Janhavi Moharil, Paul May, Daniel P. Gaile and Rachael Hageman Blair*

Belief propagation in genotype-phenotype networks

DOI 10.1515/sagmb-2015-0058

Abstract: Graphical models have proven to be a valuable tool for connecting genotypes and phenotypes. Structural learning of phenotype-genotype networks has received considerable attention in the post-genome era. In recent years, a dozen different methods have emerged for network inference, which leverage natural variation that arises in certain genetic populations. The structure of the network itself can be used to form hypotheses based on the inferred direct and indirect network relationships, but represents a premature endpoint to the graphical analyses. In this work, we extend this endpoint. We examine the unexplored problem of perturbing a given network structure, and quantifying the system-wide effects on the network in a node-wise manner. The perturbation is achieved through the setting of values of phenotype node(s), which may reflect an inhibition or activation, and propagating this information through the entire network. We leverage belief propagation methods in Conditional Gaussian Bayesian Networks (CG-BNs), in order to absorb and propagate phenotypic evidence through the network. We show that the modeling assumptions adopted for genotype-phenotype networks represent an important sub-class of CG-BNs, which possess properties that ensure exact inference in the propagation scheme. The system-wide effects of the perturbation are quantified in a node-wise manner through the comparison of perturbed and unperturbed marginal distributions using a symmetric Kullback-Leibler divergence. Applications to kidney and skin cancer expression quantitative trait loci (eQTL) data from different *mus musculus* populations are presented. System-wide effects in the network were predicted and visualized across a spectrum of evidence. Sub-pathways and regions of the network responded in concert, suggesting co-regulation and coordination throughout the network in response to phenotypic changes. We demonstrate how these predicted system-wide effects can be examined in connection with estimated class probabilities for covariates of interest, e.g. cancer status. Despite the uncertainty in the network structure, we demonstrate the system-wide predictions are stable across an ensemble of highly likely networks. A software package, geneNetBP, which implements our approach, was developed in the R programming language.

Keywords: bayesian network; belief propagation; expression QTL; gene networks; genotype-phenotype.

1 Introduction

The inverse problem of reverse engineering a network from observational data is a major challenge in Systems Biology and related fields. Networks that connect genotype to phenotype promote a deeper understanding of the complex interactions underlying disease and hold tremendous promise for personalized medicine. Phenotype-genotype network inference leverages the natural variation that arises in segregating genetic populations Benfey and Mitchell-Olds (2008), Rockman (2008). The data consists of genotypes at markers

*Corresponding author: Rachael Hageman Blair, Department of Biostatistics, State University of New York at Buffalo, 3435 Main Street, 709 Kimball Tower, Buffalo, NY 14214, USA, e-mail: hageman@buffalo.edu

Janhavi Moharil: Department of Biostatistics, State University of New York at Buffalo, 3435 Main Street, 720 Kimball Tower, Buffalo, NY 14214, USA; and Department of Chemical and Biological Engineering, University at Buffalo, 908 Furnas Hall, Amherst, NY 14260, USA

Paul May: Department of Biostatistics, State University of New York at Buffalo, 3435 Main Street, 720 Kimball Tower, Buffalo, NY 14214, USA

Daniel P. Gaile: Department of Biostatistics, State University of New York at Buffalo, 3435 Main Street, 718 Kimball Tower, Buffalo, NY 14214, USA

throughout the genome, and phenotypes, which can be broadly defined as any complex trait, e.g. clinical traits or arising from array-based profiling Jansen and Nap (2001). Nodes in the network represent measured variables in the biological system and the edges between them reflect the inferred direct and indirect relationships between them. Therefore, the topology itself can be viewed as predictive of the direct and indirect associations between variables in the network.

Structural learning of directed graphs is an *NP*-hard problem for which an approximate solution can be computationally intensive for even a small number of variables Chickering et al. (1994). In the last decade, a broad spectrum of modeling paradigms have emerged for genotype-phenotype inference. The proposed inference methods have largely focused on the structural learning aspect, which concerns the estimation of the network topology. There is a secondary layer of inference required for parameter learning, which is less emphasized. Existing approaches can be roughly categorized depending on the domain of biological variables used to make the inferences. Pairwise methods focus on relationships between pairs of phenotypes with a common quantitative trait loci (QTL) Schadt et al. (2005), Kulp and Jagalur (2006), Aten et al. (2008), Millstein et al. (2009), Neto et al. (2013). Whole-network inference takes a multivariate approach to simultaneously learning relationships between all variables in the network through a score-based greedy or sampling search over possible structures Schadt et al. (2005), Li et al. (2006), Zhu et al. (2007, 2008), Benfey and Mitchell-Olds (2008), Liu et al. (2008), Neto et al. (2008, 2010), Hageman et al. (2011b).

Recently, considerable effort has been made to address some of the shortcomings and limitations of these networks. Shortcomings include sensitivity to subtle correlation patterns in the data Li et al. (2010), controlling false positives Neto et al. (2013), influence from hidden variables and design factors Remington (2009), and poor ability to capture behavior in dynamical non-linear biological systems Blair et al. (2012). Lack of a gold-standard makes it difficult to assess the true accuracy and stability of the inferred network. Model selection or averaging based on a score or probability is used to select or summarize the network over an ensemble of candidate structures. Taken together, the interpretation of relationships in the network is challenging and should be approached cautiously, especially if used to guide future research efforts and experiments.

The inferred topology of the network typically represents the endpoint of the graphical analyses. The connections themselves provide novel insights into the existence and strength of direct and indirect relationships, but this view is limiting. One can generate topology-based hypotheses, e.g. perturbing A will effect B and C, which are binary descriptions or Boolean rules. Quantifying the system-wide effects of perturbing (inhibiting or activating) different nodes in the network cannot be discerned through the examination of the topology alone. Casting the phenotype-genotype network in an *in silico* framework facilitates this type of exploration, and is the focus of this work.

