CAUSAL INFERENCE IN BIOLOGY NETWORKS WITH INTEGRATED BELIEF PROPAGATION

RUI CHANG* and JONATHAN R KARR and ERIC E SCHADT*

Department of Genetics and Genomic Sciences
Ichan School of Medicine, Mount Sinai
NY, NY 10029, USA
**E-mail: rui.r.chang@mssm.edu*
eric.schadt@mssm.edu

Inferring causal relationships among molecular and higher order phenotypes is a critical step in elucidating the complexity of living systems. Here we propose a novel method for inferring causality that is no longer constrained by the conditional dependency arguments that limit the ability of statistical causal inference methods to resolve causal relationships within sets of graphical models that are Markov equivalent. Our method utilizes Bayesian belief propagation to infer the responses of perturbation events on molecular traits given a hypothesized graph structure. A distance measure between the inferred response distribution and the observed data is defined to assess the 'fitness' of the hypothesized causal relationships. To test our algorithm, we infer causal relationships within equivalence classes of gene networks in which the form of the functional interactions that are possible are assumed to be nonlinear, given synthetic microarray and RNA sequencing data. We also apply our method to infer causality in real metabolic network with v-structure and feedback loop. We show that our method can recapitulate the causal structure and recover the feedback loop only from steady-state data which conventional method cannot.

Keywords: Causal inference, Top-down & bottom-up modeling, Predictive network modeling, Causal network learning

1. Introduction

One of the primary objectives of biomedical research is to elucidate the networks of molecular interactions underlying complex human phenotypes such as cancer and Alzheimer's disease. Over the past few years, a wave of advanced biotechnologies has swept over the life sciences landscape to enable more holistic profiling of biological systems. Whole genome sequencing, RNA sequencing, methylation profiling, and mass spectroscopy and NMR based metabolite and protein profiling technologies have been applied to a wide range of biological problems and have contributed to discoveries relating to the complex network of biochemical processes as well as to the reconstruction of gene networks underlying living systems[1] and common human diseases.[2,3]

State-of-the-art statistical learning methods assume a Markov condition for gene network reconstructions as a way to reduce the complexity of the joint probability distribution that graphical network structures represent. It is well-known that algorithms based on Markov conditions can learn the correct causal relationships up to Markov equivalence given a large enough sample size.[4] However, because the structures represented within a given equivalence class are statistically indistinguishable from one another, it is not possible to further resolve the correct causal relationships within a class without introducing perturbations that break the symmetry giving rise to this equivalence. Given the recovery of accurate mechanistic networks

is a crucial first-step to understanding the pathophysiology of human disease and how best to diagnose and treat it, methods to accurately infer causal relationships are critical.

To at least partially address this problem, we had previously proposed a Bayesian network learning framework[1] to improve causal inference within Markov equivalence classes by integrating genotypic data associated with molecular phenotypes (e.g., expression quantitative trait loci, or eQTL) and disease traits as an asymmetric and systematic source of perturbations. This approach has been effective in helping untangle the causality in gene networks given it leverages the propagation of structural asymmetry required to break Markov equivalence.

In this paper, we propose a statistical method that infers causal structures from multivariate datasets (e.g. gene expression data), reducing the need to integrate additional data such as eQTLs to accurately infer causal relationships. Our proposed method is complementary to more integrative approaches, given it will enable accurate causal inferences in cases in which integrative approaches are not possible or not well powered. For example, even when eQTL data are available, inferring a complete causal network structure remains challenging given the conventional top-down Bayesian network learning approach decomposes the joint probability function representing a given graph into the product of local conditional probabilities based on d-separation, so that the effect of causal information stemming from eQTL-controlled root nodes will not always effectively propagate through the entire network, leaving the issue of equivalence classes unresolved at distal local structures. This issue is exacerbated as the number of nodes in the network increases, given the super-exponential rate of growth of the space of possible networks and the fact that trans-acting eQTL effects (eQTL that act on genes that are distal to the physical location of the eQTL) are difficult to detect. Therefore, the development of methods to infer causality among structures in equivalence classes remains a fundamental objective for reconstructing accurate probabilistic causal network structures. To demonstrate the utility of our causal inference procedure, we apply it to simulated gene expression data that reflects the type of noise structures common in these high-throughput biological experiments as well as the biological relationships we seek to represent. In this context we demonstrate the ability to resolve Markov equivalent structures across a variety of general assumptions regarding the nature of gene-gene interactions.

