Bayesian causal phenotype network incorporating genetic variation and biological knowledge

Jee Young Moon[1], Elias Chaibub Neto[2], Xinwei Deng[3] and Brian S. Yandell[4]

[1] Department of Statistics, University of Wisconsin, Madison, Wisconsin, USA
jymoon@wisc.edu
[2] Network Biology Department
Sage Bionetworks, Non-profit biomedical research organization, Seattle, Washington, USA
elias.chaibub.neto@sagebase.org
http://www.sagebase.org
[3] Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA
xdeng@vt.edu
http://www.stat.vt.edu/facstaff/xdeng/
[4] Departments of Statistics and Horticulture, University of Wisconsin, Madison, Wisconsin, USA
byandell@wisc.edu
http://www.stat.wisc.edu/~yandell/

Abstract A Bayesian network has often been modeled to infer a gene regulatory network from expression data. Genotypes along with gene expression can further reveal the regulatory relations and genetic architectures. Biological knowledge can also be incorporated to improve the reconstruction of a gene network. In this work, we propose a Bayesian framework to jointly infer a gene network and weights of prior knowledge by integrating expression data, genetic variations, and prior biological knowledge. The proposed method encodes biological knowledge such as transcription factor and DNA binding, gene ontology annotation, and protein-protein interaction into a prior distribution of the network structures. A simulation study shows that the incorporation of genetic variation information and biological knowledge improves the reconstruction of gene network as long as biological knowledge is consistent with expression data.

1 Introduction

A key interest in molecular biology is to understand how DNA, RNA, proteins and metabolic products regulate each other. In this regard, people have considered to construct the regulatory networks from microarray expression data with time-series measurements or transcriptional perturbations [1,2]. A regulatory network can also be constructed with genetic variation in segregating populations that

perturbs the gene expression, protein and metabolite levels. Genetic variation information can decipher genetic effects on traits and discover causal regulatory relationships between phenotypes. In addition, knowledge of regulatory relationships is available in various biological databases, which can improve the reconstruction of causal networks. This paper focuses on combining genetic variations in a segregating population and biological knowledge to improve the inference of causal phenotype networks.

Genetic variation information in a segregating population has been used to reconstruct causal phenotype networks [3,4,5] and to infer causal relationships among pairs of phenotypes [6,7,8,9,10]. Approaches based on structural equation models [11,12,13] and causal discovery algorithms [14,15] have also been proposed. A common feature of the above methods is that quantitative trait loci (QTL) mapping and phenotype network reconstruction are conducted separately. The QTL mapping without consideration of a phenotype network may generate a genetic architecture (the locations and effects of detectable QTLs) with QTLs of indirect effects. As pointed out by [16], poorly estimated genetic architectures may compromise the inference of causal relationships among phenotypes. To address this issue, several researchers [16,17] proposed to jointly infer causal phenotype networks and genetic architectures.

Various sources of biological knowledge have been incorporated with gene expression in the reconstruction of phenotype networks because it is difficult to decide the direction of gene regulation using expression data only. Transcription factors binding site information was leveraged by [18], whereas Nariai *et al.* [19] used protein-protein interaction knowledge to construct phenotype networks. Methods integrating multiple sorts of biological knowledge were proposed by [20], [21], and [22].

In this paper, we propose a Bayesian approach to jointly infer a causal phenotype network and genetic architectures with a prior distribution on network structures adjusted by biological knowledge. The joint network of causal phenotype relationships and genetic architectures is modeled as a Bayesian network adopted from [16], QTLnet. We extend the framework of QTLnet by incorporating biological knowledge into the prior distribution on network structures. This extension can enhance the predictive power of the network by capturing several fundamentals of biological knowledge [4]. The prior probability on network structures is based on the Markov random field to integrate and weight several sources of biological information allowing for flexible tuning of the analyst's confidence in different types of biological information [21]. The consideration of reliability of biological knowledge is necessary since biological knowledge can be incomplete and inaccurate. While Zhu *et al.* [4] proposed a method to incorporate genetic variation and biological knowledge to phenotype networks, their method does not consider the reliability of biological knowledge. Our proposed approach (QTLnet-prior) can integrate gene expression, genetic variation, and biological knowledge (protein-protein interaction, gene ontology annotation, and transcription factor and DNA

binding information) by weighting its reliability in the network reconstruction algorithm.

The details of our integrated framework for the joint inference of causal network and genetic architecture of correlated phenotypes are organized as follows. Section 2 describes the QTLnet method for the joint inference of causal network and genetic architecture. Section 3 presents the proposed QTLnet-prior, which is to incorporate biological knowledge into prior probability distributions over the space of network structures. A simulation study is conducted in Section 4 to compare the proposed method with several existing approaches. Finally, in Section 5 we discuss the strengths and caveats of our approach and point out future research directions.

2 Joint inference of causal network and genetic architecture

In this section we first present a standard Bayesian network for modeling expression data. Next we present an extended model, based on the homogeneous conditional Gaussian regression model, that incorporates QTL nodes into phenotype networks. We point out that, even though the directed edges in the standard Bayesian networks are often interpreted as causal relationships, in reality, they only represent conditional dependencies. Only by extending the phenotype networks with genotype nodes can we actually justify causal interpretations. Finally, we present a rationale for the joint inference of the causal phenotype network and genetic architecture, and give an overview of our joint approach for phenotype network and genetic architecture inference.

2.1 Standard Bayesian network model

A standard Bayesian network is a probabilistic graphical model whose conditional independence is represented by a directed acyclic graph (DAG). A node t in a DAG G corresponds to a random variable Y_t in the Bayesian network. A directed edge from node u to node v can supposedly represent that Y_v is causally dependent on Y_u, though an edge truly represents the conditional dependency. The local directed Markov property of Bayesian network states that each variable is independent of its non-descendant variables conditional on its parent variables,

$$Y_t \perp Y_{V \setminus de(t)} | Y_{pa(t)} \quad \text{for all } t \in V$$

where $de(t)$ is the set of descendants of t, $pa(t)$ is the set of parents of t, V is the set of all nodes in a DAG G, and $Y_{pa(t)} = \{Y_i : i \in pa(t)\}$. Assume the node index is ordered such that the index of descendants is always bigger than the index

of their parents. The joint distribution can be written to be

$$P(Y_1, \ldots, Y_T) = \prod_{t=1}^{T} P(Y_t | Y_{t-1}, \ldots, Y_1)$$
$$= \prod_{t=1}^{T} P(Y_t | Y_{pa(t)}) \quad (1)$$

where the first equality is satisfied by the chain rule and the second equality is satisfied by the local directed Markov property. Since nodes $t-1, \ldots, 1$ are non-descendants of node t and $pa(t) \in \{t-1, \ldots, 1\}$, Y_t is independent of $Y_{\{t-1,\ldots,1\} \setminus pa(t)}$ conditional on $Y_{pa(t)}$, $P(Y_t | Y_{t-1}, \ldots, Y_1) = P(Y_t | Y_{pa(t)})$.

2.2 HCGR model

The parametric family of a Bayesian network that jointly models phenotypes and QTL genotypes corresponds to a homogeneous conditional Gaussian regression (HCGR) model. Conditional on the QTL genotypes and covariates, the phenotypes are distributed according to a multivariate normal distribution, where QTLs and covariates enter the model via the mean, and the correlation structure among the phenotypes is explicitly modeled according to the DAG representing the phenotype network structure [16].