We leverage directed probabilistic graphical models (PGMs) known as Bayesian Networks (BNs), which represent the joint distribution of the variables in the model (nodes) in a compact factorization of conditional likelihoods Koller and Friedman (2009). Observing nodes or setting nodes to specified values results in probabilistic influence on the marginal distributions of other nodes in the network. The process of setting nodes to specified values is known as *absorbing evidence* into the network, and it can be viewed as a system perturbation Koller and Friedman (2009). For example, a phenotype (e.g. a gene in the network), can be inhibited by setting it to a low level of evidence in the model. Consequently, the marginal probability distributions for other nodes will change in light of this new information. Quantifying the probabilistic system-wide changes before and after evidence is entered into the network can be viewed as predictions from an *in silico* experiment.

We propose a novel paradigm for predicting and visualizing the system-wide effects of a genotype-phenotype network under perturbation. We restrict our attention to a class of mixed PGMs, known as Conditional Gaussian Bayesian Networks (CG-BNs), which jointly model quantitative (genotype) and qualitative (phenotype) variables Lauritzen (1992, 1996). The perturbations considered take the form of setting phenotype node(s) (evidence) in the network to specified values (e.g. inhibiting and activating) and quantifying the effects on all other nodes in the system. Once evidence is entered, it is propagated through the system using *belief propagation* methods, which can be viewed as a form of message passing between nodes in the network Pearl (1988). We show that the modeling assumptions adopted for genotype-phenotype inference represent a sub-class of CG-BNs, which enables exact inference in the propagation scheme. A symmetric Kullback-Leibler

divergence measure is used to quantify the change in marginal distributions after evidence is entered and propagated through the network Jeffreys (1946).

The process of perturbing and propagating, enables the treatment of a phenotype-genotype network as a computational model, which can be used to perform *in silico* experiments. Thus far, this utility of genotype-phenotype network modeling have not been considered, but can offer novel insights and ease in an otherwise difficult interpretation of networks. We apply this method to *mus musculus* expression quantitative trait loci (eQTL) kidney data from a F2 inter-cross males between inbred MRL/MpJ and SM/J strains, and to skin cancer data from a F2 backcross between inbred SPRET/Ei X FVB/N and FVB/N. We demonstrate visualizations for the absorption of evidence in single and multiple nodes. The system-wide effects, examined at the node-level, also reveal co-regulation and coordination in sub-pathways, which are largely determined by genetic variation invoked through the upstream genotypes. We showcase the flexibility of our approach through the direct estimation of probabilities related to relevant covariates and experimental factors (e.g. cancer status) for network perturbations. These system-wide effects and coordination of nodes cannot be observed through examination of network topology or through traditional conditional independence tests.

2 Materials and methods

2.1 Network modeling

The graphical model is represented as a directed acyclic graph (DAG), which can be efficiently decomposed and translated into the joint distribution of variables in the model. The nodes in the graph represent the model variables, which may be discrete (QTL) or continuous (phenotypes). The phenotypes represent any measured quantity, e.g. metabolites, gene-expression, or clinical traits. In our modeling, we further assume the phenotypes are continuous and follow a normal distribution. The distributional assumptions on the phenotypes are often natural for complex traits that are continuous, e.g. gene expression and clinical traits. These assumptions are also often adopted in QTL mapping, which relies on regression modeling of complex traits Broman and Sen (2009). We define the data as: $D=\{X_1, \dots, X_n, Q_1, \dots, Q_m\}$, where random variables X and Q represent the phenotypes and genotypes at single nucleotide polymorphism (SNP) markers, respectively.

The relationship between variables is described by a Conditional Gaussian Bayesian Network (CG-BN). In a CG-BN, discrete variables are constrained to precede the continuous variables in the network. In genotype-phenotype networks, this assumption naturally conveys the fact that genetic variation precedes phenotypic variation. For simplicity, we further assume that there are no relationships between discrete variables (no edges between them). In other words, we assume no genetic interactions (epistasis). Local relationships between continuous child nodes and parents are described using homogeneous conditional Gaussian models (HCGM) Lauritzen (1996). The conditional distribution for a phenotype $Y=X_j$ with discrete parent Q_i with genotype states (g) and continuous parent X_i ($i \neq j$) is modeled as:

$$P(Y|Q_i=g, X_i=x_i)=N(\alpha(g)+\beta(g)^T x_i, \gamma(g)), \quad (1)$$

where the mean is a regression that depends on both discrete and continuous parents, but the variance depends only on the discrete parents (genotype states). The term *homogeneity* reflects the variance dependency assumptions Lauritzen and Jensen (2001). Specifically, each local model, consisting of a child and parent nodes is described by a set of parameters. If the continuous child node has only discrete parents, then the parameters are entries in a Continuous Probability Table (CPT). If the continuous child node has only continuous parents, then the parameters are regression coefficients. If the continuous child has a mixture of discrete and continuous parents, then the parameters consist of genotype-specific regressions described in Equation 1. Parameters for the regressions can be learned using maximum likelihood.

2.2 Belief propagation

The inference problem of primary interest in this work is for the marginal distributions (known as *beliefs*) of variables in the network. Specifically, we are interested in how these beliefs will change in light of new information. These objectives are best motivated through an example. Suppose a measurement (e.g. $X_i=k$, where k is a constant) is taken from a new individual outside of the original population from which the network was fit. This new evidence can be absorbed into the network and propagated to update beliefs of other nodes in the system. In a CG-BN of a genotype-phenotype network, the updated beliefs for other phenotypes are in the form of updated parameters of a Gaussian distribution. The updated beliefs of discrete genotype nodes are a probability estimate for each of the possible genotype states ($\sum_{i=1}^k g_i=1$).

Updating beliefs after new evidence is entered in the network is achieved through belief propagation Lauritzen and Jensen (2001). The algorithm is akin to message passing, in which nodes in the network talk to each other and update their beliefs based on new messages concerning the absorbed evidence. The message passing is performed by operating on a *junction tree*, which is sometimes referred to as a cluster tree or clique graph. The junction tree is the main computational object that is leveraged in this work to initialize the beliefs, absorb evidence, and update beliefs. Our primary focus is belief propagation after absorbing evidence, and the comparison of networks to assess the effect of the new information on individual nodes and regions of the network. Therefore, we assume a starting point of a known input CG-BN, which represents the phenotype-genotype network. We leverage a stable propagation scheme for CG-BN originally proposed by Lauritzen et al. Lauritzen and Jensen (2001), and we refer the reader here for full details. Here we briefly describe the process for belief propagation in a CG-BN and the comparison of phenotype-genotype networks.