2. Developing a Method to Infer Causality from Associative Data

A new class of methods referred to as Information Geometric Causal Inference (IGCI) methods[5] defines classic measures of independence among variables in terms of orthogonal components of the joint probability distribution of these variables, as a way to leverage more information regarding the relationships among them. This approach stands in contrast to traditional approaches in which data informing on the relationships among variables of interest are extracted using only conditional independencies. In the IGCI methods, to infer whether X causes Y, orthogonality is computed between the conditional distribution $P_{Y|X}$ and P_X, which are then compared to the values computed for $P_{X|Y}$ and P_Y. If the relationship "X causes Y" is true, then the orthogonality metric is such that the causal hypothesis Y causes X is implausible. Remarkably, this asymmetry between cause and effect becomes particularly simple if X and Y are deterministically related. In the case of a nonlinear relationship between

X and Y, the nonlinearity in the function defining the relationship between the cause X and effect Y, i.e. Y=f(X), can also be considered for causal inference in the presence of additive noise.[6] The nonlinearity provides information on the underlying causal model and thus allows more aspects of the true causal mechanism to be identified. An alternative approach, referred to as functional causal modeling (a.k.a. structural causal or nonlinear structural equation modeling), involves a joint distribution function that along with a graph satisfies the causal Markov assumption.[7] This functional form can also allow one to distinguish between X→Y and X←Y.

Our proposed method sits squarely within the class of functional causal modeling approaches,[7] where we utilize the inherent probabilistic inference capability of the Bayesian network framework to generate predictions of hypothesized child (response) nodes using the observed data of the hypothesized parent (causal) nodes. By defining a distance metric in probability space that assesses how well the predicted distribution of child nodes matches the distribution of their observed values, we can evaluate the different graphical structures within an equivalence class to determine the one best supported by the data. One key advantage of this modeling approach is that it enables the propagation of the effects of a parent node to child nodes that can be greater than a path length of one from the parent, thereby making it possible to infer causality in a chain of nodes or in a more complex network structures.

To describe our approach, we begin by reformatting the general posterior probability corresponding to any given graphical structure as defined in conventional Bayesian network approaches.[8] Here we let X represent the vector of variables represented as nodes in the network; E is the evidence; D denotes the observed data; G is the graphical network structure to infer; and θ is a vector of model parameters. From this we can write the posterior probability as $P(G|D) = P(D|G)P(G)/P(D)$, where the marginal probability $P(D|G)$ can be expressed as an integral over the parameters, given a particular graphical structure G: $P(D|G) = \int_\theta P(D|G,\theta)P(\theta|G)d\theta$. Unlike the traditional Bayesian Dirichlet score, D in our case contains continuous values, and thus, the likelihood of the data is not derived from a multinomial distribution, but rather a continuous density function whose form is estimated using a kernel density estimation procedure. In addition, the parameter prior $P(\theta|G)$ does not follow a Dirichlet distribution but rather is either described by a set of non-parametric constraints in parameter space or is sampled from a uniform distribution defined in its range (the approach used herein). Given this, the above integral has no analytical solution. We will optimize the data likelihood by estimating θ using maximum-a-posteriori (MAP) estimation:

$$P(D|G) \cong P(D|G,\hat{\theta}) \tag{1}$$

where $\hat{\theta}$=argmax$_\theta\{P(D|G,\theta)P(\theta|G)\}$. We use a Monte Carlo sampling procedure to efficiently sample θ from $P(\theta|G)$, evaluating the likelihood for each parameter sample.

2.1. *Deriving a Data Likelihood Score*

To calculate and optimize the data likelihood $P(D|G,\theta)$, we incorporate belief propagation as a subroutine in the causal inference procedure to predict the marginal probabilities of all response variables given the observed data for the predictor variables for a given causal structure (G). In this instance, the marginal probability of X given G and the sampled parameter θ is calculated via belief propagation.[9] In what follows we discuss how to use the Bayesian belief inference $P(X|E,G,\theta)$ to calculate the data likelihood $P(D|G,\theta)$ given in Eq. 1, where

E and D represent the observed data on the parent and child nodes, respectively. First, to avoid confusion, we introduce the notion X_b to describe the binary variable in the probability space mapped from the continuous variable $X \in R$. Second, we rescale the original observation data so that it falls in the interval [0,1] (see discussion below). Third, we introduce a hidden variable H to fully specify the data likelihood as

$$P(D|G,\theta) = \int_H P(D|H)P(H|G,\theta) \tag{2}$$

Given G and θ, the soft evidence enters $P(X_b|E,G,\theta)$ as the observed, rescaled data D, which effectively "clamps" (or fixes) the marginal probability of the parent nodes, from which the marginal probabilities of the child nodes are predicted via belief propagation in the Bayesian network.[9] These marginal probabilities are then used to define the hidden data H, which are used to construct the marginal data likelihood in Eq. 2. In probability space, the belief inference is deterministic, i.e. given a causal structure G, a specific set of parameters θ, and evidence E, $P(X_b|E,G,\theta)$ is uniquely determined. In Eq. 2, when H=$P(X_b|E,G,\theta)$, $P(H|G,\theta)$=1 and 0 otherwise, and as a result the data marginal likelihood in Eq. 2 can be re-written as