The HCGR model is derived from a series of linear regression equations. Explicitly, let $i = 1, \ldots, n$ and $t = 1, \ldots, T$ index individuals and phenotype traits, respectively. Let $Y_t = (Y_{t1}, \ldots, Y_{ti}, \ldots, Y_{tn})'$ be expression levels of phenotype t for all individuals, and let $\epsilon_t = (\epsilon_{t1}, \ldots, \epsilon_{tn})'$ represent the associated independent normal error terms. We assume that the expression level for individual i and trait t follows the following phenotype model:

$$Y_{ti} = \mu^*_{ti} + \sum_{v \in pa(t)} \beta_{tv} Y_{vi} + \epsilon_{ti}, \quad \epsilon_{ti} \sim N(0, \sigma^2_t)$$

where β_{tv} is the partial regression coefficients relating phenotype t with phenotype v, and $\mu^*_{ti} = \mu_t + X_i \Gamma_t \theta_t$ with $\Gamma_t = \text{diag}(\gamma_t)$, where μ_t is the overall mean for a trait t, θ_t is a column vector of all genetic effects, X_i is a row vector for individual i from the design matrix X parametrized according to Cockerham's model [23] for genotypes, and γ_t is a binary vector that represents the genetic architecture of trait t with $\gamma_t = 1$ as the effect being included in the genetic architecture. It was shown by [16] that these linear regression equations set a HCGR model for phenotypes and QTL genotypes.

2.3 Systems genetics and causal inference

Systems genetics aims to understand the complex interrelations between genetic variations and phenotypes from large scale genotype and phenotype data [24].

Here we explain how the systems genetics approach allows us to infer causal networks. While causal relations between QTLs and phenotypes are justified by the randomization of alleles during meiosis and the unidirectional influence of the genotype on phenotype, causal relations among phenotypes are induced from conditional independence. The systems genetics idea is that by incorporating QTL nodes into phenotype networks we create new sets of conditional independence relationships that allow us to distinguish network structures that would, otherwise, belong to the same equivalence class.

We start with causal relations between QTLs and phenotypes. In general the genotype influences the phenotype but not the other way around. Furthermore, the genotype is randomized by the recombination of parent genetic material during the mating process. These special characteristics enable us to infer causal effects of QTLs on phenotypes since, by analogy with a randomized experiment, we have that: (1) the treatment to an experimental unit (genotype) precedes the measured outcome (phenotype), and (2) random allocation of treatments to experimental units guarantees that other common causes get averaged out. This random allocation is explicit in an experimental cross such as a backcross or intercross. While this idea can be extended to natural populations, special attention must be paid to admixture, kinship and other forms of relatedness.

Causal inference among phenotypes, on the other hand, requires the concept of conditional independence organized in DAGs composed of phenotypes and QTL nodes. In the next three paragraphs we present some definitions and results that allow us to infer phenotype-to-phenotype causal relationships.

We start with the definitions. In graph theory, a *path* is defined as any unbroken, non-intersecting sequence of edges in a graph, which may go along or against the direction of arrows. We say that a path p is *d-separated* [25,26] by a set of nodes Z if and only if: (1) p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node is in Z, or (2) p contains a collider $i \rightarrow m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z. We say that Z d-separates X from Y if and only if Z blocks every path from a node in X to a node in Y. The *skeleton* of a DAG is the undirected graph obtained by replacing its arrows by undirected edges. A *v-structure* is composed by two converging arrows whose tails are not connected by an arrow.

Equivalence concepts play a key role in learning the structure of networks from the data. Here we present three important equivalence relations for graphs or its statistical models. Two graphs are *Markov equivalent* if they have the same set of d-separation relations [27]. Two structures m_1 and m_2 for Y are *distribution equivalent* with respect to the family F if they represent the same joint distributions for Y, that is, for every θ_1, there exists a θ_2 such that $p(Y \mid \theta_1, m_1) = p(Y \mid \theta_2, m_2)$ [28]. In other words, m_1 and m_2 are distribution equivalent if the parameters θ_1 and θ_2 are simple re-parametrizations of each other. If m_1 and m_2 are distribution equivalent, then the invariance principle of maximum likelihood estimates guaran-

tees that $p(Y \mid \hat{\theta}_1, m_1) = p(Y \mid \hat{\theta}_2, m_2)$, and m_1 and m_2 cannot be distinguished using the data. In this case we say that m_1 and m_2 are *likelihood equivalent*. In a Bayesian setting we define likelihood equivalence using the prior predictive distribution, that is, an integral of the likelihood function with respect to the prior distribution. Hence, if models m_1 and m_2 are distribution equivalent, it is often reasonable to expect that $p(Y \mid m_1) = p(Y \mid m_2)$ with a proper prior on θ, and that we cannot distinguish m_1 and m_2 for any data set Y [28].

Now we state four important results regarding causal inference in systems genetics: (1) Two DAGs are Markov equivalent if and only if they have the same skeletons and the same set of v-structures [29]. (2) Distribution equivalence implies Markov equivalence, but the converse is not necessarily true [27]. (3) For a Gaussian regression model, Markov equivalence implies distribution equivalence [30]. (4) For the homogeneous conditional Gaussian regression model, Markov equivalence implies distribution equivalence [16].

Therefore, for the HCGR parametric family, two DAGs are distribution and likelihood equivalent if and only if they are Markov equivalent. This implies that we can simply check out if any two DAGs have the same skeleton and the same set of v-structures in order to determine if they are likelihood equivalent and hence cannot be distinguished using the data.

Getting back to the idea of causal inference among phenotypes, let G_Y represent a standard Gaussian regression Bayesian network of gene expression phenotypes, Y. Expression data alone can distinguish some network structures by its likelihood but there are some network structures that are not distinguishable. For example, consider the three network structures in Table 1. Models G_Y^1 and G_Y^3 have the same skeleton ($Y_1 - Y_2 - Y_3$) and the same set of v-structures (no v-structure) and, thus, are distribution/likelihood equivalent. Model G_Y^2, on the other hand, has the same skeleton but a different set of v-structures and, hence, is not distribution/likelihood equivalent to models G_Y^1 and G_Y^3. Therefore, expression data alone can identify G_Y^2 but cannot distinguish G_Y^1 and G_Y^3.

DAG structures	skeletons	v-structures
$G_Y^1 = Y_1 \rightarrow Y_2 \rightarrow Y_3$	$Y_1 - Y_2 - Y_3$	$\emptyset$
$G_Y^2 = Y_1 \rightarrow Y_2 \leftarrow Y_3$	$Y_1 - Y_2 - Y_3$	$Y_1 \rightarrow Y_2 \leftarrow Y_3$
$G_Y^3 = Y_1 \leftarrow Y_2 \rightarrow Y_3$	$Y_1 - Y_2 - Y_3$	$\emptyset$

Table 1: Models G_Y^1 and G_Y^3 are distribution/likelihood equivalent.

Adding causal QTL nodes to a phenotype network allows the inference of causal relationships between phenotypes that could not be distinguishable using expression data alone. For example, if we add a causal QTL Q_1 to Y_1 in pheno-

type networks G_Y^1 and G_Y^3 in the above example, then the corresponding extended network structures G^1 and G^3 have different v-structures as shown in Table 2.

Extended DAG structures	skeletons	v-structures
$G^1 = Q_1 \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_3$	$Q - Y_1 - Y_2 - Y_3$	$\emptyset$
$G^3 = Q_1 \rightarrow Y_1 \leftarrow Y_2 \rightarrow Y_3$	$Q - Y_1 - Y_2 - Y_3$	$Q \rightarrow Y_1 \leftarrow Y_2$

Table 2: Extended models G^1 and G^3 are no longer distribution/likelihood equivalent.

2.4 QTL mapping conditional on phenotype network structure

Single trait QTL mapping analysis may detect QTLs that directly affect the phenotype under investigation, as well as QTLs with indirect effects [16]. For example, we consider a causal phenotype network in Figure 1. Then the expected results of single trait analysis are given as in Figure 2.

