Physical network models

Chen-Hsiang Yeang

MIT Computer Science and Artificial Intelligence Laboratory

Cambridge, MA 02139, USA.

chyeang@csail.mit.edu

Trey Ideker

UCSD Bioengineering

La Jolla, CA 92093, USA.

trey@bioeng.ucsd.edu

Tommi Jaakkola

MIT Computer Science and Artificial Intelligence Laboratory

Cambridge, MA 02139, USA.

tommi@csail.mit.edu

November 24, 2003

Abstract

We develop a new framework for inferring models of transcriptional regulation. The models, which we call *physical network models*, are annotated molecular interaction graphs. The attributes in the model correspond to verifiable properties of the underlying biological system such as the existence of protein-protein and protein-DNA interactions, the directionality of signal transduction in protein-protein interactions, as well as signs of the immediate effects of these interactions. Possible configurations of these variables are constrained by the available data sources. Some of the data sources such as factor-binding data involve measurements that are directly tied to the variables in the model. Other sources such as gene knock-outs are *functional* in nature and provide only indirect evidence about the variables. We associate each observed knock-out effect in the deletion mutant data with a set of causal paths (molecular cascades) that could in principle explain the effect, resulting in aggregate constraints about the physical variables in the model. The most likely settings of all the variables, specifying the most likely graph annotations, are found by a recursive application of the max-product algorithm. By testing our approach on datasets related to the pheromone response pathway in *S. cerevisiae*, we demonstrate that the resulting model is consistent with previous studies about the pathway. Moreover, we successfully predict gene knock-out effects with high degree of accuracy in a cross-validation setting. When applying this approach to genome-wide, we extract submodels consistent with previous studies. The approach can be readily extended to other data sources, or to facilitate automated experimental design.

1 Introduction

Understanding transcriptional regulation is a leading problem in contemporary biology. The challenge arises in part from the apparent complexity of regulatory systems. Indeed, regulatory systems involve diverse physical mechanisms of control and complex inter-connected networks of molecular interactions mediating such mechanisms. To unravel such systems effectively, we need both sufficient experimental data probing relevant aspects of such systems as well as computational approaches that aid in the large scale reconstruction of the underlying

molecular biology.

It is reasonable to start with simplified models. The specificity of transcription factor-promoter bindings and cascades of protein modifications or interactions form the basis of transcriptional initiation and signal transduction pathways. These two mechanisms form a crude yet effective picture of gene regulation. Specifically, external stimuli are propagated via signal transduction pathways to activate transcription factors, which in turn bind to promoters to enable or repress transcription of the associated genes. Several genome wide datasets are already available about protein-DNA and protein-protein interactions (e.g., Deane et al. 2002, Ito et al. 2001, Lee et al. 2002, Uetz et al. 2000). Such interactions constitute the basic network of physical interactions employed in regulatory control. Consistent with previous approaches, we build on the assumption that gene regulatory processes are carried out along pathways of physical interaction networks.

Pathways of molecular interactions are necessary but not sufficient elements to understand gene regulation. Other properties along the molecular pathways need to be annotated. For example, immediate effects of molecular interactions along a pathway (activation/repression) are necessary to gauge the downstream regulatory effects. Similarly, it is necessary to understand the direction of information flow along the pathways such as directionality of protein modifications or interactions, or whether potential pathways represented in the molecular interaction network are "active" or ever used in the cell, e.g., due to protein localization. Such properties are not evident in the data pertaining directly to isolated molecular interactions. Instead, they typically have to be inferred indirectly, for example, via the effects of gene deletions. We proceed to annotate molecular interactions with such properties, and infer them from the available data.

We propose a modeling framework for inferring annotated molecular interaction networks by integrating multiple sources of data. The different types of available data either directly constrain the types of interactions present (protein-DNA and protein-protein data) or tie together possible annotations of interactions along pathways (gene deletions). In our model, the existence of interactions and their annotations are represented by variables; each annotated interaction network is represented by a particular setting of all the variables. The

available data biases the setting of the variables, therefore giving rise to the most likely hypothesis about the annotated interactions networks. The output of the physical network modeling approach is one (or multiple) annotated physical network, including the presence of specific interactions, causal directions and immediate effects of such interactions, and active pathways.

Our modeling approach is inspired by early works of Ideker et al. (Ideker et al. 2001, Ideker et al. 2002). They demonstrated that genes along the same biological pathways tend to be co-expressed, and developed an algorithm which identified active subnetworks of physical interactions that exhibit significant expression changes over a subset of conditions. Our contribution follows a number of other computational approaches proposed towards uncovering regulatory models. An early work of gene expression analysis with Bayesian networks by Hartemink et al incorporated protein-DNA binding data as a prior to the Bayesian network (Hartemink et al. 2002). A recent work of high-throughput chromatin IP experiments by Lee et al. also identified "modules" of transcription factors that co-regulate gene expression from both protein-DNA binding and gene expression data (Lee et al. 2002). A more comprehensive approach of data integration based on probabilistic relation models was proposed by Segal et al. (Segal et al. 2002). We deviate from these approaches by explicitly focusing on the problem of inferring annotated molecular interaction networks and by relying on pathways of interactions as the mechanism of regulatory control. Our modeling approach also incorporates the uncertainty about the annotated networks. This paper is an extension of our work of physical network models published in RECOMB 2003 (Yeang et al. 2003).

The organization of this paper is as follows. Section 2 introduces the concept, elements, and the assumptions about physical network models. Section 3 describes how the available data sources are associated with the variables in the model. We also provide a toy example to illustrate the formal approach. Section 4 provides the inference algorithm, with technical details left in the appendix. In section 5 we demonstrate approach first in the context of a small but well-understood subsystem in yeast (mating response pathway), followed by a genome-wide analysis. In Section 6 we discuss immediate extensions of the approach.

2 Elements of physical network models

The physical network models are based on a skeleton graph constituting the set of possible molecular interactions to be considered as well as the properties associated with the edges of this graph. The edge properties are encoded as variables; the variables we consider here include presence/absence of interactions, causal direction, and immediate activating/repressing effects of interactions. A *configuration* is a specific setting of the variables and corresponds to a particular realization of the annotated molecular interaction graph. We can bias the setting of the variables, the configuration, based on the available data sources. In this modeling approach the goal is to find the setting of the variables most consistent with the available data.

We start by discussing the key aspects of the model and our simplifying assumptions in a bit more detail.

2.1 Skeleton graph

A realization of the physical network model is an annotated graph, where the nodes are associated with genes (or their protein products) and edges correspond to pairwise molecular interactions. We consider here two types of edges: protein-DNA and protein-protein interactions. The basic network which contain possible molecular interactions is called a *skeleton graph*. Formally, the skeleton graph $G = (V, \vec{E}_G \cup \bar{E}_G)$ is defined as a directed (possibly cyclic) graph. V is the set of vertices corresponding to genes or their protein products, $\vec{E}_G$ is the set of edges corresponding to possible protein-DNA interactions, and $\bar{E}_G$ is the set of edges denoting possible protein-protein interactions. The directionality of an edge denotes the causal direction along a regulatory process. The direction of a protein-DNA edge is determined a priori. In contrast, the direction of a protein-protein edge is initially undetermined as we do not necessarily know the regulatory processes where the interaction occurs. We view a protein-protein edge as an undirected edge whose direction will be inferred from data.

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 protein-protein and protein-DNA edges. In the former case, vertices play the role of protein products, whereas in the latter case we interpret

the protein to bind to the promoter region of the affected gene. Each pathway still has a clear interpretation in this collapsed notation so long as we use the implied meaning of the nodes along the pathway.

Without any data all the interactions in the skeleton graph are possible. The graph G containing all possible physical interactions could in principle be a complete graph (where there are three edges connecting each pair of vertices). Allowing all possible interactions is computationally burdensome and unnecessary in larger systems. We restrict the set of possible interactions *a priori* by excluding physical interactions lacking confident support in the available data. We demonstrate in section 5 that the modeling results are robust against any confidence thresholds used to filter out unlikely protein-DNA interactions.

2.2 Core model attributes

The skeleton graph provides only a template of possible pairwise molecular bindings. In order to make predictions about regulatory effects, it is necessary to annotate the skeleton graph with various attributes. The attributes are selected based on their relevance to regulatory pathways (see assumptions below). The attributes will also have to stand to receive some support either directly or indirectly from the available data. Accordingly, we define the following core variables:

- $X_{\vec{E}_G} = \{x_{\vec{e}_i} : \vec{e}_i \in \vec{E}_G\}$, a collection of binary (0/1) variables denoting 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 indicating whether specific protein-protein interactions are present.

- $S_{\vec{E}_G} = \{s_{\vec{e}_i} : \vec{e}_i \in \vec{E}_G\}$ and $S_{\bar{E}_G} = \{s_{\bar{e}_i} : \bar{e}_i \in \bar{E}_G\}$ annotate pairwise interactions or edges with activation or repression. We interpret these as signs of the edges.