$$P(D|G,\theta) = P(D|H(G,\theta)) = P(D|P(X_b|E,G,\theta)) \tag{3}$$

The inner probability describes the marginal belief of the binary variable X_b in probability space to which the original continuous variable X has been mapped. This belief probability is a linear function between the child and parent marginal probabilities, multiplied by the conditional probabilities determined by sampling over the uniform distribution the parameters θ are assumed to follow:

$$P(X_b^{chd} = k) = \sum_i [P(X_b^{chd} = k|X_b^{\pi} = C_i)P(X_b^{\pi} = C_i)] \tag{4}$$

with X_b^{π} representing X_b for the parent node, X_b^{chd} for the child node, and where C_i represents the i-th configuration of the parent nodes. For example, given the hypothesis "gene A activates gene B", the marginal belief is calculated as

$$P(B) = P(B|A)P(A) + P(B|\bar{A})(1 - P(A)) = [P(B|A) - P(B|\bar{A})]P(A) + P(B|\bar{A}) \tag{5}$$

In this instance, the conditional probability distribution θ=$\{P(B|A), P(B|\bar{A})\}$ is the parameter sampled from the uniform distribution on [0,1]. We note that this belief propagation can be considered as a linear regression of the form P(B)=βP(A)+C, where β=$(P(B|A) - P(B|\bar{A}))$ and C=$P(B|\bar{A})$. Given probability measures are constrained to be between 0 and 1, $\beta \in [-1, 1]$ and C$\in$[0,1]. These constraints, which follow naturally from probability theory, give rise to the asymmetric basis of our approach for causal inference (see section 2.4). We further note that in our approach we implicitly assume binary variables in probability space, where the belief probability of a binary variable is defined as the level of belief on that variable observed in its maximal state (i.e., $P(X_b = 1)$) or minimal state (i.e., $P(X_b = 0)$). When X is equal to its minimum (or maximum) value in the real valued space D, in probability space X_b is observed in its minimum (or maximum) state, and therefore, this observed sample will correspond to $Pr(X_b = 0) = 1$ (or $Pr(X_b = 1) = 0$) in probability space. As the value of X varies between its minimum and maximum values, the belief probability of the binary mapping of this variable will vary between [0,1]. The Bayesian interpretation of the belief probability allows us to compare the inferred belief probability to the real-valued observed data by implicitly assuming

that the original data D and the marginal probabilities of H are positively correlated, though the precise kinetics of this correlation is unknown. To make such a comparison, we rescale $D \in R$ to $D \in$[0,1].

Since the exact function mapping between real values and their belief probabilities is unknown, we employ a non-parametric metric, i.e. Kullback-Liebler (KL) divergence, to compare the distribution of real observations to the distribution of predicted marginal probabilities. If the predicted and observed distribution of the child nodes match well, we can conclude that the predictions based on G and θ well reflect the observed data D, which results in a smaller value of the KL-divergence. To force the KL divergence to behave as a true probability measure, we make symmetry and normalization modifications to this function defined on D and H such that $k(D, H)$=1-exp[-(KL(D||H)+KL(H||D))/2]. The data likelihood function in Eq. 3 can be defined by any normalized monotonic decreasing function on the kernel. For model selection, we set $S = -\log(K(D, H))$ to represent the posterior score of the model, which is negatively correlated with the kernel value. To optimize the model, we maximize this score, which is equivalent to minimizing the KL-divergence.

2.2. *Piecewise Regression Fitting to D by Integrated Probability Inference*

The calculation of the KL-divergence involves comparing the real-valued observed data D to the inferred belief probabilities H given a particular causal hypothesis G and θ. The original interaction between parent(s) X_π and child X_{chd} nodes in D can be described by an arbitrary function X_{chd}=$\mu(X_\pi)$ plus some observation noise. Depending on the nature of the causal relationships to be modeled, $\mu()$ can take various forms, including linear, non-linear, monotonic, non-monotonic, concave, convex, step and periodic functions. In the biology domain, a direct causal interaction between two proteins or between a protein and DNA molecule often take the form of a hill function, a step function, or a more general non-monotonic, nonlinear function.

