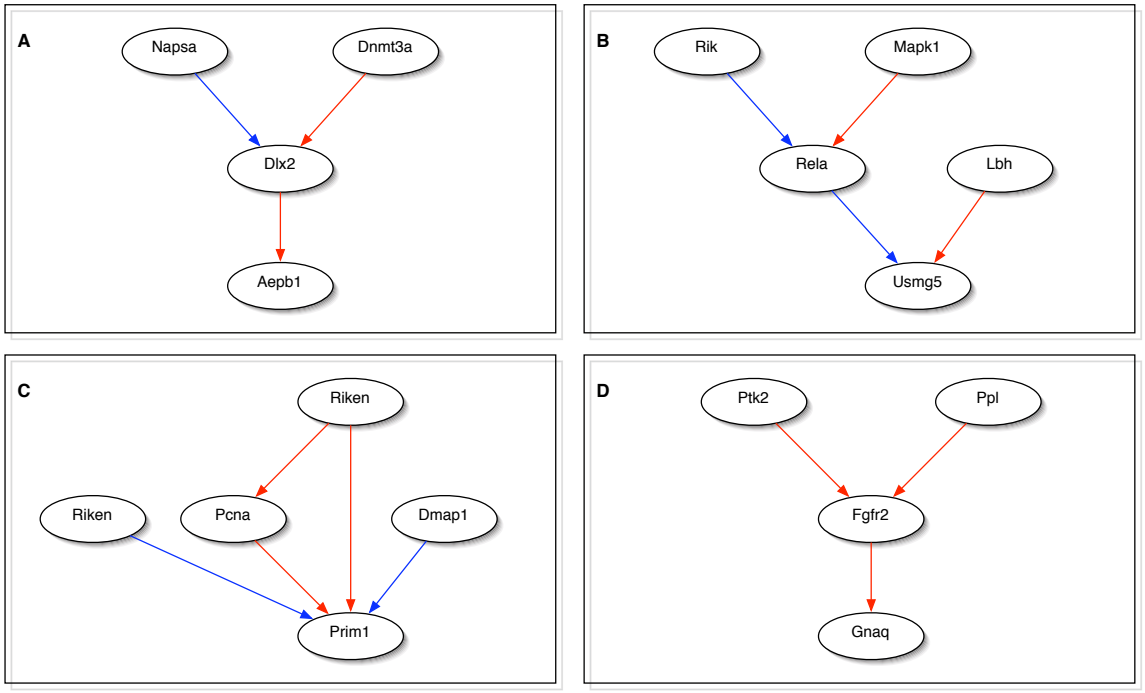

Fig. 7. Sample local regulatory networks. See text.

for analyzing regulation involving multiple genes as a directed acyclic graph. In this study we investigated the use of an information theoretic method for accurately constructing local gene regulatory network from a seed gene. Our 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.

Our simulation results suggest that reasonably accurate small networks can be constructed using our approach. Importantly, we also found that small sample size is the most important limitation on the utility of these data sets. Our study suggests that a magnitude increase in number of samples would go

a long way in identifying 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 our method is capable of recovering biologically relevant 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.

ACKNOWLEDGEMENT

We are thankful to Gary Churchill and Sharon Tsaih for their useful comments.