- $D_{\bar{E}_G} = \{d_{\bar{e}_i} : \bar{e}_i \in \bar{E}_G\}$, a collection of binary variables denoting the directions of protein-protein interactions. The direction of an edge specifies the causal direction of how the interaction is used in a

signal transduction cascade. For simplicity, we assume here that each edge in $\bar{E}_G$ has only one directional annotation, essentially reducing the protein-protein interactions to (inferred) directed edges.

We will introduce additional variables such as the selection of active pathways in the process of associating the core attributes with observed data.

2.3 Modeling assumptions and limitations

We have adopted the following assumptions to cope with the limited data.

1. We view cascades of molecular interactions as the primary causal mechanisms underlying gene regulation. Other mechanisms such as chromatin modification, alternative splicing, or signal transduction via small molecules are currently overlooked in the model.

2. We assume that the effect of gene deletions (for most genes) propagates along the molecular cascades represented in the model. We do not consider combinatorial interactions such as complex formation, parallel pathways, or competitive binding. Such issues will be addressed in later work.

3. We collapse protein-protein interactions together with protein modification (e.g., phosphorylation). Many phosphorylation reactions are transient and may have been missed in the available protein-protein data. Moreover, we consider only the case that each protein-protein interaction possesses a unique direction in the pathways that it participates. While this is clearly not always correct, the assumption serves to further constrain the model when the available data is limited.

4. All the data used in this paper are collected under normal growth conditions (rich media) in yeast. We assume that the regulatory circuitry is not radically changed following a knock-out, at least not in terms of how the effects of the deletion are propagated.

5. By limiting ourselves to the skeleton graph, we may be affected by false negatives, interactions without

any support in the available interaction data. This simplification can be problematic if an important link is missing from the physical data. We do not consider remedies to this problem.

These simplifications limit both the flexibility and interpretability of the resulting physical network models (annotated graphs). Some of the assumptions can be removed, for example, by including additional mechanisms such as chromatin remodeling. Others can be relaxed as more high-quality data becomes available. We view the current formulation as a limited instantiation of a more general physical network modeling framework.

3 Data association

The values of the variables specifying the physical network model can be inferred from the available data. To achieve this we must tie the setting of the variables to specific measurements. Some of the variables such as presence or absence of protein-DNA interactions can be directly associated with individual measurements. Others such as the signs of the edges denoting either repressive or activating interactions need to be inferred indirectly based on responses to gene deletions. We begin with a brief overview of the data association problem, then describe the association for each type of data in detail.

3.1 Overview of data association

We categorize the available data sources into two types: direct *physical* and indirect *functional* data. A physical data directly pertains to molecular interactions such as protein-protein and protein-DNA interactions. On this level of resolution there is no inherent ambiguity in interpreting binding data of this type. Indirect functional measurements, on the other hand, pertain to series of underlying molecular interactions and effects. Inherent ambiguity exists in associating the measured effects with the properties of the molecular cascades that the effects are assumed propagate along. For example, a pathway may involve both repressive and activating components but only the aggregate effect is measured. Moreover, there may be multiple pathways underlying

the transcriptional changes in response to a knock-out.

It is fairly straightforward to incorporate direct physical measurements into the model. The variable in question is tied to a noisy measurement via an observation (sensor) model. The observation or sensor model is needed to properly bias the values based on noisy measurements. In our context, we express this bias in terms of *potential functions*. For example, for a binary variable x indicating the presence/absence of a protein-DNA edge, the measurements pertaining to this variable are incorporated into the model according to

$$\phi(x) = \left[\frac{P(\text{data}|x=1)}{P(\text{data}|x=0)}\right]^x . \tag{1}$$

The likelihood ratio is defined in detail later in the paper. The value of x more consistent with the observed data yields a higher value of the potential function. Potential functions as likelihood ratios are combined by multiplication across independent observations.

The association of indirect measurements such as knock-out effects with the attributes in the model involves two steps. First, for each gene, we introduce a new unobserved variable indicating the actual knock-out effect (up/down regulated, unaffected). This variable has direct evidence in terms of the measured expression changes and thus can be linked to the data as discussed above. Second, we tie together the actual knock-out effect with candidate pathways mediating the response from the deleted gene to the gene in question. Any *active* pathway must be consistent with the actual (ideal) knock-out effect. In other words, the protein-protein and protein-DNA edges must exist and be directed along the candidate pathway, and the aggregate effect of activation/repression must match the end effect. A detailed mapping from observations to potential functions will be provided after the toy example.

3.1.1 A toy example of data association

We illustrate here the data association procedure with a toy example. The skeleton graph corresponding to this example is shown in Figure 1. For simplicity we assume that all the edges are protein-DNA interactions and so their directions are specified a priori. The only variables in the physical network model encode the

presence/absence of each edge, variables $x_1, \cdots, x_5$, and the edge signs, variables $s_1, \cdots, s_5$.

We begin by building the potential functions for biasing the presence of edges. From the error model used in protein-DNA measurements we obtain the likelihood ratio for the potential function (see later sections for a detailed procedure). We assume that the likelihood ratio is the same for all edges, and has the value 0.9, indicating a slight bias against an edge. The potential function is therefore $\phi(x_i) = (0.9)^{x_i}$.

Suppose the only data available is from g_1 deletion experiment (denoted as $g_1\Delta$), and only g_4 is significantly down-regulated in response to the knock-out. For simplicity, we assume further that the knock-out effect $(g_1, g_4, -)$ (g_1 deleted, g_4 down-regulated) is very clear and so we can take this to be the actual knock-out effect.

There are two explanations for this effect given by the two alternative paths from g_1 to g_4. For a path to explain the knock-out effect, all the edges along the path must be present and the aggregate sign along the path must be the negative of the knock-out effect (because the deletion itself is negative). These constraints can be translated into the following potential function:

$$\psi_1(x_1, \ldots, x_4, s_1, \ldots, s_4) = \begin{cases} 1 & \text{if } (x_1 = x_2 = 1, s_1 \cdot s_2 = +1) \vee (x_3 = x_4 = 1, s_3 \cdot s_4 = +1). \\ 0.01 & \text{otherwise.} \end{cases} \tag{2}$$

where $\vee$ denotes logical OR. The small value 0.01 allows for the possibility that the knock-out effect is due to causes other than those expressible in the model.

The potential functions $\phi(x_i)'s$ and $\psi_1(\cdot)$ now define a joint distribution over the settings of the variables $X = \{x_1, \ldots, x_5\}$ and $S = \{s_1, \ldots, s_5\}$:

$$P(X, S) \propto \left[\prod_{i=1}^{5} \phi_i(x_i)\right] \cdot \psi_1(x_1, \ldots, x_4, s_1, \ldots, s_4). \tag{3}$$

The most likely configurations are given by

$$\begin{aligned}(x_1,\ldots,x_5,s_1,\ldots,s_5) &= (1,1,0,0,0,+1,+1,*,*,*)\\ &\quad (1,1,0,0,0,-1,-1,*,*,*)\\ &\quad (0,0,1,1,0,*,*,+1,+1,*)\\ &\quad (0,0,1,1,0,*,*,-1,-1,*)\end{aligned} \tag{4}$$

where $*$ indicates that either value is permitted. The configurations represent the fact that either path (e_1, e_2) or (e_3, e_4) must exist with consistent aggregate signs. The configurations corresponding to the possibility that both paths explain the knock-out effect (i.e., $x_1 = x_2 = x_3 = x_4 = 1$) have lower probabilities because of the biases against including individual edges.

The more data are available, the more constraints are imposed on the possible configurations. For example, suppose we conduct an additional experiment, deleting gene g_3, and find that g_4 is down-regulated. This extra evidence will reduce the most likely configurations to

$$(x_1,\ldots,x_5,s_1,\ldots,s_5) = (0,0,1,1,0,*,*,+1,+1,*) \tag{5}$$

3.2 Data sources

We use three types of data to estimate the physical network models pertaining to *S. cerevisiae*: protein-DNA, protein-protein, and single gene knock-out expression data.

The protein-DNA interactions come from data generated from high throughput chromatin immunoprecipitation experiments (location analysis) (Lee et al. 2002). The dataset contains the binding profiles of 106 transcription factors under normal conditions. Rather than reporting a binary profile for each transcription factor (where it binds), the dataset contains a confidence value (p-value) for each potential binding event, evaluated based on a heuristic error model.

Protein-protein interaction data is available in several on-line databases, such as DIP[1] and BIND[2]. The

[1] http://dip.doe-mbi.ucla.edu/

[2] http://www.bind.ca/index.phtml?page=databases

data comes from either high-throughput experiments or small-scale assays. High-throughput experiments tend to have high error rates. For example, results of two independent yeast two-hybrid systems under the same condition may have little overlap (Ito et al. 2001, Uetz et al. 2000). Since the data is prone to errors, it is necessary to incorporate the information in a probabilistic fashion. We use the data from DIP as it provides false positive and false negative rates based on independent experiments (Deane et al. 2001). Accordingly, interactions reported in small-scale experiments or those that have been confirmed in multiple types of experiments accompany higher confidence values.

The expression data for single gene knock-outs comes from the Rosetta compendium dataset (Hughes et al. 2000). This dataset contains the expression profiles of 276 different gene deletion mutants and 24 experiments under chemical treatments. Few of the deletion mutants are double knock-outs. For the purposes of this paper we only use the single gene knock-out experiments.