One way to derive the belief inference given in Eq. 4 is to represent the relationship between the parent and child nodes as a cubic spline, which can well approximate general nonlinear relationships. However, a more straightforward alternative to splines would be regressing the marginal belief of X_{chd} onto X_π, assuming a linear relationship. If we subdivide the range of the parent nodes into L segments based upon the behavior we expect in a causal relationship between two nodes, linear regressions can be carried out in each segment.[10] In the case $|\pi|$=1, our problem is to regress onto a single variable whose range has been divided into L segments, whereas when $|\pi| >1$ we are regressing onto multiple variables divided into a $|\pi|$-dimensional grid comprised of $L^{|\pi|}$ components. Here we focus only on inferring causality between equivalent class structures in which $|\pi|$=1, but our approach easily extends to the more general case. To derive a procedure for fitting the belief inference equation to the data in a piecewise fashion, we first define some terms. Let $\mathbf{X}$ denote a vector of predictor(s)/parent node(s) (in this pair-wise causality setting $\mathbf{X}$ represents just a single variable for any given Markov equivalent structure being considered) and let Y represent the response variables. Let $D \in R^n$ represent the original noisy observed data and $D_0^l \in [0, 1]^n$ denote the rescaled observed data in the l-th segment/grid element, which is comprised of the observed values over the parent and child nodes in the given segment/grid element, i.e. D_0^l=$\{D_{\mathbf{X}}^l, D_Y^l\}$. Similarly, let the predicted data in the l-th segment be given by H^l=$\{H_{\mathbf{X}}^l, H_Y^l\}$. Given this, we pre-define a total of K bins that are evenly distributed in [0,1], I^k=$[k/K, (k + 1)/K], k = [0, K - 1]$, and then for each bin we count the number of occurrences of the inferred marginal probability of Y falling in each

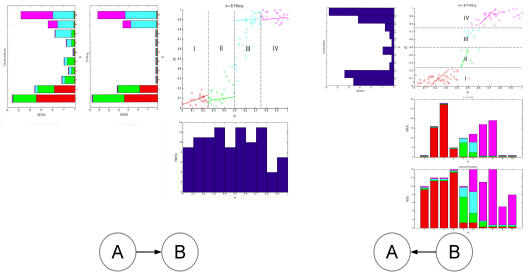


Fig. 1. Pairwise Causal Inference

of the l segments, i.e. H_Y^l falls in the k-th bin $I^k \in$[0,1]. We denote the number of occurrences for the k-th bin and the l-th segment/grid element by M_k^l. The counterpart of this number in D_Y^l, with respect to the observed data, is denoted by N_k^l. The frequency for the predicted data is then calculated as p_k^l=$M_k^l/\sum_k M_k^l$ for H_Y^l and similarly for the observed data, D_Y^l, q_k^l=$N_k^l/\sum_k N_k^l$. These counts and frequencies are used to compute the KL-divergence kernel, maximize the likelihood score, and identify the maximum LR model $\hat{\theta}$ in Eq. 1 per segment as described below. To maximize the data likelihood function $P(D|G,\hat{\theta})$ in Eq. 1 in each segment, we identify the parameter $\hat{\theta}$ that minimizes the KL-divergence for the current causal hypothesis G, which is defined as the symmetrized KL-divergence between the predicted belief and the rescaled observed data for every segment:

$$\hat{\theta}^l = \underset{\theta}{\text{argmax}}\{P(D_0^l|G,\theta)P(\theta|G)\} \propto \underset{\theta}{\text{argmax}}\{P(D_Y^l|P(Y|E=D_X^l,G,\theta))\} \propto \underset{\theta}{\text{argmax}}\{P(k(D_Y^l,H_Y^l(G,\theta)))\}$$

$$\propto \underset{\theta}{\text{argmax}}\{-\log(k(D_Y^l,H_Y^l(G,\theta)))\} \propto \underset{\theta}{\text{argmin}}\{\sum_{k=0}^{K} p_k^l \ln(p_k^l/q_k^l) + \sum_{k=0}^{K} q_k^l \ln(q_k^l/p_k^l)\} \quad (6)$$

where $P(\theta|G)$=1/M for θ sampled uniformly in [0,1]. The statistical counts of the predicted probability, p_k^l for l-th segment in k-th bin, is a function of G and θ. The optimal statistical count of the fitted model for the l-th segment and k-th bin is $\hat{M}_k^l$. The overall fitted linear regression model $(G,\hat{\theta}^l|l=1,...,L)$ with the counts across all segments in k-th bin of [0,1] is obtained by summing $\hat{M}_k^l$ over the total L segments, i.e. $\hat{\mathbf{M}}_\mathbf{k}$=$\sum_{l=1}^{L}\hat{M}_k^l$. According to Eq. 1 and Eq. 3, the final optimized estimation of the data likelihood is then equal to

$$P(D|G,\hat{\theta}) \propto -\log(1-\exp(-\frac{1}{2}\sum_{k=0}^{K}\hat{\mathbf{p}}_\mathbf{k}\ln(\hat{\mathbf{p}}_\mathbf{k}/\mathbf{q}_\mathbf{k}) + \sum_{\mathbf{k=0}}^{\mathbf{K}}\mathbf{q}_\mathbf{k}\ln(\mathbf{q}_\mathbf{k}/\hat{\mathbf{p}}_\mathbf{k}))) \quad (7)$$

For simplicity, in the experiment section below, to obtain the L segments in [0,1], we simply divide the range of each parent node evenly into L segments.

