Physical network models and multi-source data integration

Chen-Hsiang Yeang
MIT AI Lab
Cambridge, MA 02139
chyeang@ai.mit.edu

Tommi Jaakkola
MIT AI Lab
Cambridge, MA 02139
tommi@ai.mit.edu

September 30, 2002

Abstract

We develop a new framework for inferring models of transcriptional regulation. The models in this approach, which we call *physical models*, are constructed on the basis of verifiable molecular properties of the underlying biological system. The properties include, for example, the existence of protein-protein interactions and whether a DNA binding protein binds to a regulatory region of a specific gene. Each property – either implicated or hypothesized – is included as a variable in the model. The setting of all the variables defines an annotated graph representing the molecular interactions that are present. Some of the available data sources such as factor-binding data (location data) involve measurements that are directly tied to the variables in the model. Other data sources such as gene knock-outs are *functional* in nature and provide only indirect evidence about the (physical) variables. We solve this data association problem by linking each knock-out effect with a set of causal paths (molecular cascades) that could in principle explain the effect. The optimal setting of all the variables (including the selection of cascades) is found approximately via the max-product algorithm operating on a factor graph. We demonstrate that this approach is capable of predicting gene knock-out effects with high degree of accuracy in a cross-validation setting. Moreover, the approach implicates likely molecular cascades responsible for each observed knock-out effect. We can easily extend the approach to include other data sources (solve the corresponding data association problems). This includes, for example, time course expression profiles. We also discuss generalizations to represent coordinated regulation and the use of automated experiment design.

1 Introduction

Understanding transcriptional regulation is a key problem in contemporary biology. The biological system exhibiting transcriptional control needs to be understood both at the component (molecular) level as well as operationally. While a great deal is known about the components involved in regulation in general, relatively little is known about the functional behavior of the system. Model organisms and/or subsystems such as *lac operons* in *E. coli* [6] and *GAL*4 pathways in *S. cerevisiae* [7] provide clues about how gene regulation and signal transduction pathways may operate on a larger scale and in other organisms. The accumulation of high-throughput measurements including expression arrays (e.g., [3]), factor-binding profiles (location data) [5], and measurements of protein complexes [9] permit computational realization of the underlying biological mechanisms.

Any computational model of a regulatory system can be judged on the basis of its ability to explain or predict consequences of interventions such as gene knock-out effects or predict measurements carried out in the course of the natural operation of the biological system. The models determine what properties/features to explain in addition to how to explain them. A typical computational approach pertaining to gene regulation involves building a statistical model over (preprocessed) measurements in an attempt to discover (possibly causal) dependencies among such measurements. Some of the examples include Bayesian

networks [1, 2] and relational probabilistic models [8] in gene expression analysis. The variables in these models might be, for example, expression values as well as other possibly unobserved (latent) variables that are deemed useful for capturing measured dependencies among the observables. Models of this type can be used to integrate heterogeneous data sources.

Perhaps the main deficiency of such statistical models is that they require considerable effort in interpreting the results after the fact. The statistical dependencies among variables can be realized by many possible mechanisms in the process of gene regulation. For example, genes which are clustered together according to their expression profiles are possibly co-regulated by the same transcription factor, share a common cause in the regulatory network, or do not have explicit relations. To overcome this problem, we provide here an alternative approach, where the resulting description – an *annotated physical graph* – provides a clear and unambiguous interpretation of the underlying biological system. By setting the variables or attributes in these models we represent specific hypotheses concerning verifiable molecular properties. This perspective shifts the computational effort from (largely unautomated) interpretation problems to estimation problems. The estimation problems arise from the fact that many important data sources available for inferring transcriptional regulation (including gene knock-outs) do not directly measure specific molecular events but rather assess functional consequences of interventions. This data association problem is largely avoided in statistical models, where the variables are more directly tied to the observations such as expression levels. We can cast the data association problem as a standard graphical model inference problem and readily solve it.