Belief propagation and comparison for genotype-phenotype networks

1. **Formation of a junction tree:** The input object is the structure of the CG-BN (Figure 1A). In the first step, a moral graph is formed by marrying (connecting) the parents that are not already linked and eliminating the directionality of the edges to obtain an undirected graph (not shown). Secondly, a triangulated graph is obtained by adding suitable links to ensure that there are no chordless cycles of length three or more (Figure 1B) Leimer (1988). In the third step a strongly decomposable graph is obtained through the addition of edges between any discrete parents if they are connected by a path that passes only through the continuous nodes. In our example, we do this by adding a link between Q_1 and Q_3 to remove the forbidden path $Q_1-X_1-Q_3$ (Figure 1C). This ensures a strong decomposition of the undirected, marked graph, which permits closed-form likelihood estimates of parameters Leimer (1988).

Finally, a junction tree is constructed by organizing the maximal cliques as nodes on the tree and the edges are labeled by separators obtained by intersection of neighboring cliques (Figure 1D). A strong root of the junction tree is identified, such that the clique furthest away from this root has only continuous vertices beyond the separator that is not purely discrete. In our example, tree is shown in Figure 1D. The clique $Q_1-Q_2-Q_3-X_1$ is one of the possible strong roots since the clique furthest away from it, X_2-X_5 has only continuous variable, X_5 , beyond the separator X_2 .

2. **Model specification:** The parameters of the CG-BN are inferred from the data under the constraints of the topology and the Markov condition Neapolitan (2004). Each conditional density of a node X_i corresponds to a Conditional Gaussian (CG) potential ϕ_{x_i} , which is defined on the parents of $\text{pa}(X_i)$ in terms of canonical characteristics.
3. **Initialization of the junction tree:** The junction tree is initialized through the assignment of each node, X_i and Q_i , to a *universe*, V . This assignment is such that the node and parent nodes are assigned to the same universe, e.g. $(X_i \cup \text{pa}(X_i)) \subseteq V$.
4. **Absorption of evidence:** New evidence can be entered by setting phenotypes in the network to a particular value, $X_i=x_i^*$, or setting a genotype state $Q_i=g_i^*$. The evidence can pertain to a single node or multiple nodes in the network. Potentials are updated in all universes that contain the absorbed node(s) through the modification of the canonical characteristics based on the new evidence.
5. **Message passing:** Information is first propagated from the leaves of the junction tree to the strong root, and it is then distributed from the strong root back out to the leaves of the tree. Through this message

Figure 1: A schematic describing belief propagation in a genotype-phenotype network. (A) A CG-BN is provided as input. The nodes Q_i and X_i represent genotypes and phenotypes, respectively. (B) A triangulated graph obtained by marrying the parents (dotted red lines) and dropping the directionality. (C) A strongly decomposable graph obtained by adding an edge between Q_1 and Q_3 . (D) The junction tree is the main computational object for belief propagation, which is anchored at a strong root (a clique that meets decomposability criteria). Belief propagation is achieved by message passing from the strong root through the leaves, and then from the leaves back to the strong root. (E) Updated network beliefs after absorbing evidence of X_2 (green node). The node color represents the signed Jeffrey's information for the comparison of initial and absorbed marginals in d -connected nodes. The d -connected SNP markers are shown in gray.

passing, the beliefs are updated to reflect the updated probabilities after taking into account new evidence or information. Updated beliefs for discrete nodes are simply updated estimated frequencies under the new evidence. For continuous phenotypes, the updated beliefs are in terms of revised parameters for the Gaussian distribution.

6. **Comparison of networks:** The original and absorbed network can be compared node-wise by quantifying the change in marginals. For discrete nodes, this is simply observing estimated changes in genotype frequency. For continuous nodes, the comparison amounts to estimating the change in estimated Gaussian distributions. In this setting, we use a symmetric version of the Kullback-Leibler information, known as Jeffrey's information, for Gaussian distributions to compare the marginal belief in the original network $X_i^0 \sim N(\mu_0, \sigma_0^2)$ to the absorbed network $X_i^{\text{abs}} \sim N(\mu_{\text{abs}}, \sigma_{\text{abs}}^2)$ Jeffreys (1946). Jeffrey's information, which is computed for all continuous unabsorbed nodes in the network, is given as:

$$J(X_i^0, X_i^{\text{abs}}) = I^{\text{KL}}(X_i^0, X_i^{\text{abs}}) + I^{\text{KL}}(X_i^{\text{abs}}, X_i^0)$$

where

$$I^{\text{KL}}(X_i^0, X_i^{\text{abs}}) = \frac{1}{2} \left\{ \frac{(\mu_0 - \mu_{\text{abs}})^2}{\sigma_0^2} + \frac{\sigma_0^2}{\sigma_{\text{abs}}^2} - \log \left(\frac{\sigma_0^2}{\sigma_{\text{abs}}^2} \right) - 1 \right\}.$$

For ease of interpretation, the signed Jeffrey's information:

$$\text{sign}(\mu_0 - \mu_{\text{abs}}) \cdot J(X_i^0, X_i^{\text{abs}})$$

is used in our graphic displays to demonstrate the direction of change after the absorption of evidence (Figure 1E).

When tracking the changes in beliefs, we leverage of the concept of *d*-separation in Bayesian networks. *D*-separation is a criterion for deciding whether any two sets of nodes in a Bayesian network are independent, given the evidence entered in the network. More formally, two sets of nodes, *X* and *Y*, are *d*-separated by a third set, *Z*, if and only if every undirected path from *X* to *Y* is blocked thus making *X* and *Y* conditionally independent given *Z*. The nodes that are not *d*-separated are called *d*-connected Lauritzen and Spiegelhalter (1988). We measure the changes in belief only for the nodes that are *d*-connected (conditionally dependent) to the entered evidence. Nodes that are *d*-separated from absorbed evidence are not influenced, and, consequently, do not change beliefs.

2.3 Data

Belief propagation is applied to two *mus musculus* data sets to demonstrate different features and capabilities of the computational model. The belief propagation process takes as input a graphical model, or an ensemble of graphical models. Therefore, the results are conditional on the network structure. Slightly different graphical modeling approaches were used for the different data sets to accommodate the differences in the overall structure and content of the datasets. The differences in the datasets also facilitate showcasing the flexibility in the type of *in silico* experiments that can be performed using belief propagation in the context of genotype-phenotype networks.