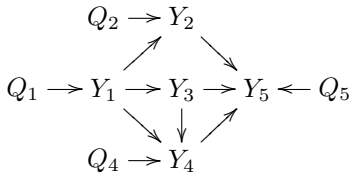


Figure 1: Example network with five phenotypes and four QTLs.

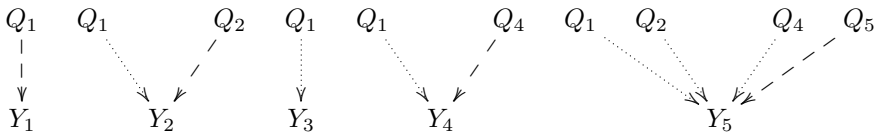


Figure 2: Output of a single trait QTL mapping analysis for the phenotypes in Figure 1. Dashed and pointed arrows represent direct and indirect QTL/phenotype causal relationships, respectively.

Now, consider QTL mapping analysis tailored to the phenotype network structure and assume, for a moment, that the phenotype network structure is known.

In this situation we can avoid detecting indirect QTLs by simply performing mapping analysis of the phenotypes conditional on their parents. For instance, in Figure 1, if we perform QTL mapping of Y_5 conditional on Y_2, Y_3 and Y_4 we do not detect Q_1, Q_2 and Q_4 because $Y_5 \perp Q_1 \mid Y_2, Y_3, Y_4$, $Y_5 \perp Q_2 \mid Y_2, Y_3, Y_4$ and $Y_5 \perp Q_4 \mid Y_2, Y_3, Y_4$. We only detect Q_5 since $Y_5 \not\perp Q_5 \mid Y_2, Y_3, Y_4$.

In practice, however, the structure of the phenotype network is unknown, and performing QTL mapping conditional on a misspecified phenotype network structure can result in the inference of misspecified genetic architectures as shown in Figure 3. The mapping analysis of a phenotype conditional on downstream phenotypes in the true network, induces dependencies between the phenotype and QTLs affecting downstream phenotypes. This leads to the erroneous inference that the phenotype includes downstream QTLs as its QTLs. However, a model with misspecified phenotype and genetic architectures will generally have a lower marginal likelihood score than the model with the correct causal order for the phenotypes and correct genetic architecture. Since in practice QTLnet adopts a model selection procedure to traverse the space of network structures, it tends to prefer models closer to the true data generating process. Simulation studies presented in [16] corroborate this point.

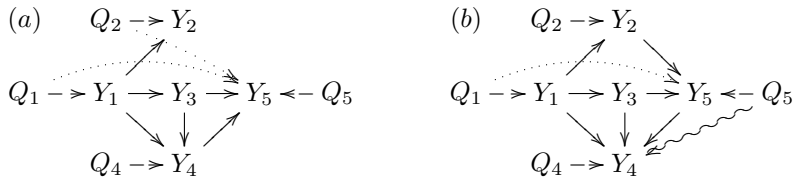


Figure 3: QTL mapping tailored to the network structure. Dashed, pointed and wiggled arrows represent, respectively, direct, indirect and incorrect QTL/phenotype causal relationships. (a) Mapping analysis of Y_5 conditional on Y_3 and Y_4 still detects Q_1 and Q_2 as QTLs for Y_5, since failing to condition on Y_2 leaves the paths $Q_1 \to Y_1 \to Y_2 \to Y_5$ and $Q_2 \to Y_2 \to Y_5$ in Figure 1 open. In other words, Q_1 and Q_2 are d-connected to Y_5 conditional on (Y_3, Y_4) in the true causal graph. (b) Mapping analysis of Y_4 conditional on Y_1, Y_3 and Y_5 incorrectly detects Q_5 as a QTL for Y_4 because in the true network the paths $Y_4 \to Y_5 \leftarrow Q_5$ and $Y_4 \leftarrow Y_3 \to Y_5 \leftarrow Q_5$ in Figure 1 are open when we condition on Y_5.

On a technical note in [16], it was shown that the conditional LOD score (logarithm of the odds favoring linkage), used for detecting QTLs according to the phenotype network structure, can be adopted as a formal measure of conditional independence between phenotypes and QTLs. Even though we restrict our attention to HCGR models, conditional LOD profiling is a general framework for the detection of conditional independencies between continuous and discrete random

variables and does not depend on the particular parametric family adopted in the modeling. Contrary to partial correlations, the conditional LOD score does not require the assumption of multi-normality of the data in order to formally test for independence and it can handle interactive covariates.

2.5 Joint inference of phenotype network and genetic architecture

As before, let G be a Bayesian network structure of phenotypes and QTLs. Let G_Y represent a standard Bayesian network of phenotypes, Y, and let $G_{Q\to Y}$ represent a graph composed of QTL nodes, phenotypes nodes and directed edges from QTL node to the phenotype node. Note that G_Y and $G_{Q\to Y}$ are subgraphs of the extended network structure G. Genetic architectures for all traits can be represented in two ways: $\boldsymbol{\gamma} = \{\gamma_t\}_{t=1}^T$ and $G_{Q\to Y}$, with γ_t defined as before. The likelihood of a Bayesian network of phenotypes and causal QTLs can be written as

$$P(Y|G, X, \theta_G) = P(Y|G_Y, G_{Q\to Y}, X, \theta_G) = P(Y|G_Y, \gamma, X, \theta_G),$$

where

$$P(Y|G_Y, \gamma, X, \theta_G) = \prod_{t=1}^{T}\prod_{i=1}^{n} N\left(\mu^{\star}_{ti} + \sum_{y_k\in pa(y_t)} \beta_{tk}\, y_{ki}\,,\, \sigma^2_t\right).$$

The marginal likelihood of phenotypes $P(Y|G, X)$ is calculated by integrating parameters θ_G out in the Bayesian network

$$P(Y|G, X) = \int P(Y|G, X, \theta_G)P(\theta_G|G)d\theta_G,$$

and can be asymptotically approximated as a function of the BIC score of network G. The posterior probability of G conditional on the data is given by

$$P(G|Y, X) = \frac{P(Y|G, X)P(G)}{\sum_G P(Y|G, X)P(G)}$$

where $P(G)$ represents the prior probability of the network structure G. In the next section we devote our attention to the specification of $P(G)$ using integrated biological knowledge.

Following [16], we adopt the QTLnet framework that jointly infers the phenotype network structure and genetic architecture. Most of the current literature in genetical network reconstruction has treated the problems of QTL inference and phenotype network reconstruction separately, generally performing genetic architecture inference first, and then using QTLs to help in the determination of the phenotype network structure [4,14]. As indicated in Section 2.4, such strategy can incorporate QTLs with indirect effects into the genetic architectures of phenotypes.

3 Causal network incorporating biological knowledge

Besides the causal QTLs, biological knowledge is another useful and important information resource to enhance the construction of a network. Such knowledge can be integrated on the top of the causal network to provide a more comprehensive picture of how genes are regulated. This integrated network could generate a new hypothesis of gene regulation, along being consistent with biological knowledge in overall.

In this section, we propose a network inference method, QTLnet-prior, from expression data with genetic variations, integrating biological knowledge. The QTLnet-prior extends the framework of QTLnet. It specifies the prior probability on network structures to integrate multiple sources of biological knowledge with flexible tuning parameters on confidence of knowledge [21]. The weighted integration of biological knowledge could produce a more predictive Bayesian network. The details of our extended framework, QTLnet-prior, are presented in Section 3.1. In Section 3.2, we sketch a Metropolis-Hastings (M-H) MCMC scheme for QTLnet-prior implementation that integrates the sampling of network structures [31,32], the QTL mapping, and the sampling of biological knowledge weights. In Section 3.3, we present how to encode biological knowledge into prior distributions over network structures.