A simple realization of the core physical model we wish to infer is an annotated graph, where the nodes are associated with genes (or their protein products) and edges correspond to types of molecular interactions. We consider here only two types of edges. Undirected edges describe protein-protein interactions whereas directed edges are used to specify which genes DNA binding proteins actually bind to. Each type of edge is in addition annotated with a sign (positive or negative), where the sign represents the immediate molecular effect of the interaction. For example, a positive directed edge may signify that the presence of a DNA binding protein is necessary for transcribing a specific gene. A negative undirected edge, on the other hand, can be used to describe the case where the associated protein complex renders one or the other protein inactive[1]. The core variables in this model correspond to the presence or absence of the edges (among the edges hypothesized to exist in the first place) as well as the signs attached to the edges. We will later introduce additional variables associated with paths (molecular cascades) that facilitate the mapping of variables to the available data sources. In the absence of any observed data, the model is a random annotated graph without any clear preference over which graph represents a likely interpretation of the biological system. The probabilistic constraints arising from available data sources are incorporated into a factor graph model. The resulting most likely configuration (an annotated physical graph) can be solved with approximate inference methods such as the max-product algorithm or variants [4]. The bulk of the effort in this paper concerns with establishing the association between the variables specifying the physical model and the available measurements.

We will use three types of data to constrain physical models of transcriptional regulation in yeast. These are protein-protein interactions derived from the YPD database[2], 161 location (factor binding) profiles of yeast transcription factors [5], as well as 300 genome-wide expression profiles of knock-out experiments [3]. The first two data sources provide direct measurements of the values of the variables in the model. To incorporate the knock-out experiments, on the other hand, we have to provide a mapping from significant effects to sets of variables in the model. This can be done by generating all sufficiently short paths in a physical graph that are capable of explaining the observed effect. Any such path provides a tentative causal explanation for the knock-out effect in terms of molecular cascades. The selection of paths that are deemed responsible for the observed knock-out effects are included as additional (biologically meaningful)

[1] In a more extensive specification protein-protein interactions are represented with directed (possibly bidirectional) edges.
[2] https://www.incyte.com/proteome/index.html

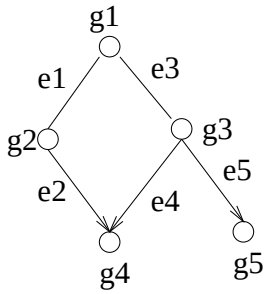

Figure 1: A simple example of physical interaction network

variables in the model.

We begin with an illustrative toy example. This is followed by a more general and formal description of the physical model and our solution to the data association problems. We subsequently evaluate the methodology in a cross-validation setting to predict the effects of gene knock-outs.

2 An illustrative example

Figure 1 shows a simple network of protein-DNA and protein-protein interactions of 5 genes. The edges in the graph represent the set of possible interactions and we wish to infer which of these edges are indeed present, and the signs of these interactions. Directed edges denote protein-DNA interactions and undirected edges signify protein-protein interactions. There are 5 edges in this network.

There are two types of variables of interest: the presence of physical (protein-DNA and protein-protein) interactions and the signs of these interactions (whether one gene has an immediate positive or negative effect on its downstream gene). More precisely, the variables include $x_1, \cdots, x_5$ and $s_1, \cdots, s_5$, where each x_i is an indicator variable encoding the presence/absence state of the physical interaction e_i and $s_i's$ are ± 1 variables providing the signs of the annotated edges. The values of the variables $x_1, \cdots, x_5$ are constrained directly by protein-protein or protein-DNA measurement(s) as well as indirectly by knock-out observations. In contrast, the signs $s_1, \cdots, s_5$ can be only inferred on the basis of knock-out effects. We must first formally tie the variables to the observed data and subsequently infer the most likely configuration of the variables in light of the available data.

Suppose now that we have protein-DNA and protein-protein interaction measurements pertaining to all the edges along with the error models characterizing the noise in the measurements. Furthermore, suppose we have observed that g_4 is down-regulated by knocking out g_1, and that there are no other significant knock-out effects.