2.3.1 Mus musculus kidney eQTL

We applied our approach to a kidney eQTL dataset from males (*n*=173) in an F2 inter-cross between inbred MRL/MpJ and SM/J strains of mice Hageman et al. (2011a). Previous eQTL analyses revealed a *trans*-band on Chr 4. Furthermore, Toll-like receptor 12 (*Tlr12*), a candidate gene on chr 4, was shown to modulate urinary albumin-to-creatinine ratio (ACR), a biomarker for kidney disease. For the graphical model, we chose to center the variable selection problem around *Tlr12* in an effort to further dissect the *trans*-band.

Variable selection was performed by filtering on significance and location of QTL, followed by an *elastic net* procedure Zou and Hastie (2005). Linkage analysis was performed for 33,872 gene expression traits using R/qtl Broman and Sen (2009), as described in Hageman et al. (2011a). A single transcript was selected to represent a gene by selecting the transcript with the maximum LOD score across the genome. The candidate variables for selection were further restricted to the 2946 genes with suggestive QTL (*p*<0.63) on chr 4. These genes were used as predictors in an elastic net regression model with *Tlr12* as the response variable. Elastic net is a shrinkage regression approach, which can accommodate high-dimensional data, and correlated predictors Zou and Hastie (2005). Models were fit in the R programming language using the *elasticnet* package, available through the CRAN repository, across a spectrum of shrinkage parameters. The optimal shrinkage parameter was identified as $\lambda=1$ using a 5-fold cross validation scheme. The resulting model shrunk all but 18 regression coefficients to zero. Four of the variables corresponding to the non-zero coefficients were eliminated due to

lack of mapping to known ontologies in KEGG or GO Kanehisa and Goto (2000), Ashburner et al. (2000). The 14 genes and their SNP markers corresponding to their QTL (LOD score >3) were included as variables for the graphical model (Tables S1–S2).

We examined two approaches to structural learning: (1) The identification of the most likely network according to the PC-algorithm Spirtes et al. (2000), and (2) a Markov Chain Monte Carlo (MCMC) approach, which samples the posterior distribution for an ensemble of likely network topologies. These two types of approaches (greedy/maximum likelihood estimates and sampling based approaches) represent the different classical approaches to structural learning for BNs Koller and Friedman (2009). In our application, we implement a version the PC-algorithm Spirtes et al. (2000) using *RHugin* package for the R programming language Konis (2011). The possible network structures were constrained to ensure discrete variables (genotypes at SNP markers) precede continuous variables (phenotypes), and edges between discrete variables was prohibited. The complexity parameter ($\alpha=0.1$) was selected based on the model's ability to support the observational data in simulation. However, we varied the threshold parameter, α , in the interval 0–0.1 to assess the stability and sensitivity of the network structure and subsequent belief propagation to the threshold value. MCMC sampling of network topologies was performed using the *qtlnet* package in R Neto et al. (2010). A Markov chain was run for 20,000 iterations, a burn-in of 200 was discarded, and the sample was thinned by re-sampling every 20 iterations. The resulting posterior sample contained 800 network structures. The model-averaged network structure was used to summarize network connections that had a posterior probability of 0.5 or higher.

2.4 *Mus musculus* skin cancer eQTL

Belief propagation was also applied to a skin tumor eQTL dataset from 71 mice in a F2 backcross between inbred SPRET/Ei X FVB/N and FVB/N mice Quigley et al. (2011). This study involved the induction of skin tumors on a cohort of 71 mice and gene expression analysis was performed on the normal skin from 71 mice, 68 benign papillomas and 60 malignant carcinomas. The raw gene expression data is available at Gene Expression Omnibus repository (accession GSE21264). The 11 phenotypes for the model were chosen from a papilloma-specific eQTL network that was previously identified including hair follicle keratins and keratin-associated proteins Quigley et al. (2011). The gene expression data for the 11 genes were standardized to have a mean 0 and standard deviation of 1. The corresponding 8 SNP markers including *Chr6* in the papilloma-specific eQTL network, were detected by a whole genome scan by implementing the QTLnet algorithm Neto et al. (2010). The 11 phenotypes and 8 SNP markers were input to the graphical model. In addition to these variables, a *Class* variable representing the tumor state (normal, papilloma or carcinoma) was introduced in the graphical model.

We inferred the causal relations between genotypes and phenotypes by implementing the QTLnet Neto et al. (2010) method. A Markov chain was run for 20,000 iterations and the network structure was sampled every 20 iterations, discarding the first 200 as burn-in. This resulted in a posterior sample with 800 network structures. Causal relations between 8 SNP markers and 11 phenotypes were detected. A version of the PC-algorithm Spirtes et al. (2000) using *RHugin* package for the R programming language Konis (2011) to learn a constrained network based on 11 phenotypes and 8 SNP markers. The causal relations inferred from QTLnet were considered as known domain knowledge and constrained as required edges. The *class* variable was also constrained as a parent of all phenotypes. The complexity parameter ($\alpha=0.1$) was selected based on the model's ability to support the observational data in simulation.

3 Results

In this work, we propose a framework for comparing genotype-phenotype networks after absorbing and propagating phenotypic evidence. Supporting visualizations are also proposed for the examination of

system-wide changes in network beliefs after absorbing a single piece or a spectrum of evidence into phenotype nodes in a CG-BN. These system-wide changes are quantified by the change in the marginal distributions of the nodes after evidence is absorbed and propagated. This approach was applied to mouse kidney eQTL data obtained from F2 inter-cross males between inbred MRL/MpJ and SM/J strains and an eQTL mouse skin cancer dataset from a F2 backcross between inbred SPRET/Ei X FVB/N and FVB/N strains.