References

1. Schadt EE, Monks SA et al *Genetics of gene expression surveyed in maize, mouse and man*, Nature, Mar 20;422(6929), pp 297-302.(2003).
2. Kulp D, Jagalur M. *Causal Inference of Regulator Target Pairs by Gene Mapping of Expression Phenotypes.* BMC Genomics. 2006; 7: 125. (2006).
3. I Tsamardinos, CF Aliferis, A Statnikov. *Algorithms for Large Scale Markov Blanket Discovery*, The 16th International FLAIRS Conference. (2003).
4. N Friedman, M Linial, I Nachman, D Peer. *Using Bayesian networks to analyze expression data*, Journal of Computational Biology, vol. 7, pp 601-620. (2000).
5. D Pe'er, A Regev, G Elidan, N Friedman. *Inferring subnetworks from perturbed expression profiles*, Bioinformatics. (2001).
6. Pe'er D, Regev A, Tanay A. *Minreg: inferring an active regulator set*, Bioinformatics 18, pp 1:S258-267. (2002).
7. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data*, Nature Genetics 34(2), pp 166-176. (2003).
8. Nir Friedman. *Inferring Cellular Networks Using Probabilistic Graphical Models*,Science, Vol 303, Issue 5659, pp 799-805 , 6 February. (2004).
9. AJ Hartemink, DK Gifford, TS Jaakkola, RA Young. *Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks*,Pacific Symposium on Biocomputing. (2001).
10. Smith VA, Jarvis ED, Hartemink AJ. *Evaluating functional network inference using simulations of complex biological systems*,Bioinformatics. (2002).
11. Perez-Enciso M, Toro MA, Tenenhaus M, Gianola D. *Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study*,Genetics, 164(4), pp 1597-606. (2003).
12. Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED. *Advances to Bayesian network inference for generating causal networks from observational biological data*, Bioinformatics, 12;20(18), pp 3594-603 (2004).
13. CH Yeang, T Jaakkola. *Physical Network Models and Multi-source Data Integration* , Proceedings of the seventh annual international conference on Computational molecular biology, pp 312 - 321. (2003).
14. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. *Combining location and expression data for principled discovery of genetic regulatory network models*, Pacific Symposium on Biocomputing. (2002).
15. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. *Computational discovery of gene modules and regulatory networks*, Nature Biotechnology, Nov;21(11), pp 1337-42. (2003).
16. A Battle, E Segal, D Koller. *Probabilistic Discovery of Overlapping Cellular Processes and Their Regulation*, Proceedings of the eighth annual international conference on Computational molecular biology, pp 167-176. (2004).
17. RB Brem, G Yvert, R Clinton, L Kruglyak. *Genetic dissection of transcriptional regulation in budding yeast*, Science, 26;296(5568), pp 752-755. (2002).
18. EE Schadt, SA Monks, SH Friend. *A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets*,Biochemical Society Transactions, 31, pp 437443. (2003).
19. Kraft P, Horvath S. *The genetics of gene expression and gene mapping*, Trends in Biotechnology, 21(9), pp 377-378.(2003).
20. Jansen RC, Nap JP. *Regulating gene expression: surprises still in store*, Trends in Genetics, 20(5), pp 223-225.(2004).
21. Doerge RW. *Mapping and analysis of quantitative trait loci in experimental populations*, Nature Reviews, Genetics,3, pp 43-52. (2002).
22. Heckerman D. *A Tutorial on Learning With Bayesian Networks*, Technical Report, Microsoft Research, MSR-TR-95-06. (1995).
23. Sen S, Churchill G. *A statistical framework for quantitative trait mapping.* Genetics 2001, 159:371-87.
24. Bing, N. and Hoeschele, I. *Genetical genomics analysis of a yeast segregant population for transcription network inference.* Genetics 2005. 170(2): 533-42.
25. Li H, Lu L, Manly KF, Chesler EJ, Bao L, Wang J, Zhou M, Williams RW, Cui Y. *Inferring gene transcriptional modulatory relations: a genetical genomics approach.* Hum Mol Genet. 2005 May 1;14(9):1119-25.
26. Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB, Schadt EE. *An integrative genomics approach to the reconstruction of gene networks in seg-*

regating populations. Cytogenet Genome Res 2004. 105(2-4):363-74.
27. Kevin P. Murphy. *The Bayes Net Toolbox for MATLAB.* Computing Science and Statistics, vol 33.
28. Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, Weiru Liu. *Learning Bayesian networks from data: An information-theory based approach.* Artificial Intelligence. Vol. 137, no. 1-2, pp. 43-90. May 2002.
29. Daphne Koller and Mehran Sahami. *Toward Optimal Feature Selection.* International Conference on Machine Learning 1996. 284-292.
30. Nir Friedman, Daphne Koller. *Being Bayesian about Network Structure.* Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference, 2000.
31. Greg Cooper and Edward Herskovits. *A Bayesian Method for the Induction of Probabilistic Networks from Data.* Machine Learning 1992. 9:309-347.
32. JM Pena, J Bjorkegren, J Tegner. *Scalable, Efficient and Correct Learning of Markov Boundaries under the Faithfulness Assumption.* Bioinformatics 2005, 21(Suppl 2):ii224-29.
33. EE Schadt and PY Lum. *Reverse engineering gene networks to identify key drivers of complex disease phenotypes.* Journal of lipid research. Vol. 47, 2601-2613, December 2006.
34. ES Lander and D Botstein. *Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps.* Genetics, Vol 121, 185-199. 1989.
35. M Lynch and B Walsh. *Genetics and analysis of quantitative traits.* Sinauer Associates, Sunderland, MA. 1998.
36. JB Margot, AE Ehrenhofer-Murray and H Leonhardt. *Interactions within the mammalian DNA methyltransferase family.* BMC Molecular Biology 2003, 4:7.
37. AJ Butte, P Tamayo, D Slonim, TR Golub and IS Kohane. *Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.* PNAS, USA. 2000 Oct 24;97(22):12182-6.
38. RB Brem and L Kruglyak. *The landscape of genetic complexity across 5,700 gene expression traits in yeast.* PNAS, USA. 2005 Feb 1;102(5):1572-1577.
39. J Pearl and TS Verma, *A Theory of Inferred Causation.* UCLA Cognitive Systems Laboratory, Technical Report (R-156).