2.3. *Pair-wise Causality Example*

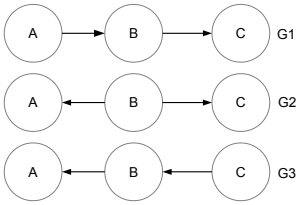
To illustrate our causal inference procedure, we apply it to a pair of variables, leaving to the following section our application to more complicated causal networks such as triple-node equivalence classes. To begin we generate synthetic data given true pair-wise relationships as depicted in Fig. 1. We assume the observed data for the parent nodes is drawn from a uniform distribution and use a hill function to describe the common interactions between the parent and child nodes. We add Gaussian noise to the synthetic data to model uncertainty inherent in

the measurement data, and without loss of generality we set L=4. In practice we must exercise some care in the selection of L, since if L is set too high, the power of the likelihood score to distinguish the true causal direction from the null hypothesis can be significantly decreased. On the other hand, if L is set too low, the fit of the data could be poor, leading to likelihood scores for the true and null causal models that may not achieve statistical significance.

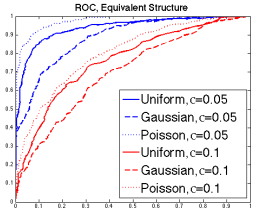
In Fig. 1, panels (a: A→B) and (b: A←B) show the fit of the regression model in the true and false causal directions, respectively. Each color denotes a different segment. The linear regression (LR) model is depicted by the line in each segment. The distribution of the predictor variables (parent nodes), is plotted in blue and the distribution of the response variables (child nodes) are shown by the stack plot, with the different colors corresponding to the different segments. In panel (a), A is the parent and its belief probability is clamped according to the observed and fitted distributions of the predicted values of the child node B. If the predicted values well match the observations, the likelihood score for (a) will be increased relative to the likelihood score for (b), which represents the opposite relationship (B as the parent node and A as the child). We note that in (b) for the 'flat' regions (I&IV) of the interaction function, the fitted LR does not cover the full range of values A can take on in these segments, which results in a truncation of the distribution of A at the ends of these two segments, resulting in a worse likelihood score compared to the true causal direction.

2.4. *Performance Analysis*

The asymmetric performance in predicting values along the true and false causal directions results from the constraints ([-1,1] in Eq. 4) defined for our pairwise causality test, which constrains the slope coefficient to fall $\in$ [-1,1] and the intercept coefficient to fall $\in$ [0,1] for each segment. These constraints enforce an asymmetry in the fit of the regression model between the true and false causal directions needed to infer the true direction. We have also assumed that any nonlinear curve can be well approximated by a piecewise linear regression (LR) model. That is, in each segment the LR is good enough when fitting along the true causal direction. When a segment is fit along the correct causal direction (i.e., the segment is assumed to lie in the dimension of the parent node and is mapped to the child node via the regression function), the length to set for the segment should be determined by the degree of noise in the data. If noise levels are low (high) along the true causal direction, then the size of the segment can be longer (shorter) with a smaller (larger) number of segments. In our case, given we constrain the slope and intercept coefficients in the LR, the belief propagation is able to fit the distribution of noisy observations well so long as the segment size is small enough. However, if the segment is fit along the wrong causal direction, the length of the segment will no longer help scatter the observed data into different segments, given the distribution of observed values of the child nodes in the flat or U shaped regions of the distribution of the parent nodes are not be completely captured, but instead are truncated as discussed above. As a result, there will be a high probability of the observed values on the child nodes falling in the same segment no matter how small the segment size is or how low the noise level is. In this case, the ideal length of the segment will be determined by the shape of the interaction function.



(a) Equivalent Structure



(b) ROC curve

Fig. 2. Synthetic Causal Network Structure

We note that while we have made the argument that smaller-sized segments will not significantly improve the fit along the wrong causal dimension, we have not ruled out the possibility that the distribution of the true parent node given the child node (wrong causal direction) is well approximated across the different segments (i.e., the truncation of the ends of the distribution discussed above goes away), when an extremely small segment size is chosen. If such a case were to arise, the likelihood score would be equally good in the true and false causal directions. Similarly, if the segment size chosen is too big, the possibility exists that even along the correct causal dimension, the coverage of the predicted distribution of the child node may not be complete, making the likelihood score in the true and false causal directions equally bad. Although there are some existing algorithms on choosing the optimal number of segments, e.g. Multivariate Adaptive Regression Splines,[13] we leave further investigation of choosing optimal segment sizes and positions as an interesting topic for future research.