On the basis of error models governing protein-DNA and protein-protein interaction data, we can define potential functions $\phi_i(x_i)$ that incorporate the evidence about the existence of the edges in light of such measurements (a more formal mapping is given later in the paper). Assume for simplicity that all x_i's are paired with identical potential functions: $\phi_i(0) = 1$ and $\phi_i(0) = 0.9$ (note that the potential functions need not be normalized).

How is the functional knock-out observation related to the values of the variables? If we hypothesize that genes are regulated through cascades of protein-protein and protein-DNA interactions, then the observed knock-out effect can be explained by partially directed paths (e_1, e_2) or (e_3, e_4) (or both). In order for a path to explain the knock-out effect, all the interactions along the path must exist, and the aggregate effect of the sign along the path must be consistent with the observed sign of the knock-out effect. These constraints can be put into a potential function:

$$\psi_1(x_1, \ldots, x_4, s_1, \ldots, s_4) = \begin{cases} 1.00 & \text{if } (x_1 = x_2 = 1, s_1 \cdot s_2 = +1) \vee (x_3 = x_4 = 1, s_3 \cdot s_4 = +1). \\ 0.01 & \text{otherwise.} \end{cases}$$

The potential function does not vanish when the constraints are violated because we may not fully trust the knock-out effect given the available error model (there may be also other paths explaining the effect but such paths would involve edges not included here). Note that x_5 and s_5 are not involved in $\psi_1(.)$ because e_5 is not on any partially directed path from g_1 and g_4.

The potential functions $\phi_i's$ and ψ_1 now define a joint distribution over the variables:

$$P(X,S) \propto \left[\prod_{i=1}^{5} \phi_i(x_i)\right] \cdot \psi_1(x_1,\ldots,x_4,s_1,\ldots,s_4).$$

where $X = \{x_1,\ldots,x_5\}$ and $S = \{s_1,\ldots,s_5\}$. This probability model is naturally viewed as a *factor graph* and inference algorithms such as *max-product* are available for finding the most likely configuration(s). The max-product algorithm tries to evaluate so called *max probabilities*, defined as $P_{max,i}(x_i) \propto max_{\{X,S\}\setminus\{x_i\}} P(X,S)$. If $P_{max,i}(1) > P_{max,i}(0)$ then necessarily $x_i = 1$ is in the most likely configuration.

In this example the most likely configurations are

$$\begin{array}{rcl}(x_1,\ldots,x_5,s_1,\ldots,s_5) & = & (1,1,0,0,0,+1,+1,*,*,*)\\ & & (1,1,0,0,0,-1,-1,*,*,*)\\ & & (0,0,1,1,0,*,*,+1,+1,*)\\ & & (0,0,1,1,0,*,*,-1,-1,*)\end{array}$$

where $*$ indicates that either value is acceptable. The configurations represent the fact that either (e_1,e_2) or (e_3,e_4) must exist with consistent aggregate signs. In this case the configurations corresponding to the situations that both paths explain the knock-out effect (i.e., when $x_1 = x_2 = x_3 = x_4 = 1$) have lower probabilities because of the slight biases arising from the individual potential functions: $\phi_i(1) < \phi_i(0)$.

It is worth noticing that albeit each edge has the same confidence value from the protein-DNA and protein-protein interaction data, their overall confidence values can be different. In this example, $x_5 = 0$ in all the most likely configurations. This is because e_5 is not involved in explaining the knock-out effect. Moreover, the more data are available, the more constraints are imposed on possible configurations. For example, suppose we conduct the experiment of deleting gene g_3 and find g_4 is down-regulated. This extra evidence can reduce the most likely configurations to

$$(x_1,\ldots,x_5,s_1,\ldots,s_5) = (0,0,1,1,0,*,*,+1,+1,*)$$

3 Physical models

Our physical model can be represented as a collection of attributes or variables pertaining to verifiable molecular properties of the biological system such as protein-DNA binding events and formation of protein complexes. The variables need not to be (currently) directly observable and may involve collective properties such as signal transduction pathways. The main requirement is that the variables have to be tied to (in principle) verifiable properties. Any particular setting of such variables gives rise to an annotated physical graph representing interactions that are present. In contrast to dependency models, physical models not only explain observed dependencies but also articulate clear hypotheses about the underlying biological mechanisms.