3.1 Model

Extended model Denote by G a Bayesian network structure of phenotypes and QTLs. G consists of a phenotype network (G_Y) and genetic architectures for phenotypes ($G_{Q\to Y}$). Let Y be expression data, X be genetic variations, and W represent weights of biological knowledge. The QTLnet framework presented in Section 2 assumes intrinsically a uniform prior over network structures. Here, biological knowledge determines a prior probability on phenotype network structures, G_Y. Additionally, we specify a prior distribution on the weights of biological knowledge in order to control the consistency between expression data and knowledge. Because the prior information can be inaccurate or incompatible with the expression data, it is important to quantify its uncertainty. We write the extended model as follows.

$$\begin{aligned} P(G,W|Y,X,B) &\propto P(Y|G,W,X,B)P(G,W|X,B) \\ &= P(Y|G,X)P(G,W|X,B) \\ &= P(Y|G,X)P(G_Y,W|X,B)P(G_{Q\to Y}|X,B) \\ &= P(Y|G,X)P(G_Y,W|B)P(G_{Q\to Y}|X) \\ &= P(Y|G,X)P(G_Y|B,W)P(W|B)P(G_{Q\to Y}|X) \end{aligned} \quad (2)$$

where $P(Y|G,W,X,B)$ is the marginal likelihood of the traits given the network structure G and can be simplified to be $P(Y|G,W)$, and $P(G,W|X,B)$ is a prior

probability of a network and weights given marker information and biological knowledge. In the second equality relation, the prior probability $P(G, W|X, B)$ can be decomposed into $P(G_Y, W|X, B)$ and $P(G_{Q\to Y}|X, B)$ by assuming the joint independence between a phenotype network G_Y along with the weights W and genetic architectures $G_{Q\to Y}$ given marker information X and biological knowledge B. The third equality is provided by the fact that $P(G_Y, W|X, B) = P(G_Y, W|B)$ because the genetic markers are not included in the structure of phenotype network G_Y, and $P(G_{Q\to Y}|X, B) = P(G_{Q\to Y}|X)$ because the biological structure B is about the phenotype network structure. The extended model in (2) shows that prior distributions on phenotype network structure $P(G_Y|B, W)$, biological knowledge weights $P(W|B)$, and genetic architectures $P(G_{Q\to Y}|X)$ are needed to be specified. We will describe how to set $P(G_Y|B, W)$, $P(W|B)$, $P(G_{Q\to Y}|X)$ in the following.

Prior on phenotype network structures $\boldsymbol{P(G_Y|B, W)}$ Incorporation of *a priori* biological knowledge into a prior on network structures can lead to discriminate Bayesian networks of the same likelihood [21,33]. If G^1 and G^2 have the same likelihood ($P(Y|G^1) = P(Y|G^2)$) but have different prior probabilities ($P(G^1) \neq P(G^2)$), the posterior probabilities would become different ($P(G^1|Y) \neq P(G^2|Y) \propto P(Y|G^2)P(G^2)$). For example, if it is known *a priori* that $t \to v$ is more likely than $t \leftarrow v$, the posterior prefers $p(t \to v|Y) \propto p(Y|t \to v)p(t \to v)$ over $p(t \leftarrow v|Y) \propto p(Y|t \leftarrow v)p(t \leftarrow v)$.

Various types of information can supplement the learning of a network of gene expression. We can encode this supplementary information into unequal priors on network structures. A transcription factor binding location can be used to prefer the direction from a transcription factor to the target gene [34]. Pathway information can also guide to infer directions among phenotypes [21]. Regulation inference [35,36,37] from knock-out data and protein-protein interaction [38] can be used as a prior for network structure. We will describe how to encode this information in Section 3.3. Since QTLnet is a Bayesian approach, we can flexibly incorporate various sources of biological knowledge through constructing meaningful priors for the network structures.

Now, it remains to set the prior distribution on phenotype network structure G_Y with respect to biological knowledge B. It is known that a graphical model (Markov random field) for an undirected graph has a Gibbs distribution and vice versa [39,40]. As a Markov random field's distribution is factored by its cliques $\prod_{C:cliques} \phi(Y_C)$, a Bayesian network's distribution is factored by its parent-child relations $\prod_t P(Y_t|Y_{pa(t)})$. Hence, it is natural to assume the Gibbs distribution for the prior on DAG structures [20]. Imoto *et al.* [20] and Werhli and Husmeier [21] used the Gibbs distribution as the prior distribution over network structures to integrate biological knowledge. We adapted the prior formulation over network structures in [21] as follows. First, the *energy* of a phenotype network G_Y relative

to the biological knowledge B is defined to be

$$\mathcal{E}(G_Y) = \sum_{i,j=1}^{T} |B(i,j) - G_Y(i,j)| \tag{3}$$

where B is an encoding to describe biological knowledge ranging from 0 to 1, G_Y is represented by an adjacency matrix of network structure, $G_Y(i,j) = 1$ means the presence of the directed edge from node i to j and $G_Y(i,j) = 0$ means the absence of the directed edge from i to j.

The energy $\mathcal{E}(G_Y)$ acts as a distance measure between biological knowledge and a network structure G_Y. For a fixed biological knowledge matrix B, there are network structures close to the biological knowledge B which agree with the knowledge well and hence have small energy, while there are network structures distant from B which disagree with the knowledge and hence have large energy. Therefore, the prior distribution on network structures can be constructed in terms of energy $\mathcal{E}(G_Y)$ to be adjusted by biological knowledge B.

$$P(G_Y|B,W) = \frac{\exp(-W\mathcal{E}(G_Y))}{Z(W)}, \quad G_Y \in \text{DAG}.$$

For a fixed W, network structures with small energy will have higher prior probabilities than network structures with large energy. W, the weight of biological knowledge B, is introduced to tune the confidence of biological information since biological information can be inaccurate or incompatible with expression data. As $W \to 0$, the influence of *a priori* knowledge gets negligible and the prior distribution of network structure is assumed to be almost uniform. On the contrary, as $W \to \infty$, the prior on network structure peaks at the biological knowledge. $Z(W)$ denotes the normalizing constant $\sum_{G_Y \in \text{DAG}} \exp(-W\mathcal{E}(G_Y))$.

Multiple sources of biological knowledge can be integrated into a prior on network structures with different weights.

$$P(G_Y|B,W) = \frac{\exp(-\sum_k W_k \mathcal{E}_k(G_Y))}{Z(W)}, \quad G_Y \in \text{DAG}$$

where B_k is an encoding matrix of biological knowledge from source k, W_k indicates a weight of B_k relative to the data, $W = (W_1, \ldots, W_k)$, and $Z(W)$ is the summation of the numerator over all DAGs.

Prior on biological knowledge weights $P(W|B)$ We specify the prior probability distribution on each biological knowledge weight W to be an exponential distribution, $W \propto \exp(-\lambda W)$, with the rate parameter $\lambda = 1$. The exponential distribution is chosen to control the case when biological knowledge disagrees with expression data, so that it can easily reduce the contribution of negative biological knowledge.

Prior on genetic architectures $P(G_{Q\to Y}|X)$ We assume a prior on genetic architectures to be a uniform distribution. Several alternative specifications can be found in Bayesian QTL mapping such as [41] and [42].

3.2 Sketch of MCMC

A main challenge in the reconstruction of networks is that the graph space grows super-exponentially with the number of nodes. An exhaustive search approach over all network structures is impractical even for small networks. Hence, heuristic approaches are needed to efficiently traverse the graph space. We adopt a Metropolis-Hastings (M-H) MCMC scheme that integrates the sampling of network structures [31,43], the QTL mapping, and the sampling of biological knowledge weights W. The MCMC scheme iterates between accepting a network structure G and accepting k weights $W_1, \cdots, W_k$ corresponding to k types of biological knowledge.

