Identifying the Signaling Cascades and Regulatory Mechanisms that Control Stress Responses

Anthony Gitter

CMU-CS-12-119

May 2012

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Ziv Bar-Joseph, Chair
Chris Langmead
Eric Xing
Judith Klein-Seetharaman, University of Pittsburgh
David Heckerman, Microsoft Research

Submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

This research was sponsored by the National Institutes of Health (NIH) under grant numbers 1RO1-GM085022 and 1U01HL108642-01 and by the National Science Foundation (NSF) under a Graduate Research Fellowship, CAREER award 0448453, and grant number DBI-0965316.

Keywords: Transcriptional regulation, time series gene expression, protein-protein interactions, biological networks, redundancy, osmotic stress, influenza, factor graph

For Rebecca and William

Abstract

Adaptation to diverse and ever-changing environmental conditions is vital to the survival of all organisms. From single-celled organisms reacting to changes in the chemical makeup of their surroundings to human cells fighting off infection, there are many global similarities across stress responses. In general, sensory proteins detect environmental perturbations and, via signaling cascades, alert specific transcription factors to adjust gene regulation and counteract negative effects of the stress. In this thesis, we present the challenges that arise when trying to understand such responses and propose computational methods for developing end-to-end models of stress response.

One primary goal when modeling the reaction to environmental perturbations is to determine the sensory proteins (sources) and transcription factors (targets) that form the endpoints of the directed signaling pathways. Many previous approaches rely on gene deletions for this task; however, we show that this strategy is unreliable due to widespread redundancy in transcriptional regulatory networks, which can mask the effects of a knockout. Instead, we propose to utilize condition-specific dynamic gene expression data to identify the transcription factors that control the divergence points in groups of gene expression profiles. We then construct a network of undirected physical protein interactions, the backbone of signaling pathways, and search for an optimal orientation of the network that connects the sensory proteins, which are already known in many conditions of interest, and the predicted active transcription factors.

Analysis of yeast signaling pathways reveals that our predicted interaction orientations are generally consistent with known annotations but also contain novel orientations that are biologically valid. Through a detailed analysis of yeast hyperosmotic stress, we demonstrate our method's ability to construct accurate end-to-end models and identify not only the transcription factors that are active in the response, but also when they are active and how they receive messages from upstream sensors. We also discuss the challenges of scaling to human interaction networks and how to overcome them. Comparative analysis of several strains of influenza demonstrates how our models can be used to identify genes with clinical relevance in the immune response to pathogens. Lastly, we explore alternative computational models for stress response that have a global probabilistic interpretation.

Acknowledgments

The proceeding work would not have been possible without the guidance and mentoring of my advisor Ziv Bar-Joseph. Ziv has been an inspirational role model and helped build the strong foundation that will drive what I expect to be a prosperous research career. I am grateful for everything I learned from him, from specific biological concepts to broad perspectives on how to be a successful researcher.

I am privileged to have David Heckerman, Judith Klein-Seetharaman, Chris Langmead, and Eric Xing as my thesis committee members. As cross-disciplinary experts in both computer science and biology, they provided insightful feedback that substantially improved my thesis. Outside the context of my thesis, my internship with David and time spent as a TA for Chris allowed me to grow as a researcher and educator.

I would also like to thank the past and present members of Ziv's Systems Biology Group, including Jason Ernst, Xin He, Peter Huggins, Hai-Son Le, Henry Lin, Yong Lu, Saket Navlakha, Yanjun Qi, Marcel Schulz, Aaron Wise, Shan Zhong, and Guy Zinman. Through our conversations and collaborations they made numerous important contributions to my projects. As co-authors of the recent and upcoming publications that compose parts of my thesis, Naama Barkai, Miri Carmi, Will Devanny, Oriol Fornes, Anupam Gupta, Michael Klutstein, Baldo Oliva, Zehava Siegfried, and Itamar Simon deserve much credit. In addition, I thank Jennifer Listgarten and Jonathan Carlson for their support during my Microsoft Research internship, Bhaskar DasGupta for many helpful algorithmic discussions, and Ted Ross for assistance with the influenza research.

There is not enough that can be said about the love and support I have received from my family. My wife Rebecca and son William mean the world to me. The happiness they bring me and strength they provide enabled me to rejoice in the triumphs and overcome the obstacles that arose throughout my graduate research. My mom, dad, and brother Dan have always been there for me. They empowered me and set me on a path to accomplish my goals, academic and otherwise. Mom and Dad Stirratt, James, John, and Carolyn have fully embraced me as one of their own, and I cherish having them in my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Perturbation of the cellular environment typically incites a vast and complex reaction that involves a multitude of proteins operating together in a sophisticated manner. Although the widespread availability and falling costs of microarray technologies along with the rise of RNA-Seq have made it easier to quantify the transcriptional aspects of a cellular response, measurements of gene expression alone represent only a limited glimpse of the processes employed by a cell in order to adapt to an external or environmental change. To fully explain and ultimately control the response to environmental stress, it is necessary to construct end-to-end models of both the signaling and regulatory mechanisms that are activated.

As will be described in greater detail in Section 4.1, many previous attempts to connect signaling pathways with transcriptional regulatory networks rely on gene deletions to determine the endpoints of the pathways. The knocked out genes are taken as the upstream sources, and the genes that are differentially expressed following a knockout (KO) are considered to be the downstream targets that are to be linked to the sources via cascades of physical interactions (i.e. protein-protein and/or protein-DNA binding interactions). However, there are several problems inherent in techniques that depend on gene knockouts. Not only are many genes essential (~ 1100 in yeast [63]) and therefore unable to be considered as sources in these models, but a gene's role in the condition of interest may be indiscernible from a knockout in a normal growth condition. Hillenmeyer *et al.* [78] found that although $\sim 80\%$ of yeast gene deletions in rich medium yield no phenotypic consequence, 97% of deletions demonstrated a change in growth fitness in one or more of 1144 chemical and environmental stress conditions. This staggering difference suggests that any method that relies on knockout data must perform the knockouts in the condition of interest, thus requiring substantial experimental effort for each new condition studied. Due to

the prohibitive costs of profiling numerous deletion strains in each condition, the resulting measurements and models are almost always static. However, the stress responses themselves are typically complex, dynamic processes [59] that must be studied over time to be fully understood.

Another more troubling problem faced by knockout-based approaches is that sophisticated backup mechanisms in the regulatory network can obscure the true role of transcription factors (TFs). In a comprehensive analysis of the agreement between binding and knockout experiments, 269 budding yeast TFs were knocked out one at a time [84], and the differentially expressed gene targets were compared to the protein-DNA binding data generated previously for 188 of those TFs [75]. It was determined that only 3% of bound genes were affected by the deletion, and similarly only 3% of knockout-affected genes were bound by the corresponding TF. While this overlap is statistically significant, the percentage is surprisingly low. On one hand, our analysis of the knockout-affected genes that are not directly bound shows that many of these events are explained by indirect effects (Section 2.4), supporting the similar goal of knockout-based algorithms. However, such methods will fail to correctly account for the other direction — bound genes that are not differentially expressed after the binding TF is deleted. As discussed in Section 2.3, many of these cases are in fact manifestations of redundancy in the regulatory program. Thus, even though the TFs are oftentimes actually controlling their bound genes, single knockouts do not detect their influences. Redundancy in biological networks may also contribute to a phenomenon observed in humans. Perturbation of the human proteins that directly interact with proteins of an infecting influenza virus generally does not lead to phenotypic change [178] (see also Section 5.2.2) even though these proteins are critical to the infection response.

To overcome the obstacles described above, we utilize condition-specific time series gene expression data instead of knockouts to infer the signaling and regulatory mechanisms at work during a stress response. Our approach, the Signaling and Dynamic Regulatory Events Miner (SDREM), assumes that many of the upstream receptors and sensory proteins are already known from signaling databases, host-pathogen interaction experiments, or other data sources. We then use the dynamic expression data to identify the TFs that are most likely to be actively driving the observed transcriptional response based on their connectivity to the known sources and their gene binding profiles. Our algorithm ultimately yields a complete picture of the stress response, including directed signaling cascades from sensory proteins to the TFs, the times at which those TFs are actively regulating their bound genes, and the primary expression profiles that characterize the affected genes.

Identifying the signaling pathways that connect the sensory proteins and target TFs

is quite a challenging problem in itself. Whereas the directionality of the physical interactions that compose regulatory networks, namely protein-DNA binding interactions, is always TF to gene, the protein-protein interactions (PPIs) that form the backbone of signaling networks are typically reported as undirected relationships (e.g. [51, 113]). To reconstruct the directed paths between sources and targets that compose signaling networks, we need to infer an orientation for undirected PPI networks (Figure 1.1). This is a difficult problem because there are many paths that can link two proteins in the interaction network, and these paths frequently disagree about the directionality of PPI. Fortunately, we can rely on a few established assumptions to simplify the problem. First, it is likely that biological responses are controlled by reasonably short signaling cascades so we can only search for length-bounded paths. Pathways in signaling databases such as KEGG [99] and the *Science Signaling* Database of Cell Signaling [70] on average contain only 5 edges between a target and its closest source [65], and previous signaling pathway prediction methods have focused on pathway segments of only 3 to 4 edges [18]. Second, we have varying degrees of confidence in the available interaction data (for example, small-scale versus high-throughput experiments [204]) and, as we show, focusing on the more confident edges leads to better pathways. Finally, in many cases there are overlapping parallel pathways linking sources and targets [37, 160, 175] so selecting an orientation that generates multiple possible pathways may produce better reconstruction results. These assumptions motivate formulating a graph theory problem to orient PPI networks, and we develop several approximation algorithms for this problem. Our orientation algorithms have been used not only as a component of SDREM but also as a precursor to other biological network analysis including the prediction of missing edges in signaling pathways (Section 3.5) and the analysis of topological redundancy [1].

Given our technique for connecting sources and targets in a signaling network by globally orienting all PPI, we return to the problem of inferring the mechanisms involved in the response to an environmental perturbation. Our strategy for building such end-to-end models is to iteratively search for TFs that can explain the dynamic gene expression data and then determine if those putative active TFs could possibly be influenced by directed signaling pathways that begin at the known source proteins. We extended the Dynamic Regulatory Events Miner (DREM) [50] to combine the condition-specific gene expression data with largely condition-independent protein-DNA binding data and infer which TFs may be responsible for the observed dynamic differential gene expression. The TFs identified by DREM are treated as potential targets of the pathways in the signaling network. Our aforementioned physical interaction network orientation algorithm predicts directed signaling cascades connecting the sensory proteins and targets, and the oriented network influences the next round of DREM. Figure 1.2 summarizes this iterative process.

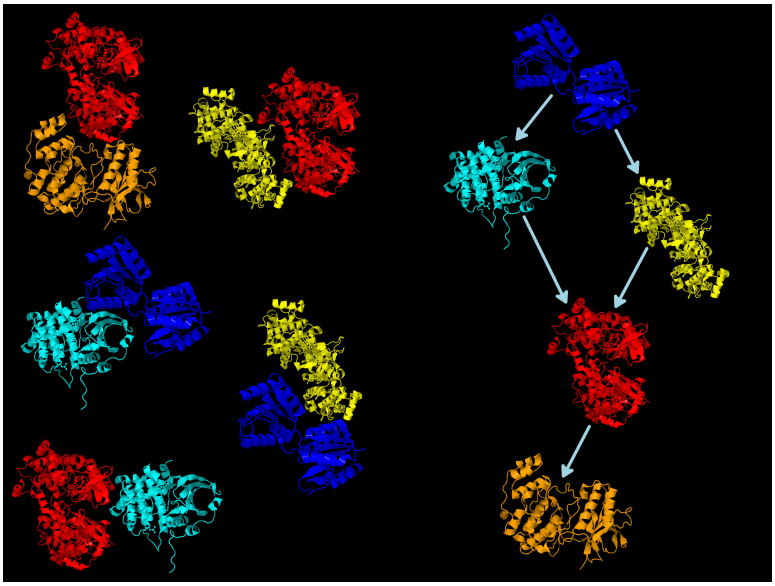

Figure 1.1: PPI are commonly reported as undirected interactions in PPI databases (left). However, by inferring the orientation of these interactions we can gain insight into the functional role they play in signaling pathways (right). Protein structures are from the Protein Data Bank [173].

SDREM has been quite successful when applied to several yeast stress responses. Models of the hyperosmotic stress and rapamycin stress responses agree with known pathway annotations, and novel osmotic stress predictions were verified experimentally (Section 4.3). Despite SDREM's practical successes, the yeast study did reveal limitations that we subsequently improved. Namely, the complexity of the network orientation problem inhibits SDREM from scaling to accommodate mammalian data and it lacks a global probabilistic interpretation. We address the scalability issues with algorithmic extensions and by leveraging new biological resources. Conceptually there is little difference when moving from yeast to mammalian datasets, but scalability becomes a limiting factor due to the larger transcriptome and proteome. Additional proteins means there are many more potential signaling pathways from the source proteins to the TFs, and the difficulty in the network orientation problem grows with both the number of potential pathways and the number of disagreements about the orientation of each edge. Thus, we enhanced SDREM by parallelizing strategic parts of the algorithm and approximating the set of all potential source-target paths. In addition, when analyzing human responses, we integrated additional data sources in order to impose new constraints on the optimal orientation, which

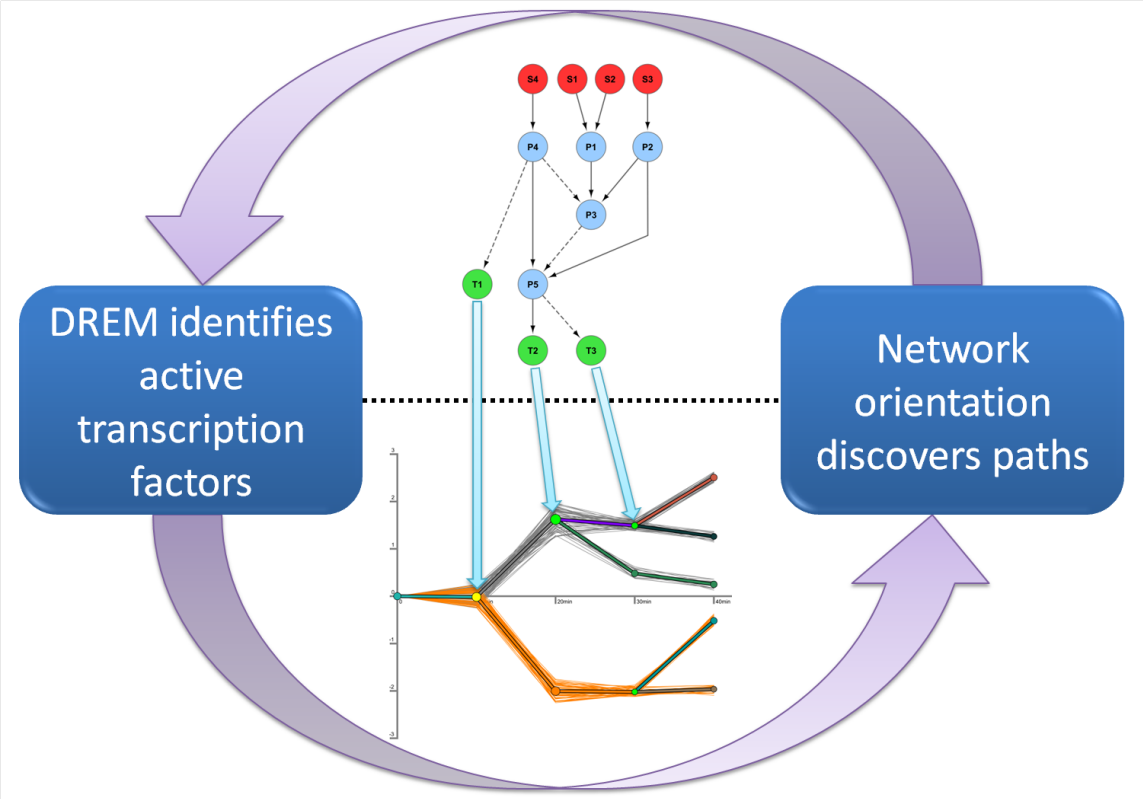

Figure 1.2: Iterative application of DREM and the network orientation algorithm generates an end-to-end model of the stress response. Predicted directed signaling pathways (top) from sources (red nodes) to TFs (green nodes) explain how those TFs are responsible for divergence in dynamic gene expression profiles (bottom). The signaling pathway images were generated with Cytoscape [177].

reduce the size of the search space. Genome-wide RNA interference (RNAi) screens [28, 100, 107] are an excellent complementary data source that provide information about whether individual nodes in the signaling network are relevant to the environmental perturbation. RNAi data allows the orientation algorithm to search for paths that include both high-confidence PPI and genes for which there is prior evidence of their involvement, thus yielding more trustworthy predictions. We also use post-translational modification data [141] to predetermine the orientation of PPI whenever possible.

The extended version of SDREM enables the exploration of clinically relevant human stress responses. Specifically, we investigate viral infection because such responses oftentimes initiate at host-pathogen PPI [33, 46, 56, 148], which form the source nodes in the host PPI network. The immune response to highly pathogenic strains of influenza is a major global health concern and also well-suited for SDREM due to the abundant data available. Thus, our initial human models target H1N1 influenza infection, which we give context to by comparing them to SDREM models of other respiratory viruses. Furthermore, we present techniques for prioritizing SDREM's predictions for experimental validation based on which genes or pairs of genes are most likely to be hits in RNAi or genetic interaction screens, respectively.

Lastly, we return to a theoretical drawback of SDREM, the fact that it is not a fully probabilistic model. The two phases of the algorithm certainly do influence one another but have their own likelihood or objective functions that are optimized independently. Therefore, we developed a unified graphical model to replicate SDREM's functionality. This model preserves the DREM component of SDREM, which is already a type of graphical model, and complements it with variables and functions over those variables that represent the signaling pathways and the orientation of their member PPI.

The thesis is structured in the following manner. Chapter 2 explores the role of transcriptional redundancy and the apparent discrepancy between genome-wide TF binding and knockout data in yeast, which motivates our decision to not use knockout data to obtain sources and targets. This work is described more fully in [67]. Chapter 3 presents the PPI network orientation algorithm and is based on [65] and [145] (in preparation). Chapter 4 presents SDREM [64] (in preparation), our iterative approach for linking signaling networks and dynamic transcriptional regulation, and demonstrates its ability to accurately reconstruct stress responses in yeast. Chapter 5 discusses how to scale SDREM to human datasets, the new insights it provides into influenza infection, and the unified graphical model. We conclude in Chapter 6 and examine future directions, both computational and biological.

Chapter 2

Backup in regulatory networks

Several methods suggested using knockouts as the starting points for reconstructing signaling and regulatory networks [155, 158, 217]. The surprisingly low percentage of genes that are both bound by a transcription factor and differentially expressed when it is knocked out (only 3% in genome-wide yeast studies [75, 84]) prompted us to explore whether such knockout-dependent techniques are suitable in the general setting. To examine possible complications inherent in knockout-based computational methods and determine whether the yeast expression and TF-gene binding interaction datasets are indeed complementary, we undertook a systems approach by studying the dependence of their agreement on the TFs' homology relationships and on the protein interaction network context of the TFs. As we show, both play a major role in the low overlap. Accounting for these contexts increases both the percentage overlap and its significance, indicating the difference may be explained by backup mechanisms employed when cells lose specific TFs. Because of these effects of redundancy in transcriptional regulatory networks, we conclude that knockouts alone are inadequate for determining the sources and targets of signaling pathways activated during a stress response.

2.1 Related work

Although regulatory backup mechanisms and robustness to perturbations have been previously studied, one major contribution of our work is the finding that shared PPI play a role in redundancy in addition to sequence similarity. Kafri *et al.* studied how partial coregulation and regulatory motif overlap affect paralogs' ability to backup one another [95] as well as the degree distribution of compensatory duplicate gene pairs in PPI net-

works [96]. Other work examined how topological properties of biological networks such as feedback loops affect robustness [118]. An alternative explanation for the discrepancy between TF binding and knockout effects is that eukaryotic TF binding can be nonspecific and often nonfunctional, as was suggested by an information-theoretic study of binding motifs [215].