Our framework comprises three parts: a partially directed graph representing the set of possible physical interactions, the set of variables whose values determine a physical model, and the construction of a joint distribution over the variables by incorporating observed measurements as evidence. We describe each part in detail with the emphasis on data association.

3.1 Graph representation

Graphs provide a natural representation of possible physical interactions. Here $G = (V, \vec{E}_G \cup \bar{E}_G)$ defines as a partially directed (possibly cyclic) graph. V is the set of vertices corresponding to genes or their protein products, $\vec{E}_G$ is the set of directed edges corresponding to possible protein-DNA interactions, and $\bar{E}_G$ is the set of undirected edges denoting the possible protein-protein interactions. In this simple representation we do not distinguish between the DNA sequence, mRNA template,or the protein product of a gene. Two genes g_1 and g_2 can be therefore linked by both an undirected edge and up to two directed edges. In the former case, vertices play the role of protein products whereas in the latter case we refer to the protein binding to the promoter region of the corresponding gene.

The edges in this simple graph G denote possible pairwise physical interactions and G may be a complete graph (where there are three edges connecting each pair of vertices). It is often possible to restrict the possible interactions a priori, e.g., by excluding protein-DNA interactions without sufficient support from the location data. The graph representation can be extended to a hypergraph in order to represent multi-way interactions, for example, the formation of a complex or coordinated regulation. The hyper-edges in this case correspond to sets of vertices. We will limit the approach here to pairwise interactions.

3.2 Variables

We select only variables that encode meaningful (verifiable) biological properties. It is sensible to include only those variables that stand to receive some support either directly or indirectly from the available data. For example, we include activation delays only when relevant time course profiles are available. The variables are associated with features in the physical graphs such as vertices, edges, hyper-edges, paths, or clusters.

We focus here on a model which incorporates three types of data: location analysis data of protein-DNA interactions, protein-protein interactions, and the mRNA expression levels of gene knock-out experiments. The relevant variables of the regulatory model are in this case:

- $X_{\vec{E}_G} = \{x_{\vec{e}_i} : \vec{e}_i \in \vec{E}_G\}$, a collection of binary (0/1) variables pertaining to the presence or absence of protein-DNA interactions.
- $X_{\bar{E}_G} = \{x_{\bar{e}_j} : \bar{e}_j \in \bar{E}_G\}$, an analogous collection of binary variables denoting whether protein-protein interactions are present
- $S_{\vec{E}_G} = \{s_{\vec{e}_i} : \vec{e}_i \in \vec{E}_G\}$ and $S_{\bar{E}_G} = \{s_{\bar{e}_i} : \bar{e}_i \in \bar{E}_G\}$ which provide the signs $(+1/-1)$ of the interactions represented by the edges.

3.3 Potential functions

We formalize here how the variables can be tied to the observations through potential functions. The joint distribution over all the variables can be then defined as a product of the potential functions similarly to the toy example.

3.3.1 Potential functions for protein-protein and protein-DNA data

To build potential functions on the basis of the location and protein-protein interaction data we introduce two sets of measurement variables: $Y_{\vec{E}_G} = \{y_{\vec{e}_i} : \vec{e}_i \in \vec{E}_G\}$ is a collection of real valued DNA binding affinity measurements and $Y_{\bar{E}} = \{y_{\bar{e}_j} : \bar{e}_j \in \bar{E}_G\}$ provides the observed protein-protein binding affinities[3] For a given dataset, the values of the variables in $Y_{\vec{E}_G}$ and $Y_{\bar{E}_G}$ are fixed.

[3]If these are available. Alternative heuristic error models may have to be used in the absence of such measurements.