3. Inferring Causality Among Markov Equivalent Structures

We next formally tested our causal inference procedure on synthetic data simulated to represent relationships that are common in biological systems, and on data generated from a dynamical systems model to recover metabolic network models. For the simulation experiment, we generated a large number synthetic datasets based on different nonlinear functions from the more difficult triple-node structures that are Markov equivalent. To infer known metabolic networks, we applied our method to yeast data generated from a metabolic model to demonstrate the ability to recover known metabolic networks.

3.1. *Markov Equivalent Structures*

For this problem, we test our method using the nonlinear functions $y{=}ax^3{+}\sin(k\pi x)$ and $z{=}by^2{+}\sin(k\pi y)$ to model equivalent structures. These functions represent a flexible framework for representing nonlinear biological relationships such as activation/inhibition and feedback control relationships, with the parameters a, b and k controlling the nonlinear features of such relationships: a,b$\in$[-2,+2] and k$\in$[0,2]. To demonstrate the broad applicability of our method to general nonlinear data, we allow the parameters to vary across a wide range of values. For the simulation component of our study, we generated 1000 datasets for each of the scenarios depicted in Fig. 2, with each dataset comprised of 100 samples (a typical sample size in biological experiments). The data were simulated based on the ground truth structure G_1 shown

in Fig. 2. The data were simulated from different distributions: Uniform(U), Gaussian(G) and Poisson(P), to mimic microarray and RNA-sequencing gene expression data. For each simulation, the parent node A (x) was sampled from the U, G and P distributions, and the child nodes B (y) and C (z) were then generated according to the above nonlinear function. Before adding Gaussian noise, we firstly rescale the observation of A, B and C into [0,1]. Next, Gaussian noise (0, σ^2) was also simulated to reflect technological variation inherent in the types of measures made in biology; Finally, this noise was added to the values of A, B and C and the added value is rescaled to [0,1] to complete the generation of our observation data D. The graphical structures depicted in Fig. 2a are Markov equivalent and so in the context of conventional Bayesian networks they all give rise to the same data likelihood, and thus, are statistically indistinguishable from one another. We evaluated the learning performance by generating receiver operator characteristic (ROC) curves (Fig. 2b). Here we see that our method infers the correct causal relationships among the Markov equivalent structures, given the explicit assumptions made on the nature of the interaction between the parent and child nodes and on the distribution of these nodes.

3.2. *Inferring Metabolic Signaling Pathways in Yeast*

We next applied our method to metabolic data generated from a yeast model to assess whether we could recover a known metabolic pathway, trehalose biosynthesis. Trehalose functions as a carbohydrate reservoir and has has recently been shown to play a crucial role in stabilizing proteins and cellular membranes under stress conditions such as heat shock. The metabolic pathway that produces trehalose is believed to regulate glucose uptake, particularly when the cell exists in a high-stress environment. It has also been shown that trehalose 6-phosphate (T6P), an intermediate of trehalose biosynthesis, plays a key role in the control of glycolytic flux. The kinetic values of this dynamical model have been identified experimentally[12] and are represented in the BioModels Database[11] (BIOMD0000000380).

We generated data for the trehalose biosynthetic pathway (in-silico) simulating the kinetic model for this model represented in the BioModel database. This model is comprised of a cycle of reactions between 11 different metabolites: external glucose (GLX), cellular glucose (GLC), glucose 6-phosphate (G6P), fructose 6-phosphate (F6P), glucose 1-phosphate (G1P), uridine triphosphate (UTP), uridine diphosphate (UDP), uridine diphosphate-glucose (UDG), diphosphate (Pi), trehalose 6-phosphate (T6P), and trehalose (TRH). These metabolites can be divided into two groups, the primary metabolites (M=7) whose concentrations vary as a result of reactions in this pathway, and extracellular glucose and boundary metabolites whose concentrations are fixed but they impact the reaction rates. Fig. 3 depicts the core causal signaling network we attempted to recover with our causal inference procedure. Represented in this network is a v-structure and feedback loop, structures containing Markov equivalent components that cannot be unambiguously resolved using classic Bayesian network approaches. To infer causality along each undirected (bold) edge, we generated a dataset by sampling 100 (a typical sample size in biological experiments) starting concentrations of extracellular glucose[12] (changing the medium) X^0_{glx} from the interval [0,100] (To ensure the system exhibits nonlinearity response to perturbations, we choose a wide range for the extracellular