In the other direction, it is well-established that indirect knockout effects can be explained via physical interaction networks (e.g. [213, 217]). Previous analysis of the genome-wide yeast datasets, in which pathways of protein-DNA binding interactions were allowed as supporting evidence for the knockout effects, did not incorporate PPI networks and resulted in negligible improvements to the overlap and its significance [84].

2.2 Cleaning the data

To lessen the extent to which experimental and biological noise affected the disagreement between the knockout and binding data, we cleaned the data sets in several ways. We first removed genes that were affected by the knockout of a large number of TFs. We termed these "general KO genes" because they are likely responding to the general stress of the knockout experiments rather than the specific TF deletions and thus are not expected to be bound by the deleted TFs. In addition, we restricted the set of TF binding targets to those with sequence motifs conserved in two other species [133]. Many of the original 203 TFs do not have a known or conserved motif and were removed from subsequent analysis. After these cleanup steps, the agreement between the two datasets increases to 6.7% of binding data and 4.5% of knockout data (p-value of 2.29E-133 versus the original p-value of 2.10E-114). In addition to the cleanup, we performed several other exploratory analyses including varying the p-value thresholds, calculating overlaps via rank-based tests in order to remove the threshold dependency, and accounting for possible condition-specific TF activity [67]. None of these investigations could substantially account for the disagreement in the binding and knockout data, indicating that it cannot be explained by issues related to the analysis of the data but is rather likely to represent a specific biological phenomenon.

The overlaps and their significance are calculated in the following manner. For a given TF t, we define the set of genes significantly bound by t (at some predetermined p-value threshold) to be G_B and the set of genes significantly affected by the knockout of t as G_K.

The binding overlap B and knockout overlap K are:

$$B = \frac{|G_B \cap G_K|}{|G_B|}$$
$$K = \frac{|G_B \cap G_K|}{|G_K|}$$

We use the hypergeometric distribution, also known as the one-tailed version of Fisher's exact test, to calculate a p-value for the overlap of the binding and knockout targets:

$$p = \sum_{o=|G_B \cap G_K|}^{\min(|G_B|,|G_K|)} \frac{\binom{|G_K|}{o}\binom{|G_A|-|G_K|}{|G_B|-o}}{\binom{|G_A|}{|G_B|}}$$

where $\binom{n}{k}$ is the choose function of n and k, G_A is the set of all possible genes targets in the binding or knockout datasets, and o is the size of the overlap.

2.3 Redundancy explains binding interactions absent from the knockout data

2.3.1 Genome-wide effects of redundancy

We next tested whether redundancy can help explain the small overlap observed. Following Kafri *et al.* [95] we used BLASTP [3] to identify gene pairs with varying levels of homology. We divided the set of TFs into four groups: those with a paralogous TF with an E-value of E-20 or less, between E-20 and E-10, between E-10 and E-3, and those with no homolog at E-3 or less. TFs with the most similar paralogs had no overlap between their binding and knockout data. In contrast, those with the least similar paralogs had an overlap of more than 12%, nearly two folds higher than the average overlap. The other groups followed a similar trend where the overlap increased as the similarity to the closest paralog decreased (Figure 2.1). To further test our finding that redundancy impacts the expression outcome we used Pfam [52], which focuses on the binding domain only, as a measure of similarity and obtained similar division into four groups. As with the BLASTP value, for groups with similar paralogs the overlap was lower than for those with more distant homologs (4% versus 10%).

Another feature that may impact how well one TF can compensate for the loss of another TF is shared protein-protein interactions. We divided each of the homology groups

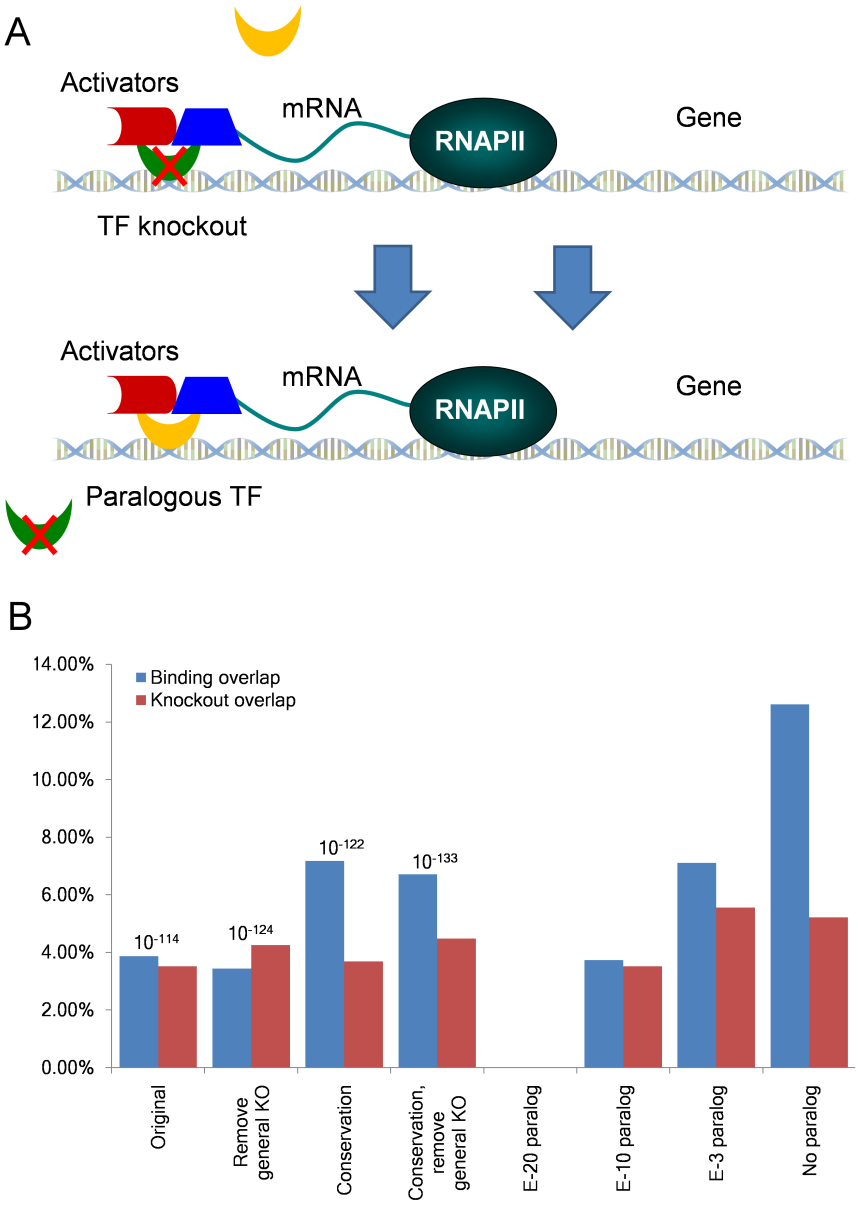

Figure 2.1: Improved overlap between binding and knockout experiments. A) A schematic view of our analysis. Both sequence homology and shared interactions may lead to one TF compensating for another. The yellow TF can replace the green TF when it is knocked out and is able to recruit the transcription machinery. B) The binding and knockout overlap for various subsets of the data. The p-value of the overlap is given above the columns.

defined above based on the percentage of protein interaction partners the TF shares with another TF in that homology group using a large set of literature-curated PPI [169]. Similar to the trend we saw using sequence homology, within each group the overlap decreases as the percentage of shared PPIs increases (Figure 2.1 and Table 2.1). For TFs with the least similar homologs and the fewest shared interactions, we observed an overlap greater than 13%. At all BLASTP E-value thresholds TFs that shared a larger portion of PPI with their paralog had lower binding overlap. This indicates that putative paralogs with many common PPI are better able to compensate for the deletion effects (Table 2.1).

Table 2.1: Analysis of overlap based on paralogs and shared PPIs. Transcription factors were divided into four groups based on their most similar TF homolog as determined by the BLASTP E-values. These sets were further divided based on the percentage of PPI a TF shared with its paralog. TFs with a putative paralog that share at least 20% PPI are more likely to be redundant and thus exhibit lower overlap.

BLASTP E-value	**Shared PPI**	**Binding overlap (%)**	**Knockout overlap (%)**
E-3 $<$ E-value $\leq$ E1	PPI $<$ 20%	13.37	5.37
E-3 $<$ E-value $\leq$ E1	PPI $\geq$ 20%	3.30	2.13
E-10 $<$ E-value $\leq$ E-3	PPI $<$ 20%	7.51	5.85
E-10 $<$ E-value $\leq$ E-3	PPI $\geq$ 20%	0.00	0.00
E-20 $<$ E-value $\leq$ E-10	PPI $<$ 20%	4.20	3.00
E-20 $<$ E-value $\leq$ E-10	PPI $\geq$ 20%	2.77	7.48
E-value $<$ E-20	PPI $<$ 20%	0.00	0.00
E-value $<$ E-20	PPI $\geq$ 20%	0.00	0.00

2.3.2 Experimental validation

To further validate our results regarding the backup mechanisms employed in regulatory networks we collected expression data from three double knockout experiments involving pairs of factors we predicted could compensate for the loss of each other (Fkh1-Fkh2 [230], Ace2-Swi5 [205], Yhp1-Yox1 [163], all from the E-20 set, Table 2.2). We also carried out new experiments for an additional pair (Pdr1-Pdr3, also in the E-20 set, see [67] for experimental details). As predicted, when the paralogous partner is not present to compensate for the effect of a single knockout, the overlap of the knockout and binding data increases significantly. In fact, with the exception of *ACE2*, *FKH2*, and *SWI5*, we found that single knockouts of TFs with a strong putative paralog do not affect the expression

levels of a significant number of genes they bind due to their partner's compensation. In sharp contrast, when both a TF and its paralog are deleted, the backup mechanism is eliminated and a significant number of bound genes are differentially expressed (Table 2.3). Even for *ACE2*, *FKH2*, and *SWI5* the effects of a double knockout are more pronounced. Likewise, the overlap between genes affected by the single knockout of *YHP1* and *YOX1*, two cell cycle transcription factors, and the genes bound by these factors is 0% and 3% respectively (neither is significant). In contrast, the overlap for the double knockout and the binding targets of *YHP1* and *YOX1* is 14% and 25%, respectively. Similar results were obtained for the other double knockouts we collected (Table 2.3). For our *PDR1-PDR3* double knockout experiment we again observed a large increase in the percentage of overlap for *PDR1* compared to the single knockout experiment. The overlap increased from 1% (not significant) to 6% (p-value of 4.40E-6). For *PDR3* we saw only a small increase (Table 2.3). Thus, these experiments support our claim of backup provided by these pairs of factors and can also provide clues to the mechanisms utilized as we discuss below.

Table 2.2: BLASTP E-values and shared PPI of putative paralogs examined in double knockout experiments. Strong sequence similarity and a high percentage of shared PPI were used to identify putatively redundant TF pairs that were likely to compensate for each other's deletion. Fkh1 and Fkh2 do not have any PPI in common, but we consider them because of their sequence similarity and previously reported evidence of redundancy [81]. The BLASTP E-value is not symmetric, and the lower of the two values is reported here.

TF1	**TF2**	**BLASTP E-value**	**TF1 PPI shared by TF2 (%)**	**TF2 PPI shared by TF1 (%)**
Pdr1	Pdr3	9E-139	25	25
Fkh1	Fkh2	6E-115	0	0
Ace2	Swi5	2E-66	36	21
Yhp1	Yox1	4E-47	100	33

2.3.3 Mechanisms leading to TF redundancy

A subset of the homologous TFs we identified bind to an overlapping group of targets, and thus it is not surprising that knocking out one of them has small effect on the expression of its targets. One such example is the two homologous transcription factors involved in methionine metabolism, Met31 and Met32 [23]. These TFs have a large overlapping set of target genes ($> 60\%$), and neither has any target genes that are differentially expressed

Table 2.3: Double knockouts of paralogous TFs significantly affect bound genes. Functional redundancy explains the apparent lack of response that results from a single knockout of a TF with a strong paralog. When the paralog is deleted concurrently, bound genes are significantly affected. The original binding dataset was used for these results.

Genes bound by	**TF(s) deleted**	**p-value**	**Binding overlap (%)**	**Knockout overlap (%)**	**Bound genes**	**Deletion-affected genes**
Pdr1	Pdr1 and Pdr3	4.40E-6	18.71	5.75	139	452
Pdr3	Pdr1 and Pdr3	0.860	4.26	0.44	47	452
Pdr1	Pdr1	0.852	0.72	1.18	139	85
Pdr3	Pdr3	1.00	0	0	47	13
Fkh1	Fkh1 and Fkh2	3.46E-5	5.02	15.38	239	78
Fkh2	Fkh1 and Fkh2	3.38E-9	8.43	19.23	178	78
Fkh1	Fkh1	1.000	0	0	239	21
Fkh2	Fkh2	3.69E-2	1.12	18.18	178	11
Ace2	Ace2 and Swi5	4.47E-8	10.08	14.12	119	85
Swi5	Ace2 and Swi5	3.44E-12	10.84	21.18	166	85
Ace2	Ace2	8.69E-8	6.72	25.00	119	32
Swi5	Swi5	1.84E-3	1.81	30.00	166	10
Yhp1	Yhp1 and Yox1	7.81E-5	7.55	14.29	53	28
Yox1	Yhp1 and Yox1	3.52E-8	8.86	25.00	79	28
Yhp1	Yhp1	1.00	0	0	53	42
Yox1	Yox1	0.316	1.27	3.33	79	30

after deletion. Another example is the two forkhead transcription factors, Fkh1 and Fkh2. These only bind a partial overlapping set of target genes. However, it has been shown [81] that the binding of Fkh1 to Fkh2 targets is enhanced in the absence of Fkh2 and vice versa suggesting that compensation can occur beyond the common targets, as predicted by our findings.

This type of compensation may happen due to competition between the two TFs that is resolved in the absence of one of them. Another possibility is that the activity of one TF is enhanced in the absence of its homolog due to a feedback mechanism between the two TFs [95]. In order to check this idea, we looked at the expression levels of the TFs believed to be compensating for the knockout (most similar based on BLASTP). As expected, we have not found any example in which the expression level of the homologous TF was significantly decreased. However, a significant increase was observed in only a few cases. Thus, it appears that these changes are mainly driven by post-transcriptional events,

perhaps by the protein interaction networks mentioned below.

2.4 Protein interaction networks provide physical support for knockout effects

To help explain why genes affected by TF knockout are not bound by the TF we constructed a network that includes both PPI and protein-DNA edges. We considered a gene affected by the knockout to be explained by the network if 1) the TF directly binds the gene or 2) there is a path leading from the TF to another TF that directly binds the gene. For the indirect result we varied the maximum path length (number of edges from the initial TF to the last TF). Using a path length of 2 leads to an overlap of 22% while significantly increasing the p-value of the overlap (from 2.29E-133 to 1.27E-211). Path lengths greater than 2 increased the percentage of the overlap but reduced the p-value due to the large number of paths that did not explain an indirect knockout effect (Figure 2.2). Randomization tests and further analysis using different sets of PPI data confirmed the significance of the increase in overlap due to the PPI network.

Our results reaffirm that physical interaction networks can be used to explain TF knockout effects in the absence of direct binding and demonstrate that backup mechanisms play an important role in regulatory networks and cannot be ignored. Strikingly, for the TFs that we predicted are most likely to have a redundant paralog, no bound genes were affected by TF deletion. Because functional binding may be obscured in knockout data, accurate reconstruction of signaling and regulatory networks must exploit alternate data sources. In Chapter 4 we show how a combination of static TF binding data and dynamic gene expression data from wild type cells can be used to identify functional TF binding and avoid the complications of redundancy.

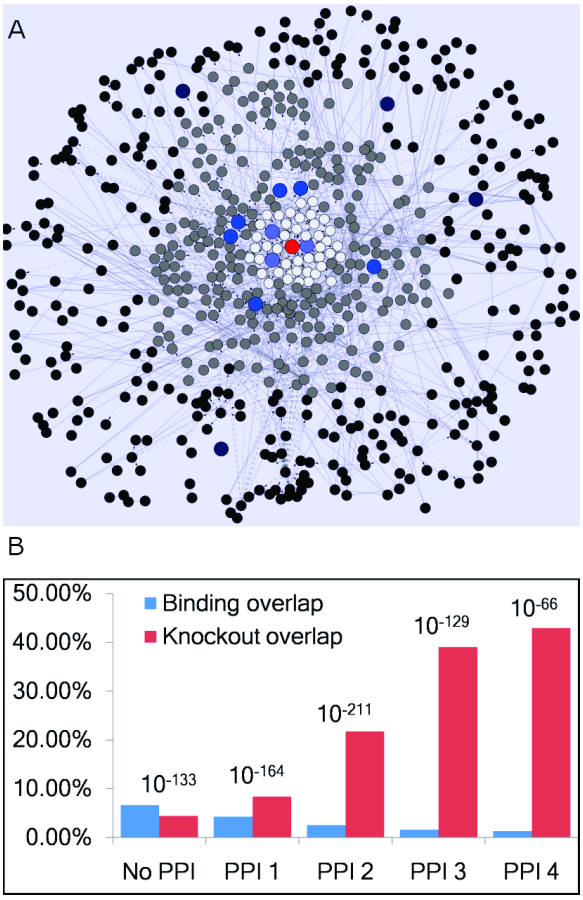

Figure 2.2: Influence of physical interaction networks. TFs that do not directly bind a gene can exert influence via pathways of PPI and protein-DNA interactions. A) A network consisting of Skn7 (red), its knockout targets (shades of blue), and 20% of all other yeast genes selected at random (shades of gray). Genes are arranged around Skn7 according to the shortest number of interaction edges needed to reach them. The black and dark blue nodes correspond to genes that are 3 or more interactions away, the medium gray and medium blue genes are 2 interactions away, and the light gray and light blue genes are a single interaction from Skn7. 85% of Skn7's knockout-affected genes are either directly bound by Skn7 or another TF that can be reached via paths of length 1 or 2. B) As longer paths in the network are examined, a much higher percentage of the knockout-affected genes are connected to the deleted TF. The p-value of the overlap is given above the columns, which indicate percent overlap.

Chapter 3

Discovering signaling pathways

Before exploring the more challenging problem of learning both the TFs active in a stress response and the directed signaling pathways that influence them, we first consider the special case where both the sources and targets in the signaling network are known. In this section, we formalize the PPI network orientation problem, introduce several approximation algorithms, and show that they perform well from both a theoretical and biological perspective.

3.1 Related work

Although much attention has been given to the signaling pathway prediction problem, nearly all previous work does not consider the orientation of the paths and simply selects subsets of edges, yielding undirected predictions. One of the earliest undirected pathway prediction algorithms was NetSearch [188]. NetSearch enumerated linear pathways and ranked all putative pathways by clustering the gene expression profiles of pathway members and generating hypergeometric distribution-based scores. Because linear paths do not fully capture the complexity of signaling networks, Scott *et al.* [176] used a color-coding technique to search for paths and higher order structures (trees and parallel paths) in a weighted protein interaction graph. Lu *et al.* [132] presented a randomized divide-and-conquer algorithm that, like Scott *et al.*, supported complex non-linear pathways structures. PathFinder [18] integrated multiple data sources to extract association rules describing protein function in known signaling pathways and then used these rules, along with additional expression data, to detect new pathways of interest in the network. Whereas many previous methods searched for source-target pathways individually, Zhao

et al. [228] formulated a linear program to identify a single global signaling subnetwork that satisfies various constraints. We refer to their technique as the unoriented edge selection algorithm. Recognizing the tradeoffs between local and global search approaches, Yosef *et al.* [221] presented an algorithm that combined the two objectives and could be tuned to give preference to one or the other on a particular run. While all of these methods led to useful findings, none of them generates directed pathways. As we show in the Section 3.3, by ignoring the edge orientations these methods lose important information that improves pathway reconstruction and thus contain far fewer known signaling pathways in their predictions.