We decompose each data source into sets of pairwise observations. For example, factor A binds to promoter B, protein A binds to protein B, deleting gene A either up/down regulates gene B or leaves it unaffected. Strictly speaking such pairwise observations are not necessarily independent (experimental variations may results in correlated outcomes). We nevertheless adopt the independent assumption but note that the assumption can be easily removed if such correlating events can be identified.

3.3 Potentials and protein-DNA data

The chromatin IP (location analysis) experiments provide a p-value for each pair of transcription factor and an intergenic region. This confidence value reflects the relative abundance of the promoter DNA segments enriched by chIP in comparison to unenriched segments. We show here how the confidence values can be transformed into reasonable potential functions.

Let $Y_{\vec{E}_G} = \{y_{\vec{e}_i} : \vec{e}_i \in \vec{E}_G\}$ denote the variables corresponding to actual measurements in location analysis (not p-values). The values of these variables are fixed by the data. The potential function associated with a

specific binding event $\phi_{\vec{e}_i}(x_{\vec{e}_i}; y_{\vec{e}_i})$ is defined based on the following likelihood ratio:

$$\phi_{\vec{e}_i}(x_{\vec{e}_i}; y_{\vec{e}_i}) \quad = \quad \left[\frac{P(y_{\vec{e}_i}|x_{\vec{e}_i}=1)}{P(y_{\vec{e}_i}|x_{\vec{e}_i}=0)}\right]^{x_{\vec{e}_i}}. \tag{6}$$

Note that the potential function is only a function of $x_{\vec{e}_i}$ (whether binding took place) since $y_{\vec{e}_i}$ is fixed by the available data. The difficulty here is that we do not have access to the likelihood ratio indicated above but only the p-values evaluated (currently) based on a heuristic error model. We proceed to recover the likelihood ratio from the p-value and the associated sample sizes.

We assume that the p-value comes from testing the null hypothesis H_0 ($x_{\vec{e}_i} = 0$, the binding does not occur) against the alternative hypothesis H_1 ($x_{\vec{e}_i} = 1$, the binding occurs). Under certain regularity conditions the asymptotic sampling distribution of the log likelihood ratio statistic is χ^2 with degrees of freedom equal to one, assuming the difference in the degrees of freedom in the two hypotheses is one. We assume that the extra degree of freedom in the alternative hypothesis comes from the unknown binding affinity. The value of the log-likelihood ratio statistic can be recovered simply by

$$\mathcal{L} = 2\log\frac{P(y_{\vec{e}_i}|H_1)}{P(y_{\vec{e}_i}|H_0)} = F^{-1}(1-p) \tag{7}$$

where p is the reported p-value and $F(.)$ is the cumulative χ^2 distribution with one degree of freedom. The probabilities $P(y_{\vec{e}_i}|H_0)$ and $P(y_{\vec{e}_i}|H_1)$ in the log-likelihood ratio statistic are, however, fitted to the data. To recover a likelihood ratio that is appropriate for biasing the binding event one way or the other, we interpret $P(y_{\vec{e}_i}|x_{\vec{e}_i})$ as a Bayesian marginal likelihood that has a simple asymptotic approximation (Schwarz 1978): for example

$$P(y_{\vec{e}_i}|x_{\vec{e}_i} = 1) \approx P(y_{\vec{e}_i}|H_1)e^{-(d_1/2)\log(n)} \tag{8}$$

where $P(y_{\vec{e}_i}|H_1)$ involves a maximum likelihood fit; d_1 is the number of degrees of freedom in H_1, and n is the sample size from which the p-value was computed (the number of replicates; $n = 3$ in our case). Thus

$$\frac{P(y_{\vec{e}_i}|x_{\vec{e}_i} = 1)}{P(y_{\vec{e}_i}|x_{\vec{e}_i} = 0)} \approx \frac{P(y_{\vec{e}_i}|H_1)}{P(y_{\vec{e}_i}|H_0)}e^{(d_0-d_1)/2\log(n)} = e^{\mathcal{L}/2-\frac{1}{2}\log(n)}. \tag{9}$$

where $n = 3$. To get this result we have employed several asymptotic approximations that are not valid in the current setting. However, we believe that the resulting potential function provides a reasonable mapping from the available heuristic p-value to likelihood ratios, in the absence of verified training cases. When sufficient number of verified cases are available, we can directly estimate a parametric model $P(p_{\vec{e}_i}|x_{\vec{e}_i} = 1)$, where the p-values are treated as observations.

3.4 Potentials and protein-protein data

The available protein-protein interactions come from multiple sources and lack a coherent overall error model governing the experimental outcomes. Interactions reported from small-scale experiments or those supported by multiple experiments are generally more reliable than high-throughput assays. The DIP dataset can be subdivided into subsets of small-scale experimental outcomes and those verified in multiple experiments. We use empirical results of Deane et al. (Deane et al. 2001) to estimate the error rates for each protein pair.

Deane et al. performed two independent tests, ERP and PVM, on DIP and its subsets to gauge the level and type of errors introduced. EPR examines the distribution of the Euclidean distances between expression profiles of protein pairs, and uses this information to estimate the false positive rate of a set of protein pairs. PVM, on the other hand, verifies whether paralogs of a protein pair also bind and give corresponding confidence measures. Each pair (g_1, g_2) of proteins has four binary labels: whether it appears in the DIP dataset (f_1), whether it is reported from multiple sources (f_2), whether it is validated in the PVM test (f_3), and whether it appears in small-scale experiments (f_4). Each labeling is treated as an independent piece of evidence concerning whether g_1 and g_2 interact, represented here with a binary 0/1 variable b. EPR analysis provides $P(b = 1|f_1 = 1) = 0.5$, $P(b = 1|f_2 = 1) = 0.85$, whereas PVM gives $P(f_3 = 1|b = 0) = 0.05$, $P(f_3 = 1|b = 1) = 0.5$. The protein pairs with labeling $f_4 = 1$ (i.e., reported in small-scale experiments) were taken as correct interactions so that $P(b = 1|f_4 = 1) = 1$. We will relax this slight by using $P(b = 1|f_4 = 1) = 1 - \epsilon$ for small $\epsilon > 0$.

Deane et al's analysis did not report $P(b = 1|f_1 = 0)$, $P(b = 1|f_2 = 0)$, nor $P(b = 1|f_4 = 0)$. We assume

here that the absence of a protein pair from a given subset provides no additional information about the interaction. We therefore omit any evidence from f_1, f_2 or f_4 when their values are 0. More precisely, we assume that $P(b = 1|f_i = 0) = P(b = 0|f_i = 0)$ for $i = 1, 2, 4$.

We can now specify the potential function $\phi_{\bar{e}_i}(x_{\bar{e}_i}; y_{\bar{e}_i})$ for selecting protein-protein edges. Here $y_{\bar{e}_i}$, similarly the set $Y_{\bar{E}} = \{y_{\bar{e}_j} : \bar{e}_j \in \bar{E}_G\}$, represents the available information about protein-protein interactions. The potential function is based on a likelihood ratio as before, where the likelihood ratio is given by

$$\frac{P(y_{\bar{e}_i}|x_{\bar{e}_i} = 1)}{P(y_{\bar{e}_i}|x_{\bar{e}_i} = 0)} \equiv \frac{P(f_1, f_2, f_3, f_4|b = 1)}{P(f_1, f_2, f_3, f_4|b = 0)}. \tag{10}$$

Assuming the labels are conditionally independent given the known state of interaction or b,

$$P(f_1, f_2, f_3, f_4|b) = \prod_{i=1}^{4} P(f_i|b). \tag{11}$$

and by using Bayes law to transform the components into the form available from Deane et al's analysis, we finally get

$$\frac{P(f_1, f_2, f_3, f_4|b = 1)}{P(f_1, f_2, f_3, f_4|b = 0)} = \frac{P^3(b = 0)P(b = 1|f_1)P(b = 1|f_2)P(f_3|b = 1)P(b = 1|f_4)}{P^3(b = 1)P(b = 0|f_1)P(b = 0|f_2)P(f_3|b = 0)P(b = 0|f_4)}. \tag{12}$$

where $P(b), P(b|f_1), P(b|f_2), P(b|f_4)$ and $P(f_3|b)$ can be substituted with empirical values obtained from EPR and PVM tests.

3.5 Potentials and single gene knock-outs

We aim to explain only significant activation/repression responses of genes following a knockout. The significance is based on the available error model for the expression measurements. The rationale for this restriction is that there are many possible explanations for non-responses, some of which may not have been included in the model (e.g., currently we do not incorporate combinatorial effects). In contrast, a significant response is more likely to be consistent with (at least) one cascade.

Let $\mathcal{K}_p$ be the index set of significant knock-out effects in the available data, those with p-values lower than a given threshold (specific thresholds discussed below). We derive here potential functions only for observations

indexed by $\mathcal{K}_p$. The significant effects according to the error model are not assumed to be necessarily correct. Each pairwise effect is first tied to an unobserved variable that represents actual (as opposed to measured) knock-out effect. The actual knock-out effects are subsequently associated with variables along candidate pathways.