1. Sample a new phenotype network structure G_Y^{new} from a network structure proposal distribution $R(G_Y^{new}|G_Y^{old})$.
2. Given the phenotype network structure G_Y^{new}, sample a new genetic architecture $G_{Q\to Y}$ from an architecture proposal distribution $R(G_{Q\to Y}^{new}|G_{Q\to Y}^{old})$.
3. Accept the new extended network structure G^{new} composed of G_Y^{new} and $G_{Q\to Y}^{new}$ given the biological knowledge weights W with a probability

$$A_G = \min\{1, \frac{P(Y|G^{new},X)P(G_Y^{new}|B,W)P(G_{Q\to Y}^{new}|X)}{P(Y|G^{old},X)P(G_Y^{old}|B,W)P(G_{Q\to Y}^{old}|X)} \times \frac{R(G_Y^{old}|G_Y^{new})R(G_{Q\to Y}^{old}|G_{Q\to Y}^{new})}{R(G_Y^{new}|G_Y^{old})R(G_{Q\to Y}^{new}|G_{Q\to Y}^{old})}\}.$$

4. For each biological knowledge k,
   (a) Sample a new weight W_k^{new} of biological knowledge k from a weight proposal distribution $R(W_k^{new}|W_k^{old})$.
   (b) Accept the new biological weight W_k^{new} given the phenotype network G_Y with a probability

$$A_{W_k} = \min\{1, \frac{P(G_Y|W_k^{new},W_{-k}^{old},B)}{P(G_Y|W^{old},B)}\frac{P(W_k^{new}|B)}{P(W_k^{old}|B)}\frac{R(W_k^{old}|W_k^{new})}{R(W_k^{new}|W_k^{old})}\}.$$

5. Iterate the steps 1-4 until the chain converges.

In step 1, a new phenotype network structure is proposed by a mixture of single edge operations (single edge addition, single edge deletion, single edge reversal) and edge reversal moves [32]. It has been shown that edge reversal moves significantly improve the convergence of MCMC sampler.

In step 2, genetic architectures of phenotypes can be sampled conditional on its phenotypic parents. One way is a Bayesian QTL mapping proposed in Yi *et al.* [41] for each phenotype. Another way is the interval mapping of QTL for each phenotype conditional on its phenotypic parents. The interval mapping of QTL is a fast algorithm approximating the Bayesian mapping of QTL though it might fail to satisfy the irreducibility of the Markov Chain. We use the interval mapping for practical reasons.

In step 3, the computation of the ratio of marginal likelihood $P(Y|G,X)$, or Bayes factor, can be approximated by the difference of BIC scores [44] when the sample size is big,

$$\frac{P(Y|G^{new},X)}{P(Y|G^{old},X)} \approx \exp(-\frac{1}{2}(BIC_{G^{new}} - BIC_{G^{old}})).$$

In step 4, we need to compute

$$\frac{P(G_Y|W^{new})}{P(G_Y|W^{old})} = \frac{\frac{\exp(-W^{new}\mathcal{E}(G_Y))}{Z(W^{new})}}{\frac{\exp(-W^{old}\mathcal{E}(G_Y))}{Z(W^{old})}}$$

where $Z(W) = \sum_{G_Y \in \text{DAG}} \exp(-W\mathcal{E}(G_Y))$ is a normalizing constant. Note that it is not feasible to compute the exact $Z(W)$ due to the exclusion of cyclic networks. We approximate the normalizing constant by the summation over directed graphs with restriction on the number of parents, e.g., 3 as adopted by [21].

After running a MCMC chain, we need to efficiently summarize the chain for the inference of a network structure. The choice by the highest posterior network structure might not produce a convincing model because the graph space grows rapidly with the number of phenotype nodes and the most probable network structure might still have a very low probability. Therefore, instead of selecting the network structure with the highest posterior probability, we perform Bayesian model averaging [45] over the causal links between phenotypes to infer an averaged network. Explicitly, let Δ_{uv} represents a causal link between phenotypes u and v, that is, $\Delta_{uv} = \{Y_u \rightarrow Y_v\ ,\ Y_u \leftarrow Y_v\ ,\ Y_u \not\rightarrow Y_v$ and $Y_u \not\leftarrow Y_v\}$. Then

$$\begin{aligned} P(\Delta_{uv} \mid Y, X) &= \sum_G P(\Delta_{uv} \mid G, Y, X)\, P(G \mid Y, X) \\ &= \sum_G \mathbb{1}\{\Delta_{uv} \in G\}\, P(G \mid Y, X). \end{aligned}$$

The averaged network is represented by the causal links with maximum posterior probability or with posterior probability above a predetermined threshold, e.g., 0.5.

3.3 Summary of encoding of biological knowledge

In Section 3.1, we construct a prior distribution on network structures in terms of energy $\mathcal{E}(G_Y)$ relative to biological knowledge B. Now we describe how to encode a biological knowledge matrix B from several biological information. When

there is no available knowledge, we would put every element in B as $1/2$. Since every DAG has the same energy, the probability of a network structure conditional on W is $1/K$ with K as the number of all DAGs. We will look at several ways in which biological knowledge can be used to encode B - transcription factor and DNA binding [34], protein-protein interaction [46], and gene ontology annotations [47].

Transcription factor and DNA binding Bernard and Hartemink [34] suggested an approach to convert p-values, quantifying how well a transcription factor binds to putative target genes, into a posterior probability for the presence and directionally of an edge in a Bayesian network. Following [34] we assume that the p-value follows a truncated exponential distribution with mean λ when the TF binds to DNA ($G(i,j) = 1$) and a uniform distribution when the TF does not bind to DNA ($G(i,j) = 0$).

$$P_\lambda(P_{ij} = p|G(i,j) = 1) = \frac{\lambda e^{-\lambda p}}{1 - e^{-\lambda}},$$
$$P_\lambda(P_{ij} = p|G(i,j) = 0) = 1.$$

The presence of an edge before observing any biological data is assumed to be $P(G(i,j) = 1) = 1/2$ so that without any biological data, the probability of the presence of the edge only depends on the expression data. By the Bayes' rule, the probability of presence of an edge after observing a p-value is

$$P_\lambda(G(i,j) = 1|P_{ij} = p) = \frac{\lambda e^{-\lambda p}}{\lambda e^{-\lambda p} + (1 - e^{-\lambda})}.$$

Here λ is assumed to be uniformly distributed over the interval $[\lambda_L, \lambda_H]$ and the integration over λ is performed to get the probability of the presence of an edge,

$$P(G(i,j) = 1|P_{ij} = p) = \frac{1}{\lambda_H - \lambda_L} \int_{\lambda_L}^{\lambda_H} \frac{\lambda e^{-\lambda p}}{\lambda e^{-\lambda p} + (1 - e^{-\lambda})} d\lambda.$$

This can be solved numerically, i.e. $\lambda \in [0, 10000]$. We would get the estimate of $B(i,j) = P(G(i,j) = 1|P_{ij} = p)$.