Relatively few methods have been developed to try to explicitly address the edge orientation problem. In [137] the authors defined the Maximum Tree Orientation (MTO) problem, which focused on reachability. They considered a source-target pair to be satisfied as long as any single path of arbitrary length connected them. As a result, cycles in the PPI network could be contracted and the problem was equivalent to orienting a tree. While this variant of the edge orientation problem can be approximated well, such a structure cannot give preference to short paths or high-confidence edges and also ignores redundant pathways. Extensions to this work improve the theoretical guarantees and are able to handle mixed graphs, in which some edges have predefined orientations [58]. Liu *et al.* [131] predicted directed signaling pathways in multiple species. However, because their method relies on specific protein domain interactions, it does not scale to the entire proteome. Indeed, as the authors noted, coverage, the fraction of interactions in the test set for which predictions could be made, was less than 50% at the thresholds they used. Probabilistic graphical models have also been used to orient edges when trying to explain knockout effects via a physical interaction network consisting of PPI and protein-DNA interactions [217]. The Physical Network Models (PNM) algorithm constructs a factor graph and applies belief propagation to infer both PPI directionality and regulatory effect (inhibition or activation). While this approach works well for relatively small networks and short pathways, it does not scale well [65]. SPINE [155] adopts the PNM formulation but expresses the problem as an integer program. However, SPINE only focuses on identifying activation and repression regulatory effects of either proteins or edges and does not attempt to orient the network. Conversely, our goal is to determine directionality in PPI signaling networks where the positive and negative regulatory effects upon genes are not the primary concern.

3.2 Theoretical aspects of network orientation

3.2.1 Formalizing the Maximum Edge Orientation (MEO) problem

We assume we are given a weighted undirected graph $G = (V, E)$ which represents our current knowledge of protein interactions. We are also given a maximum path length k and source-target pairs of the form $< s_i, t_i >$ such that $s_i \in S \subseteq V$ and $t_i \in T \subseteq V$. Our goal is to orient edges $e = (u, v) \in E$ from u to v or from v to u such that the weight of all satisfied paths between sources and targets with length at most k is maximized. Each simple path takes the form $p = (v_1, v_2), (v_2, v_3), \ldots, (v_l, v_{l+1})$ where $v_1 = s_i$, $v_{l+1} = t_i$, and $l \leq k$ for some pair $< s_i, t_i >$. A path is satisfied in a given network orientation if and only if for every edge (v_j, v_{j+1}) along the path the edge is oriented from v_j to v_{j+1} in the network. Multiple paths may exist between a single source-target pair as long as paths with the same source and target have at least one disjoint edge (as mentioned above, parallel pathways are very common). After orientation there may be directed source-target paths in the graph that contain more than k edges, but they are not incorporated into the objective function.

All vertices and edges in the graph have real-valued weights denoted $w(v)$ and $w(e)$ respectively. We set the weights of all vertices (proteins) to 1 for the time being but relax this restriction in Sections 4.2.3 and 5.1.1. The edge weights are assigned based on the confidence in each protein interaction, which in our implementation depends on the type of experimental support provided for that edge. Weights represent our confidence in the presence of the edge or in the involvement of a gene in the response, and the weight of an entire path p is

$$w(p) = \prod_{v \in p} w(v) \prod_{e \in p} w(e)$$

Since we use weights in the range $[0, 1]$ to represent edge confidence, this definition of path weight causes long paths to have lower weights than short paths. Thus, the objective in the Maximum Edge Orientation (MEO) problem is to maximize the function:

$$\sum_{p \in P} I_s(p) w(p)$$

where P is the set of all unique paths between sources and targets with length at most k and $I_s(p)$ is an indicator function that has the value 1 if path p is satisfied.

Although we currently assume that edge weights are symmetric, one simple yet powerful generalization is to allow asymmetric edge weights when there is a prior belief that one orientation of an edge is more likely than the other. Incorporating such information

involves using the appropriate direction-specific weight for each edge when calculating $w(p)$ during path enumeration, but does not require any adjustments to the proposed MEO approximation algorithms.

3.2.2 MEO is NP-hard

Before discussing MEO NP-hardness, inapproximability, and approximation algorithms, we introduce the concept of an r-approximation. For an instance p of a maximization optimization problem, if the optimal value of the objective function is OPT(p) and an approximation algorithm guarantees a value of at least APX(p), we say the algorithm guarantees an r-approximation where

$$r = \frac{\text{APX}(p)}{\text{OPT}(p)}$$

Similar to Medvedovsky *et al.* [137], we prove that MEO is NP-hard for any $k \geq 2$ by reduction from Maximum Directed Cut (MAX-DI-CUT) [74]. Given a directed graph $G = (V, E)$, the objective of MAX-DI-CUT is to partition the vertices V into sets A and B, where $A \subseteq V$ and $B = V \setminus A$, such that the number of directed edges that begin in A and end in B is maximized. To reduce a MAX-DI-CUT instance $G = (V, E)$ to MEO, we add a new node C and construct an undirected graph $H = (V', E')$, where $V' = V \cup \{C\}$ and $E' = (v', C)$ for all $v' \in V$ (Figure 3.1). All edges and vertices in H are given a weight of 1 so that for all p, $w(p) = 1$. For every directed edge (u, v) in the MAX-DI-CUT instance, we create a source-target pair $< u, v >$ in the MEO instance.

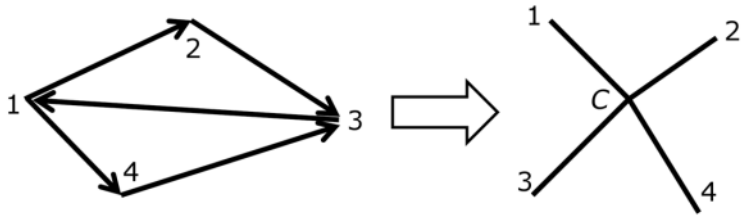

Figure 3.1: An example of the MAX-DI-CUT to MEO transformation. The MEO graph has the same vertices as the MAX-DI-CUT graph plus an additional center vertex, to which all other vertices are connected. The MAX-DI-CUT edges are used to define the MEO source-target pairs.

Observe that there is a one-to-one mapping between an orientation O of H and a cut $A \subseteq V$ of G. Any orientation of the MEO instance that achieves a score m can be used to construct a solution to the MAX-DI-CUT problem that places m directed edges across the cut. In the orientation, if an edge (v', C) is oriented toward C, then place the corresponding vertex v in the set A. For all edges (v', C) oriented away from C, include v in the set B. All paths in the MEO instance consist of two edges $(v'_1, C), (C, v'_2)$. Thus, if a path is satisfied the orientation of these edges must be directed $v'_1 \rightarrow C$ and $C \rightarrow v'_2$. As a result, in every satisfied path the vertex v_1 in G corresponding to the source v'_1 will be in the set A and every vertex v_2 in G corresponding to the target v'_2 will be in the set B. In other words, the directed edge (v_1, v_2) will be across the cut in G (Figure 3.2). Because source-target pairs were derived from the directed edges in G, we know that there is a unique directed edge (v_1, v_2) in G that corresponds to the source-target pair. In addition, there is only one path connecting a particular source-target pair in H. It follows that for every satisfied source-target path, the corresponding directed edge will begin in A and end in B so that if there are m satisfied paths in H there will be m edges across the cut in G.

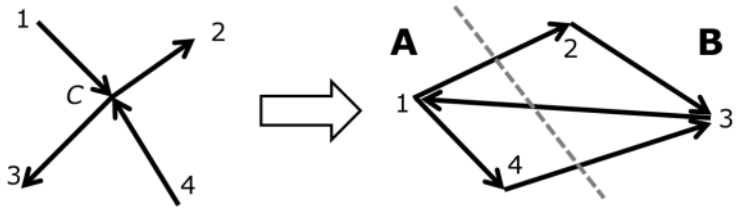

Figure 3.2: Mapping an orientation of the MEO instance back to a directed cut. An orientation in the MEO problem uniquely defines a cut in the MAX-DI-CUT instance. The number of satisfied paths in the MEO instance is identical to the number of directed edges from A to B.

Similarly, any partitioning of the vertices in G will yield a unique orientation in H. For every vertex v in A, orient (v', C) toward C. For every vertex v in B, orient (v', C) toward v'. Using this procedure, any cut of m edges will produce an orientation with m satisfied paths because each directed edge across the cut will correspond to a source-target pair and the path connecting that pair will have its first edge oriented toward C and its second edge oriented away from C toward the target. Consequently, the number of edges across the cut in the optimal solution to the MAX-DI-CUT problem is equal to the objective function score of the optimal MEO orientation.

Because the problems have the same optimal solution and an orientation that achieves

a score m can be used to construct a vertex partitioning that places m directed edges across the cut, an algorithm that achieves an r-approximation for MEO can achieve an r-approximation for MAX-DI-CUT as well. MAX-DI-CUT cannot be approximated within 12/13 [77], therefore MEO is inapproximable within 12/13. The reduction only requires paths of length 2 so this result holds for any $k \geq 2$. However, MEO is even harder for larger (yet still reasonable) values of k. We can reduce MAX-3-SAT, which is harder to approximate [77], to MEO with $k \geq 5$ yielding the stronger inapproximability bound of 7/8 for this range of k (proof omitted). Because MEO is NP-hard even for small k, we developed approximation algorithms for orienting the graph with varying theoretical guarantees and running times.

3.2.3 Approximation algorithms

The simplest approximation algorithm randomly assigns an orientation to each edge in the graph. For a particular path, let the orientation an edge takes when the path is satisfied be the optimal orientation for that edge with respect to the path. After a random orientation, each edge in a particular path will be optimally oriented with probability $\frac{1}{2}$. Because the path contains at most k edges and all edges are oriented independently, the probability that a given path is satisfied is

$$P(I_s(p) = 1) = \prod_{e \in p} P(I_o(e, p) = 1) = \prod_{e \in P} \frac{1}{2} \geq \left(\frac{1}{2}\right)^k$$

where $I_o(e, p)$ is an indicator function that takes value 1 if the edge e is optimally oriented for path p. Thus, the expected value for a path is $E[p] \geq w(p) \left(\frac{1}{2}\right)^k$ and by linearity of expectation the random orientation yields a $\frac{1}{2^k}$-approximation. In practice, for all of our approximation algorithms we deterministically fix the orientation of any edges that are used in the same direction by all paths that contain them and only randomly orient the remaining edges. This can only improve the likelihood that a particular path is satisfied, thus the approximation guarantee is not affected.

Although the MEO problem is a maximization problem, an MEO instance can be transformed to a weighted MIN-k-SAT [22, 105] instance. Weighted MIN-k-SAT is an optimization version of the traditional Boolean satisfiability (SAT) problem in which weighted disjunctive clauses with at most k literals are given and the objective is to find the assignment to all variables that minimizes the sum of the weights of the satisfied clauses. For each edge (u, v) in the MEO graph, the MIN-k-SAT instance will have a corresponding edge variable x_{uv}. The goal is to orient the edge by assigning a value of 1 $(u \rightarrow v)$ or 0

$(v \rightarrow u)$ to that edge. We first enumerate all simple paths of length at most k via depth first search. Then for each path, we construct a disjunctive clause that has the same weight as the path. The edge variables in the clause are given by the edges used by the path. If a path uses an edge in its canonical positive orientation $(u \rightarrow v)$, the negation of the edge variable appears in the clause. Otherwise the edge variable appears in the clause but is not negated. Observe that there is a one-to-one mapping between clauses that are satisfied and paths that contain at least one edge oriented in the wrong direction and will not be satisfied. The constructed MIN-k-SAT instance therefore aims to minimize the sum of the weights of the paths that are not satisfied (which, of course, maximizes the sum of those satisfied).

Figure 3.3 illustrates the transformation for an instance with two paths: $p_1 = (1,3), (3,4), (4,6)$ with $< s_1, t_1 >=< 1, 6 >$ and $p_2 = (5,4), (4,3), (3,2)$ with $< s_2, t_2 >= < 5, 2 >$. All vertices have been assigned an index, and the canonical positive orientation of each edge is the orientation toward the vertex with the larger index. Because p_1 uses all edges in the positive direction, all edges variables in clause 1 are negated. Thus, if any of these three edges are oriented in the negative direction (toward the lesser index), clause 1 will be satisfied and the objective function will be penalized by $w(p_1)$.

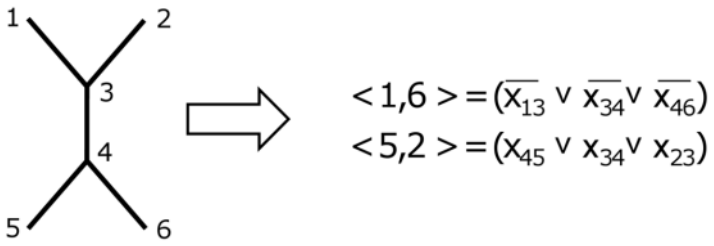

Figure 3.3: Formulating an MEO instance as a MIN-k-SAT problem. Each path connecting a source-target pair becomes a disjunctive clause. The literals in the clause are given by the edges in the path.

The constructed MIN-k-SAT instance can be solved using an algorithm by Bertsimas *et al.* [22]. The MIN-k-SAT instance is formulated as an integer program and then relaxed as a linear program (LP). The authors present a dependent randomized rounding scheme for transforming the LP solution into variable assignments for the MIN-k-SAT problem. We use lp_solve [44], an open-source LP solver based on the revised simplex method, in our implementation of the MIN-k-SAT-based approximation algorithm.

While the optimal solution for the weighted MIN-k-SAT problem will provide the optimal solution to our problem, the $\frac{2^{(k-1)}}{2^k-1}$-approximation ratio for the specific algorithm by Bertsimas *et al.* does not hold for MEO. This is due to our transformation of the MEO maximization problem into a minimization problem; the optimum of the weighted MIN-k-SAT instance is the sum of the weights of all paths *minus* the optimum of the MEO instance.

Rather than minimizing the weights of paths that are not satisfied as in the MIN-k-SAT-based approximation, it is more straightforward to directly maximize the weights of satisfied paths by using *conjunctive* clauses. The trade-off is that the resulting optimization problem is more difficult to approximate than MIN-k-SAT. The transformation is similar that used in the MIN-k-SAT-based algorithm except that edge variables used in the positive canonical direction by a path are positive in the conjunctive clause and vice versa. Figure 3.4 shows the transformation using the previously introduced MEO example.

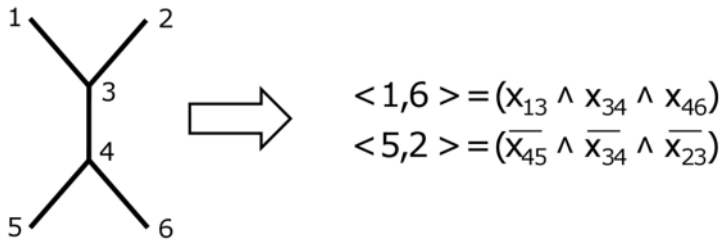

Figure 3.4: Transforming an MEO instance into MAX-k-CSP. Each path connecting a source-target pair is mapped to a conjunctive clause. As in the MIN-k-SAT transformation, the literals in the clause are given by the edges in the path.

Optimizing the weights of the satisfied conjunctive clauses is an instance of MAX-k-AND, which is also referred to as MAX-k-CSP (constraint satisfaction problem) because the more general MAX-k-CSP can be approximated as well as MAX-k-AND [199]. The state of the art MAX-k-CSP approximation [167] does not yield an explicit approximation ratio. However, previous work by Charikar *et al.* [32] provides a $O\left(\frac{k}{2^k}\right)$-approximation ratio for general k, and even better special case solutions for k equal to 2, 3, and 4 exist as well [73, 124, 231]. Because the MAX-k-CSP reduction is approximation-preserving, these general and special case theoretical guarantees apply directly to the MEO problem as well, improving the $\frac{1}{2^k}$-approximation guarantee obtained via random orientation.

Although they provide theoretical guarantees, the above MAK-k-CSP approximations

are all based on semidefinite programming. Consequently, they do not scale well on large instances (e.g. genome-wide protein-protein interaction networks) and are not typically used in practice. Therefore, to solve the MAX-k-CSP reduction we use toulbar2 [174], a branch and bound-based solver, which was by far the best performing solver in the MAX-CSP portion of the Third International CSP Solver Competition.

The solution returned by any of the algorithms described above can typically be improved by using it as the starting point for a local search instead of taking it directly as the final orientation. Specifically, local search in the MEO problem involves iteratively finding the edge that will yield the greatest improvement in the objective function if its orientation is changed and flipping that edge's direction. While helpful in practice, local search does not improve the theoretical guarantees of any of the algorithms.

3.2.4 Algorithms outperform approximation guarantees

To evaluate our orientation algorithms from a theoretical perspective, we examined the objective function values achieved in practice with respect to the approximation guarantees by using a real interaction network [187] and simulated source-target pairs. Edge weights in the PPI network were computed using both the confidence in the experimental systems used to detect the interaction and the number of separate publications that report the interaction. For each interaction between proteins $P1$ and $P2$, the probability their interaction is a true positive is given by the formula

$$P(interact(P1, P2)) = 1 - \prod_{i \in I_{P1,P2}} (1 - c(i))$$

where i is a member of the set $I_{P1,P2}$, all of the distinct (based on experiment type and PMID) instances of that interaction in the PPI dataset, and $c(i)$ is the confidence in the class of experiments to which i belongs. The confidence assigned to each type of experiment can be found in [65]. We set the maximum path length to 5 (allowing for 6 proteins in each pathway), which is longer than the 3 to 4 edges preferred by previous pathway prediction algorithms [18, 155]. We randomly selected 5 unique sources and 10 unique, distinct targets for each test case leading to 50 source-target pairs per instance.

To compute an upper bound on the optimal score for each instance, we formulate MEO

as the following integer program

$$\begin{aligned} \max \quad & \sum_{p_j \in P} w(p_j) p_j \\ \text{subject to} \quad & p_j \leq e_i \quad \forall e_i \in E_j^+ \\ & p_j \leq 1 - e_i \quad \forall e_i \in E_j^- \\ & e_i, p_j \in \{0, 1\} \quad \forall e_i, \forall p_j \end{aligned}$$

where P is the set of all simple source-target paths with length at most k, e_i are the edge variables, $w(p_j)$ is the weight of path p_j, E_j^+ is the set of all edges used in their positive canonical direction in path p_j (as defined in the MIN-k-SAT algorithm description), and E_j^- is the set of all edges used in their negative canonical direction in path p_j. If any edge in the set E_j^+ has the value 0, which corresponds to being oriented in the negative direction, the path cannot be satisfied and must have the value 0 as well. Likewise, if any edge in the set E_j^- has the value 1 the path cannot be satisfied and must have the value 0.

This formulation provides an exact representation of MEO. Consequently, the integer program's optimal solution is equal to the maximum MEO objective function value. The optimal solution to the LP relaxation of this integer program provides an upper bound of the optimal MEO score. This is because the optimal orientation corresponds to an integer solution to the integer program. That integer solution is a valid solution to the LP, which means the maximum LP value cannot be lower than the value obtained by using that solution. This upper bound can be used to obtain a lower bound on the performance of the algorithms since the ratio of their objective value achieved to the upper bound could be even larger if the actual optimal score replaced the upper bound in the ratio. Recall that larger ratios correspond to better approximations.

Figure 3.5 shows the fraction of the upper bound achieved by the algorithms on instances with simulated sources and targets. Note that even for a fixed number of sources and targets, the number of possible paths in the network varies greatly due to network topology. We observe that for those instances that yield fewer paths, the best approximation algorithm either achieves the optimal value or finds an orientation with value greater than 99% of the upper bound. Even in the worst case we encountered, the best ratio achieved is greater than 0.7, which is far better than the $\frac{k}{2^k} = \frac{5}{32} \approx 0.16$ best known theoretical guarantee of the MAX-k-CSP algorithm.