We introduce here three types of variables:

- $K = \{k_{ij} : (i,j) \in \mathcal{K}_p\}$ is a collection of the discrete variables of actual pairwise single knock-out effects taking values in $\{-1, 0, +1\}$. 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 is unaffected by the knock-out.

- $\Sigma = \{\sigma_{ija} : (i,j) \in \mathcal{K}_p, \pi_a \in \Pi\}$ is a collection of binary (0/1) path selection variables, where Π is the set of all valid paths in G (the notion of a valid path will be clarified later). σ_{ija} denotes whether path π_a is active causal explanation of knock-out effect k_{ij}.

- $O = \{o_{ij} : (i,j) \in \mathcal{K}_p\}$ denote the measurements of gene expression levels in knock-out experiments.

The actual knock-out effect k_{ij} is tied to the measurement o_{ij} via a potential function ϕ_{ij} analogously to the protein-DNA and protein-protein interaction data:

$$\phi_{ij}(k_{ij}; o_{ij}) \propto \left[\frac{P(o_{ij}|k_{ij})}{P(o_{ij}|k_{ij} = 0)}\right]. \tag{13}$$

where the likelihood ratios are derived from the available error model. We explain each such knock-out effect with a cascade of molecular interactions. For example, if two genes g_i and g_j are connected via a path π in the skeleton graph, the path is directed from g_i to g_j, and the aggregate sign along π agrees with $k_{ij} \in \{-1, 1\}$, then the path is said to explain k_{ij}. A valid path has to satisfy several additional constraints.

For π to qualify as an explanation for k_{ij}, it must satisfy the following conditions:

1. The end nodes of π are g_i and g_j.

2. The last edge in π is a protein-DNA interaction.

3. All the 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. If intermediate genes along π have been knocked out, they also exhibit a knock-out effect on g_j.

The first condition reiterates the assumption that a cascade of physical interactions explain regulatory effects. The second condition defines the last step of gene regulation to be transcriptional 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 combinatorial effects and missing data. A path which satisfies these conditions is a candidate explanation for k_{ij}. We say that k_{ij} is explained by the physical model if there exists at least one path which satisfies these conditions. These conditions can be modified to incorporate simple notions of coordinate regulation.

Conditions 1, 2, and 5 can be verified without knowing how the edges are annotated, merely assuming that they can exist (are present in the skeleton graph). Therefore, for each k_{ij} we can identify a set of connecting paths $\Pi_{ij} = \{\pi_1, \cdots, \pi_n\}$ which satisfy these conditions. Π_{ij} contains all candidate paths which could in principle explain k_{ij}. A candidate path, if selected, imposes three types of additional constraints on the variables: conditions 3 and 4 together with the fact that all the edges must exist along the path.

Let $\pi_a \in \Pi_{ij}$ be a candidate explanatory path of k_{ij}, $E_a = \{e_a \in \pi_a\} = \vec{E}_a \cup \bar{E}_a$ denote the protein-DNA and protein-protein edges along π_a, $X_a = \{x_e : e \in E_a\}$, $S_a = \{s_e : e \in E_a\}$ be the edge presence and signs along π_a, and $D_a = \{d_e : e \in \bar{E}_a\}$ the directions of protein-protein edges along π_a. Then π_a explains k_{ij} if the following conditions hold:

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

- $\forall e \in \bar{E}_a$, $d_e = \hat{d}_e$ ($\hat{d}_e$ is determined by our definition of path direction).

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

$$\psi_{ija}(X_a, S_a, D_a, k_{ij}) = \prod_{e \in E_a} I(x_e = 1) \cdot I(\prod_{e \in E_a} s_e = -k_{ij}) \cdot \prod_{e \in \bar{E}_a} I(d_e = \hat{d}_e) \tag{14}$$

where $I(.)$ is a 0/1 indicator function.

When there are multiple candidate paths connecting g_i and g_j, we require that the conditions along at least one of the paths suffice to explain k_{ij}. Encoding these OR-like constraints in a single potential function is cumbersome. Instead, we introduce auxiliary path selection variables and factorize the potential function into terms corresponding to single paths. Recall that σ_{ija} denotes the selection variable of path π_a, $\sigma_{ija} = 1$ if π_a is used to explain k_{ij}, and zero otherwise. Physically, σ_{ija} represents whether the pathway π_a plays a regulatory role in the context of the specific experiment. The potential function ψ_{ija} in equation 14 is augmented with variable σ_{ija}:

$$\psi_{ija}(X_a, S_a, D_a, k_{ij}, \sigma_{ija}) = \epsilon_1 + (1 - \epsilon_1) \cdot I(\sigma_{ija} = 1) \cdot \psi_{ija}(X_a, S_a, D_a, k_{ij}) \tag{15}$$

The potential function does not vanish even when the constraints are violated so as to allow other causes (those not included in the model) to explain the knock-out effect. Our experimental results are not sensitive to the specific value of ϵ_1.

We construct a potential function term ψ_{ij}^{OR} to specify the condition that at least one candidate path is selected to explain k_{ij} if k_{ij} is explained. Similar to other potential functions, ψ_{ij}^{OR} is a "soft" logical OR:

$$\psi_{ij}^{OR}(\sigma_{ij1}, \cdots, \sigma_{ij|\Pi_{ij}|}) = \epsilon + (1 - \epsilon)\big(1 - \prod_a I(\sigma_{ija} = 0)\big) \tag{16}$$

Combining equations 15 and 16, the potential function associated with each pairwise knock-out effect is as follows. Let $E_{ij} = \cup_{\pi_a \in \Pi_{ij}} E_a$, $X_{ij} = \cup_{\pi_a \in \Pi_{ij}} X_a$, $S_{ij} = \cup_{\pi_a \in \Pi_{ij}} S_a$, $D_{ij} = \cup_{\pi_a \in \Pi_{ij}} D_d$, and $\Sigma_{ij} = \{\sigma_{ija} : \pi_a \in \Pi_{ij}\}$, then

$$\psi_{ij}(X_{ij}, S_{ij}, D_{ij}, \Sigma_{ij}, k_{ij}) \quad = \quad \psi^{OR}(\sigma_{ij1}, \cdots, \sigma_{ij|\Pi_{ij}|}) \cdot \textstyle\prod_a \psi_{ija}(X_a, S_a, D_a, \sigma_{ija}, k_{ij}). \tag{17}$$

ψ^0_{ij} returns a relatively high value if there exists at least a path which can explain k_{ij} provided selected paths all satisfy the conditions of explanation. Moreover, the returned value is higher if there are more paths which explain the knock-out effect Note that this bias encourages parallel pathways as explanations (the effect is mediated by multiple alternative pathways).

4 Probability model and inference

We can now combine the potential functions into a joint distribution over the variables in the physical model so that the probability value reflects the degree of support that each possible physical model (annotated graph) has in the available data. Specifically:

$$P(X_{\vec{E}_G}, S_{\vec{E}_G}, D_{\bar{E}_G}, X_{\bar{E}_G}, S_{\bar{E}_G}, K, \Sigma \,|\, Y_{\vec{E}_G}, Y_{\bar{E}_G}, O_K) \propto$$
$$\prod_{\vec{e}_i \in \vec{E}_G} \phi_{\vec{e}_i}(x_{\vec{e}_i}; y_{\vec{e}_i}) \cdot \prod_{\bar{e}_j \in \bar{E}_G} \phi_{\bar{e}_j}(x_{\bar{e}_j}; y_{\bar{e}_j}) \cdot \prod_{(i,j) \in \mathcal{K}_p} \phi_{ij}(k_{ij}; o_{ij}) \psi_{ij}(X_{ij}, S_{ij}, \Sigma_{ij}, k_{ij}). \quad (18)$$

The potential functions are combined in a product form because we assume that individual measurements of interactions or knock-out effects are statistically independent. This may not be a realistic assumption; for example, readings of adjacent spots on microarrays may be correlated. However, it is a sensible simplification to start with since adequate models of dependencies in the measurements are not yet available.

The joint distribution in Eq. (18) is naturally viewed as a specific class of graphical models, namely *factor graphs* (see, e.g., Kschischang et al. 2001). A factor graph is simply a probability model, where the joint distribution is represented as a product of potential functions called factors. Factor graphs are visualized as undirected graphs, where the variables appear as circles and factors as squares. An edge is drawn between a variable and a factor whenever the factor depends on the variable. This representation is convenient for guiding inference calculations.

To find the most likely realization of the physical model, we find the most likely settings of the variables in the joint distribution (MAP configuration). Since the joint distribution involves a large number of inter-

dependent variables, we settle for finding an approximate MAP configuration with distributed message passing algorithms such as the *max-product* (see, e.g., Kschischang et al. 2001, Weiss et al. 2001, Wainwright et al. 2002). These algorithms find a locally consistent setting of the variables, where the notion of locality is tied to the factor graph, but are not necessarily guaranteed to find the exact MAP configuration. Some guarantees already exist (Weiss et al. 2001, Wainwright et al. 2002), however.