Protein-protein interaction Since protein-protein interaction does not have the directionality, we put the same probability on both directions. If we do not consider the diverse reliabilities of protein-protein interaction experiments, we put $B(i,j) = B(j,i) = \delta > 1/2$ when we find any interaction. If we consider the diverse reliabilities and have gold and negative standards for protein-protein interactions, we can use the Bayes classifier proposed by Jansen *et al.* [46] to combine heterogeneous data. Suppose there are L data sets of protein interactions

and each data set has a different false positive rate. We can calculate the posterior odds of an interaction from observations $f_1, \ldots, f_L$ from L data sets,

$$\begin{aligned} O_{posterior} &= \frac{P(pos|f_1, \ldots, f_L)}{P(neg|f_1, \ldots, f_L)} = O_{prior} \times LR \\ &= \frac{P(pos)}{P(neg)} \times \frac{P(f_1, \ldots, f_L|pos)}{P(f_1, \ldots, f_L|neg)}. \end{aligned}$$

In a set of protein pairs with the same observation values $f_1, \ldots, f_L$, we get the relative occurrence rates of protein pairs in the positive gold standard and in the negative gold standard. The likelihood ratio is the ratio of two relative occurrence rates and the prior odds can be defined by an expert. The encoding of B can be obtained by transforming the posterior odds into posterior positive rate,

$$B(i,j) = B(j,i) = \frac{O_{posterior}}{1 + O_{posterior}}.$$

When the posterior odds is 1, $B(i,j) = B(j,i) = 1/2$. As the posterior odds increases, $B(i,j) = B(j,i)$ increases.

Gene ontology The Gene Ontology (GO) [48] is a well controlled vocabulary of terms describing the molecular functions, biological processes, and cellular components of a gene. A large fraction of genes are annotated with GO terms. The distance between two genes can be defined in terms of their GO annotations. One well defined distance is Lord's similarity [47]. This measure takes into account the hierarchy of GO ontology and GO term occurrences in the myriad of genes. If two genes share a more specific GO term, in the below of the GO hierarchy, they are more likely to be similar. However, even if the shared GO terms lie in the same level of the hierarchy, the similarity of two genes can differ by how informative the term is. Suppose there are one hundred genes annotated with term c_1 and there are one thousand genes annotated with term c_2. Then the chance that gene g_1 and gene g_2 share the term c_2 is higher than the chance sharing the term c_1. Therefore, term c_1 is more informative. In consideration of the GO hierarchy and frequency of a GO term, the information content $IC(c)$ for a GO term c is defined to be the negative logarithm of the number of times the term or any of its descendant terms occurs in the myriad of genes divided by total GO term occurrences. The root of the hierarchy will have zero information content while the leaf of the hierarchy will have high information content. Once the information content $IC(c)$ for each node in the GO ontology is set up, we can define GO term similarity and gene similarity. The similarity between two GO terms is defined to be the maximum information content among the shared parents of the two terms, which is

$$sim(c_1, c_2) = \max_{c \in (pa(c_1) \cap pa(c_2))} IC(c).$$

Then, the similarity between two genes can be defined to be the average similarities of pairs of GO terms between two genes, which is

$$sim(g_1, g_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} sim(c_{1,i}, c_{2,j})}{nm}.$$

This Lord's measure can be used as an encoding of B if it is rescaled to be in the interval $[0, 1]$.

4 Simulations

We performed a simulation study for comparing the proposed method (QTLnet-prior) with three other methods - QTLnet [16], WH-prior [21], and Expression. Table 3 gives a summary of these four methods in terms of using the genetic variation information and biological knowledge. The QTLnet was implemented using R/QTLnet, the QTLnet-prior was implemented with prior setting on R/QTLnet, the WH-prior was programmed as in [21] with a modification of approximating the marginal likelihood with the BIC score instead of using the BGe score [49]. The Expression was programmed by modifying R/QTLnet excluding QTL mapping.

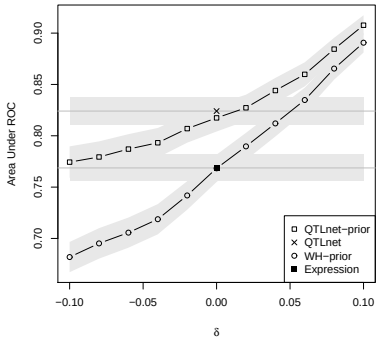
Method	Use Genetic Variation Information	Use Biological Knowledge
QTLnet-prior	YES	YES
QTLnet	YES	NO
WH-prior	NO	YES
Expression	NO	NO

Table 3: Four methods which differ in the use of genetic variation information and biological knowledge.

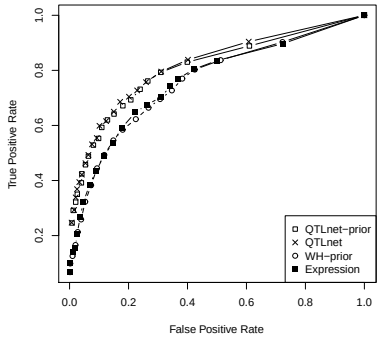
We simulated expression data and *a priori* knowledge matrix according to the network topology in Figure 1 for 100 times. To generate expression data based on the network in Figure 1, the genetic information was simulated first. The genetic map had 5 chromosomes of length 100cM with 10 equally spaced markers in each chromosome and the markers were simulated for 500 mice in an F2 population using R/qtl [50]. We assumed QTL Q_t is located in the middle of chromosome t. Then, each expression data set of F2 population was realized with different genetic effects and partial regression coefficients between phenotypes. Genetic additive effects were sampled from a uniform distribution $U[0, 0.5]$ and dominance effects were sampled from $U[0, 0.25]$. The partial regression coefficients β_{uv} were sampled from $U[-0.5, 0.5]$. The residual phenotypic variance was 1. Biological knowledge matrix B was generated for several cases. $B(t, u)$ was

generated from one of two $[0, 1]$-truncated normal distributions $N_{\pm}(0.5 \pm \delta, 0.1)$ [51]. $B(t, u)$ of true edge was generated from N_{+} and $B(t, u)$ of false edge was generated from N_{-}. The parameter δ controls the accuracy of the prior knowledge. We examined eleven cases of different accuracies of prior knowledge, $\delta = \pm 0.1, \pm 0.08, \pm 0.06, \pm 0.04, \pm 0.02, 0$. In the extreme case when $\delta = 0.5$, the prior knowledge almost correctly reflects the network structure while when $\delta = -0.5$, the prior knowledge reflects the network structure almost in the opposite way. When $\delta = 0$, the information is generated with no distinction between true and false edges. In each simulated data, we ran a Markov chain Monte Carlo for 30300 iterations, discarded the first 300 iterations, sampled every 10 iterations, and generated 3000 samples.

We assessed these four methods by using receiver operator characteristic (ROC) curves of the proportion of recovered and spurious edges. Bigger areas under the ROC curve generally indicate better performance, as the area represents the probability that the classifier ranks true edges higher than false edges [52]. The ROC curves are obtained from the set of proportions of recovered edges and spurious edges for various posterior probability thresholds ranging from 0 to 1.



(a) The areas under ROC curves of QTLnet-prior, QTLnet, WH-prior, and Expression. The areas under ROC curves of QTLnet-prior and WH-prior are plotted against the accuracy of biological knowledge, δ. Since QTLnet and Expression do not incorporate biological knowledge, they are plotted in a single point each ($\times$, ■). The shaded area indicates the standard error of the area under ROC curve.



(b) The ROC curves of QTLnet-prior and WH-prior are drawn when noninformative biological knowledge ($\delta = 0$) is incorporated. They are compared with the ROC curves of QTLnet and Expression which do not incorporate biological information.

Figure 4: ROC curves

First, we evaluate the effect of incorporation of genetic variation information. The effect of QTL mapping can be tested by comparing QTLnet-prior and WH-prior. QTLnet-prior has more power in recovering the network structure than WH-prior in Figure 4a and we can conclude that QTL mapping increases the power. This can be explained by that QTL mapping increases the differences in likelihood between true model and wrong model. Even when prior probability is entailed by negative knowledge, the likelihood increase by QTL mapping can overcome the prior probability.