The benefit of local search varies greatly by algorithm and by the number of paths. As expected, the random orientations without local search perform much worse than the orientations after search. For the smaller instances and one larger instance with roughly 50000

paths, the MIN-SAT algorithm obtains an excellent orientation without search. However, in the worst instance local search improves the MIN-SAT score nearly twofold. Of all three algorithms, MAX-CSP is the top performer without local search, and search does little to improve its orientations. This is not surprising because its underlying solver already uses an internal search-based strategy.

Interestingly, all three algorithms achieve quite similar ratios after local search across all instances we tested, to the extent that their respective points on the plot oftentimes overlap. This suggests that in practice the local search itself is more important when finding an optimal orientation than the actual algorithm used to obtain the starting point for the local search.

3.3 Evaluating algorithms using gold standard pathways

To confirm that the orientations produced by our algorithms not only achieve good approximation ratios but also produce biologically meaningful results, we compared the paths in the oriented networks with all yeast signaling pathways from KEGG [99] and the *Science Signaling* Database of Cell Signaling [70] using actual sources and targets to define the endpoints of the paths in our objective function. Only proteins without parent nodes in the diagram were chosen as sources. Any protein that was downstream of the sources was allowed to be a target, although preference was given to those proteins without children in the graph (see [65] for details of the gold standard, sources, and targets). To assess the soundness of the assumptions upon which our MEO formulation is based, we included two other methods for discovering pathways in the evaluation. The first is the reachability-focused MTO algorithm [137]. The second is the undirected edge selection algorithm by Zhao *et al.* [228]. Zhao *et al.* directly evaluated their technique against other notable undirected signaling pathway prediction algorithms [18, 176, 188] and showed that it compares favorably. Thus we consider it to be representative of the general class of undirected methods.

Because the oriented networks can contain thousands of paths connecting the source-target pairs, we needed a method for identifying which paths are most likely to be biologically meaningful. We tested several such methods including path weight; min, max, and average edge weight; min, max, and average edge use; and min, max, and average node degree. Edge use is the number of times an edge is a member of satisfied paths. Vertex degree is the sum of the in and out degrees. In our evaluation, we ranked all paths returned by the orientation algorithms using these criteria and calculated how many of the top 100 paths with 5 edges (containing exactly 6 proteins) are at least partially present in

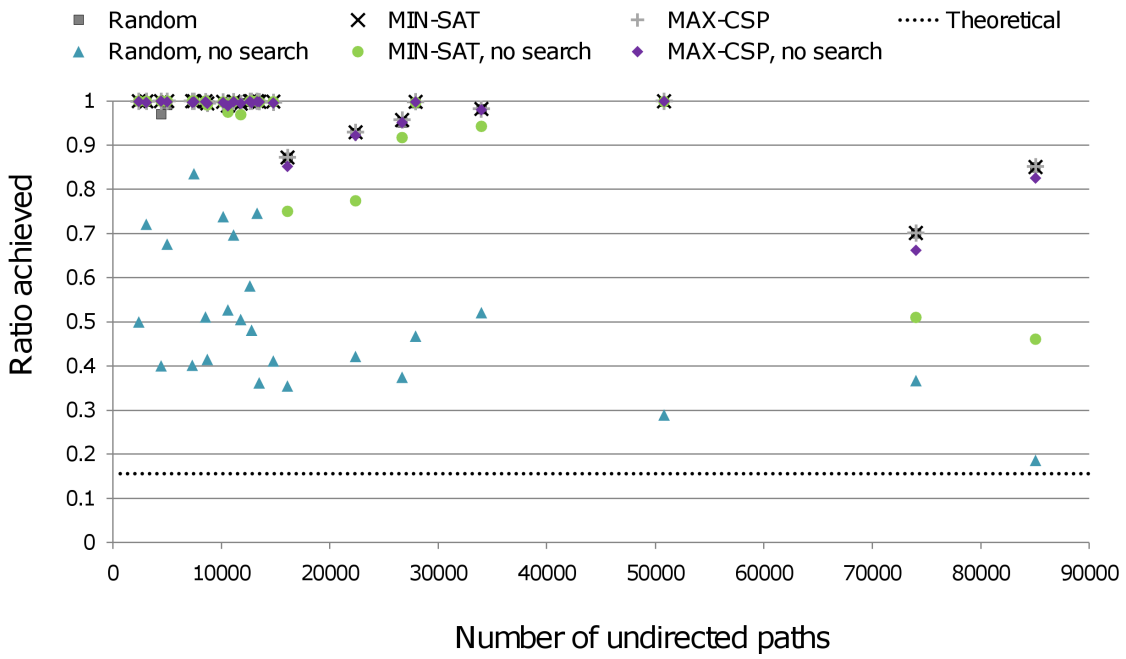

Figure 3.5: Fraction of the objective function upper bound achieved on instances with simulated sources and targets. After local search, all approximation algorithms perform much better than the MAX-k-CSP theoretical guarantee on instances with simulated source-target pairs and find orientations whose objective function values are virtually indistinguishable. The number of undirected paths includes all paths from a source to a target before the network is oriented. The y-axis plots the ratio achieved by each algorithm, which is the score of the orientation returned by the algorithm divided by the upper bound on the optimal objective function value. For each instance there are six points (one for each algorithm with and without local search) that have the same x-coordinate, the number of undirected paths, and different y-coordinates, the ratios achieved. Instances have been ordered along the x-axis by the number of distinct source-target paths in the network before orientation, which is a coarse indication of the difficulty of the instance.

a gold standard pathway. Partially present means that at least 4 of the 6 proteins are found *consecutively* in both the gold standard and a satisfied path returned by the algorithm. Table 3.1 summarizes the results of this evaluation. 40% of the top ranked paths discovered by the local search algorithm (following random orientation) are partially present in the gold standard when sorting by minimum edge use. Note that since the pathway databases are incomplete, the number of biologically valid pathways discovered is even larger (see Section 3.3.2).

Table 3.1: Number of top-ranked predicted paths that correspond to known signaling pathways. For each of the algorithms, all satisfied paths with exactly 5 edges (6 proteins) were ranked by various criteria. The table shows the number of the top 100 ranked paths that partially matched gold standard pathways.

Algorithm	**Path weight**	**Max edge weight**	**Avg edge weight**	**Min edge weight**	**Max edge use**	**Avg edge use**	**Min edge use**	**Max deg.**	**Avg deg.**	**Min deg.**
Random + search	37	11	36	34	0	0	40	10	0	0
MIN-SAT	2	0	2	1	0	0	0	1	0	0
MIN-SAT + search	33	9	32	28	0	0	40	10	0	0
MAX-CSP	14	7	14	16	0	0	16	3	0	0
MAX-CSP + search	7	5	6	7	0	0	16	3	0	0
MTO	3.2	3.2	3.2	3.2	3.0	3.0	3.0	3.0	2.8	3.2
Unoriented edge selection	20	20	20	20	20	20	20	20	20	20
Oriented baseline	9.5	4.3	9.8	7.5	0.4	0.2	3.2	4.6	0	0

We found that path weight, average and minimum edge weight, and minimum edge use are useful criteria for ranking pathways for most algorithms whereas vertex degree is a poor ranking criterion. Of the three edge use-based metrics, the minimum edge use is consistently the most informative. This demonstrates that predicted pathways that contain only edges that are critical to a large number of other satisfied paths correspond to the gold standard better than pathways that contain some edges that belong to many other paths and some edges that are isolated. The average and minimum vertex degree criteria yield top-ranked paths that generally do not match known signaling pathways because they consist only of paths that contain the highest-degree protein, Hek2, which is not known to be involved in our gold standard signaling pathways.

3.3.1 Orientation improves pathway identification

Surprisingly, although all three of our approximation algorithms achieved similar fractions of the upper bound on simulated instances (Figure 3.5), the fastest method we presented, random orientation followed by local search, is able to recover a far greater number of gold standard pathways in its top ranked paths than the CSP-based algorithm for all criteria used and performs as good as or better than MIN-SAT with search in all cases. Therefore, even though the MIN-SAT and MAX-CSP algorithms are interesting from a theoretical perspective, there is little reason to prefer them in practice over the random orientation with local search, which is much faster and can handle larger values of k. The benefits of local search are highlighted by the MIN-SAT algorithm, which performs drastically better when local search is applied. Unlike our algorithms, MTO and unoriented edge selection do not produce more biologically meaningful results after local search [65].

On average MTO finds only three pathways that partially match the gold standard no matter what ranking criteria is used. This reflects the different objective of MTO. Because it attempts to connect source-target pairs with paths of arbitrary length, very few of the resulting paths are reasonably short. In fact, in many runs we found that the MTO-oriented network did not even contain 100 source-target paths with exactly 6 proteins, whereas our algorithms find thousands of such paths. For the minimum edge use ranking criteria, our random orientation with search discovers thirteen times as many known pathways as MTO.

Our evaluation also highlights the weaknesses of the undirected edge selection algorithm, which can only identify 20 paths in the gold standard regardless of the ranking criteria used. This is only half of what our random orientation with search discovers when ranking by minimum edge use, and demonstrates that crucial network edges can be overlooked when subnetworks are selected without regard to edge orientation. In fact, the unoriented edge selection method discarded so many of these relevant edges that it found less than 100 source-target paths containing at most 6 proteins, which is why its evaluation was not affected by the ranking criteria used. These results strongly indicate that the unique edge orientation constraint utilized by our algorithms helps improve the quality of the pathways these methods recover.

As a control, we also calculated how many gold standard pathways could be recovered by random orientations *without* local search, which we refer to as "Oriented baseline" in Table 3.1. We found that on average no more than 10% of the top ranked pathways were present in a gold standard pathway for any of the ranking criteria, which is much lower than the results when random orientations are followed by local search.

3.3.2 Literature search validates additional orientations

Given the success of the methods in recovering known pathways, we asked whether the novel pathways that ranked highly according to our criteria may also be correct and represent information that is missing from current databases. We divided the pathways predicted by our random orientation with local search algorithm into three groups and analyzed the top 20 pathways in each group using the path weight for ranking. The first (Figure 3.6A) contains pathways of 5 or 6 proteins that were present, in their entirety, in the signaling databases. The second (Figure 3.6B) are pathways predicted by our method that consist of exactly 6 proteins and partially overlap a known pathway. For these we asked whether the additional interactions may represent known or sensible extensions to the pathway that were not previously known or were not recorded in the databases. The third (Figure 3.6C) are pathways discovered by our method that do not match any known pathways in the databases. For these we asked whether they represent known pathways not in the databases or novel hypotheses that make sense biologically.

In all three figures, we merged overlapping linear paths discovered by our algorithm. Our algorithm's predictions can be easily merged in this manner to form larger signaling networks because each edge is oriented uniquely in all paths. This feature of our orientation algorithm demonstrates its advantages over undirected methods. In undirected approaches, although edges in a single predicted path have an implicit orientation because information is known to flow from source to target, these local orientations are not globally consistent across all predictions. Thus the predictions may either be considered in isolation or merged into less informative undirected networks (e.g. the pheromone response predictions by Scott *et al.* [176]).

The paths in Figure 3.6A can be found exactly as predicted in various mitogen-activated protein kinase (MAPK) pathways. Sln1→Ypd1→Ssk1→Ssk22→Pbs2 is a component of the high osmolarity glycerol (HOG) pathway. The filamentous growth pathway contains the cascade Msb2→Cdc42→Ste20→Ste11→Ste7. The remaining paths that begin at Rga1 or Ste50 and extend to Dig1, Dig2, Fus3, Ste7, and Ste12 are members of the pheromone signaling pathway.

For the partial match pathways (Figure 3.6B) we found evidence that many of their edges missing from the databases are in fact valid and that our algorithm discovered previously unknown variants of common signaling pathways. Some of these paths in the pheromone signaling pathway contain the edge Ste11→Ste5. In the evaluation summarized in Table 3.1, this edge was considered a mistake since in the gold standard it was oriented in the opposite direction. That orientation is based on a model in which Ste5, after being recruited by Ste4, mediates Ste20 phosphorylation of Ste11 by facilitating the com-

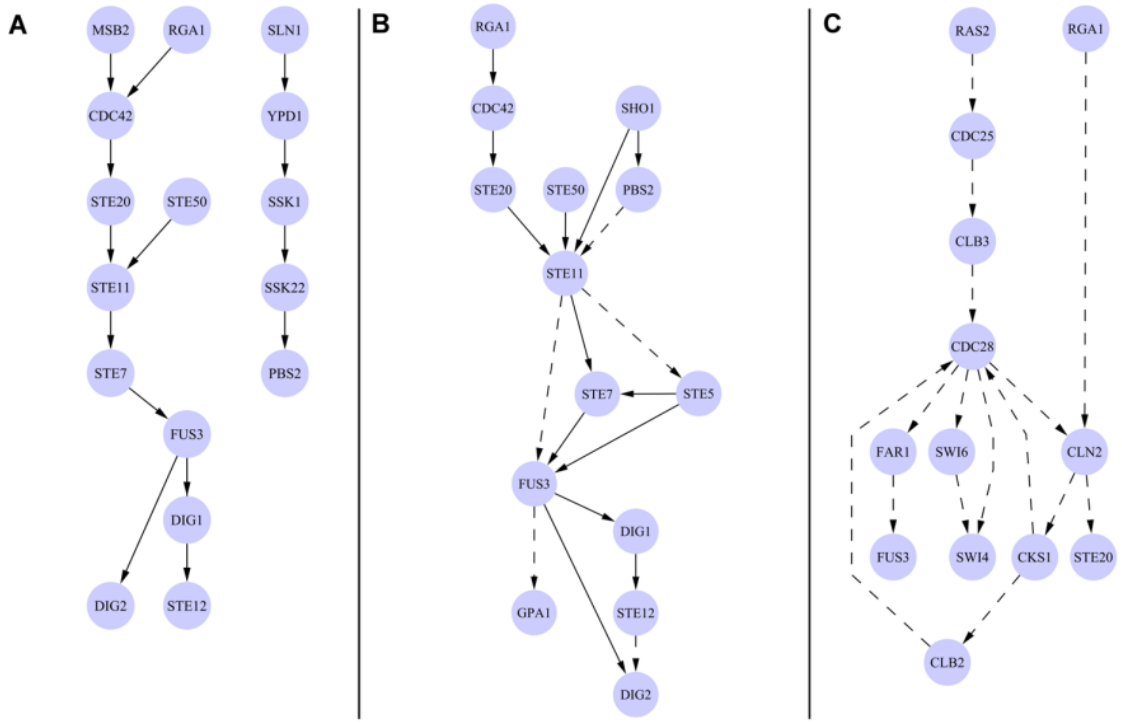

Figure 3.6: The top-ranked pathways discovered by the random orientation plus local search algorithm. Solid edges were present in the gold standard and dashed edges were absent or oriented in the opposite direction. A) Pathways that are completely contained within a known gold standard pathway. B) Pathways that partially overlap a gold standard path but contain new edges as well. C) Pathways that do not have any edges in common with our set of gold standard pathways. Images do not contain all of the top-ranked paths per category but rather a highly overlapping subset.

plex formation via its scaffolding function. However, it was shown recently that Ste5 and Ste11 already form a tight complex in the cytosol, in fact with the highest affinity (50nM) as compared to all other pairwise interactions between Ste5, Ste7, Ste11, and Fus3 [134]. Thus, our predicted Ste11→Ste5 edge is also valid. This interaction is included in a number of paths because there is redundancy in the function of some components downstream of this edge so several of the partial matches are in fact complete matches.

Another predicted interaction that disagrees with the direction in the gold standard database is Pbs2→Ste11. However, Pbs2 is a scaffold protein that simultaneously binds the osmosensor receptor Sho1, the upstream MAPK kinase kinase (MAPKKK) Ste11, and the downstream MAPK Hog1 [226]. Thus, even though Ste11 acts on Pbs2, its scaffolding function makes the edge direction ambiguous because formation of the signaling complex at Sho1 is required, and Sho1 and Pbs2 have therefore been termed "coscaffolds". Thus, drawing the edge in both directions is reasonable.

A particularly interesting prediction is the edge Fus3→GPA1, which was not found in the gold standard database. GPA1 is the G→ protein that is activated by pheromone stimulation of the membrane receptors which are G protein coupled receptors. Thus, GPA1 is located close to the top input level of the pathway and is a critical step in mediating the sequence of six consecutive intracellular events leading to Ste12 activation. Recently, it was found that there is a feedback loop from Fus3 (the kinase that directly activates Ste12) to GPA1 to Ste4 (another subunit in the heterotrimeric G protein complex), which is phosphorylated by Fus3 and negatively regulates the pathway [138]. Thus, the predicted Fus3→GPA1 edge is supported by this experimentally demonstrated feedback loop.

While most of the orientation results either agree with the gold standard orientation or with recent studies, we found two cases where the orientation determined by the algorithm is likely wrong. The first is the Ste11→Fus3 edge where both partners are part of the same macromolecular complex but the logic progression of the signal requires another partner in the complex. The second is the Ste12→Dig2 edge where again a third protein is involved in the communication of signal. Thus, in both cases the complex membership may confuse the algorithm by creating "shortcuts" that are not biologically meaningful.

Due to the strict requirement that all directed edges must be present consecutively in a *single* gold standard pathway in order to be considered a complete match, some of our partial match pathways actually agree with the gold standard on all individual edge orientations. For instance, one top ranked path Sho1→Ste11→Ste7→Fus3→Dig1→Ste12 predicts the correct orientation for each edge. However, it is not labeled a complete match because the Sho1→Ste11 edge is a member of the gold standard HOG pathway whereas the other four edges are found consecutively in the gold standard pheromone signaling pathway.

In our analysis of the predicted paths that do not overlap with any of the database-derived pathways, we found many edges that are either known or raise interesting biological hypotheses. Figure 3.6C depicts 9 of these paths that are cell cycle-related. Three of the cell cycle pathways originate at Rga1, a regulatory protein important for cytokinesis (end of M) and bud site formation. It is known to interact with Cln2 [7]. Cks1 activates Cdc28 [193] and sends the M cyclin Clb2 to degradation [97]. Cks1, Cln2, and Cdc28 form a complex [60], and Cdc28 complexes phosphorylate many proteins in the G1/S transition, including Swi4 [5] and Swi6 [62] in a regular cell cycle, Ste20 [214] in a mating response and during filamentous growth, and Far1 [24] in response to alpha factor. Another cascade starts with Ras2 and Cdc25 instead of Rga1. These proteins work together and are important for the exit from a G0 state [53]. Along with Cdc28 they allow the G1/S transition by increasing Cln2 levels [45]. Both Clb2 and Clb3 regulate Cdc28 activity and are expressed in the G2 and late S phases, respectively [83]. In summary, there is strong evidence for 10 of the predicted orientations that are missing from the gold standard, demonstrating that the true accuracy among our top-ranked pathways is greater than that indicated by our gold standard evaluation (Table 3.1).

3.4 Motivation for orienting all protein-protein interactions

In some cases it may be ideal to leave certain PPI in the network undirected. However, in practice, orienting the entire network does not affect our ability to correctly discover signaling pathways due to the nature of the interaction datasets we use. In general, when a complex interacts with some external protein, all (or most) members of the complex are shown as interacting with that protein in PPI databases. This is a consequence of the high throughput studies (for example pull-down assays) that often cannot distinguish between direct and indirect interactions. Thus, any orientation of the internal edges between complex members is appropriate because external proteins that interact with the complex are connected to both endpoints of the internal edges.

Several of our predicted cell cycle paths (Figure 3.6C) demonstrate how our orientation of edges in a complex can correspond to the biological truth. Clb2 and Clb3 each form a complex with Cdc28, yet orienting the edges Clb2→Cdc28 and Clb3→Cdc28 is justified because these two proteins are also reported to activate Cdc28. In addition, the edges Cdc28→Cln2→Ste20 represent the Cdc28-Cln2 complex mediating Ste20 even though this predicted path does not contain a direct Cdc28→Ste20 edge.