The potential function $\phi_{\vec{e}_i}(x_{\vec{e}_i}; y_{\vec{e}_i})$ pertaining to the direct evidence about a protein-DNA interaction $\vec{e}_i$ is proportional to the ratio of the conditional probabilities derived from the error model:

$$\phi_{\vec{e}_i}(x_{\vec{e}_i}; y_{\vec{e}_i}) \quad = \quad \left[\frac{P(y_{\vec{e}_i}|x_{\vec{e}_i}=1)}{P(y_{\vec{e}_i}|x_{\vec{e}_i}=0)}\right]^{x_{\vec{e}_i}}. \tag{1}$$

where $\phi_{\vec{e}_i}(x_{\vec{e}_i}; y_{\vec{e}_i})$ is a function of $x_{\vec{e}_i}$ only since the value of $y_{\vec{e}_i}$ remains fixed. The potential function $\phi_{\bar{e}_i}(x_{\bar{e}_i}; y_{\bar{e}_i})$ of an undirected edge can be defined analogously. We assume here for simplicity that the error models for protein-DNA and protein-protein interaction data are given and discuss practical remedies in the context of evaluating the methodology.

3.4 Potential functions from knock-out data

We begin by introducing variables corresponding to idealized knock-out effects. $K = \{k_{ij}\}$ is a collection of the discrete variables of pairwise single knock-out effects whose domains are $\{-1, 0, +1\}$. k_{ij} denotes the effect of knocking out gene g_i on gene g_j. $k_{ij} = -1$ if g_j is down-regulated, $+1$ if g_j is up-regulated, and 0 if g_j remains unaffected by the knock-out. These variables can be directly tied to the observed knock-out measurements, or $O_K = \{o_{k_{ij}}\}$, analogously to those discussed in the previous section. The resulting potential functions are

$$\phi_{ij}(k_{ij}; o_{k_{ij}}) \propto \left[\frac{P(o_{k_{ij}}|k_{ij})}{P(o_{k_{ij}}|k_{ij}=0)}\right]. \tag{2}$$

The actual knock-out effect is associated with multiple core attributes in the model. This association amounts to explaining each knock-out effect through a cascade of physical interactions available in the model (here signed protein-DNA and protein-protein interactions). We must first decide what aspects of the knock-outs we attempt to capture. While any significant knock-out effect (a gene is up or down-regulated) can be easily attributed to a cascade of physical interactions, unaffected genes are much more difficult to explain as other causes may be at play. To a first approximation we attempt to explain only likely up/down regulations.

The potential function associated with a knock-out effect k_{ij} reflects the constraint that a cascade in the physical model has to explain k_{ij}. For a path in G to qualify for explaining k_{ij}, the path, denoted here as π, must satisfy:

1. The end nodes of π are g_i and g_j.
2. The last edge in π is directed (protein-DNA interaction).
3. All the directed edges in π are in the forward direction (from g_i to g_j).
4. The signs of the edges along π are consistent with the sign of the knock-out effect.
5. The length of π is less than a pre-defined upper bound.
6. Intermediate genes along π either have knock-out effects on g_j or have no knock-out measurements.

The first condition manifests the assumption of using a cascade of physical interactions to explain gene regulation. The second condition is based on the accepted assumption that the last step of gene regulation is transcription control. The third condition ensures that the path has a causal interpretation. The fourth condition is evident as stated and the fifth one excludes unreasonably long cascades. The last condition requires that each interaction along a path is a necessary component for gene regulation with the exception of missing data. A path which satisfies these conditions is able to explain the knock-out effect k_{ij}. k_{ij} is

explained by the physical model if there exists at least one path which satisfies these conditions. These conditions would have to be modified slightly to incorporate the notion of coordinate regulation.

The above conditions impose constraints on the presence of edges and the signs of edges. Let $\Pi_{ij} = \{\pi_1, \cdots, \pi_n\}$ denote the candidate paths connecting g_i and g_j which satisfy conditions 1, 2, 3, 5, and 6. Condition 4 remains inapplicable until we incorporate the signs of edges along the paths. We need to introduce here an auxiliary variable that is used to select the path that explains the knock-out effect. Specifically, $\Sigma = \{\sigma_b : b \in K\}$ is the collection of path selection variables for explaining knock-out effects. Each $\sigma_b \in \Sigma$ is indexed by a pairwise knock-out effect $b \in K$, and its value is the index of the path which explains the knock-out effect b.