(a) True Network (b) Undirected Skeleton

Fig. 3. Core Causal Signaling Pathway for Trehalose Synthesis in Yeast

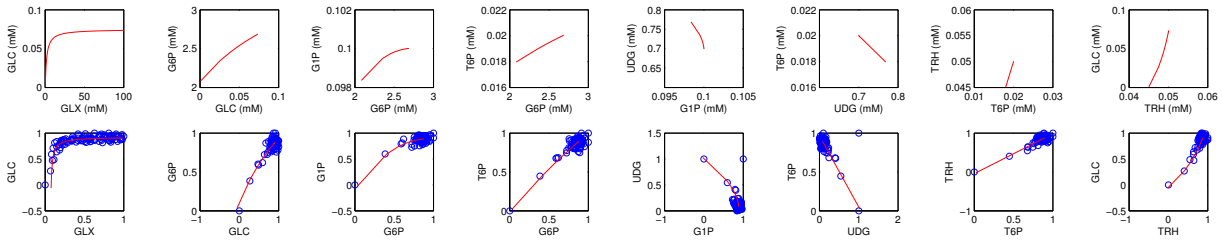
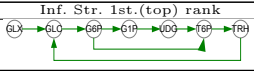
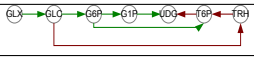


Fig. 4. Steady-state concentration given variations in starting extracellular glucose concentrations

glucose level). For each starting condition we let the dynamic system representing the trehalose biosynthetic pathway evolve to its new steady-state, generating a vector of steady state values for every primary metabolite, for each of the 100 starting conditions. This procedure resulted in a 100×7 data matrix of primary metabolite steady-state concentrations. In addition, we simulated Gaussian noise and added these noise components to the data matrix. Our final observation dataset is shown in Fig. 4. Each subplot describes the relationship of the steady-state concentrations between two (undirected) neighbor nodes in the pathway given the 100 different external glucose concentration starting conditions. The upper row shows these steady-state values before adding noise, while the lower row shows the rescaled values with the noise terms added, which represent the data used for the causal inference.

Given the connectivity structure of this network, we sought to resolve the edge direction by applying our method by calculating the causal structure score (Eq. 7) for each of the possible causal configurations. Given there are a total 8 edges in this network and that each edge can be oriented in one of two possible directions, there are 256 possible causal configurations to consider. The causal structure with the highest score was selected as the most likely causal structure supported by the data. In table 1 we list the top three inferred causal structures and their causality scores. We can see that our inferred top structure is the true causal network in 3. We note that the correct causal structure was inferred by considering the global structure of this network, as opposed to resolving the structure using pairwise causal relationships. One of the unique features of the modeling approach we developed is the ability to propagate information through the entire network. As a result, our global causal inference approach can leverage the correctly inferred causal relationships at between a given pair of nodes to infer the appropriate causal relationships among other nodes. This feature of our modeling approach is demonstrated by the causal inference of the feedback loop, i.e. TRH→GLC. With existing causal inference procedures, the inferred causal relationship would be estimated as GLC→TRH, whereas our method appropriately leveraged the global structure to correctly infer this edge, given the fitness of GLC, G1P and UDG is improved when we consider the

Inf. Str. 1st.(top) rank	GLX,TRH→GLC	GLC→G6P	G6P→G1P	G1P→UDG	G6P,UDG→T6P	T6P→TRH	Total Score
	3.41012	4.47932	5.3111	4.2021	4.3202	4.21400	25.93696
Inf. Str. 2nd. rank	GLX→GLC	GLC,T6P→G6P	G6P→G1P	G1P,T6P→UDG	TRH→T6P	GLC→TRH	Total Score
	3.40264	4.47932	5.27015	4.08511	4.04915	4.04646	25.33285
Inf. Str. 3rd. rank	GLX→GLC	GLC→G6P	G6P→G1P	G1P,T6P→UDG	G6P→T6P	GLC→TRH	Total Score
	3.40264	4.564723	4.99965	4.09345	4.16278	4.04646	25.26973

feedback in the top structure, compared to the other competing structures.