In fact, allowing our algorithms to leave certain edges undirected is not a viable option

given our problem formulation. Orienting an undirected edge can only reduce the number of satisfied paths in the network and correspondingly lower the objective function. Thus the optimal solution for all instances would be to not orient any edges. One reasonable way to overcome this would be to penalize solutions for including undirected edges, but setting the parameter that controls the tradeoff between the objective function's penalty term and path weight term would require much more training data (i.e. known signaling pathways) than is currently available.

3.5 Predicting missing signaling pathway edges

Our analysis of yeast MAPK signaling pathways demonstrated that using the MEO formulation to orient PPI networks with respect to source and target proteins can adeptly reconstruct the known pathways. However, one drawback of MEO is that our ability to recover these pathways is limited by the coverage and accuracy of the underlying PPI network. Even for model species, only a fraction of true physical interactions are known [76, 87]. Furthermore, interactions may be condition- and tissue-specific [26], but current experimental methods often focus on one condition and one cell type [113]. Missing even a small number of crucial edges in the PPI network can substantially affect the accuracy of the predicted pathways if those edges are involved in multiple source-target paths.

As reviewed in [181, 183], many methods have been proposed to computationally predict PPI. These techniques leverage a variety of data sources including protein structure, orthology, gene expression, literature mining, sequence, or a combination of heterogeneous features to learn a predictive model or classifier. Network-only approaches range from completing defective cliques [224] to embeddings of the network to find non-interacting but adjacent proteins in the new space [36, 115] to analyses based on the shared topology or the distance between two candidate proteins [146]. None of these approaches, however, focuses on specific pathways and they do not leverage known sources and targets to make pathway-aware predictions. Further, most other approaches use local cues of similarity, whereas our approach attempts to optimize a global distance function. There has also been theoretical work on predicting "shortcut edges" in graphs to minimize the average shortest path distance amongst all nodes in the graph [139] or the diameter of the graph [43, 126]. These works also do not exploit sources and targets and thus it would be difficult to interpret their predictions from the perspective of pathway requirements.

Therefore, we developed an algorithm that seeks to add k directed interactions to the oriented PPI network so as to maximally decrease the shortest path distances between sources and targets. The original formulation (Shortcuts) adds these so-called "shortcut"

edges in the network without explicitly constraining the length of the best source-target path. We compare this approach with a variant (Shortcuts-X) in which path lengths are restricted as they are in MEO.

3.5.1 Formalizing the Shortcuts problem

The input to the Shortcuts and Shortcuts-X formulations resembles that of MEO in many ways. We assume we are given a *directed* protein interaction network $G = (V, E)$, in which all edges either have a predetermined biological orientation or have been oriented algorithmically. Each edge is again weighted by a value $\in [0, 1]$ denoting our confidence in the interaction. We also assume we are given a set of sources S and targets T, the same sources and targets used to orient the network if it was not fully directed initially. Our goal is to predict missing (directed) edges that lie centrally "in-between" the sources and targets. These edges putatively belong to the pathway, but are not present in current databases. Formally, in the Shortcuts problem we wish to add k directed edges to E to minimize

$$\sum_{t \in T} \sum_{s \in S} d(s, t)$$

i.e. the average shortest path distance between the pairs.

To measure the distance $d(u, v)$ between proteins u and v in the weighted network, we use the shortest path metric (as opposed to other distance measures, such as those based on random walks [121, 198]) because the shortest path represents a direct and specific series of high-likelihood signaling events. Note that the shortest path between two nodes in a weighted graph can be very long (either because the diameter is long or if many edges along the path are of high confidence). As observed in Chapter 1, this may not be biologically reasonable since pathway targets are typically no more than 5 edges away from their closest sources. Thus, we also propose a hop-restricted version of our problem. Let $d_r(s_i, t_i)$ be the shortest path distance between s_i and t_i that uses at most r edges ($d_r(s_i, t_i) = \infty$ if no such path exists). As before, in Shortcuts-X we add k directed edges to E to minimize

$$\sum_{t \in T} \sum_{s \in S} d_r(s, t)$$

i.e. the length-restricted average shortest path distance. In [145] we prove that both edge prediction problems formulated above are NP-hard using reduction from exact cover by three sets and explore additional variants of the objective where the same source does not need to regulate all targets, but every target is regulated by at least one source.

Given these hardness results, we consider a heuristic greedy algorithm for our collection of edge prediction problems. The greedy algorithm selects k edges to add iteratively; in each step, it predicts a single edge that maximally reduces the objective function. However, unlike the greedy step in the MEO algorithm that chooses one of m edges to flip, the greedy Shortcuts algorithm must evaluate all possible $n(n-1)-m$ non-existent edges (excluding self-loops), where n is the number of proteins. To make the search more efficient we pre-compute the distances from each source to every other node and from every node to each target, which allows us to recompute the shortest path lengths from each source to each target quickly when evaluating a possible edge addition. Similar optimizations can be made for the hop-restricted version [145].

3.5.2 Extending the yeast HOG pathway

To evaluate our edge prediction algorithm and compare the objective function variants, we focus on the yeast osmotic stress response. This response is primarily mediated by the high osmolarity glycerol (HOG) pathway, whose core component is the MAPK Hog1. Its main physiological function is to counteract the effects of increased osmolarity such as water loss and cell shrinking [79]. Whereas the orientation evaluation in Section 3.3 included all yeast MAPK pathways and relied solely on the KEGG [99] and *Science Signaling* Database of Cell Signaling [70] resources for sources and targets, we now incorporate sources and targets from HOG literature [42, 150, 152]. In addition, when orienting the network we only required edge weights for the known PPI, which were previously derived from BioGRID [187]. However, BioGRID does not provide a way to weight possible connections between pairs of proteins that are not known to physically interaction. To weight such potential edges we would like to leverage other data sources (such as expression, sequence, and literature evidence) when making when assessing their likelihood. To naturally integrate these resources into our framework, we turn to the STRING [190] database, which catalogs PPI as well as the above relationships between proteins pairs. Instead of making predictions from amongst all possible edges (all pairs of proteins that are not in the set of STRING PPI), we only consider an edge if it exists within the set of STRING potential edges, edges with some known functional relationship. Each PPI and potential edge is weighted by STRING with a confidence value in $[0, 1]$, which we explicitly set to w_{uv} when considering the benefit of an edge. By using these data types and weights together, we can pinpoint putative interactions that have evidence from a wide variety of biological sources as well as evidence from the network. We oriented the STRING network using the random algorithm with local search and predicted the 10 edges that most improve source-target connectivity in the HOG pathway. To verify that the network ori-

entation algorithm would provide a high-quality initial directed network, we computed the percentage of KEGG and *Science Signaling* HOG pathway edges that were oriented correctly. Of the 16 KEGG edges, 9 existed in the STRING PPI network and 7 of these (77.8%) were oriented correctly. Similarly, of the 42 *Science Signaling* edges, 29 existed in the STRING PPI network and 18 of these (62.1%) were oriented correctly.

Table 3.2 presents the top 10 predictions made with the Shortcuts objective function, many are known physical interactions missing from STRING. Two edges (the first and eighth predictions) have direct evidence of physical interaction according to BioGRID but were not present in the STRING network. The second and tenth predictions lie within the STRING binding edges (and thus represent physical interactions), but were either oriented in the opposite direction or were left out of the oriented network. These correct predictions demonstrate that our approach can correct for limitations of the edge orientation. Prp19→Sto1 was originally oriented Sto1→Prp19, but our algorithm suggests that that this edge was either oriented incorrectly or is bidirectional. The orientation algorithm did not find any length-bounded paths that include the edge Reg1→Tpk1 so this edge was excluded from the oriented network. Although in general biological pathways are short, this prediction exemplifies an exception where considering longer pathways through the edge Reg1→Tpk1 improves the source-target connectivity.

For the following three predictions, we verified both the physical interaction between the two nodes and the directionality (which is not possible for edges validated with the undirected BioGRID database). The sixth prediction (Msn4→Msn2) involves two general stress TFs that play a substantial role in the HOG pathway [30]. Harbison *et al.* [75] showed that Msn4 indeed binds the *MSN2* gene in the succinic acid stress condition. This study did not profile Msn4 DNA binding in osmotic stress, but it is plausible that this stress-activated TF could bind *MSN2* in other conditions as well. The seventh prediction (Hog1→Cin5) was recently shown by Pokholok *et al.* [162] to occur in osmotic stress. We discuss the fourth prediction (Tpk2→Sok2) at length in the next section.

Overall, 7 of the top 10 predictions have support for direct physical binding in the cell. In addition, the fifth prediction was not directly supported in the literature but warrants further study. Both Reg1 and Msn4 have been shown to physically associate with the 14-3-3 proteins Bmh1 and Bmh2 [98] but have not yet been shown to directly interact with one another. Proteins with a common physical interaction partner may be more likely to directly interact themselves than proteins with other types of functional connections (e.g. genetic interactions) [10, 146, 224].

Table 3.3 presents the top 10 predictions made when using the Shortcuts-X objective function, which attempts to model more biological constraints by imposing a path length-restriction on the source-target paths. Remarkably, the top three predictions (Hog1→Msn2,

Table 3.2: Top 10 predictions using the Shortcuts objective. The original value of the objective function (score) was 12.91. The Src and Tgt columns indicate the direction of the predicted edge. The markers (s) and (t) imply that the protein was an original HOG source or target, respectively. The weight of the edge comes from STRING. Predictions for which there is evidence of direct, physical interaction are highlighted in gray with comments.

#	Src	Tgt	Score	Weight	Comments
0	—	—	12.91	—	Original score
1	Hkr1(s)	Syf1	11.63	0.998	Physical interaction in BioGRID [PCA high-throughput]
2	Prp19	Sto1	10.13	0.999	Oriented in opposite direction; BioGRID [Affinity Capture-MS]
3	Ssk1	Sho1	9.12	0.999	Only indirect interaction reported; two different HOG input paths
4	Tpk2	Sok2(t)	8.19	0.996	We studied experimentally (see Section 3.5.3)
5	Reg1	Msn4(t)	7.35	0.999	Indirect partners; both physically interact with Bmh1/2 [98]
6	Msn4(t)	Msn2(t)	6.63	0.999	Msn4 binds Msn2 in succinic acid [75]
7	Hog1	Cin5(t)	6.06	0.872	Hog1 binds Cin5 in osmotic stress [162]
8	Bem2	Cdc42(s)	5.72	0.998	Physical interaction reported in BioGRID [Biochemical activity]
9	Msb3	Yap6(t)	4.93	0.915	Only indirect interaction reported
10	Reg1	Tpk1	4.77	0.999	STRING binding edge but left out of orientation

Hog1→Msn4, and Hog1→Cin5) represent best-case predictions: The two genes/proteins involved are known to physically interact, the directionality is correct, and the interaction is highly relevant to osmotic stress response. In particular, Hog1→Msn2 and Hog1→Msn4 are core HOG pathway interactions that are well-characterized [30] and appear in gold standard databases [99], but lack evidence for physical binding in STRING. Indeed, the MAPK Hog1 is central to the HOG response program, and its activation of downstream TFs is a critical component of the response. The other two validated predictions involve HOG pathway members as well. Sho1 is a transmembrane osmosensor, and its branch of activation of Hog1 is known to be mediated by interaction with Cdc42 [194]. The Sho1→Cdc42 interaction is also present as part of the related starvation subpathway of MAPK in KEGG. Similarly, the tenth prediction (Ste50→Cdc42) is between two members of the Sho1 HOG pathway input branch [42]. Overall, of the 659719 STRING potential edges considered, only 0.0011% are in KEGG, and thus the fact that three of the top

ten predicted edges can be validated using KEGG is highly significant (p-value 8.96E-14, Fisher's exact test).

Table 3.3: Top 10 predictions using the Shortcuts-X objective.

#	Src	Tgt	Score	Weight	Comments
0	—	—	18.24	—	Original score
1	Hog1	Msn2(t)	15.93	0.968	Hog1 activates Msn2 in osmotic stress [30]; KEGG
2	Hog1	Msn4(t)	14.34	0.962	Hog1 activates Msn4 in osmotic stress [30]; KEGG
3	Hog1	Cin5(t)	12.76	0.872	Hog1 binds Cin5 in osmotic stress [162]
4	Hkr1(s)	Ste20	11.96	0.802	Only indirect interaction reported
5	Sln1(s)	Ptc1	11.31	0.968	Only indirect interaction reported
6	Msb3	Yap6(t)	10.82	0.925	Only indirect interaction reported
7	Sho1	Cdc42(s)	10.08	0.965	Cdc42 required for Sho1-activation of Hog1 [194]; KEGG
8	Sln1(s)	Sho1	9.72	0.959	Only indirect interaction reported; two different HOG input paths
9	Cla4	Swi4	9.32	0.983	Only indirect interaction reported
10	Ste50	Cdc42(s)	8.64	0.989	Oriented in opposite direction; BioGRID [Complex, Y2H]

Other predictions whose physical interaction could not be validated also involve pairs of HOG pathway members. Some predictions occur between the two independent upstream input branches in the pathway (e.g. Ssk1→Sho1 and Sln1→Sho1) or between upstream proteins and proteins that are very far downstream (e.g. Sln1→Ptc1). From an algorithmic standpoint, these edges do indeed provide faster diffusion of signal from sources to targets; however, they may not represent direct interactions that occur in the cell. In contrast, the Hkr1→Ste20 prediction is a shortcut within the Sho1 input branch, which contains the cascade Hkr1→Sho1→Ste20 [42]. Note that several of these predicted edges (e.g. Ssk1→Sho1) have very high weights from STRING reflecting their strong functional dependencies, which makes them more likely to be selected by our algorithm.

In general, the hop-restricted algorithm tends to select central nodes through which much signal flows (e.g. Hog1). The non-hop-restricted algorithm may induce alternative longer paths that circumvent these hubs. This may explain the different strengths we observed for Shortcuts and Shortcuts-X above. Shortcuts made more predictions whose physical binding could be verified than Shortcuts-X (7 versus 5). However, Shortcuts-X made more condition-relevant predictions between two HOG pathway members (8 versus

3). As intended, the hop-restricted variant does yield more condition-specific results. Note that for both objectives, not every edge had the highest possible confidence in STRING (0.999). Indeed, several predictions were made despite lower evidence, which suggests that their addition strongly reduced source-target distances.

3.5.3 Tpk2's interaction with Sok2

To demonstrate our approach's ability to make novel, biologically meaningful predictions we selected Tpk2→Sok2 for experimental validation. This edge was the second uncharacterized prediction made when using the Shortcuts algorithm (Table 3.2), but is more suitable for experimental follow up than the other higher-ranking prediction, Ssk1→Sho1. Ssk1 and Sho1's relationship is well-studied, and these proteins are known to participate in the distinct input branches of the HOG pathway making them unlikely to physically interact.

Verifying a directed protein-protein interaction at the mechanistic level requires extensive experimentation and is beyond the scope of this work. However, gene knockouts can be used to establish condition-specific causal relationships between two proteins in a putative signaling pathway. If Tpk2 controls the TF Sok2 in osmotic stress, *TPK2* deletion should affect Sok2's regulatory activity and its bound gene targets. Because many interactions along signaling pathways occur post-translationally, we would not expect the *SOK2* gene to be differentially expressed in the *tpk2*Δ mutant even if Tpk2 does activate or inhibit Sok2 at the protein level. Instead we determine the degree to which the deletion alters Sok2's function as a transcriptional regulator. We used microarrays to quantify the effects of *TPK2* deletion in sorbitol, a hyperosmotic medium (see [145] for experimental details). In order to isolate the specific effects of the deletion in osmotic stress and ignore general knockout effects, we removed genes that do not respond to osmotic stress in wild type cells [59] from our analysis. As predicted, the knockout significantly affected genes bound by Sok2 (p-value 9.40E-3 using Fisher's exact test) — 37 of Sok2's 168 binding targets (22%) [133] were differentially expressed. One would not expect that all Sok2's target genes are affected by the knockout because Sok2 is still connected to the sources via alternate parallel pathways. The knockout alone cannot confirm whether the Tpk2→Sok2 interaction is direct or indirect, but clearly establishes that there is a functional connection between these proteins that is active in osmotic stress. Moreover, the orientation of the predicted Tpk2→Sok2 edge is correct because if Sok2 were upstream of Tpk2 in the pathway, its bound genes would be unaffected by *TPK2* deletion.

Related literature gives additional context to this prediction and suggests the Tpk2→Sok2 interaction may in fact be direct. Tpk1, Tpk2, and Tpk3 form the catalytic subunit of pro-

tein kinase A (PKA), the complex at the heart of the Ras/cAMP/PKA signaling pathway [225]. Through interactions with its many substrates, PKA is involved in general stress response, metabolism, growth, ribosome biogenesis, and various other biological processes [225], including osmotic stress response. PKA's involvement in the osmotic stress response is parallel to the HOG pathway [165]. Msn2, Msn4, and Sok1, which along with Hot1 are considered to be the primary HOG pathway TFs [30], are each affected by PKA in osmotic stress [69, 165]. Decreased PKA activity modulates the repressive effects of Sok1 in this condition. This behavior is complementary to Hog1's phosphorylation of Sko1, which also alleviates Sko1 repression of its target genes [165]. While Tpk2's role in osmotic stress is well-established, Sok2 is not considered to be a core HOG pathway TF, but was rather assumed to be controlled by the primary TFs [152]. However, genetic screens illustrate that its role in the osmotic stress response may be larger [78, 222] and our own computational analysis supports this role (Section 4.3.1).

Our *TPK2* knockout establishes a functional link between Tpk2 and Sok2 in which Sok2 is downstream of Tpk2. A previous genetic interaction reported by Ward *et al.*, who suggested that PKA may directly phosphorylate Sok2, supports this directionality and relationship [208]. Subsequent experiments confirmed that active PKA phosphorylates Sok2 when glucose is the carbon source [180]. However, this link does not appear in other conditions. For example, Sok2 was found to function in a pathway parallel to PKA [156] and Tpk2 [135] in pseudohyphal growth and adhesive growth, respectively. In addition, Tpk2 does not interact with Sok2 in a mutant yeast strain that is sensitive to exogenous cAMP [157]. These findings highlight the importance of pathway-specific predictions of missing interactions as opposed to general protein interaction predictions.

Coupled with previous evidence that PKA can directly phosphorylate Sok2, our knockout results suggest that the proposed Tpk2→Sok2 interaction warrants further detailed experimental validation. Because the Tpks have distinct sets of substrates [166] despite their high sequence similarity, confirmatory future work must also include experimentally establishing which of the PKA subunits phosphorylates Sok2.

Chapter 4

Signaling and Dynamic Regulatory Events Miner (SDREM)

In Chapter 3 we demonstrated that our MEO formulation and random orientation with local search were highly successful in generating biologically valid directed pathways when given the endpoints involved in the stress response. For many conditions of interest, the upstream sensory proteins that initiate the response have indeed been discovered; however, it is difficult to experimentally determine which targets (TFs) are active downstream. Here we present the Signaling and Dynamic Regulatory Events Miner, an iterative algorithm that uses temporal condition-specific gene expression data and a network of condition-independent physical interactions to infer the set of TFs that are both actively controlling genes and well-connected to the upstream sources. By applying our algorithm to well-studied yeast stress responses we are able to quantitatively assess its ability to recover known pathways, demonstrate its utility for expanding pathways via novel predictions, and characterize its strengths and weaknesses.