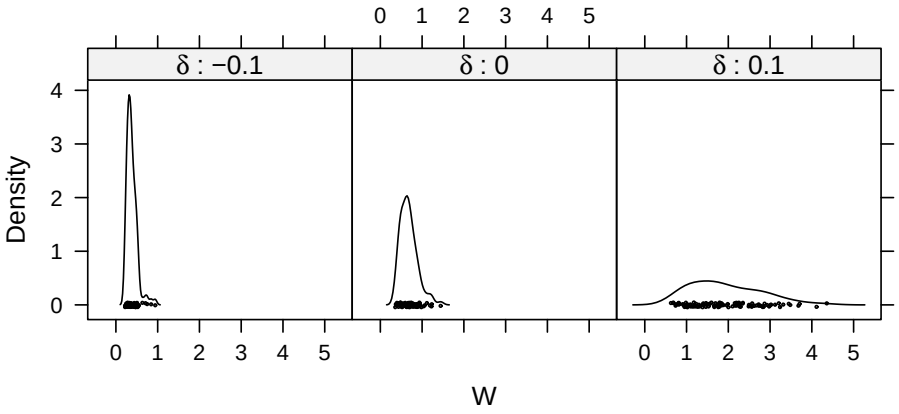


Figure 5: The distribution of median weight W of posterior sample by QTLnet-prior inference. Each panel shows the median W distribution when biological knowledge is defective ($\delta = -0.1$), noninformative ($\delta = 0$), and informative ($\delta = 0.1$).

Second, we evaluate the effect of incorporation of biological knowledge. In Figure 4a, when δ is positive, QTLnet-prior performs better than QTLnet and WH-prior performs better than Expression, whereas when δ is negative, QTLnet-prior performs worse than QTLnet and WH-prior performs worse than Expression. With a positive δ, as the accuracy of knowledge increases, QTLnet-prior and WH-prior benefit by the prior knowledge incorporation. However, a negative δ, indicating that the knowledge disagrees with the true network structure, makes QTLnet-prior and WH-prior be harmed by prior knowledge incorporation. The decreased performances in QTLnet-prior and WH-prior bring in the attention whether W can effectively control the influence of negative knowledge. Figure 5 shows that the median of W in the posterior sample is close to 0 with negative knowledge. It implies that the weight W effectively controls the use of negative knowledge but not completely. In comparison with QTLnet and Expression, the reduced performance of QTLnet-prior and WH-prior can be explained by the remaining uncontrolled ef-

fect of prior probability incorporating negative knowledge. When noninformative knowledge is incorporated, there is no significant difference in area under ROC curve between QTLnet and the QTLnet-prior (p-value=0.73) and between the Expression and the WH-prior (p-value=0.99) as shown in Figures 4a and 4b.

5 Conclusion

We have developed a network inference method (QTLnet-prior) to incorporate genetic variation information and biological knowledge. Genotypes are known to control phenotypes but not the other way and thereby can help to distinguish phenotype network structures. Biological knowledge can improve the clustering and directional inference between phenotypes. The simulation study shows that the proposed method can improve the reconstruction of network by integrating genetic variation information and biological knowledge as long as knowledge agrees with data. When knowledge does not agree with data, the weight of prior knowledge controls the contribution of prior on the likelihood of data to some extent but not completely, which results in decreased performance.

When we interpret the inferred networks, we need to be cautious. Even though, in theory, the incorporation of causal QTLs allows us to distinguish network structures that would otherwise be likelihood equivalent, in practice some of the detected expression-to-expression causal relationships might be invalid. The problem is that the inferred expression network represents a projection of real causal relationships that might take place outside the transcriptional regulation level. For instance, the true causal regulations could be due to transcription factor binding, direct protein-protein interaction, phosphorylation, methylation, etc, and might not be well reflected at the gene expression level. The incorporation of diffused biological knowledge, mined from different levels of biological regulation, could potentially improve the reconstruction of gene-expression regulatory networks. In any case, the inference of these networks can still play an important role in generating hypothetically possible causal relations.

There are several factors that could change the inference by QTLnet-prior. One is the prior distribution specification. We have used the Gibbs distribution as a prior distribution on network structures ($P(G_Y|B,W) = \exp(-W\mathcal{E}(G_Y))/Z(W)$) with an absolute distance measure ($\mathcal{E}(G_Y) = \sum_{i,j=1}^{T} |B(i,j) - G_Y(i,j)|$) to incorporate biological knowledge and the exponential distribution for the weight of biological knowledge ($W \propto \exp(-\lambda W)$) with the rate parameter ($\lambda = 1$) in (2). However, we could consider different choices of network structure distributions, measures to incorporate information, weight distributions, and hyperparameters. Another factor is the sample size of expression data. As the sample size increases, the contribution of biological knowledge will be generally reduced. This puts unequal contribution of expression data and biological knowledge to the reconstruction of network, even though biological knowledge B can also be obtained from number of experiments as discussed in [21]. Finally, the encoding of biological

knowledge plays an important role. We have proposed to use the encoding for transcription factor and its targets by [34], protein-protein interaction by [46], and gene ontology annotations by [47]. These encodings are mainly about direct relationships in separate biological regulation levels. As discussed in the previous paragraph, this diffused biological knowledge could improve the Bayesian network reconstruction.

There are shortcomings of QTLnet-prior framework inherited from QTLnet. One of the assumptions of QTLnet is no latent variables. Latent variables can make it impossible to find the marginalized model in the class of DAG as shown in [53] and can induce erroneous relations. Suppose there are three nodes y_1, y_2, y_3, and y_1 and y_2 have a common parent c_1, and y_2 and y_3 have a common parent c_2. If the common parents c_1 and c_2 are not observed, we get the independence relations that $y_1 \perp y_3$ and $y_1 \not\perp y_3 \mid y_2$. Then we mistakenly infer that y_1 and y_3 are parents of y_2. One approach to overcome this problem is to consider the more general class of ancestral graphs, which takes care of latent variables. Ancestral graphs open up the possibility of latent variables while they do not explicitly include the latent variables in the network structures [53].

A persistent challenge in Bayesian network analysis is to cope with large networks since the DAG space size grows super-exponentially with the number of nodes. Approaches based on Markov blankets with and without restrictions on the number of parent nodes have been proposed [54,55,56]. Jaakkola *et al.* [57] approximated the Bayesian network problem to a linear programming problem. Tamada *et al.* [58] developed a parallel algorithm that infers subnetworks restricted on a Markov blanket and merges the subnetworks. Likewise, in phylogeny estimation, the supertree reconstruction from small trees has been studied [59]. We think the rigorous development of super Bayesian network methodology to integrate small subnetworks is a promising direction to infer a large network since the inference of small subnetworks is computationally inexpensive and multiple subnetworks can be parallelized for computation. In the era of vast biological data and knowledge in various aspects, integrating them reasonably in a large scale can be an interesting topic for future research.

Acknowledgements

The authors wish to thank NIH/NIDDK 58037 (JYM, BSY) and 66369 (JYM, BSY) and NIH/NIGMS 74244 (JYM, BSY) and 69430 (JYM, BSY) and NCI ICBP U54-CA149237 (ECN) and NIH R01MH090948 (ECN) for supporting this work. Alan D. Attie and Mark Keller motivated the work through our collaboration on causal models for diabetes and obesity.

References

[1] Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**, 601–620.

[2] Gardner, T. S., di Bernardo, D., Lorenz, D., and Collins, J. J. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **301**, 102–105.

[3] Zhu, J., et al. (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenetic Genome Research*, **105**, 363–374.

[4] Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., Bumgarner, R. E., and Schadt, E. E. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet*, **40**, 854–861.

[5] Winrow, C. J., et al. (2009) Uncovering the genetic landscape for multiple sleep-wake traits. *PLoS ONE*, **4**, e5161.