4. Conclusion and Discussion

In the life and biomedical sciences the technology now exists to score molecular and higher order phenotypes and genotypes on a massive scale, producing rich patterns of associations among molecular and higher order features that have the potential to elucidate the complexity of living systems. However, missing in biology is knowledge of the comprehensive set of pathways that operate in living systems, the structure of these pathways, how they interact with each other, how they change over time in response to different biological contexts etc. Even what are considered as canonical pathways are routinely shown to be incomplete and even inaccurate in different contexts. Therefore, methods that can help infer the causal relationships among the vast sea of phenotypes that can be scored are needed to better focus the type of hypotheses that can be experimentally pursued in a laboratory setting. Here we have attempted to address one significant limitation towards this end by developing a method to infer causality from correlation-based data by utilizing a Bayesian belief inference framework that is capable of distinguishing between Markov equivalent structures. By assuming different functional forms of the interactions that are possible among molecular features, observed data can be fit to probabilistic models of these relationships to assess which model best predicts the observed data. Our method is able to achieve good power in resolving causality by appropriately constraining the parameters of the probabilistic model in a way that allows the putative deterministic relationship between two variables to be assessed in a probabilistic framework. We applied our algorithm to multiple synthetic gene expression and RNA sequencing datasets to demonstrate that our approach can accurately infer causality under different biologically realistic assumptions regarding interaction types and noise structures.

Perhaps the biggest advantage of our approach is that it enables causal inference in a more complex network setting compared to previous methods that are limited to assessing pairwise causal relationships. This generality is achieved by leveraging the marginal probabilistic inference in a Bayesian network setting. We believe this advantage has significant importance given conventional top-down Bayesian network approaches can be systematically combined with our causal inference approach to form a integrated learning-inference framework. That is, our approach can enable a unified bottom-up and top-down modeling approach. Of course there are a number of ways in which the modeling approach we propose can be improved beyond our initial proof concept. Because our approach infers causality by maximizing the data likelihood that is based on a symmetrized KL-divergence measure between predicted and observed probabilities, implemented using a piece-wise linear regression framework, optimizing

the selection of the segment size and number is necessary to achieve maximal power and accuracy. Further, the causality inference can be embeded in structure search engine to search for optimal causality given initial graph. Finally, integrating our approach with a conventional structure-based learning approach, e.g. Bayesian network, has the potential to provide a very flexible framework that can model biological systems in a more comprehensive and accurate fashion, providing a way to incorporate bottom up modeling in a top-down framework to maximally leverage not only existing data, but knowledge derived from such data.

5. Acknowledgement

We thank the grant R01MH097276 of The National Institute of Mental Health (NIMH) and R01AG043076 of the Nation Institute of Aging (NIA) at National Institute of Health for their support to this work. JRK was supported by a James S. McDonnell Foundation Postdoctoral Fellowship Award in Complex Systems.

References

1. Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, (17 others). (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics* **37** 710 - 717.
2. Emilsson V, Thorleifsson G, Leonardson AS, (29 others), Stefansson K, Schadt EE. (2008) Genetics of gene expression and its effect on disease. *Nature* **452**:423-428.
3. Chen Y, Zhu J, Lum PY, Yang X, (16 others), Schadt EE. (2008) Variations in DNA induce changes in molecular network states that in turn lead to variations in obesity and related metabolic traits. *Nature* **452**:429-435.
4. Nir Friedman, Daphne Koller. (2003) Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning* Volume 50, Issue 1-2.
5. Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, Bernhard Schölkopf. (2012) Information-geometric approach to inferring causal directions. *Artificial Intelligence* Volumes 182-183.
6. Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, Bernhard Schölkopf. (2008) Nonlinear causal discovery with additive noise models. *Neural Information Processing Systems*
7. Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, Joris Mooij. (2012) On Causal and Anticausal Learning. *International Conference on Machine Learning*
8. David Heckerman. (1996) A Tutorial on Learning With Bayesian Networks. *Technical Report.*
9. Judea Pearl. (2009) *Causality: Models, Reasoning, and Inferenc; 2nd Edit.* Cambridge Univ. Press.
10. Trevor Hastie, Robert Tibshirani, Jerome Friedman. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction; 2nd Edit.* Springer
11. Li Chen, Donizelli Marco, Rodriguez, Nicolas, Dharuri Harish, Endler Lukas, Chelliah Vijayalakshmi, Li Lu, He Enuo, Henry Arnaud, Stefan Melanie, Snoep Jacky, Hucka Michael, Le Novere Nicolas, Laibe Camille. (2010) BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology* Volume 4-1.
12. Kieran Smallbone, Naglis Malys, Hanan L. Messiha, Jill A. Wishart, Evangelos Simeonidis, Chapter eighteen - Building a Kinetic Model of Trehalose Biosynthesis in Saccharomyces cerevisiae, In: Daniel Jameson, Malkhey Verma and Hans V. Westerhoff, Editor(s), *Methods in Enzymology, Academic Press, 2011, Volume 500, Pages 355-370.*
13. *Jerome H Friedman. (1991) Multivariate Adpative Regression Splines. The Annals of Statistics Volume 19, No. 1,1-141.*