4.1 Related work

There are many approaches for integrating signaling and regulatory networks that rely on knockout data, some of which were previously discussed in Section 3.1. Both PNM [217, 218] and SPINE [155] explain knockout cause-effect pairs via chains of physical interactions. One innovative application of PNM used deletion buffering events, genes that are typically differentially expressed in a stress condition but unaffected by the condition after a knockout, to define the cause-effect pairs [213]. Peleg *et al.* [158] proposed a

"network-free" approach to the problem of explaining knockout cause-effect pairs, which does not depend on enumerating pathways between knockouts and their targets in the physical network. Instead, their algorithms operate on a functional network, which contains edges between deleted genes and their affected targets, and only annotate the edges in the physical interaction network as an optional post-processing step. Nested Effects Models [136] and their extensions, including those that account for network dynamics [6], are another popular approach for analyzing knockouts and other types of perturbations. All of these knockout-dependent methods are potentially vulnerable to the effects of redundancy in regulatory networks described in Section 2.3. Vinayagam *et al.* avoid the complications of knockouts by orienting PPI with respect to shortest paths between all membrane receptors and TFs [203], but their approach relies on general topological features and does not reveal the pathways or regulators most relevant to a specific response

Other methods have used additional types of perturbations as starting points, including data from genetic screens. However, as we demonstrate in Section 5.2.2, even if we ignore the effects of redundancy, screens can be ill-suited for defining the source nodes in the network because they may not detect the most upstream members of the response pathways. Motivated by the vast discrepancy between hits in genetic screens and genes that are differentially expressed in response to a stimulus, ResponseNet [219] combines these two types of data to generate integrated signaling and regulatory networks for a condition of interest. A related approach for combining these complementary datasets uses the genetic hits and differentially expressed genes as the relevant "terminal" nodes in the network [89]. Connections between these nodes were identified using a prize-collecting variant of the Steiner tree problem, which does not yield a fully oriented network as our method does. This algorithm was also successfully applied using proteins with differentially phosphorylated sites in place of the genetic screen hits. Nevertheless, this algorithm and others derived from the Steiner tree problem [12, 221] do not model redundant and parallel pathways to the target nodes.

In addition to techniques for integrating signaling and regulatory networks, several approaches have been proposed to reconstruct dynamic regulatory networks [50, 57, 128, 185, 209] (see [66] for a review). These either focus on the regulatory network component exclusively, and thus do not explain how the TFs regulating the response are activated, or only utilize known (database-derived) pathways. For instance, Wei and Li [209] apply a hidden spatial-temporal Markov random field to incorporate upstream pathways into their analysis of dynamic gene expression data. The graphical model determines which genes in the expression dataset are temporally differentially expressed. In addition to the edges representing temporal dependencies, variables are connected in the Markov random field if their corresponding genes are neighbors in some pathway such that adjacent genes are

more likely to have similar expression states. However, their method relies on fixed known pathways as opposed to inferring which pathways are involved. Dependence upon known pathways limits the applications of this class of algorithms to well-studied networks and species and prevents them from providing new predictions regarding the members and interactions in the signaling network portion of the integrated model.

Boolean networks [2, 34] are another rich class of models that can be used to examine the dynamics of integrated signaling and regulatory networks. In a typical formulation, nodes in these networks represent mRNAs and proteins, and each node has a binary state indicating its presence or activity [2]. States are updated over time as dictated by Boolean logic applied to the states of parent nodes. Extensions allow distinct time scales for different types of biological interactions and processes (e.g. models in which post-translational modifications and protein complex formation occur more quickly than transcription, translation, and degradation) and asynchronous updates, leading to even more realistic simulations of network dynamics [34]. As a result, Boolean networks represent a more detailed model of dynamic biological networks than SDREM. SDREM identifies important proteins, the pathways involving them, and the timing of TF regulatory activity but not the temporal activity of signaling proteins and the logical control governing these activities.

However, SDREM is more widely applicable to lesser known pathways and better suited for discovering which proteins are relevant to a stress response with minimal condition-specific data and prior knowledge. Constructing a Boolean network is typically a manual process [2, 9, 34] that requires an extensive literature search (and sufficient literature describing the pathway of interest) to specify the genes, proteins, and Boolean functions. There are methods for learning the logical functions of a Boolean network using techniques from model checking, but these still require that the nodes and edges are known in advance [120]. Approaches for learning the Boolean network's topology from biological data rely on gene expression to provide the activity of nodes, implicitly assuming that a protein must be differentially expressed in order to control its target gene or protein [143]. This unrealistic assumption leads to networks in which edges represent logical relationships but not direct physical interactions, thereby removing the detailed simulation capabilities and advantages of the manually constructed Boolean networks.

4.2 Reconstructing dynamic networks and orienting interaction networks

We developed a new method for integrating time series expression data with static protein-protein and protein-DNA interactions to infer dynamic regulatory networks and the sig-

naling pathways that activate these networks. In addition to the general interaction data, SDREM uses condition-specific time series data and a small set of proteins that are known to sense the environmental stress, interact with the infecting agent, or play some other role in initiating the response as input. In many cases such proteins are either known [99] or can be experimentally determined (e.g. in the response to viral infection [33, 46, 56, 148]).

4.2.1 SDREM overview

SDREM builds upon two previously developed methods, the Dynamic Regulatory Events Miner (DREM) [50] and our network orientation procedure (Chapter 3). DREM uses protein-DNA binding interactions and time series gene expression data to reconstruct dynamic regulatory networks by identifying bifurcation events, places in the time series where a set of genes that were previously co-expressed diverges. DREM annotates these split events with TFs that are predicted to regulate genes in the outgoing upward and/or downward paths allowing us to associate temporal information (the timing of the splits) with the static protein-DNA interaction data. An input-output hidden Markov model (IOHMM) [19], which unlike traditional HMMs also includes additional observed (in our case static) input data that can influence transition probabilities, is the underlying probabilistic graphical model. In DREM, protein-DNA interactions serve as the static input data that influence transitions between hidden states. An L_1-regularized logistic regression classifier is trained at all expression profile bifurcations to assign transition probabilities to genes based on the set of TFs that bind them. DREM searches the state space of possible splits in gene expression profiles to predict a compact set of diverging regulatory paths and the TFs that control them. It was successfully applied to reconstruct networks in a large number of species including yeast [50], *Escherichia coli* [48], fly [196], and human [72]. Although DREM identifies the active TFs, it does not explain what activated these TFs or consider whether the TFs are consistent with the signaling pathways involved in the response.

The second component is the network orientation algorithm presented in Chapter 3. The network orientation algorithm can complement DREM by linking the identified TFs to the source proteins in order to explain their activation. However, to accurately combine the two we need to address several computational challenges. First, DREM is a probabilistic model whereas the network orientation method solves a combinatorial optimization problem. Thus, values computed in one model cannot be directly transferred to the other. In addition, DREM is unable to account for the network connectivity of the TFs (i.e. prefer TFs that are well-connected to the upstream sources) because it considers all TFs to be equally likely to be active in the response. Similarly, some active TFs in the

DREM model may be implicated more strongly than others in the oriented network model, but the original TF enrichment scores DREM calculates cannot consider TF priors and are not compatible with the network orientation objective function.

To address these issues we developed SDREM, which iteratively combines the two methods. Figure 1.2 presents a high-level overview of SDREM. SDREM uses a modified version of DREM to infer the TFs that regulate genes as part of the response as well as the time at which this regulation takes place. The identified TFs become targets for the network orientation algorithm. The oriented network is then used to determine which of the target TFs are supported by the discovered signaling pathways. TFs that cannot be explained by the signaling network are penalized such that they are less likely to be selected in the subsequent DREM analysis. This process repeats until convergence, which leads to the final pathways and regulatory network.

In practice, unifying DREM and the network orientation algorithm requires overcoming the challenges described above. To address the issue of different model types we implemented a strategy that allows SDREM to incorporate prior (continuous-valued) information about the TFs during the analysis of the gene expression data (Section 4.2.2). In order to compute these TF activity priors, we developed a new method that assigns a posterior score for each target TF based on its dominance in the oriented network with respect to random targets (Section 4.2.3). To allow information flow in the other direction (from DREM to the network orientation method) we extended DREM so that it outputs an activity score for each TF at each regulatory path split. We further modified the orientation algorithm to use these scores to prioritize the targets. These new scores provide a set of TFs that are believed to be active as well as a quantitative measure of their activity level.

4.2.2 DREM extensions

To link the two methods we first extended DREM in order to make it suitable for our iterative approach. Originally DREM only accepted either binary input for TF-gene binding interactions or ternary input (-1, 0, 1) if the TFs are known to be activators or repressors. We generalized this to allow continuous TF activity priors. Initially all priors are set to 0.5, but in subsequent iterations modified priors are derived from the oriented network as described below.

The activity priors influence the transition probabilities in the IOHMM as well as the activity scores that DREM now calculates (see below). The activity score, which tells how well a TF explains bifurcation events in the gene expression data, is calculated for each TF at each bifurcation point in the gene expression profiles. A TF that explains bifurcation

events well is believed to play in active role in the organism's response to the stress. The activity score for TF t at a given bifurcation event e (also known as a split in the gene expression profiles) is defined as the likelihood ratio

$$score(t, e) = \frac{P(a = 1 \mid G_t)}{P(a = 0 \mid G_t)}$$

where $a = 1$ means the TF is active in the stress condition and G_t represents the expression profiles of the set of genes bound by TF t that are on the path into the split e. By applying Bayes rule and assuming that the expression profiles of bound genes are independent (a simplifying assumption that is unlikely to be true in real data) we obtain

$$\begin{aligned}
\frac{P(a = 1 \mid G_t)}{P(a = 0 \mid G_t)} &= \frac{\frac{P(G_t \mid a=1)P(a=1)}{P(G_t)}}{\frac{P(G_t \mid a=0)P(a=0)}{P(G_t)}} \\
&= \frac{P(G_t \mid a = 1)P(a = 1)}{P(G_t \mid a = 0)P(a = 0)} \\
&= \frac{\left(\prod_{g_i \in G_t} P(g_i \mid a = 1)\right) P(a = 1)}{\left(\prod_{g_i \in G_t} P(g_i \mid a = 0)\right) P(a = 0)}
\end{aligned}$$

Initially we place a uniform prior on all TFs, $P(a = 0) = P(a = 1) = 0.5$. In subsequent iterations, the prior is influenced by the network orientation such that TFs that are well-connected in the network have a larger prior.

To estimate the remaining probabilities, the set of bound genes G_t is divided into two sets, those genes that are assigned to the primary path out of the split $(g_i = 1)$ and those assigned to the secondary path(s) out of the split $(g_i = 0)$. The set of genes assigned to the primary path is denoted G_P and the set of genes on the secondary path is G_S. The primary path is the path out of the split followed by the majority of genes bound by the TF. In the case of a tie, the path with the fewest genes (regulated by any TF, not just t) that is involved in the tie becomes the primary path. All other paths out of the split are designated secondary paths and are considered as a single group. There will always be at least one secondary path because the TF activity score is only calculated at nodes in the model that

have two or more children. After splitting genes by the path they take, the score becomes

$$\frac{\left(\prod_{g_i \in G_t} P(g_i \mid a=1)\right) P(a=1)}{\left(\prod_{g_i \in G_t} P(g_i \mid a=0)\right) P(a=0)}$$

$$= \frac{\left(\prod_{g_i \in G_P} P(g_i=1 \mid a=1)\right)\left(\prod_{g_i \in G_S} P(g_i=0 \mid a=1)\right) P(a=1)}{\left(\prod_{g_i \in G_P} P(g_i=1 \mid a=0)\right)\left(\prod_{g_i \in G_S} P(g_i=0 \mid a=0)\right) P(a=0)}$$

$$= \frac{P(g_i=1 \mid a=1)^{|G_P|} P(g_i=0 \mid a=1)^{|G_S|} P(a=1)}{P(g_i=1 \mid a=0)^{|G_P|} P(g_i=0 \mid a=0)^{|G_S|} P(a=0)}$$

We assume that all bound genes respond to TF activity in the same manner and estimate that 80% of genes that are bound by a TF that is active in the stress condition are affected by the binding based on the activity of known stress TFs (Section 4.3.8). In other words, $P(g_i = 1 \mid a = 1) = 0.8$ and $P(g_i = 0 \mid a = 1) = 0.2$. When the TF is not active, i.e. $(a = 0)$, the probability that a gene will be affected by the binding is given by the background distribution. The background distribution is the percentage of *all* genes (not just the set bound by t) along each path out of the split. $P(g_i = 1 \mid a = 0) = \frac{|O_P|}{|O_P \cup O_S|}$, where O_P is the set of all genes that follow the primary path out of the split and O_S is the set of all genes on a secondary path out of the split. Note that O_P and O_S exclude genes that are on another path and do not enter the split. Likewise, $P(g_i = 0 \mid a = 0) = \frac{|O_S|}{|O_P \cup O_S|}$. Thus, the final activity score is

$$score(t, e) = \frac{0.8^{|G_P|} 0.2^{|G_S|} P(a=1)}{\left(\frac{|O_P|}{|O_P \cup O_S|}\right)^{|G_P|} \left(\frac{|O_S|}{|O_P \cup O_S|}\right)^{|G_S|} P(a=0)}$$

The TF activity score at a particular iteration of SDREM is the maximum score achieved over all possible bifurcation events e. Because TF activity scores can take arbitrarily large values, we normalize them before incorporating them into the network orientation objective function. Activity scores are normalized by taking a TF's percentile in the randomized distribution (below) and multiplying by k, the maximum path length in the network orientation.

To determine the significance of a specific activity score we use a randomization method. We run DREM multiple times (10 for all analyses here) with protein-DNA binding data that has been randomized. This generates a distribution of random TF activity scores from which we select the top 50 activity scores from each randomized run. All TFs with *real* activity scores in the 50th or greater percentile in this distribution are considered active, and these TFs are used as targets during the subsequent network orientation.

4.2.3 Network orientation algorithm modifications

We extended the network orientation algorithm so that in addition to the edges weights, it incorporates the target (TF) weights from DREM. Specifically, the modified path weight used in the MEO objective function is:

$$w(p) = w(t) \prod_{v \in p} w(v) \prod_{e \in p} w(e)$$

where p is a source-target path, t is the target on that path, v is a vertex on the path, and e is an edge on the path. For all vertices $w(v) = 1$ in the yeast analysis, and $w(t)$ is the normalized version of the activity score from DREM that ranges from 0 to k. We set $k = 5$ and based on the results in Section 3.3 we use random orientations followed by local search.

To calculate connectivity scores for the targets, which are used as priors in DREM, we add random targets. Each random target has $w(t) = 1$, and the number of random targets is equal to the number of real targets from DREM. The top $5T$ paths, where T is the number of real and random targets, are considered top-ranked paths and satisfied paths are ranked by path weight. A target's score is the sum of all path weights of satisfied top-ranked paths that end at that target. These target connectivity scores are averaged over 10 runs for the real targets, and the scores of the random targets in all 10 runs are used to create a target connectivity score distribution. Node connectivity scores are obtained similarly using a separate set of 10 orientations that do not include random targets. The node connectivity score calculation sums over all satisfied paths that include the node as opposed to paths that end at a particular target. The fraction of the top $5T$ paths that contain a particular node is the node's connectivity score.

Activity priors for the next iteration of DREM are then increased or decreased according to both the target and node connectivity scores. A TF's activity prior is increased if it meets either of two criteria: its target connectivity score is greater than 80% of the scores in the random distribution or its node connectivity score is at least 0.01. The new prior of these well-connected TFs is the average of 1.0 and the old prior. Any target that did not meet these criteria has its prior halved unless it would become less than the minimum value, 0.01, in which case the prior is set to 0.01. The binding priors of all other TFs that were not identified as active targets are not changed. Note that nodes with connectivity scores ≥ 0.01 also define the set of nodes that we predict to be signaling proteins, and in some cases TFs can be included in our model because of their node connectivity score even though they are not active on a regulatory path. Sensitivity analysis indicates that SDREM is robust to variations in the values of these and other parameters (Section 4.3.8).

4.3 Yeast stress response

To test SDREM, we first applied it to study the response of *Saccharomyces cerevisiae* cells to high osmolarity, which is primarily controlled by the HOG pathway (the same pathway we targeted for edge prediction in Section 3.5.2). We collected general protein-DNA [133] and protein-protein [187] interaction data. In addition, we used condition-specific protein-DNA binding data for Hot1 and Sko1 [30], source proteins (Cdc42, Msb2, Sho1, Sln1, and Ste50) from the *Science Signaling* Database of Cell Signaling [70], and two complementary time series gene expression datasets. The first expression dataset [172] measures gene expression up to 15 minutes leading to our short model. The second [59] is used to construct the long model (up to 90 minutes) because it includes the recovery phase of the response.

4.3.1 Osmotic stress models

We display the resulting networks in two parts corresponding to the signaling and regulatory components of the reconstructed networks. TFs serve as the interface between these two models, and some of the connections between the two components are highlighted in Figure 4.1. In the regulatory network part of the short model, there are 10 distinct paths controlled by a multitude of TFs (Figure 4.1A). Figure 4.1B presents the high-confidence paths leading from the sources to targets in the protein interaction network, including the inferred PPI orientation. The model predicts that proteins along these paths play an important role in the osmotic stress response. We emphasize that all targets in this network are the same active TFs that can be found along the short model regulatory paths (Figure 4.1A).

In a recent study of the transcriptional network activated by Hog1 [30], Capaldi *et al.* described how Hog1 directly controls the TFs Hot1, Msn2, Msn4, and Sko1 as part of the core component of the hyperosmotic response. SDREM successfully recovered this core component of the HOG response pathway (Figure 4.1C) even though none of these proteins were given as input and only *MSN2* is differentially expressed. The PPI Hog1-Hot1 and Hog1-Sko1 were both correctly oriented toward Hot1 and Sko1 respectively. Our interaction dataset lacked direct interactions from Hog1 to Msn2 and Msn4. However, our model incorporates Hog1's control over these TFs via indirect interactions through Sko1, showing its robustness to missing physical interaction data. The nodes and edges immediately upstream of Hog1 (Figure 4.1B) are consistent with HOG pathway literature as well. The edges Ste50→Ste11, Sho1→Ste11, Sho1→Pbs2, Ste11→Pbs2, and Pbs2→Hog1 compose the majority of the Sho1 input branch of the HOG pathway [110].