We use σ_{ij} to denote the path selection variable for the knock-out effect k_{ij}. Let $E_a = \{e \in \pi_a\}$ denote the (directed and undirected) edges along π_a, $X_a = \{x_e : e \in E_a\}$ and $S_a = \{s_e : e \in E_a\}$ be the presence and sign variables of the edges along π_a, $E_{ij} = \cup_{\pi_a \in \Pi_{ij}} E_a$, $X_{ij} = \cup_{\pi_a \in \Pi_{ij}} X_a$, and $S_{ij} = \cup_{\pi_a \in \Pi_{ij}} S_a$. Then π_a explains k_{ij} if we select π_a ($\sigma_{ij} = a$) and the following conditions hold:

- $\forall e \in E_a$, $x_e = 1$.
- $\prod_{e \in E_a} s_e = -k_{ij}$.

The potential function encoding these conditions can be expressed as follows:

$$\psi_{ija}(X_a, S_a, k_{ij}) = \begin{cases} 1 & \text{if } (\bigwedge_{e \in E_a} x_e) \wedge I(\prod_{e \in E_a} s_e = -k_{ij}), \\ \epsilon & \text{otherwise.} \end{cases} \tag{3}$$

where $\wedge$ denotes logical AND and $I(.)$ is the indicator function. The potential function does not vanish even when the constraints are violated so as to allow us to refrain from explaining some of the knock-out effects.

The potential function associated with a pairwise knock-out effect is obtained by combining the potential functions along each candidate path:

$$\psi^0_{ij}(X_{ij}, S_{ij}, \sigma_{ij}, k_{ij}) = \sum_{a=1}^{|\Pi_{ij}|} I(\sigma_{ij} = a)\psi_{ija}(X_a, S_a, k_{ij}). \tag{4}$$

$\psi^0_{ij}(.)$ returns a relatively high value if there exists a path which can explain k_{ij} provided that the path is selected.

Since we are currently explaining only significant knock-out effects (i.e., excluding unaffected genes), we modify the potential function slightly to incorporate this choice a priori:

$$\psi_{ij}(X_{ij}, S_{ij}, \sigma_{ij}, k_{ij}) = I(k_{ij} \neq 0)\psi^0_{ij}(X_{ij}, S_{ij}, \sigma_{ij}, k_{ij}) + I(k_{ij} = 0). \tag{5}$$

$\psi_{ij}(.)$ returns a relatively high value if either there is a significant knock-out effect between g_i and g_j and the model explains this knock-out effect, or there is no significant knock-out effect between g_i and g_j.

4 Inference of model attributes

We can combine the potential functions into a joint distribution over all the core and auxiliary variables

$$\begin{aligned} P(X_{\vec{E}_G}, S_{\vec{E}_G}, X_{\bar{E}_G}, S_{\bar{E}_G}, K, \Sigma; Y_{\vec{E}_G}, Y_{\bar{E}_G}, O_K) \quad \propto \quad & \prod_{\vec{e}_i \in \vec{E}_G} \phi_{\vec{e}_i}(x_{\vec{e}_i}; y_{\vec{e}_i}) \cdot \prod_{\bar{e}_j \in \bar{E}_G} \phi_{\bar{e}_j}(x_{\bar{e}_j}; y_{\bar{e}_j}) \cdot \\ & \prod_{k_{ij} \in K} \phi_{ij}(k_{ij}; o_{k_{ij}}) \cdot \prod_{k_{ij} \in K} \psi_{ij}(X_{ij}, S_{ij}, \sigma_{ij}, k_{ij}). \end{aligned} \tag{6}$$

which can be interpreted as a factor graph [4]. A factor graph is an undirected bi-partite graph with two types of nodes: variables whose values we are interested in, and factors (potential functions) exhibiting constraints between the variables. The edges exist between each variable and a factor that constrains its value.