We note that our use of approximate methods for finding the most likely physical model here is analogous to the use of greedy search algorithms in clustering, estimation of Bayesian networks, or relational models from the available data. Greedy search algorithms offer little or no guarantees of finding the best solution but are nevertheless useful in practice.

We provide here a brief overview of the max-product algorithm in this context; more details can be found in the references provided or in the supplemental material (Yeang et al. 2003).

In order to find a MAP configuration it is sufficient to evaluate so called *max-marginals* for each variable:

$$P_x^{\max}(x) = \max_{\mathbf{U}\backslash\{x\}} P(x, \mathbf{U}\backslash\{x\}). \tag{19}$$

where U stands for all the variables in the model. If the MAP configuration is unique, then the max-marginals have unique maximizing arguments. Otherwise there are multiple MAP configurations, and they may need to be uncovered recursively.

The max-product algorithm is a local propagation algorithm that evaluates approximate max-marginals for each variable. We can understand the algorithm as follows. Each factor or potential function in the joint distribution ties together the values of the variables involved. The max-product algorithm determines, for each potential, what information needs to be shared between the variables in order to evaluate their max-marginals, assuming that the associated variables are independent of each other in the absence of the potential in question. All the operations are local in this sense. The algorithm is guaranteed to find the correct max-marginals whenever the factor graph has a tree-like structure; stronger guarantees are available[20], particularly for modified versions of the algorithm[19].

In our setting, some of the potential functions (those associated with knock-out effects) involve many binary variables. The potential functions have special structure, however, that can be exploited so that the application of the algorithm is efficient. We use the resulting approximate max-marginals as proxies for the correct max-marginals.

The resulting approximate max-marginals may sometimes fail to identify a unique MAP configuration. In other words, some of the max-marginals may not have unique maximizing arguments. In this case, we apply the following recursive procedure. At each iteration, we run the max-product algorithm conditioned on the fixed values obtained from previous steps. Variables are fixed if their inferred max-marginal probabilities yield a unique max argument. If there are still undetermined variables, then we choose one such variable, fix it to one of the degenerate values, and continue the iteration. The algorithm stops when all variables have been set.

In a relatively well-constrained case (for example, the pheromone response network in section 5.1), the recursive search is capable of identifying all (approximate) MAP configurations. At each iteration, the algorithm branches out with all degenerate values of the selected variable. The resulting MAP configurations are represented as a decision tree, where internal branches represent values that were fixed, and each leaf corresponds to a full configuration.

The problem of enumerating all MAP configurations becomes intractable as the number of variables greatly exceeds the number of available constraints. This is the case in the genome-wide analysis. Instead of generating all MAP configurations, we revise the recursive algorithm to decompose the physical network into subnetworks such that the variables in different subnetworks do not interfere given the constraints from the available data. The MAP configurations can be therefore expressed as a product of subconfigurations in subnetworks. The details of the subnetwork decomposition algorithm are described in the supplemental material (Yeang et al. 2003).

5 Empirical results

The data we used to evaluate the physical models framework consisted of a) high throughput chIP data of protein-DNA interactions in *S. cerevisiae* (Lee et al. 2002), b) protein-protein interactions from the YPD and DIP databases, and c) mRNA expression of knock-out experiments (Hughes et al. 2000). We selected protein-DNA pairs with p-value cutoff of 0.001 and significant pairwise knock-out effects with p-value cutoff 0.02. These threshold values naturally change the input to the model but, as demonstrated below in a cross-validation setting, the quality of the resulting predictions remains largely intact over a range of threshold values.

Our empirical results focus on the following points. First, we demonstrate that the MAP configuration(s) provide explanations for a number of knock-out effects. Second, we show in a cross-validation setting that the resulting physical models are able to predict out of sample knock-out responses. Third, we demonstrate that the predictive power is robust to various parameter choices as well as the addition of random edges to the skeleton graph. Fourth, we show that the models help identify knock-out effects that cannot be explained with the currently available molecular interactions. After the quantitative tests, we extract subnetworks from the inferred models. We first uncover subnetworks whose attributes are uniquely determined by the data and then decompose the remaining model into products of subnetworks. We compare the subnetworks to yeast biology.

We begin with a small scale analysis involving the yeast mating response, followed by a genome-wide analysis.

5.1 Mating response pathways

We selected 46 genes involved in pheromone response pathways, those discussed in Hartemink et al. 2002, and an additional set of 14 genes bound by STE12 and which have significant changes in STE12Δ (STE12 knock-out). We extracted 37 protein-DNA interactions from the location analysis data and 30 protein-protein

interactions from the yeast database (YPD)[3]. 13 genes on the list are deleted in the compendium dataset: STE2, STE4, STE18, FUS3, STE7, STE11, STE5, STE12, KSS1, STE20, SST2, SIN3, TUP1. There are 149 pairwise knock-out interactions generated from these experiments. The list of physical and knock-out interactions is provided in the supplementary website (Yeang et al. 2003).

The yeast mating response network has been analyzed in a great detail (see, e.g., Schrick et al. 1997). Here we give a very brief overview of the major pathway. Pheromone molecules are bound by receptors (STE2, or STE3) at the cellular membrane. STE2 receptor is linked to the G-protein complex (GPA1, STE4, STE18), which activates the signal transduction cascade of MAP kinases (STE20 $\rightarrow$ STE11 $\rightarrow$ STE7 $\rightarrow$ FUS3/KSS1 $\rightarrow$ STE12). STE12 subsequently turns on mating-specific genes[4]. Some mating-specific genes are also activated by the transcription factor MCM1. In addition, KSS1 enables STE12 to activate genes involved in filamentous growth.

Potential functions for explaining knock-out effects and the joint probability function over these variables were constructed as described in section 3. Here we restrict the path length ≤ 5.

We measure the flexibility of the physical network model by verifying how many of the knock-out effects can be explained. An explanation means that for each path which connects the knock-out pair and is selected according to the MAP configuration ($\sigma_a = 1$), the variables along the path satisfy conditions 1-6 in section 3.6. Since potential functions encode soft constraints about knock-out effects, it is possible that some of these constraints are not satisfied in a MAP configuration. Among the 149 knock-out pairs in 13 experiments, there are 106 pairs in 9 experiments which are connected via candidate paths of length ≤ 5 in the skeleton network. All those 106 pairs are successfully explained by all the resulting 4 equivalent MAP configurations. Moreover, only 21 knock-out interactions are trivially explained by direct protein-DNA binding by STE12. All other effects have to be explained by paths longer than 1. This illustrates the advantage of incorporating pathway information over explanations based on protein-DNA interactions alone.

[3]https://www.incyte.com/tools/proteome/databases.jsp

[4]http://mips.gsf.de/proj/yeast/CYGD/db/index.html

We then verified the predictive accuracy of the model in a leave-out cross validation sense. We randomly held out a fixed number of knock-out pairs when constructing the potential functions and running the inference algorithm. For each leave-out pair, we then examined whether all the connecting valid paths in all MAP models yielded the sign change consistent with its knock-out effect. Table 1 shows the results of leave-n-out cross validation, where n equals to 1, 5 and 20. For the number of leave-out pairs > 1, the error rate was computed by dividing the total number of mistakenly predicted held-out pairs by the total number of held-out pairs in all trials. For instance, in the leave-20-out experiment over 200 trials, there are $20 \times 200 = 4000$ possible held-out pairs (many of them are repeated) and 236 pairs are inaccurately predicted over the 200 trials. The reported error rate is therefore $236/4000 = 5.9\%$. The knock-out pairs considered in the cross validation experiments are those connected via valid paths in the skeleton graph. The low error rate in each experiment indicates that the algorithm can predict the knock-out effects with high degree of accuracy. This is to be expected since there are sufficient number of knock-out experiments perturbing a small network, and the information about each knock-out interaction is distributed among multiple interactions along pathways.

Although the cross-validation outcome is encouraging, the results may have been artifacts of a particular setting of the model parameters/thresholds. We provide here a sensitivity analysis to exclude this possibility.

We considered the following adjustable parameters: the maximum length of candidate paths, thresholds on p-values of selecting candidate protein-DNA and knock-out pairs, and the error probabilities used as soft constraints in the potential functions (ϵ_1 in equation 15, results not shown for brevity). In addition, we also performed the cross validation tests by adding random edges (both protein-DNA and protein-protein interactions) to the skeleton graph. We performed ten random trials for each fixed number of added edges and reported the average test accuracy over the ten trials. Figure 2 shows the leave-one-out test accuracy rates across a wide range of these parameters. The test accuracy here is normalized by the number of knock-out effects that the inferred model could in principle explain, i.e., the number of knock-out pairs connected via valid paths (this is a function of the number of edges they contain). The default values of these parameters

are: location p-value threshold = 10^{-3}, knock-out p-value threshold = 0.02, $\epsilon_1 = 0.3$, and the maximum path length = 5. Robustness tests were carried out by varying one parameter and fixing all others at their default values.

It is clear that test errors are robust against the location and knock-out p-value thresholds as well as the potential function values. In contrast, test errors are sensitive to the path length upper bound. The models constructed from short paths (length < 3) cannot accurately predict knock-out effects since short paths receive little support from other knock-out effects. The test errors are also robust against the addition of random edges: it increases only from 2% to 7% even when the number of random edges added to the skeleton graph is approximately equal to the original size of the skeleton graph.