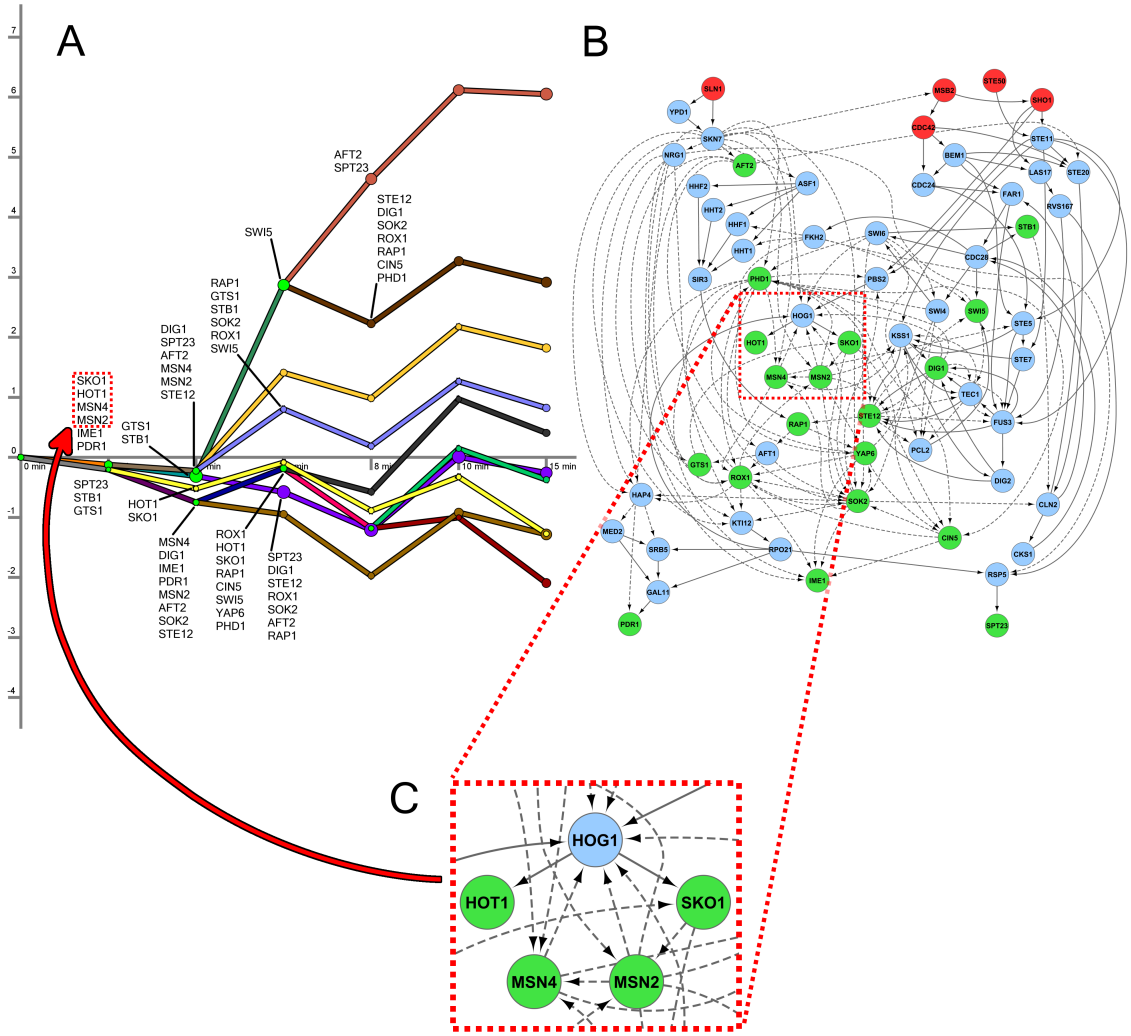

Figure 4.1: Short osmotic stress model. A) The regulatory part of the model contains 10 paths (clustered gene expression profiles). The x-axis displays when gene expression was measured. The y-axis shows log2 fold change in expression. The nodes following bifurcation events are annotated with the TFs that are predicted to control this split. TFs are only shown the first time they are active along a regulatory path. B) This subset of the oriented PPI network contains three types of nodes: upstream proteins used as sources (red), predicted signaling proteins (blue), and active TFs from DREM (green). Dashed edges are protein-DNA interactions and solid edges are oriented PPI. C) An enlarged view of a subsection of the PPI network identified shows that the core transcriptional unit of the HOG pathway was recovered. These TFs were inferred in the regulatory component of the model, and the network displays SDREM's explanation of how they are activated.

To assess the accuracy of the other predicted target TFs and signaling proteins, it is necessary to consider known HOG pathway models as well as other relevant osmotic and general stress proteins that lie outside the HOG pathway. We compiled a gold standard of established HOG pathway members derived from KEGG [99], the *Science Signaling* Database of Cell Signaling [70], and recent HOG literature and reviews [42, 79, 80, 110, 171]. Four of the seven TFs and six of the thirty other signaling proteins in the gold standard were correctly identified by SDREM (p-values of these overlaps are 7.70E-3 and 1.11E-8, respectively, using Fisher's exact test), indicating that the HOG pathway does compose a significant portion of the short model. To account for other proteins involved in the response, we constructed a set of osmotic stress-related genes by incorporating a genetic screen [78] and searching the literature. Many of our predictions that are not present in canonical HOG models are indeed supported by these additional data sources. Twelve of the 19 target TFs (63%) and 27 of the 39 predicted signaling proteins (69%) were found to be associated with osmotic stress.

The model reconstructed from the longer time series dataset is presented in Figure 4.2. As expected from the fact that it captures the recovery phase and more transcriptional events (Figure 4.2A), the long model identified 28 active TFs compared to the 19 active TFs in the short model. Many of these additional TFs were determined to be active at the 30 and 45 minute time points indicating their role in restoring gene expression levels to steady-state.

Although the two expression datasets were collected in rather diverse experimental settings and each contains many unique differentially expressed genes, there was very good agreement between the networks reconstructed by SDREM. Specifically, 16 of the 19 (84%) TFs identified in the short model were also identified in the long model including the four core HOG TFs. The osmotic stress evidence supports 13 TFs (46%) and 17 signaling proteins (74%) identified in the long model. As with the short model, the overlaps between the SDREM predictions in the long model and the gold standard were significant, with p-values of 0.0161 for the TF overlap and 2.55E-8 for the signaling proteins. In the long model, the network orientation procedure again correctly orients the PPI Hog1-Hot1 and Hog1-Sko1 (Figures 4.2B and 4.2C). Thus, both models point to the ability of our algorithm to correctly identify HOG pathway members and osmotic stress responders while at the same time reconstructing the networks by which they are activated.

4.3.2 Validating predicted osmotic stress transcription factors

While many proteins in the SDREM-reconstructed networks were supported by the gold standard databases, they also included novel predictions. To validate these predictions we

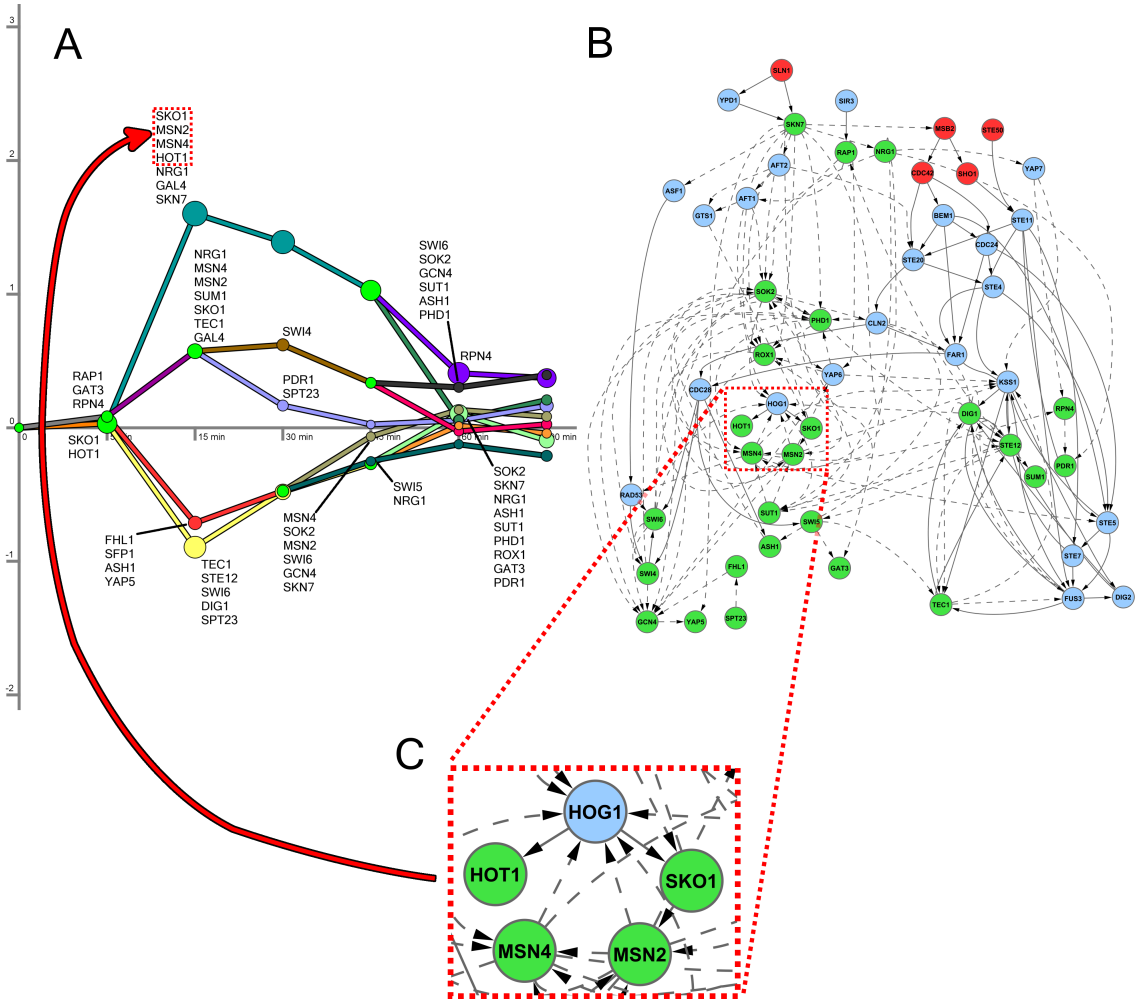

Figure 4.2: Long osmotic stress model. A) The regulatory model for the long osmotic stress expression data contains 9 paths. The initial splits overlap with those in the short model in terms of the TFs predicted to control them. B) The sources, signaling proteins, and active TFs in the long model. Again, there is a large overlap with the signaling model from the short time series dataset. The figure does not include the targets Gal4 and Sfp1 because they are connected to the sensory proteins via intermediate nodes whose scores fell below our threshold for inclusion in the model. Nevertheless, these targets are still well-connected to the source nodes. C) The primary TFs of the osmotic stress response are recovered in the long model as well. Hog1 and Sko1 are shown a second time along the uppermost regulatory path to emphasize the connection between the signaling and regulatory components.

performed a number of follow up experiments. The first set of experiments focused on TFs that were predicted to regulate either the response (in both models) or the recovery (in the long model). We thus selected four TFs from the short and long models — Cin5, Gcn4, Rox1, and Spt23 — that are all absent from the HOG gold standard as well as Hog1 as a control.

We used fluorescence microscopy to determine whether these proteins were differentially localized following sorbitol treatment at the times predicted by our models (see [64] for the experimental methodology). Cin5, Hog1, and Rox1 displayed significant nuclear localization patterns following treatment with sorbitol (p-values of 1.87E-11, 2.67E-7, and 1.02E-15, respectively, using a one-tailed t-test) as predicted by SDREM (Figure 4.3) and in accordance with Hog1's known rapid import into the nucleus in osmotic stress[79]. In contrast, we did not observe a significant change in localization for Gcn4 or Spt23.

In addition to microscopy, we also performed fluorescence-activated cell sorting (FACS) analysis to determine whether protein levels of the four TFs and Hog1 increased following sorbitol treatment. The levels of Gcn4 and Rox1 were found to increase significantly (p-values 6.98E-4 and 5.29E-4, respectively, using a one-tailed t-test) at times consistent with SDREM's predictions (Figure 4.4). The FACS experiments validated not only the osmotic stress relevance of the SDREM predictions Rox1 and Gcn4, but also the timing of their involvement. The elevated Rox1 protein levels were detected 30 minutes after treatment, supporting SDREM's predictions that it is active from 8 minutes onward in the short model and as late as 45 minutes in the long model. Gcn4's differential protein expression was detected 1 hour after treatment, consistent with the prediction that Gcn4 is active at the latest divergence point in the long model. Hog1, whose protein expression is stable after sorbitol treatment [210], served as a negative control and was not significantly affected (p-value 0.185). In summary, we validated that four of our five predicted osmotic stress-activated regulators (including the control Hog1) are indeed activated following treatment with sorbitol.

4.3.3 Knockouts support signaling protein predictions

To validate predicted proteins that are not TFs (which we term signaling proteins), we used knockout expression experiments. Because SDREM produces an oriented network, each signaling protein has a well-defined set of TFs that are downstream of it in the signaling cascades. By comparing the genes predicted to be regulated by these downstream TFs with those affected by the deletion, we can determine whether the KO effects agree with the proposed SDREM models.

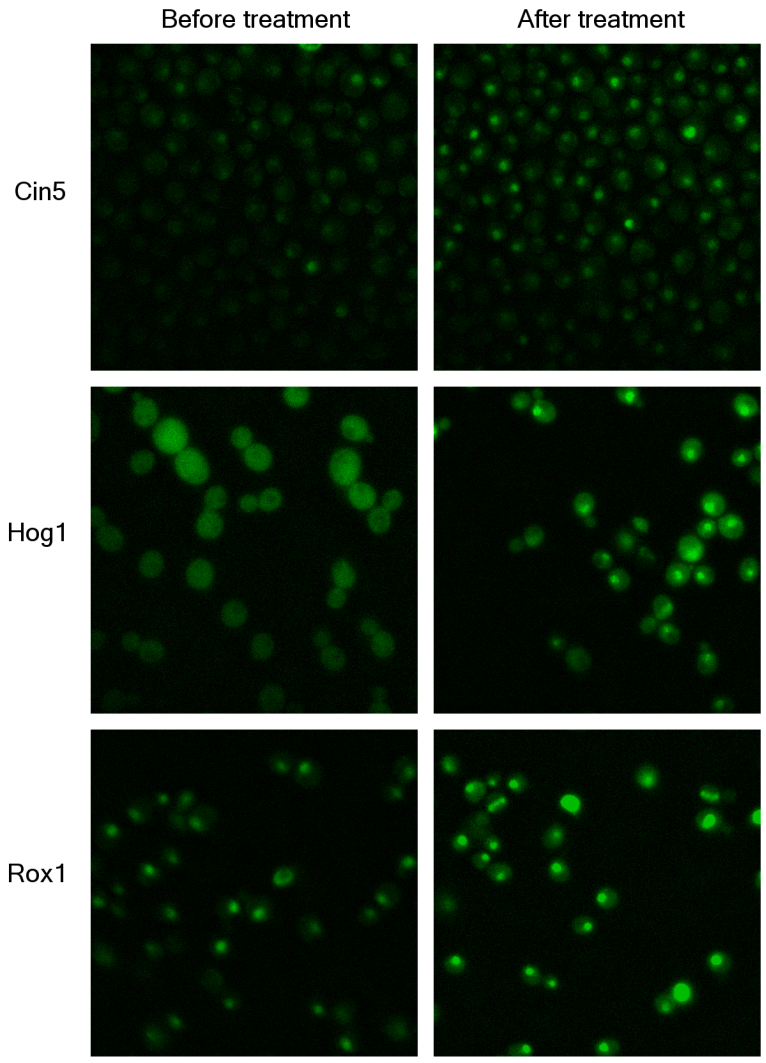

Figure 4.3: Differential nuclear localization after treatment with sorbitol. Each row corresponds to localization of the predicted osmotic stress responder before and after sorbitol treatment. The images were taken 50 minutes after treatment for Cin5, 21 minutes for Hog1, and 26 minutes for Rox1. P-values of the differential localization are 1.87E-11 for Cin5, 2.67E-7 for Hog1, and 1.02E-15 for Rox1.

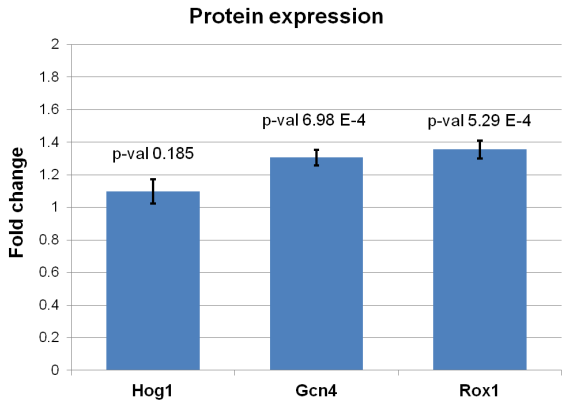

Figure 4.4: Differential protein expression after treatment with sorbitol. FACS reveals increased protein levels for Gcn4 and Rox1. The y-axis is the protein level ratio relative to the level before sorbitol treatment. The error bars show the standard deviation of the protein level ratios over all replicates.

We selected six genes that SDREM determined to be involved in separate high-confidence paths: the nucleosome assembly factor *ASF1*, the cell polarity-related *BEM1*, the MAPK *FUS3*, Mediator complex member *GAL11*, the cyclin *PCL2*, and the actin-associated *RVS167* (Figure 4.5A). These six genes were selected because they are absent from the HOG gold standard, nonessential, members of many high-confidence pathways, and predicted to belong to different levels of the signaling network hierarchy. Microarrays were used to profile wild type and knockout strains treated with sorbitol. Significance analysis of microarrays [200] was used to identify significantly differentially expressed genes.

We compared the differentially expressed genes with the short and long models to determine whether the knockout-affected genes significantly overlapped the genes assigned to the regulatory paths in the SDREM models. Such an overlap would suggest that the deleted gene affects the regulatory activity of the TFs that control the regulatory path, which putatively places the deleted gene upstream of those TFs in the signaling network. In order to ensure that any observed overlaps could be attributed to the osmotic stress response and recovery as opposed to the general stress response [59], we analyzed only osmotic stress-specific genes. For the short model, we found that there was significant overlap (p-value < 0.05 using Fisher's exact test with Bonferroni correction) for five of the six deletions: *ASF1*, *BEM1*, *GAL11*, *PCL2*, and *RVS167* (Figure 4.5B). Seven of the ten paths in the regulatory network were significantly associated with at least one KO experiment (p-value $<$ 1E-5 when compared to enrichment of random paths). Similar results

were obtained for the long model, where *BEM1*, *FUS3*, *GAL11*, and *RVS167* knockouts significantly overlapped one or more regulatory paths (Figure 4.6). Together, we found significant overlap for all six genes in at least one of the two models, although the support for *FUS3* and *PCL2* was weaker than the others.

We highlight the *ASF1* knockout to explicitly demonstrate how the overlap with the SDREM regulatory paths confirms Asf1's osmotic stress involvement and the inferred network orientation. Asf1 is downstream of the source Sln1 and upstream of numerous transcription factors in the oriented network, including the crucial HOG pathway TFs Hot1 and Sko1 (Figure 4.5C). Our model predicts that *ASF1* deletion is likely to partially affect many of these TFs and consequently perturb the genes (and regulatory paths) they control in the osmotic stress response. Indeed, we find that differentially expressed genes in the *asf1*Δ mutant significantly overlap with regulatory path 1 in the short model (Figure 4.5B) and all 7 TFs predicted to control this path's split from path 2 (Figure 4.5D) are downstream of Asf1 (Figure 4.5C), supporting the SDREM model.

In addition to Asf1, we found several other cases where the loss of a signaling protein affects paths controlled by the downstream TFs in our oriented network. One such example involves Bem1. The genes that are differentially repressed after *BEM1* deletion in sorbitol significantly overlap path 7 in the long model (Figure 4.6B), a path on which genes are repressed at 5 minutes and then gradually recover after 15 minutes. SDREM predicts five TFs that are actively controlling genes on this path — Ste12, Tec1, Swi6, Dig1, and Spt23 — and all five are indeed downstream of Bem1 in the oriented network (Figure 4.6A).

The genes affected by the *GAL11* KO further validate our predictions. Differentially expressed genes in the *gal11*Δ mutant significantly overlap with five paths in the short model (Figure 4.5B). Of these, all but path 8 are controlled in part by Pdr1, the only TF downstream of Gal11 in the short model network (Figure 4.5A), early in the response. In fact, Pdr1 is directly bound by Gal11 in the oriented network. Rvs167 is upstream of 15 TFs in the short model, which explains why its deletion affects so many regulatory paths (Figure 4.5B). The majority of the TFs controlling these paths are downstream of Rvs167 in the oriented network. For instance, six of the seven TFs controlling path 1's split from path 2 are downstream of Rvs167. Additional examples exist as well, and as a whole our knockouts support our predictions in both the short and long models.