[6] Schadt, E. E., et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, **37**, 710–717.

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

[8] Chen, L. S., Emmert-Streib, F., and Storey, J. D. (2007) Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome biology*, **8**, R219.

[9] Millstein, J., Zhang, B., Zhu, J., and Schadt, E. (2009) Disentangling molecular relationships with a causal inference test. *BMC Genetics*, **10**, 23.

[10] Chaibub Neto, E., Keller, M. P., Broman, A. T., Attie, A. D., and Yandell, B. S. (2010) Causal model selection tests in systems genetics. Tech. Rep. 1157, Department of statistics, University of Wisconsin-Madison.

[11] Li, R., Tsaih, S.-W., Shockley, K., Stylianou, I. M., Wergedal, J., Paigen, B., and Churchill, G. A. (2006) Structural model analysis of multiple quantitative traits. *PLoS Genet*, **2**, e114.

[12] Liu, B., de la Fuente, A., and Hoeschele, I. (2008) Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics*, **178**, 1763–1776.

[13] Aten, J. E., Fuller, T. F., Lusis, A. J., and Horvath, S. (2008) Using genetic markers to orient the edges in quantitative trait networks: The NEO software. *BMC Systems Biology*, **2**, 34.

[14] Chaibub Neto, E., Ferrara, C. T., Attie, A. D., and Yandell, B. S. (2008) Inferring causal phenotype networks from segregating populations. *Genetics*, **179**, 1089–1100.

[15] Valente, B. D., Rosa, G. J. M., de los Campos, G., Gianola, D., and Silva, M. A. (2010) Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, **185**, 633–644.

[16] Chaibub Neto, E., Keller, M. P., Attie, A. D., and S., Y. B. (2010) Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *Ann. Appl. Stat.*, **4**, 320–339.

[17] Hageman, R. S., Leduc, M. S., Korstanje, R., Paigen, B., and Churchill, G. A. (2011) A Bayesian framework for inference of the genotype-phenotype map for segregating populations. *Genetics*, **187**, 1163–1170.

[18] Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., and Miyano, S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**, II227–II236.

[19] Nariai, N., Kim, S., Imoto, S., and Miyano, S. (2004) Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pac Symp Biocomput*, pp. 336–47.

[20] Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2004) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *J Bioinform Comput Biol*, **2**, 77–98.

[21] Werhli, A. V. and Husmeier, D. (2007) Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, **6**, 15.

[22] Christley, S., Nie, Q., and Xie, X. (2009) Incorporating existing network information into gene network inference. *PLoS ONE*, **4**, e6799.

[23] Kao, C.-H. and Zeng, Z.-B. (2002) Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics*, **160**, 1243–1261.

[24] Nadeau, J. H. and Dudley, A. M. (2011) Systems genetics. *Science*, **331**, 1015–1016.

[25] Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference*. Morgan Kaufmann.

[26] Pearl, J. (2000) *Causality: Models, Reasoning and Inference*. Cambridge University Press.

[27] Spirtes, P., Glymour, C., and Scheines, R. (2000) *Causation, Prediction, and Search*. The MIT Press, second edn.

[28] Heckerman, D., Meek, C., and Cooper, G. (2006) A Bayesian approach to causal discovery. Holmes, D. and Jain, L. (eds.), *Innovations in Machine Learning*, vol. 194 of *Studies in Fuzziness and Soft Computing*, pp. 1–28, Springer Berlin / Heidelberg.

[29] Verma, T. and Pearl, J. (1990) *Equivalence and synthesis of causal models. In Readings in Uncertain Reasoning (G. Shafer and J. Pearl eds.)*. Morgan Kaufmann.

[30] Heckerman, D. and Geiger, D. (1996) Likelihoods and parameter priors for Bayesian networks. Tech. Rep. MSR-TR-95-94, Microsoft Research.

[31] Madigan, D. and York, J. (1995) Bayesian graphical models for discrete data. *Int. Stat. Rev.*, **63**, 215–232.

[32] Grzegorczyk, M. and Husmeier, D. (2008) Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, **71**, 265–305.

[33] Zhu, J., Wiener, M. C., Zhang, C., Fridman, A., Minch, E., Lum, P. Y., Sachs, J. R., and Schadt, E. E. (2007) Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol*, **3**, e69.

[34] Bernard, A. and Hartemink, A. J. (2005) Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data. *Pacific Symposium on Biocomputing 2005*, pp. 459–470.

[35] Peleg, T., Yosef, N., Ruppin, E., and Sharan, R. (2010) Network-free inference of knockout effects in yeast. *PLoS Comput Biol*, **6**, e1000635.

[36] Yeang, C. H., Ideker, T., and Jaakkola, T. (2004) Physical network models. *Journal of Computational Biology*, **11**, 243–262.

[37] Ourfali, O., Shlomi, T., Ideker, T., Ruppin, E., and Sharan, R. (2007) SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, **23**, I359–I366.

[38] Imoto, S., Kim, S., Goto, T., Miyano, S., Aburatani, S., Tashiro, K., and Kuhara, S. (2003) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J Bioinform Comput Biol*, **1**, 231–52.

[39] Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **6**, 721–741.

[40] Kindermann, R. and Snell, J. L. (1980) Markov random fields and their applications. *Contemporary Mathematics*, **1**, 142.

[41] Yi, N. J., Yandell, B. S., Churchill, G. A., Allison, D. B., Eisen, E. J., and Pomp, D. (2005) Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, **170**, 1333–1344.

[42] Yi, N. J., Shriner, D., Banerjee, S., Mehta, T., Pomp, D., and Yandell, B. S. (2007) An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. *Genetics*, **176**, 1865–1877.

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

[44] Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

[45] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382–417.

[46] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003) A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*, **302**, 449–453.

[47] Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

[48] Ashburner, M., et al. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

[49] Heckerman, D., Geiger, D., and Chickering, D. M. (1995) Learning Bayesian networks - the combination of knowledge and statistical-data. *Machine Learning*, **20**, 197–243.

[50] Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, **19**, 889–890.

[51] Geier, F., Timmer, J., and Fleck, C. (2007) Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Systems Biology*, **1**, 11.

[52] Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recogn. Lett.*, **27**, 861–874.

[53] Richardson, T. and Spirtes, P. (2002) Ancestral graph Markov models. *The Annals of Statistics*, **30**, 962–1030.

[54] Riggelsen, C. (2005) MCMC learning of Bayesian network models by Markov blanket decomposition. Gama, J., Camacho, R., Brazdil, P., Jorge, A., and Torgo, L. (eds.), *Machine Learning: ECML 2005*, vol. 3720 of *Lecture Notes in Computer Science*, pp. 329–340, Springer Berlin / Heidelberg.

[55] Schmidt, M., Niculescu-Mizil, A., and Murphy, K. (2007) Learning graphical model structure using L1-regularization paths. *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, pp. 1278–1283, AAAI Press.

[56] Perrier, E., Imoto, S., and Miyano, S. (2008) Finding optimal Bayesian network given a super-structure. *Journal of Machine Learning Research*, **9**, 2251–2286.

[57] Jaakkola, T., Sontag, D., Globerson, A., and Meila, M. (2010) Learning Bayesian network structure using LP relaxations. *Journal of Machine Learning Research - Proceedings Track*, **9**, 358–365.

[58] Tamada, Y., Imoto, S., and Miyano, S. (2011) Parallel algorithm for learning optimal Bayesian network structure. *Journal of Machine Learning Research*, **12**, 2437–2459.
[59] Bininda-Emonds, O. R. P., Gittleman, J. L., and Steel, M. A. (2002) The (Super)tree of life: Procedures, problems, and prospects. *Annual Review of Ecology and Systematics*, **33**, 265–289.