Quantitative tests results suggest the inferred models can both fit existing data and predict new knock-out effects according to related constraints. To further validate the results, we directly compared the inferred models with the current knowledge about the yeast mating pathway. We first applied the max-product algorithm once and identified the variables whose values were uniquely determined by the max-marginal probabilities. Figure 3 shows the physical subnetwork annotated with these attributes. It is visualized using *Cytoscape*[5], a freeware developed by Ideker et al. Solid lines correspond to protein-DNA and dash lines represent protein-protein interactions. The directions of protein-DNA arrows are given in the data, while the arrows of protein-protein edges are inferred from the model. Edge signs are color-coded with black (positive) and light grey (negative).

The inferred results have substantial overlap with the biology of yeast mating response. First, all protein-DNA edges emanating from STE12 have positive signs. This result confirms the activating role of STE12. Second, the inferred directions of protein-protein interactions (STE18,STE4), (STE4,STE5), (STE5,STE11), (STE7,FUS3), (STE7,KSS1), (FUS3,STE12) and (KSS1,STE12) are all consistent with the directions of the MAP kinase signal transduction pathway. These directions are reported because they yield consistent explanations for the depressions of the STE12-regulated genes in deletion experiments STE5Δ, STE11Δ and

[5]http://www.cytoscape.org.

STE7Δ.

However, some inferred attribute values contradict with current biology. The inferred direction of (STE7,STE11) is the opposite of the signal transduction direction. This is because STE11 also binds to KSS1 and FUS3, and the knock-out effects in STE7Δ experiment are explained by paths STE7 → STE11 → KSS1/FUS3 → STE12 → downstream genes. Finally, the edges (FUS3,STE12), (STE7,FUS3), (STE11,FUS3) have negative signs because a few STE12-controlled genes are up-regulated in FUS3Δ experiment. Since FUS3 is a MAP kinase which activates STE12, these responses are not caused by the phosphorylation cascade. Without additional information, assigning negative signs to these edges is a consistent explanation. The contradiction between suggests potentially new pathways/mechanisms. For example, the up-regulated genes in FUS3Δ experiment may connect to FUS3 via a negative feedback mechanism. Moreover, although FUS3 and KSS1 (indirectly) activates genes of different functions (mating response and filamentous growth respectively), both kinases also function in complementary fashion. Single deletions of FUS3 Δ and KSS1 Δ yield only few changes in STE12-controlled genes, while the double deletion FUS3 Δ KSS1 Δ generates significant impacts. Thus the pathways containing FUS3 or KSS1 in fact violate condition 6 in section 3.5 (the deletion of all intermediate genes should exhibit significant effects).

We then applied the max-product algorithm recursively to enumerate all MAP configurations. Since this small network is constrained by 13 knock-out experiments, there are only 4 MAP configurations. All degeneracies occur at edge signs, and these configurations can be expressed as products of subconfigurations of two subnetworks. Figure 4 shows these decomposed subconfigurations. Subnetwork 1 reflects the ambiguity of the sign of protein-protein interaction (STE12,MCM1). Some genes which are down-regulated in STE12Δ experiment are bound jointly by STE12 and MCM1 (with protein-DNA interactions). Thus their knock-out effects in STE12Δ can be explained either by the direct protein-DNA bindings of STE12 or the paths mediated by MCM1. Since MCM1Δ experiment is unavailable (in fact deleting MCM1 is lethal in yeast), we speculate that both paths STE12 → downstream genes and STE12 → MCM1 → downstream genes are active pathways.

The product of signs of (STE12,MCM1) and (MCM1,downstream genes) is fixed while individual signs are not. Subnetwork 2 reflects the ambiguity of the sign of protein-protein interaction (GPA1,STE11). This edge is essential for explaining the up-regulation of genes in SST2Δ and STE2Δ experiments. However, the same group of genes are also down-regulated in STE11Δ experiment, and these down regulations can be explained by paths from STE11 to the affected genes. Therefore, the aggregate signs along the paths from STE11 to these genes are positive, and the aggregate signs along the paths SST2 $\rightarrow$ GPA1 $\rightarrow$ STE11 and STE2 $\rightarrow$ GPA1 $\rightarrow$ STE11 are negative.

5.2 Genome-wide analysis

The location analysis data covers the binding profiles of 106 transcription factors on 6135 genes (Lee et al. 2002). By choosing p-value threshold 0.001, we extracted 5485 protein-DNA pairs from the data. There are 14876 protein-protein interactions of yeast proteins reported in the DIP database. The compendium dataset contains 271 single gene deletion experiments (Hughes et al. 2000). By choosing the p-value threshold to be 0.02, we extracted 23766 pairwise knock-out effects.

Although the size of the entire physical network is 300 times larger compared to the pheromone response network, the fraction of knock-out effects which can possibly be explained by this network is smaller. By restricting the length ≤ 3, only about one twentieth (1091 out of 23766) of all knock-out pairs are connected via valid paths. This small fraction may be due to limitations of the model (a longer path length or additional mechanisms may be required) or limitations of the available data (protein-DNA and protein-protein interactions are not complete). In this paper, we restrict path length to ≤ 3 for computational efficiency and focus only on the knock-out pairs which are connected in the skeleton graph via short paths. The list of physical and knock-out interactions is provided in the supplementary website (Yeang et al. 2003).

We first report results from quantitative analysis. We applied the max-product algorithm once to infer the variables whose values are uniquely determined by the data. The resulting model induces a much smaller

network: it contains 128 genes and 142 physical interactions, and explains 194 knock-out interactions. We then recursively instantiated one MAP configuration. This configuration explains 986 knock-out interactions. Some of the additional 792 knock-out pairs explained by the non-unique part of the network may be due to the artifacts of the recursive instantiation. However, by randomize the instantiation of 100 MAP configurations, we found that most of those knock-out pairs (984 out of 1091) are explained by all the MAP configurations.

There are 105 knock-out pairs connected by valid paths but are not explained by any specific MAP configuration. This is because the edge signs involved yield contradictions. For instance, in Figure 5 the sign of edge (GLN3,GCN4) is known to be negative from the invariant (uniquely determined) part of the network. Since GLN3 $\rightarrow$ GCN4 $\rightarrow$ MET4 $\rightarrow$ MET14 is the only pathway connecting GLN3 and GCN4 to MET14, the knock-out effects (GLN3,MET14,-) and (GCN4,MET14,-) cannot be simultaneously explained. We note that finding contradictions of this type cannot be done simply on the basis of the available molecular interaction data.

We summarize the types of protein-protein interactions involved in explaining knock-out effects in Table 2. There are roughly similar to the numbers of protein-protein interactions from small-scale and high-throughput experiments. However, since there are far more interactions reported in high-throughput data, the usage is strongly biased toward interactions from small-scale experiments. The p-value of the corresponding hypergeometric test is $< 4.8 \times 10^{-17}$.

Figure 6 illustrates the subnetworks from uniquely determined variables. The graph semantics is identical to Figure 3, and red shaded nodes denote genes for which we have knock-out data. The graph is composed of the following subnetworks:

- The subnetwork of the pheromone response pathway is partially retrieved by the algorithm. Because the maximum path length is shorter, the part involving STE20, STE8 and STE4 is not covered.
- GCN4 activates a number of genes involved in amino acid synthesis (Natarajan et al. 1999). The evidence comes from both protein-DNA bindings and the knock-out experiment GCN4Δ. In addition, GLN3 binds

to GCN4 promoter and deleting GLN3 up-regulates some genes controlled by GCN4. Thus the edge (GLN3,GCN4) has a negative sign. ARG80 also regulates a few genes involved in arginine synthesis.

- SWI5 activates a number of genes involved in cell cycle control.

- SWI4, SWI6, MBP1 and CLB2 are all involved in cell cycle control, particularly genes expressed at G1/S phases (Simon et al. 2001). Some genes are affected in both SWI4Δ and SWI6Δ experiments, and some are affected in both SWI4Δ and MBP1Δ experiments. Contrasting previous knowledge that these factors are activators, protein-DNA edges emanating from them have both positive and negative signs.

- YAP1 regulates genes involved in oxidative stress response (Cohen et al. 2002). HIR2 regulates histones HHT1 and HHF2. MAC1 regulates FRE1 and FTR1 involved in iron utilization (Andrews et al. 1999).

We could no longer enumerate all MAP configurations since the number of variables far exceeds the number of constraints. Instead, we applied the recursive algorithm described in section 4 (see also Yeang et al. 2003) to decompose the model into submodels. Within each submodel, we can change the values of undetermined variables independently with respect to other submodels. We identified 22 submodels which contain more than 5 undetermined variables. Figure 7 shows three of these submodels that are biologically significant. We illustrate the knock-out effects explained by these subnetworks and summarize the biology of the subnetworks whenever relevant information is available from YPD and MIPS databases. A complete list of submodels is provided on the supplementary website (Yeang et al. 2003).