In the kidney eQTL data, variable selection using QTL significance and an elastic net algorithm yielded a tractable subset of genes and SNPs from which we inferred a CG-BNs to represent the genotype-phenotype network. MCMC- and greedy search-based structural learning was both used to infer CG-BNs, which represent the input and starting points for belief propagation. On the other hand, the skin cancer eQTL data is a more complicated design that suffers from a small sample size coupled with extreme heterogeneity within and between the healthy, papilloma, and cancer tissues. Consequently, a different approach was needed for network inference in light of weak, yet suggestive effect sizes Quigley et al. (2011). Variable selection was not required, as we were able to focus in on a previously identified correlation network in papillomas that consisted primarily of genes in the keratin family. Network inference required a two-stage approach where the parent SNPs were determined using QTLnet, which along with *Class*, was used to constrain the PC-algorithm. Therefore, both data sets require different learning paradigms, which can be readily accommodated in the belief propagation approach.

The initial beliefs (marginal) represent the system in an initial state before absorbing evidence. We examine the change in these marginal beliefs after setting the node(s) in the networks to specified levels of evidence. For ease of interpretation, we restrict our applications to absorption for one or two nodes in the network. We measure the system-wide effects by tracking and comparing the model parameters of the marginal distributions, of nodes d -connected from absorbed evidence, before and after evidence is entered and propagated. Comparisons between initial and absorbed marginals were made using a signed Jeffrey's information (symmetric Kullback-Leibler divergence) to quantify the node-wise effect of evidence absorption. The sign of the information indicates an increase or decrease in the mean of the marginal, whereas the magnitude of the effect indicates the distance between marginals.

In the kidney data application, *Tlr12* was selected for evidence absorption because of its position in the network, and previous results, which show that *Tlr12* modulates the albumin-to-creatinine ratio, a marker for chronic kidney disease Hageman et al. (2011a). Evidence for *Tlr12* over a spectrum of values ± 2 standard deviations from its marginal mean of 0.06 was absorbed and propagated through the network. The changes in beliefs were tracked for the 14 nodes d -connected from *Tlr12*. Visualizations are presented which convey the system-wide changes in the d -connected nodes for a single piece of evidence, *Tlr12* = -1 (Figure 2A), and across a spectrum of evidence (Figure 2B). Specifically, the snapshot of behavior (Figure 2A) reveals the system response for a single piece of evidence, the the white nodes represent the d -separated nodes, which are independent to the absorbed evidence due to their network position. Given the evidence of *Tlr12* = -1, the genotype at the d -connected SNP marker Q.chr4 (shown in gray) is most likely homozygous (SS) with a probability of 0.79. This piece of evidence represents a single point in Figure 2B (*Tlr12* = -1), which tracks the profile of the belief changes across a spectrum of evidence for *Tlr12*.

Figure 2A–B suggests that different groups of genes are coordinated in their response to change in *Tlr12*. Specifically, groups of genes within the network are repressed (blue nodes in Figure 2A) or activated (red nodes in Figure 2A) when evidence for *Tlr12* is set and propagated through the network. Among the 13 d -connected phenotypes, seven of them showed a decrease in their marginal mean (blue), while an increase was observed for the remaining six nodes (red). Co-regulation is observed in the snapshot (Figure 2A), whereas examining the spectrum of information (Figure 2B) reveals that this co-regulation remains intact (coordinated) across a spectrum of values. Ontology information through Mouse Genome Informatics (MGI) implicates some of the coordinated genes to share common functional roles. Specifically, $\{Rbbp4, Stx12, Ak2, Hmgc1\}$ which are coordinated in their response to *Tlr12* changes (Figure 2A–B), are involved either in AMP/ADP/ATP metabolism or protein biosynthesis/transport Bult et al. (2008).

In a second experiment, using the kidney data, evidence was absorbed for two nodes, *Ak2* and *Ptp4a2*, which varied over a grid of values within ± 2 standard deviations from their marginal means. The beliefs for

Figure 2: Visualization of change in beliefs after absorbing evidence for a single node $Tlr12$ (green node) in the *mus musculus* kidney data. (A) The system-wide effects, measured by change in node-specific beliefs, after a single piece of evidence is absorbed ($Tlr12 = -1$). The node color represents the signed Jeffrey's information for the comparison of initial and absorbed marginals in d -connected nodes. (B) The system-wide effects, measured by change in node-specific beliefs, across a spectrum of absorbed evidence at equally spaced points in the interval $[\mu_{Tlr12} \pm \sigma_{Tlr12}]$.

the 13 continuous and two discrete d -connected nodes were tracked and compared to the unabsorbed marginals. The updated network beliefs are visualized in Figure 3A after instantiating the evidence at $Ak2 = -1.76$ and $Ptp4a2 = -0.746$. The SNP marker Q.chr11 is d -separated (shown in white) given the evidence of $Ak2$ and $Ptp4a2$. Under these lines of evidence, the genotype at Q.chr4 is most likely homozygous MM with a probability of 0.56, while the genotype at Q.chr15 is most likely to be heterozygous with a probability of 0.55. Coordination and co-regulation within the network are also evident (Figure 3A), and the patterns differ for the single absorption case (Figure 2A–B). Monotonicity is observed in the surface plots of the signed Jeffrey's information for individual nodes after absorbing over a grid of evidence (Figure S1). Surface plots for select nodes (Figure 3B), *Mecr* and *Zbtb8a*, demonstrate the nonlinear monotonic response. Furthermore, the monotonicity supports the notion that the coordinated nodes in sub-pathways observed in the static snapshot, e.g. $\{Rbbp4, stx12, Hmgc1\}$ and $\{Mecr, Tlr12, Zbtb8a, Slc5a9\}$, are consistent across a grid of absorbed evidence.