The remaining problem is to find the most likely configuration of the values of all the variables in this factor graph representation. This MAP configuration can be computed approximately by the *max-product* algorithm or its refinements [4]. In brief, the max-product algorithm is a special case of the belief propagation algorithms for graphical model inference. Each node iteratively passes messages to its neighbors, where a message contains the information about this node and all previous messages coming to this node (except the destination node of the message). The algorithm stops when all messages in the graph converge. Given the functional forms of the potential functions we have described earlier, this message passing algorithm can be implemented efficiently (we can evaluate expectations over the potential functions relative to a factorized distribution). Of course, the fact that we can run the inference algorithm efficiently provides no guarantees about the quality of the solution. Certain guarantees are known for max product, however (see, e.g., [10]).

The max-product algorithm may converge to max-marginals that are uninformative (do not specify which values some variables should take in the most likely configuration(s)). In such cases we perform an additional recursive search by fixing some of the variables and running the max-product in stages. More details about the inference algorithms and their application in this context will be provided in a longer version of the paper.

5 Empirical results

We evaluate the framework using three datasets in budding yeasts: location analysis data about protein-DNA interactions [5], protein-protein interaction data manually pulled out from the YPD database, and mRNA expression of knock-out experiments from the Rosetta compendium data [3]. To simplify the task we focus on genes involved in the pheromone response pathway. We select protein-DNA interactions and pairwise knock-out effects whose p-values ≤ 0.001. The resulting model contains 46 genes, 34 directed edges, 30 undirected edges and 164 pairwise knock-out effects[4].

The heuristic error model developed by Hughes et al. [3] is applied to location and knock-out data. False positive p-values are derived according to this error model. As a simple (and incorrect) use of the error model, we take the p-values to represent the probabilities $P(\text{measurement}|\text{interaction does not exist})$ in location and knock-out datasets. The datasets do not provide information about $P(\text{measurement}|\text{interaction exists})$. In this preliminary evaluation, we set these to arbitrary fixed values (0.002) for all confident edges.

Protein-protein interaction data is obtained from the YPD database [5]. The degree of confidence in each interaction is not provided in the database. Here we set the potential functions of all implicated protein-protein interactions to $\phi_i(x) = 2.0I(x = 1) + I(x = 0)$ to reflect the high degree of false positives in the dataset.

The variables in the model used in the evaluation include indicator variables about the presence or absence of physical interactions, signs of those interactions, actual (idealized) pairwise knock-out effects, as well as path selection variables for explaining the knock-out effects. The joint probability over these variables was constructed by combining the potential functions as described earlier. The max-product algorithm was applied to obtain the (max) marginals for each variables and the configurations were further enumerated via iterative search when necessary. In this example, there are only 8 degenerate configurations.

By restricting the path length ≤ 5, 136 out of 164 knock-out pairs are connected via valid paths (the

[4]The genes and interactions can be downloaded from http://www.ai.mit.edu/people/chyeang/pheromone.ps

[5]https://www.incyte.com/proteome/index.html

Table 1: Cross validation on knock-out pairs

# hold-outs	# trials	% error
1	136	0.74 %
5	500	0.68 %
20	200	12.22 %

paths which satisfy conditions 1, 2, 3, 5, 6 in section 3.4). It turns out all those 136 pairs are consistent with the MAP configurations obtained from the max-product and the iterative search. To validate the inference results, we randomly hold out a number of knock-out pairs when constructing the joint distribution, and use the resulting MAP configurations to explain the hold-out knock-out pairs. Table 1 shows the results of leave-n-out cross validation, where n equals to 1, 5, and 20. The results indicate that the algorithm is very robust against random removal of information along the pathway. This is to be expected since the information about a knock-out interaction is distributed among multiple interactions along pathways. In contrast, if we systematically hide all effects regarding a particular knock-out experiment, then the small number of other available knock-out experiments no longer suffices to constrain the variables enough to predict the effects.

6 Extensions

There are a number of ways in which we can extend our framework. We provide here a few examples.