- Network 1 contains protein-DNA interactions (SWI4,SOK2), (SOK2,YAP6), (SOK2,CUP9), (SOK2,MSN4), (SOK2,HAP4), and protein-DNA interactions emanating from hubs SOK2, CUP9, YAP6, MSN4 and HAP4. This structure explains many responses in SWI4Δ experiment. SOK2 is a hub which relays the influence of SWI4 on the downstream genes. We have not found functional relations between those hub genes in the YPD and MIPS databases.

- Network 2 contains a protein-protein interaction pathway TUP6 → SSN6 → NRG1 and protein-DNA interactions from NRG1 to a few genes. The protein-protein bindings of (TUP1,SSN6) and (SSN6,NRG1) are essential for explaining some up-regulated genes in TUP1Δ and SSN6Δ experiments. NRG1 is a transcription repressor involved with glucose metabolism. Its repression function is activated by recruiting the SSN6-TUP1 complex (Park et al. 1999).

- Network 3 contains protein-protein interaction (ARG80,ARG81) and protein-DNA interactions from ARG81 to arginine biosynthesis genes. It is responsible for explaining some knock-out effects in ΔARG80 experiment. ARG80 and ARG81 are known to form a complex in regulating arginine biosynthesis genes (Rijcke et al. 1992). The inferred model confirms this result.

6 Conclusions

We have proposed a new modeling framework for reconstructing annotated molecular networks – physical network models – from multiple sources of data. Physical models are based on explicit descriptions of candidate properties and causal mechanisms to be inferred from the available data. In this paper we developed a specific realization of the general methodology, where causal regulatory mechanisms were taken to be cascades of molecular interactions, paths in the associated graph.

The resulting annotated molecular interactions networks were shown to be highly predictive about knock-out effects in a cross-validation setting involving the yeast mating pathway. In a genome-wide analysis, a much smaller fraction of knock-out effects could be explained by short paths (of lengths ≤ 3); the results reflect primarily the lack of systematic knock-out data to constrain annotations of interaction networks. Inferred subnetworks from the genome-wide analysis were nevertheless largely consistent with the current understanding of transcriptional regulation in yeast.

There several immediate extensions of the basic approach including 1) computational experiment design

and 2) coordinate regulation. First, the genome-wide models are currently under-constrained due to the lack of available knock-out data. We can prioritize new gene deletion experiments according to their ability to distinguish between competing network models. Such calculations are supported by physical models since they also capture the uncertainty about possible network models. The experiment design approach described here is currently in use. Second, our approach in this paper does not model coordinate binding or activation of transcription factors; each transcription factor is considered in isolation. However, many transcription factors operate in concert in regulating associated genes (e.g., [10]). Coordinate regulation can be modeled within the physical modeling framework by articulating a set of candidate mechanisms.

Acknowledgement

The authors gratefully acknowledge discussions with our colleagues from MIT Whitehead Institute, Artificial Intelligence Laboratory and Laboratory for Computer Science: Owen Ozier, Richard Young, David Gifford and Tomas Lozano-Perez. Special thanks to the Young laboratory at Whitehead Institute for providing the location analysis data. Tommi Jaakkola acknowledges support from the Sloan foundation in the form of the Sloan Research Fellowship. The work was also partially funded by grants from DARPA and NIH.

References

[1] Andrews, N., Fleming, M., Gunshin, H. 1999. Iron transport across biologic membranes. *Nutrition Review*, 57(4):114–123.

[2] Cohen, B., Pilpel, Y., Mitra, R., Church, G. 2002. Discrimination between paralogs using microarray analysis: application to the yap1p and yap2p transcriptional networks. *Molecular Biology of the Cell*, 13(5):1608–1614.

[3] Deane, C., Salwinski, L., Xenarios, I., Eisenberg, D. 2002. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular Cell Proteomics*, 1(5):349–356.

[4] Hartemink A., Gifford, D., Jaakkola, T., Young, R. 2002. Combining location and expression data for principled discovery of genetic regulatory network models. In *Pacific Symposium of Biocomputing*, 437-449.

[5] Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., Armour, C., Bennett, H., Coffey, E., Dai, H., He, Y., Kidd, M., King, A., Meyer, M., Slade, D., Lum, P., Stepaniants, S., Shoemaker, D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S. 2000. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126.

[6] Ideker, T., Thorsson, V., Ranish, J., Chirstmas, R., Buhler, J., Eng, J., Bumgarner, R., Goodlett, D., Aebersold, R., Hood, L. 2001. Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science*, 292:929–934.

[7] Ideker, T., Ozier, O., Schwikowski, B., Siegel, A. 2002. Discovering regulatory and signalling circuits in molecular integration networks. *Bioinformatics*, 18 Suppl 1:S233-240.

[8] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, 98(8):4569–4574.

[9] Kschischang, F., Frey, B., Loeliger, H. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519.

[10] Lee, T., Rinaldi, N., Robert, F. Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thimpson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, B., Ren, B., Wyrick, J., Tagne, J., Volkert, T., Franekel, E., Gifford, D., Young, R. 2002. A transcriptional regulatory network map for *Saccharomyces Cerevisiae*. *Science*, 298:799–804.

[11] Natarajan, K., Meyer, M., Jakson, B., Slade, D., Roberts, C., Hinnebusch, A., Marton, M. 2001. Transcriptional profiling shows that gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Molecular Biology of the Cell*, 21(13):4347–4368.

[12] Park S., Koh, S., Chun, J., Hwang, H., Kang, H. 1999. Nrg1 is a transcriptional repressor for glucose repression of sta1 gene expression in *Saccharomyces Cerevisiae*. *Molecular Cell Biology*, 19(3):2044–2050.

[13] Rijcke, D., Seneca, S., Punyammalee, B, Glansdorff, N, Crabeel, M. 1992. Characterization of the DNA target site for the yeast ARGR regulatory complex, a sequence able to mediate repression or induction by arginine. *Molecular Cell Biology*, 12(1):68–81.

[14] Schrick, K., Garvik, B., Hartwell, L. 1997. Mating in *Saccharomyces Cerevisiae*: the role of the pheromone signal transduction pathway in the chemotropic response to pheromone. *Genetics*, 147(1):19–32.

[15] Schwarz, G. 1978. Estimating the dimension of a model. *Annals of statistics*, 6(2):461–464.

[16] Segal, E., Barash, Y., Simon, I., Friedman, N., Koller, D. 2002. From promoter sequence to expression: a probabilistic framework. In *The Sixth Annual International Conference on Research in Computational Molecular Biology*.

[17] Simon, I., Barnett, J., Hannett, N., Harbison, C., Rinaldi, N., Volkert, T., Wyrick, J., Zeitlinger, J., Gifford, D., Jaakkola, T., Young, R. 2001. Serial regulation of transcriptional regulators in the yeast cell. cycle. *Cell*, 106:697–708.

[18] Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces Cerevisiae*. *Nature*, 403:623–627.

[19] Wainwright, M., Jaakkola, T., Willsky, A. 2002. Exact map estimates by (hyper)tree agreement. In *Advances in Neural Information Processing Systems*.

[20] Weiss, Y., Freeman, W. 2001. On the optimality solutions of the max-product belief propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory*, 47:736–744.

[21] Yeang, C., Jaakkola, T. 2003. Physical network models and multi-source data integration. In *The Seventh Annual International Conference on Research in Computational Molecular Biology*.

[22] Yeang, C., Ideker, T., Jaakkola, T. 2003. Supplementary website. `http://www.ai.mit.edu/people/tommi/suppl/jcb03/`.

Notation summary

We summarize the notations used in section 3 in Table 3 for readers' benefit.

Figure 1: A simple example of a physical interaction network.