The resulting surfaces can provide additional information about belief changes in individual nodes when evidence is absorbed for a pair of nodes. For instance, a surface plot of *Mecr* (Figure 3B) shows that Jeffrey's signed information has a peak value 2.17 at $Ak2 = -1.76$ and $Ptp4a2 = 1.85$, and then decreases gradually as evidence for $Ak2$ is increased and that for $Ptp4a2$ is decreased. On the contrary, the Jeffrey's signed information for *Zbtb8a* has a peak at $Ak2 = -1.76$ and $Ptp4a2 = -1.84$, decreases sharply, and plateaus for most of the domain close to 0 before a sharper decrease to -0.45 at $Ak2 = 1.74$ and $Ptp4a2 = 1.85$. Unlike *Mecr*, the surface of *Zbtb8a* is not discontinuous since the Jeffrey's signed information remains 0 until the evidence is set to and beyond ± 1 standard deviation from the marginals. Collectively, these results suggest that while *Mecr* is very

Figure 3: Visualization of change in beliefs after absorbing evidence for two nodes *Ak2* and *Ptp4a2* (green nodes) in the *mus musculus* kidney data. (A) The system-wide effects, measured by change in node-specific beliefs, after a single piece of evidence is absorbed (*Tlr12* = -1). The node color represents the signed Jeffrey's information for the comparison of initial and absorbed marginals in d -connected nodes. The most likely genotype state after absorption is indicated for d -connected SNP markers. (B) The system-wide effects, measured by change in node-specific beliefs, across a grid of absorbed evidence $[\mu_{Ak2} \pm \sigma_{Ak2}] \times [\mu_{Ptp4a2} \pm \sigma_{Ptp4a2}]$.

sensitive to changes to *Ak2* and *Ptp4a2*, the beliefs of *Zbtb8a* are not affected until the evidence of *Ak2* and *Ptp4a2* reaches more extreme values.

To assess the stability of the predicted system-wide changes after evidence absorption over different network structures in the kidney data, we (1) varied the significance level α in the PC-algorithm, and (2) used a MCMC-based approach to collect a sample of likely network structures. Varying the PC-algorithm parameter $\alpha \in [0, 1]$ led to 38 different network structures in the ranging from dense to sparse in terms of connectivity. In order to assess the sensitivity of the results to structural learning, a spectrum of evidence for *Tlr12* and tracked the sign of Jeffrey's information for all d -connected nodes to *Tlr12* in each of the 38 networks. We found that the signs (representing the direction of change in the marginal means) were consistent for all 13 d -connected nodes to *Tlr12* across all network structures for the spectrum of evidence absorbed. (Supplemental Table 3). Likewise, in the MCMC-based approach we obtained 24 network structures summarizations by varying the posterior probability cut-off from 1 to 0.5 in the model averaging. Evidence was then absorbed for *Tlr12* and the number of networks in which the Jeffrey's signed information for the d -connected nodes had positive or negative signs were recorded. All d -connected nodes had a consistent direction of change in their marginal means across all the network structures for each evidence absorbed (Supplemental Table 4).

Absorbing evidence in *Tlr12* below the marginal mean causes a decrease in the marginal mean of *Mecr* indicating inhibition of fatty acid synthesis and in turn activation of *Wdtd1* which plays a role in negative regulation of fatty acid biosynthesis Suh et al. (2007) and vice-versa. Decreasing the evidence in *Tlr12* also seems to inhibit sodium dependent glucose transport as evidenced by a decrease in Jeffrey's signed information of *Slc5a9* and activate sodium and chloride dependent glycine transport indicated by the Jeffrey's signed information of *Slc6a9*.

The cancer network was examined across various levels of evidence of *Krt71*. This gene and others in the network were selected based previous implications of involvement in an association network Quigley et al. (2011). The class-specific data provided the unique opportunity to model *Class* as a variable in the network. The constraint that it is upstream and connected to every phenotype reflects its use as a covariate in the local models. In this setting, the system-wide effects inferred through *in silico* experiments include the prediction of class (healthy papilloma, and tumor) probability. Estimation of class probability over a spectrum of evidence for *Krt71* (Figure 4B) supports previous findings of monotonic increases in expression over the tumor, papilloma, and healthy classes, respectively. Moreover, the changes coordinate strongly with *Krt25* (Figure 4B–C), which was previously shown to exhibit a large and coordinated effect. Unlike the kidney network, the activity of the keratin gene network appears to be largely coordinated and sole activated in its response to various levels of evidence in *Krt71*. The effect on the path connecting *Krt27*, *Krt25*, *Krt33a*, and *Krt31* is more pronounced (Figure 4C).

Figure 4: Visualization of change in beliefs after absorbing evidence for *Krt71* (green nodes) in a skin cancer network. (A) The prediction of class (healthy, papilloma, or tumor) across a grid of absorbed evidence $[\mu_{Krt71} \pm \sigma_{Krt71}] \times [\mu_{Krt71} \pm \sigma_{Krt71}]$. (B) Predicted system-wide effects, measured by change in node-specific beliefs, across the same grid of absorbed evidence. (C) The system-wide effects, measured by change in node-specific beliefs, after a single piece of evidence is absorbed (*Krt71*=0.98). The node color represents the signed Jeffery's information for the comparison of initial and absorbed marginals in *d*-connected nodes.

4 Discussion

Leveraging natural variation to infer genotype-phenotype networks has been an extremely active area of research Rockman (2008). Considerable attention and research effort has been primarily directed toward the structural learning problem of detecting direct and indirect relationships between phenotypes, using genotypes at QTL as a causal anchor Schadt et al. (2005), Kulp and Jagalur (2006), Aten et al. (2008), Millstein et al. (2009), Neto et al. (2013), Schadt et al. (2005), Li et al. (2006), Zhu et al. (2007, 2008), Benfey and Mitchell-Olds (2008), Liu et al. (2008), Neto et al. (2008, 2010), Hageman et al. (2011b). Structural learning has been used to generate hypotheses and design future studies based on edge connections, but the inferred topology typically represents a premature endpoint in the analyses. PGMs, by design, lend themselves to probabilistic reasoning and querying, which to our knowledge, has not been leveraged in genotype-phenotype networks, or gene-networks.

The focus of this work is the unexamined problem of utilizing a learned network as a predictive computational model. To achieve this, we transform the phenotype-genotype network into an *in silico* PGM, which can operate as a *black box model*. The inputs to the black box are prescribed values (evidence) for single or multiple phenotypes in the network. Within the black box is the network itself, which is assumed to be fixed. Belief propagation is the computational engine, which is used to update and calibrate the network based on evidence inputs. The output is the predicted beliefs (marginals) for given inputs. The comparison of the initial and absorbed states of the network provides novel information about the system-wide effects of new evidence on the network, which cannot be gained through examination of the topology alone.