Biological experiments are typically costly and time consuming. Systematizing the experimental effort with the help of computational techniques can be important. The use of the physical models is especially natural in this context. In our framework, new experiments can serve two major purposes. The existing datasets are unlikely to impose sufficient constraints to yield a unique physical interpretation. New experiments become necessary in order to further distinguish between degenerate models. We can view a model as a system which responds to inputs (environmental or internal perturbations) by producing a set of observable outputs. Ideally, we would like the predicted outputs of the degenerate models to be all distinct. When this is not possible, we want the experimental outcomes to evenly divide the space of degenerate models. This leads to an entropic criterion for selecting the experiments. We are developing approximate methods to evaluate such an information criterion efficiently. New experiments can also verify or falsify existing (unique) models. Any inferred interactions not yet known to exist can be verified experimentally. Moreover, we can use gene knock-outs to fill in information along each explaining cascade about the knock-out effects of the intermediate genes. This can confirm or falsify the inferred function of the path.

All the physical interactions in the examples given above have been pairwise interactions. In an actual biological system, interactions may involve multi-protein complexes (e.g., holoenzyme in yeast) or coordinated binding of multiple proteins. Our interest is in particular in capturing coordinated means by which proteins can regulate a single gene. If these proteins are transcription factors and the activity of the downstream gene is the mRNA expression level, then this multi-way interaction is a hybrid of protein-DNA and protein-protein interactions. In other cases, the combinatorial effect is due to protein-protein interactions alone. We can use a hyper-graph to represent the regulatory network involving this type of multi-way interactions. As before we can incorporate variables specifying the presence or absence of hyper-edges. We can also generalize the notion of the edge sign to a hyper-edge. Here the "sign" specifies instead a combinatorial (logic) function (e.g., AND) of how coordination is required for a regulatory effect. This approach can be particularly useful in interpreting double knock-out experiments.

We can also incorporate other types of functional data to further constrain the model. Unlike knock-

out expression data in which causes (the deleted genes) and effects (the affected genes) are clear, causal relations are often difficult to resolve in most expression datasets. In time course profiles, however, the order of the measurements does restrict possible causal interpretations. We can incorporate time course profiles as evidence in our framework both in terms of trying to infer additional attributes (time lags of interactions) as well as to explain appropriately chosen time lag correlations on the basis of common ancestors in the physical graph. Such association of observations to sets of variables is analogous to the knock-out case.

7 Conclusion

We have developed a new framework for inferring genetic regulatory networks from multiple sources of data. Our approach differs from many previous methods (statistical dependency models) in terms of requiring readily interpretable and verifiable models of underlying biological mechanisms. Our experimental results are encouraging. Cross validation experiments on a reduced regulatory subsystem indicate that the presence and sign of protein-DNA and protein-protein interactions can be accurately predicted under this framework. The framework can be naturally extended to model other characteristics of the regulatory network such as coordinated effect of multiple transcription factors or even to resolve hidden causes of responses to environmental perturbations.

References

[1] Friedman, N. et al (2000). *Using Bayesian networks to analyze expression data*, *RECOMB Proceedings.*

[2] Hartemink, A.J. et al. (2001). *Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks*, *PSB Proceedings.*

[3] Hughes, T.R. et al. (2000). *Functional discovery via a compendium of expression profiles*, *Cell*, 102:109-126.

[4] Kschischang, F.R. et al. (2001). *Factor graphs and the sum-product algorithm*, *IEEE transactions on information theory*, 47(2):498-519.

[5] Lee, T.I. et al. (2002). *A transcriptional regulatory network map for Saccharomyces cerevisiae*, *Science*, in press.

[6] Lewin, B. (2000). *Genes VII*, Oxford University.

[7] Lohr, D. et al. (1995). *Transcriptional regulation in yeast GAL gene family: a complex genetic network*, *FASEB Journal*, 9:777-787.

[8] Segal, E. et al. (2001). *Rich probabilistic models for gene expression*, *ISMB Proceedings.*

[9] Uetz, P. et al. (2000). *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae*, *Nature*, 403:623-627.

[10] Wainwright, M.J. et al. (2002). *Tree consistency and bounds on the performance of the max-product algorithm and its generalizations. LIDS Technical Report, Laboratory for Information and Decision Systems, MIT*, Cambridge, MA.