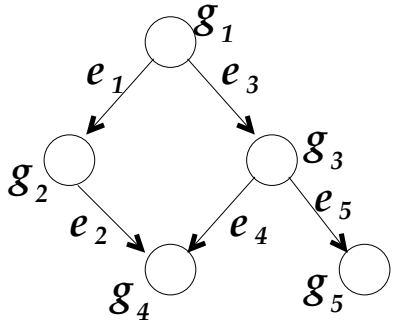

Figure 2: Sensitivity analysis on test accuracy

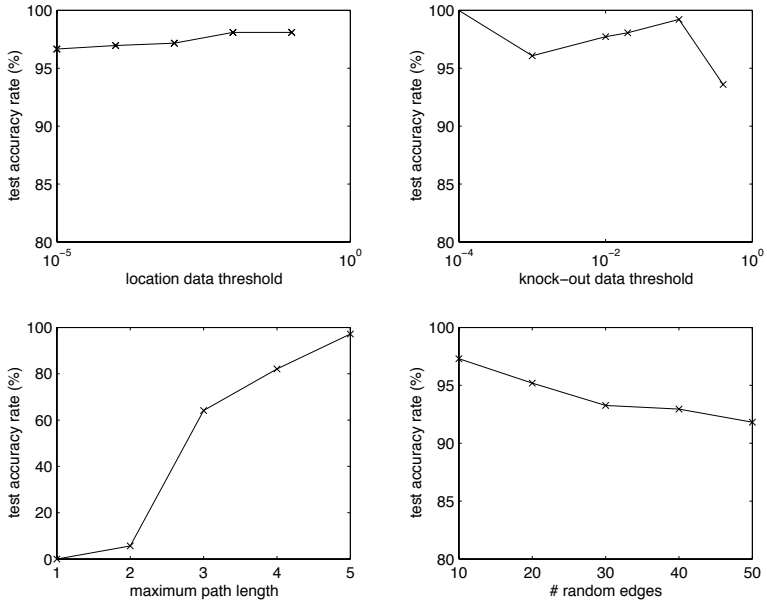

Figure 3: Invariant part of yeast mating response network

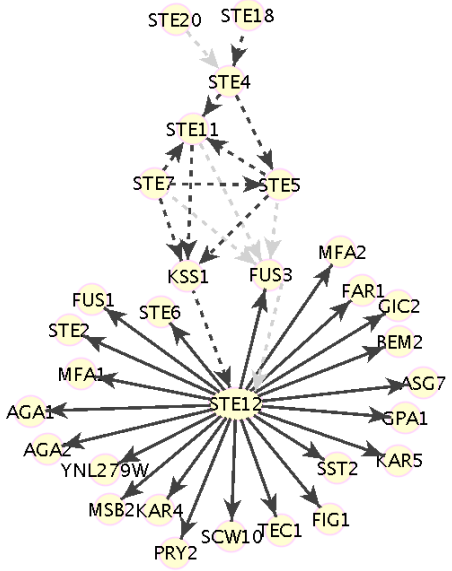

Figure 4: Variant part of yeast mating response network

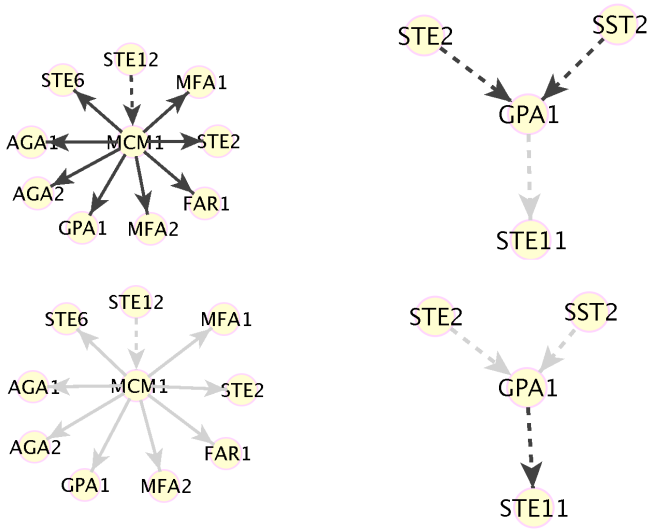

Figure 5: Contradictory knock-out effects

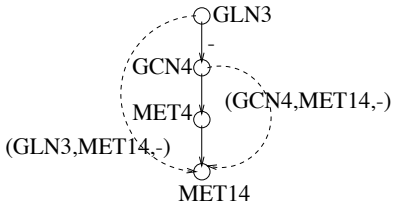

Figure 6: Physical network model uniquely determined from Rosetta data

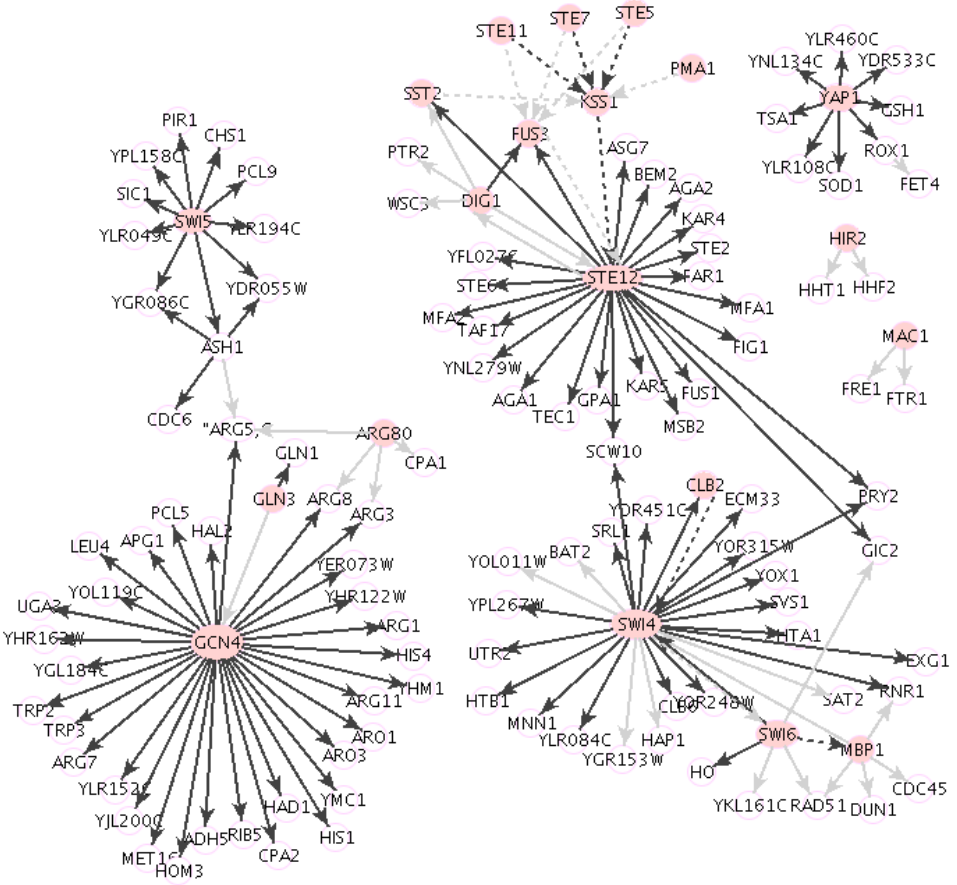

Figure 7: Decomposed subnetworks

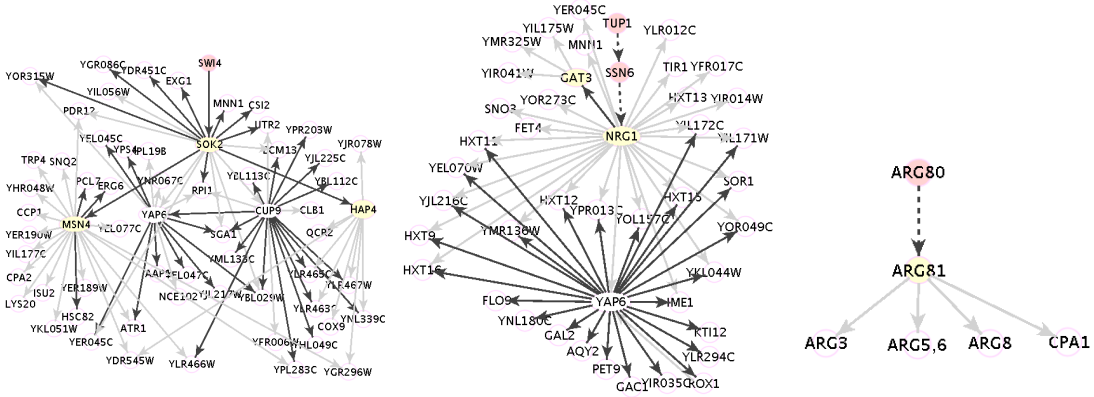

Table 1: Cross validation on knock-out pairs

# hold-outs	# trials	% error
1	106	2.83 %
5	200	3.5 %
20	200	5.9 %

Table 2: Protein-protein interactions involved in explaining knock-out effects

Type	Not used	Used ≤ 10 times	Used > 10 times
High-throughput	11650	140	23
Small-scale	2647	93	19

Table 3: Notation summary

Notation	Definition
$\vec{e}_i$	a protein-DNA edge
$x_{\vec{e}_i}$	presence of edge $\vec{e}_i$
$y_{\vec{e}_i}$	measurement of $x_{\vec{e}_i}$
$\phi_{\vec{e}_i}$	potential function of protein-DNA measurement
$\bar{e}_i$	a protein-protein edge
$x_{\bar{e}_i}$	presence of edge $\bar{e}_i$
$y_{\bar{e}_i}$	measurement of $x_{\bar{e}_i}$
$\phi_{\bar{e}_i}$	potential function of protein-protein measurement
f_1	indicator that a protein pair appears in DIP
f_2	indicator that a protein-protein interaction is reported in multiple sources
f_3	indicator that a protein-protein interaction is validated in PVM test
f_4	indicator that a protein-protein interaction appears in small-scale experiments
b	indicator that a protein-protein interaction occurs
k_{ij}	knock-out effect from gene i to gene j
σ_{ija}	indicator that path π_a is active in explaining k_{ij}
o_{ij}	measurement of k_{ij}
ϕ_{ij}	potential function of knock-out effect measurement
ψ_{ija}	potential function of knock-out effect explanation with a single path
ψ_{ij}^{OR}	potential function for path selection
ψ_{ij}	potential function of knock-out effect explanation
$P_x^{max}(x)$	max-marginal probability of x