We have demonstrated belief propagation in a network inferred from a *mus musculus* population using eQTL data. Beliefs before and after absorption of evidence for one and two nodes were compared and quantified using a signed and symmetric version of Kullback-Leibler divergence, known as Jeffrey's information Jeffreys (1946). We limited our experiments to a single or double perturbation due to the lack of visualization for model interpretation in higher dimension (greater than two perturbations) over spectrums of evidence. However, computationally it is feasible to consider more two perturbations in a network, as it requires the same forward and backwards pass from the strong root of the junction tree. To this end, visualization is possible for multiple perturbations (≥ 3) that are each considered for single pieces of evidence (e.g. three knock-downs, or more).

Utilizing a signed information criteria reveals potential coordination and co-regulation in regions of the network, suggesting the existence of branches and sub-pathways. Moreover, these patterns were consistent across a spectrum of values, suggesting sub-pathways within the network (Figure 2B). Increasing the overall number of SNPs in the model improves the ability to detect sub-pathways in the model. Equivalently, targeting nodes for absorption that have more d -connected SNPs can reveal more sub-pathways. This suggests that genetic variability gained through d -connected SNPs drives the observed coordination of sub-pathways.

The coordination and co-regulation was also apparent in the absorption of two nodes, *Ptp4a2* and *Ak2*. In the case of two nodes, the examination of sub-pathways requires examining trends in nonlinear surfaces, which are more difficult to quantify. For example, the static snapshot (Figure 3A) suggests coordination in three sub-pathways. On the other hand, the surface plots (Figure 3B) reveal monotonic increases and decreases in Jeffrey's information, but the potential coordination of *Wdtd1*, *Cyp4a31*, and *Slc6a9* is not obvious to the naked eye. Taken together, we found the visualizations complementary and essential to our understanding of the predicted changes. While the absorption of evidence in more than two nodes is possible in terms of computation, the visualization is not, and consequently the model interpretability will suffer.

Modeling a *class* variable in the cancer network, enables summarizations that extend beyond the standard genotype and phenotype variables (Figure 4A). Understanding the predicted probabilities of covariate states under different levels of evidence provides additional information that can be used to better understand the system. The use of covariates can be expanded to include experimental factors such as tissue type, disease state, sex, and other. Building in covariates is a simple process that can be achieved, for example, by hard-wiring (forcing) edges in a network, or through constraining inferred edges between covariate and phenotype to be ordered. Within the R programming language, there is tremendous flexibility in the implementation of

constraints for structural learning of a Bayesian Network Højsgaard et al. (2012). Since the network structure is *input* to the belief propagation routine, the structural learning process is rather independent and can be determined by the user. Our software, `geneNetBP`, can facilitate the structural learning, with or without constraints, by leveraging existing R packages, but also accepts an adjacency matrix as expert input.

We leveraged a CG-BN framework for genotype-phenotype inference for several reasons. BNs that are uniform with respect to the variable type, i.e. all continuous or all discrete nodes, are more popular. Exact algorithms for parameter inference are most developed in discrete BNs Lauritzen (1996). On the other hand, CG-BNs contain discrete and continuous variables, and the inference and belief propagation is more complex, prone to instability, and far less studied Lauritzen (1992), Lauritzen and Jensen (2001). We restrict the class of phenotypes to continuous complex traits, although we demonstrate that discrete covariates (e.g. class) can be accommodated in the CG-BN paradigm. Furthermore, we utilized the PC-algorithm and MCMC to infer a highly likely structure(s). However, structures learned through other methods, or using expert knowledge, can be used as input and parameterized as a CG-BN.

Network size is a major consideration. It has been shown that graphical reasoning requires large sample sizes in linkage studies Li et al. (2010). We recommend restricting the size of the network to a moderate number of variables (e.g. <30), and promote caution in the placement and balance of node type. In particular, inferences in the CG-BN paradigm introduce additional sensitivity to the balance of discrete and continuous variables, and their placement in the DAG. For example, in an F2 intercross, there are three measured genotype states, AA, Aa, aa, for each discrete node. If a continuous node has two SNP parents, then there are eight combinations of genotypes between them (AAAA, AAaa, Aaaa, aaaa, aaAA, aaAa, AaAa, Aaaa), which would need to be conditioned on. The local probability model is a Conditional Gaussian with both mean and variance depending on the stratification of genotype states of the parent nodes (Equation 1). Therefore, we believe it is also prudent to limit the number of discrete parents (SNPs) to support reliable inference.

A development version of the software package, `geneNetBP`, is available through R-forge (<http://genenetbp.r-forge.r-project.org>) to automate and visualize this process, for a given data set. `GeneNetBP` is compatible with R/qtl output, but depends on `RHugin`, which is powered by a free version of the commercial Hugin software (www.hugin.com) known as Hugin lite Broman and Sen (2009), Konis (2011). Hugin lite imposes limitations on the network size in terms of states and the number of samples, but provides a stable propagation scheme Lauritzen and Jensen (2001). Larger networks can be investigated with Hugin expert commercial software. In terms of analysis, `geneNetBP` facilitates that smooth transition from QTL mapping to graphical modeling to system-wide prediction after absorbing network evidence. Specifically, `geneNetBP` can be used to absorb and propagate phenotypic evidence through a given CG-BN representation of a genotype-phenotype network, compute the Jeffrey's Information across the network, and provide visualizations for interpretation.

In summary, belief propagation is a promising avenue for *post hoc* analyses of an inferred phenotype-genotype network. Graphical modeling approaches to the interpretation of the data offer valuable perspective, which hold promise for personalized medicine. We have proposed a paradigm and supporting software which enables computation and visualization of *in silico* predictions of the system-wide response of perturbation. This work represents a first step toward alleviating longstanding issues associated with model interpretation of genotype-phenotype networks. Insights provide a new layer of information, which may guide future research efforts and experiments.

Acknowledgments: RHB, JM, and PM were supported through NSF DMS 1312250.

References

- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matrese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000): "Gene ontology: tool for the unification of biology," *Nat. Genet.*, 25, 25–29.

- Aten, J., T. Fuller, A. Lusi and S. Horvath (2008): "Using genetic markers to orient the edges in quantitative trait networks: the NEO software," *BMC Syst. Biol.*, 2, 34.
- Benfey, P. N. and T. Mitchell-Olds (2008): "From genotype to phenotype: systems biology meets natural variation," *Science*, 320, 495–497.
- Blair, R. H., D. J. Kliebenstein and G. A. Churchill (2012): "What can causal networks tell us about metabolic pathways?" *PLoS Comput. Biol.*, 8, e1002458.
- Broman, K. W. and S. Sen (2009): *A guide to QTL mapping with R/qlt*, Springer, New York.
- Bult, C. J., J. T. Eppig, J. A. Kadin, J. E. Richardson, J. A. Blake and Mouse Genome Database Group (2008): "The mouse genome database (mgd): mouse biology and model systems," *Nucleic Acids Res.*, 36, D724–D728.
- Chickering, D. M., D. Geiger and D. Heckerman (1994): "Learning bayesian networks is np-hard," Technical Report, Citeseer.
- Hageman, R. S., M. S. Leduc, C. R. Caputo, S.-W. Tsaih, G. A. Churchill and R. Korstanje (2011a): "Uncovering genes and regulatory pathways related to urinary albumin excretion," *J. Am. Soc. Nephrol.*, 22, 73–81.
- Hageman, R. S., M. S. Leduc, R. Korstanje, B. Paigen and G. A. Churchill (2011b): "A bayesian framework for inference of the genotype–phenotype map for segregating populations," *Genetics*, 187, 1163–1170.
- Højsgaard, S., D. Edwards and S. Lauritzen (2012): *Graphical models with R*, Springer Science & Business Media.
- Jansen, R. C. and J. Nap (2001): "Genetical genomics: the added value from segregation," *Trends Genet.*, 17, 388–391.
- Jeffreys, H. (1946): "An invariant form for the prior probability in estimation problems," *Proc. Roy. Soc. Lond. A. Mat.*, 186, 453–461.
- Kanehisa, M. and S. Goto (2000): "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, 28, 27–30.
- Koller, D. and N. Friedman (2009): *Probabilistic graphical models: principles and techniques*, The MIT Press, Cambridge, MA.
- Konis, K. (2011): "Rhugin," R Package Version, 7–5.
- Kulp, D. C. and M. Jagalur (2006): "Causal inference of regulator-target pairs by gene mapping of expression phenotypes," *BMC Genomics.*, 7, 125.
- Lauritzen, S. L. (1992): "Propagation of probabilities, means, and variances in mixed graphical association models," *J. Am. Statist. Assoc.*, 87, 1098–1108.
- Lauritzen, S. L. (1996): *Graphical models*, Oxford University Press, Oxford.
- Lauritzen, S. L. and F. Jensen (2001): "Stable local computation with conditional gaussian distributions," *Stat. Comput.*, 11, 191–203.
- Lauritzen, S. L. and D. J. Spiegelhalter (1988): "Local computations with probabilities on graphical structures and their application to expert systems," *J. Roy. Stat. Soc. B Met.*, 50, 157–224.
- Leimer, H.-G. (1988): "Triangulated graphs with marked vertices," *Ann. Discrete Math.*, 41, 311–324.
- Li, R., S.-W. Tsaih, K. Shockley, I. M. Stylianou, J. Wergedal, B. Paigen and G. A. Churchill (2006): "Structural model analysis of multiple quantitative traits," *PLoS Genet.*, 2, e114.
- Li, Y., B. M. Tesson, G. A. Churchill and R. C. Jansen (2010): "Critical reasoning on causal inference in genome-wide linkage and association studies," *Trends Genet.*, 12, 438–498.
- Liu, B., A. de la Fuente and I. Hoeschele (2008): "Gene network inference via structural equation modeling in genetical genomics experiments," *Genetics*, 178, 1763–1776.
- Millstein, J., B. Zhang, J. Zhu and E. E. Schadt (2009): "Disentangling molecular relationships with a causal inference test," *BMC Genet.*, 10, 23.
- Neapolitan, R. E. (2004): *Learning bayesian networks*, Pearson Prentice Hall, Upper Saddle River, NJ, USA.
- Neto, E. C., C. T. Ferrara, A. D. Attie and B. S. Yandell (2008): "Inferring causal phenotype networks from segregating populations," *Genetics*, 179, 1089–1100.
- Neto, E. C., M. P. Keller, A. D. Attie and B. S. Yandell (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.
- Neto, E. C., A. T. Broman, M. P. Keller, A. D. Attie, B. Zhang, J. Zhu and B. S. Yandell (2013): "Modeling causality for pairs of phenotypes in system genetics," *Genetics*, 193, 1003–1013.
- Pearl, J. (1988): *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann.
- Quigley, D., M. To, J. Kim, K. Lin, D. Albertson, D. Sjolund, J. Perez-Losada, and A. Balmain (2011): "Network analysis of skin tumor progression identifies a rewired genetic architecture affecting inflammation and tumor susceptibility," *Genome Biol.*, 12, R15.
- Remington, D. L. (2009): "Effects of genetic and environmental factors on trait network predictions from quantitative trait locus data," *Genetics*, 181, 1087–1099.
- Rockman, M. V. (2008): "Reverse engineering the genotype-phenotype map with natural genetic variation," *Nature*, 456, 738–744.
- Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs and A. J. Lusis (2005): "An integrative ge-nomics approach to infer causal associations between gene expression and disease," *Nat. Genet.*, 37, 710–717.
- Spirtes, P., C. N. Glymour and R. Scheines (2000): *Causation, prediction, and search*, volume 81, MIT press, Cambridge, MA.

- Suh, J., D. Zee, R. McKay, J. Seo, Z. Salo, R. Li, M. Wang and J. Graff (2007): “Adipose is a conserved dosage-sensitive antiobesity gene,” *Cell Metab.*, 6, 195–207.
- Zhu, J., M. C. Wiener, C. Zhang, A. Fridman, E. Minch, P. Y. Lum, J. R. Sachs and E. E. Schadt (2007): “Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations,” *PLoS Comput. Biol.*, 3, e69.
- Zhu, J., B. Zhang, E. N. Smith, B. Drees, R. B. Brem, L. Kruglyak, R. E. Bumgarner and E. E. Schadt (2008): “Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks,” *Nat. Genet.*, 40, 854–861.
- Zou, H. and T. Hastie (2005): “Regularization and variable selection via the elastic net,” *J. Roy. Stat. Soc. B*, 67, 301–320.

Supplemental Material: The online version of this article (DOI: 10.1515/sagmb-2015-0058) offers supplementary material, available to authorized users.