Although we were able to use the oriented network to explain many of the effects we observed when predicted signaling proteins were deleted, in some cases the abundance of paths involving the deleted node impaired these efforts. Especially for proteins like Bem1 that are further upstream in the signaling network and directly interact with the sensory proteins (Figure 4.5A), there are many TFs that are downstream of them in the network. Thus, there is ambiguity when determining exactly how the deletion impacted

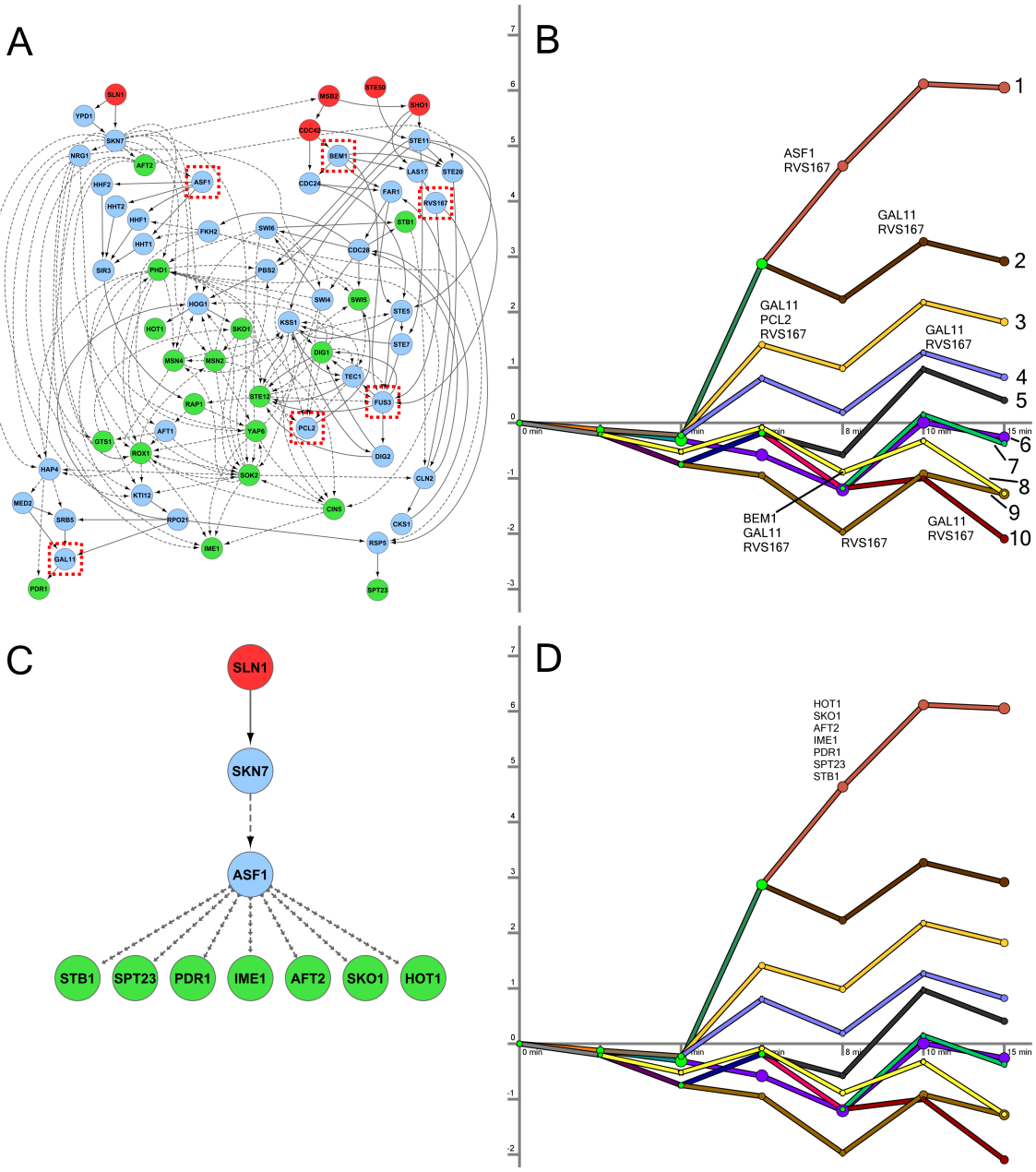

Figure 4.5: Knockouts affecting the short model. A) Knocked-out genes are highlighted with red boxes. B) Five knockouts significantly affected the genes assigned to the numbered regulatory paths. C) The subnetwork affected by the *ASF1* deletion. Only the relevant subset of the downstream TFs is shown and the edges connecting Asf1 to the TFs are omitted for clarity. D) The seven TFs predicted to control path 1's split from path 2 are displayed above path 1 and are all downstream of Asf1 in the oriented network.

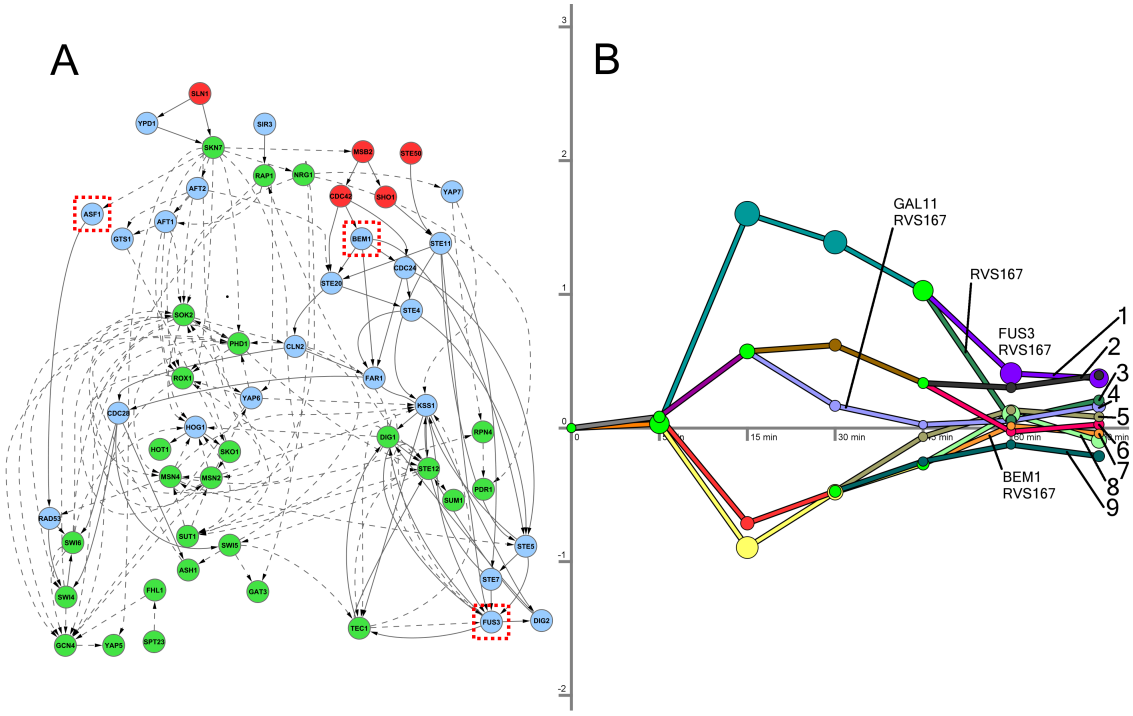

Figure 4.6: Knockouts affect downstream genes in the long model. A) The position of the deleted genes that are also in the long model. B) Four knockouts significantly affected the genes assigned to the regulatory paths. Numbered paths are annotated with the knockouts where we found significant overlap between path members and knockout-affected genes after filtering general stress genes.

gene expression because any of these TFs could have been affected by the deletion, but for any given TF there are typically other parallel paths that do not involve the deleted node. Furthermore, any errors in the network orientation can impair our ability to explain the observed knockout effects.

In general, the differentially activated genes after a knockout overlapped the upper regulatory paths and repressed genes overlapped lower paths. We can explain this phenomenon in many cases, but it is nevertheless counterintuitive. One would expect to see more cases where the positive regulators downstream of the deleted protein are deactivated after the knockout, which causes the genes on that path to be differentially repressed instead of activated.

4.3.4 Putative HOG pathway members

By excluding general environmental stress response genes from the regulatory pathway overlap analysis, we have some evidence that the six genes we deleted are osmotic stress-specific responders. Since not all osmotic stress-related proteins are activated via the HOG pathway, we performed additional single and triple knockout experiments of upstream proteins in this pathway (*HOG1*, *SSK2-SSK22-SHO1*, and *SSK2-SSK22-STE11*) to assess the HOG pathway membership of our predictions. The triple KOs were designed to disrupt both input branches of the HOG pathway, with *SSK2-SSK22* deletion severing the Sln1 branch and *SHO1* or *STE11* KO cutting off the Sho1 branch. The *hog1*Δ strain yielded the largest set of differentially expressed genes, and the *ssk2*Δ *ssk22*Δ *ste11*Δ mutant had a greater effect than *ssk2*Δ *ssk22*Δ *sho1*Δ, consistent with prior knockouts [154].

After removing the environmental stress response genes from all sets of differentially expressed genes, we calculated the overlaps between the HOG pathway knockouts and the deletions of our predicted genes. Five of the six predictions — *ASF1*, *BEM1*, *GAL11*, *PCL2*, and *RVS167* — significantly overlapped with the *HOG1* KO-affected genes. The overlaps were strongest for *GAL11* and *RVS167*, indicating that they may in fact belong to the HOG pathway. *BEM1*, *GAL11*, and *RVS167* also showed significant overlaps with both triple knockouts, whereas *ASF1* and *FUS3* overlapped with only one of the two.

Analysis of time series expression data from a *hog1*Δ strain [172] was used to further explore the potential HOG pathway membership of all SDREM model members. To reflect the *HOG1* deletion, Hog1 was removed from the PPI network when analyzing this data. Comparing the SDREM models of the wild type and *hog1*Δ mutant osmotic stress response (Table 4.1) indicates which predictions are putative HOG pathway members and which are other osmotic or general stress-related proteins. In addition to Hog1, there are 31 proteins that SDREM identifies solely in the wild type response, including other core HOG pathway proteins such as Hot1 and Pbs2. Many of the proteins that we selected for experimental validation — Asf1, Cin5, Gal11, Pcl2, and Rvs167 — are also uniquely predicted in the wild type model. Absence from the *HOG1* deletion model suggests that some of these proteins could be participants in the HOG pathway.

Very few proteins are predicted to be active solely in the knockout strain. Of these, Gcn4 is interesting because this TF is included in the (wild type) long model and its protein level was found to increase following osmotic stress treatment in our FACS analysis. This suggests that it is indeed activated in the stress response but via a pathway parallel to the HOG pathway. The remaining set of proteins that are predicted in both short models (wild type and *hog1*Δ) include MAPKs from other pathways, Fus3 and Kss1, that are unaffected by the absence of Hog1 as expected. In addition, the HOG pathway TFs Msn2 and Msn4

are predicted to still be active in the *hog1*Δ expression data, consistent with previous reports that Msn2 and Msn4 activity and nuclear import are controlled by the HOG and at least one other pathway [30].

4.3.5 Further support for validated predictions

We have shown that three of the four predicted TFs we investigated experimentally — Cin5, Gcn4, and Rox1 — localized to the nucleus and/or increased in expression in response to osmotic stress. Previous work provides further support for some of these findings and indicates that this activation may be important for overcoming sorbitol-induced stress. For example, *cin5*Δ mutants have been found to exhibit growth sensitivity to osmotic shock, and Cin5 induction peaks 30 to 60 minutes after exposure to moderate NaCl-induced stress [149]. Gcn4 has also been shown to play a role in salt-induced stress. Following NaCl exposure, mutations that incite Gcn4 activity also increase sensitivity to salt [68]. Osmotic stress mRNA synthesis analysis also reported Gcn4 as a regulator of salt stress genes [140].

A few of the signaling proteins we validated using knockouts were similarly identified as playing diverse roles in the osmotic stress response. Single and double knockouts revealed that Asf1 operates together with Rtt109 and in parallel with Arp8 to reassemble chromatin following hyperosmotic stress-induced transcription [104]. Bem1's involvement in the HOG pathway is tightly coupled with Cdc42, which was selected as a source protein in our study, and Ste20, a kinase recovered in both the short and long models. Binding domain mutations revealed that both Bem1 and Cdc42 independently contribute to Ste20's function in the HOG pathway. Whereas single Bem1 or Cdc42 binding domain mutations yielded only partial defects in osmoresistance, a double mutation generated a much stronger phenotype [211]. In both of our network models, we recover the correct orientations of the Bem1$\rightarrow$Ste20 and Cdc42$\rightarrow$Ste20 PPI. Genes affected by the *RVS167* knockout in sorbitol had the strongest overlaps with the regulatory paths and *HOG1* deletion. Under normal growth conditions, *rvs167*Δ mutants display slight deregulation of the actin cytoskeleton. However, in the presence of NaCl, the actin cytoskeleton of the mutant strain is completely deregulated and exhibits many abnormalities [16].

Although our single knockout only weakly confirmed Pcl2's involvement in the HOG pathway, a study by Lee *et al.* [122] provides insight into this result. They found that a mutant strain in which *PCL2* was deleted (similar to the KO strain we used) was able to colonize in a high salt environment, but a quintuple deletion of Pcl1,2-type cyclins (*pcl1*Δ *pcl2*Δ *clg1*Δ *pcl5*Δ *pcl9*Δ) failed to grow on this medium. Redundancy among these cyclins obscured the salt sensitivity phenotype in the single deletion. The fact that

Table 4.1: HOG-dependence of short model predictions. The three columns display which proteins are predicted when SDREM is run on the wild type short expression data, the *hog1*Δ data, or both variants. The five sources are not included in the lists.

Wild type only	***hog1*Δ only**	**Both**
Asf1	Gcn4	Aft2
Cdc28	Mbp1	Aft1
Cin5	Sum1	Bem1
Cks1	Sut1	Cdc24
Cln2	Xbp1	Dig1
Fkh2		Dig2
Gal11		Far1
Hap4		Fus3
Hhf1		Gts1
Hhf2		Kss1
Hht1		Msn2
Hht2		Msn4
Hog1		Nrg1
Hot1		Phd1
Ime1		Rap1
Kti12		Rox1
Las17		Skn7
Med2		Sko1
Pbs2		Sok2
Pcl2		Ste11
Pdr1		Ste12
Rpo21		Ste5
Rsp5		Ste7
Rvs167		Tec1
Sir3		Yap6
Spt23		Ypd1
Srb5		
Stb1		
Ste20		
Swi4		
Swi5		
Swi6		

our algorithm correctly recovered Pcl2 as an osmotic stress participant despite the weak support in its single knockout affirms our strategy to rely on dynamic gene expression data instead of knockouts for model inference. Interestingly, it was also reported that the Pcl2-Pho85 kinase phosphorylates Rvs167 [122]. Much like *RVS167* deletion strains, the quintuple Pcl1,2-type cyclin deletion exhibited abnormalities in the actin cytoskeleton that were more pronounced in the presence of salt.

4.3.6 Rapamycin response

While we have primarily focused on the osmotic stress response, we also used SDREM to study the target of rapamycin (TOR) response pathway in yeast to demonstrate SDREM's flexibility and generality. Although yeast contains two complexes, TORC1 and TORC2, in which the Tor proteins are members, only TORC1 is inhibited by the drug rapamycin [225]. Thus, we used the five TORC1 complex members as the sources in our TOR pathway modeling: Kog1, Lst8, Tco89, Tor1, and Tor2 [225]. Tor2 is only a TORC1 complex member in the absence of Tor1, but we include both proteins as sources. TORC1 has been shown to respond to not only rapamycin but also caffeine [117], nitrogen source quality [225], and other stimuli.

The TOR response expression data [202] contained measurements at 20, 30, 60, 90, 120, and 180 minutes. Unlike the long osmotic stress expression dataset, the genes differentially expressed in the TOR response generally remained activated or repressed for the full 3 hours and did not return to steady state during this period (Figure 4.7A). Along with the extensive TF-gene binding data from cells grown in rich media [133], SDREM was also provided rapamycin-specific data for 14 TFs previously implicated in the TOR response [75].

Despite the prior evidence for these TFs' TOR involvement, conventional TOR pathway representations contain very limited knowledge of the downstream TFs. One model [225] contains only Gln3, Msn2, Msn4, and Sfp1, and the *Saccharomyces* Genome Database (SGD) shows no TFs annotated with the Gene Ontology (GO) [8] term 'TOR signaling cascade'. In contrast, SDREM predicts that 23 TFs are active regulators in the TOR pathway (Figure 4.7A), and of these only Sfp1 is a member of the previous TOR models. Nevertheless, we found support for 17 of these predictions (74%) in the two aforementioned TOR pathway models, rapamycin screens [31, 78, 216], a set of genes curated by SGD that have a rapamycin resistance phenotype, and/or previous literature.

SDREM identifies 25 additional proteins that connect TORC1 to the downstream TFs (Figure 4.7B). Of these, 14 (56%) are present in the extended gold standard or were found

to have possible links to the TOR pathway in a literature search. Altogether, the overlap between SDREM's TOR predictions and the collection of TOR- or rapamycin-relevant genes is significant (p-value 2.55E-3 using Fisher's exact test). Therefore, even though very few predictions were present in the two canonical TOR models and many known TOR members were not recovered, SDREM accurately identifies an extended TOR pathway representation. The SDREM model includes many proteins that are traditionally primarily associated with other signaling pathways but are affected by rapamycin, for example Dig1, and explains how they may in fact be involved in the rapamycin response.

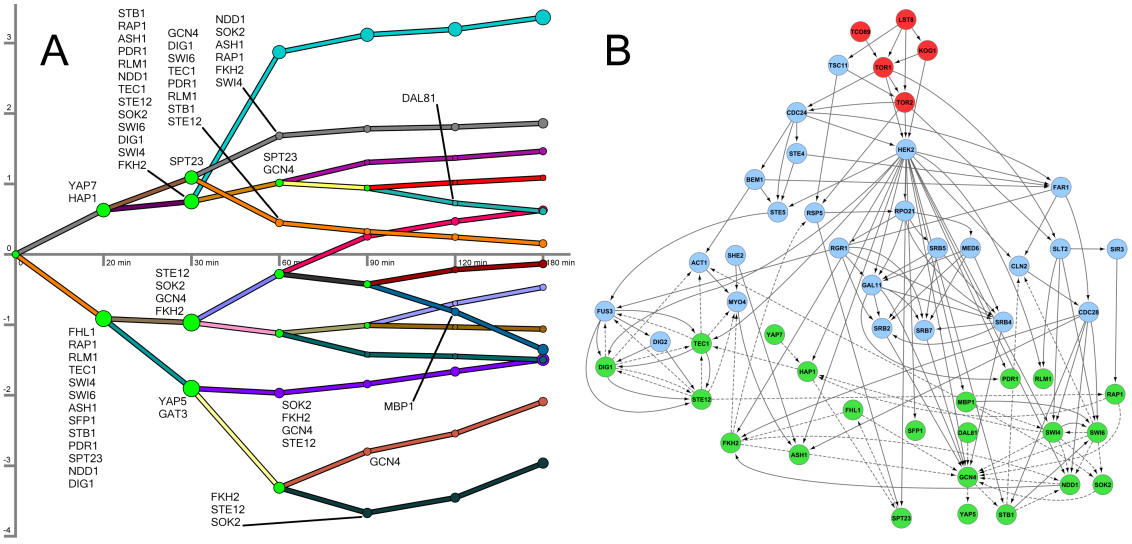

Figure 4.7: Rapamycin model. A) The rapamycin response model contains 15 regulatory paths. Unlike the long osmotic stress model, the differentially expressed genes remain highly or lowly expressed after the initial shock for the duration of the experiments. B) The sources, signaling proteins, and active TFs in the rapamycin model. The TF Gat3 does not appear in the figure because it does not directly interact with other predicted pathway members. Rather, it is influenced by upstream proteins via paths containing other proteins that were not deemed to be core members of the response.

4.3.7 SDREM improves upon previously suggested methods

While SDREM relies in part on two previously developed methods, integrating them is not a trivial task. Computationally, the two methods represent very different types of

computational models (probabilistic graphical models and combinatorial optimization). Biologically, the two methods target different types of networks. Indeed, neither component of SDREM, on its own, can accurately recover the osmotic stress response network. The most glaring example of this is that Hog1, the central component of the osmotic stress response, would have been missed by DREM alone. Hog1's DNA binding is generally indirect; therefore, DREM does not detect it as a significant direct regulator in the osmotic stress response. Moreover, DREM can identify which TFs are important in the stress response but cannot explain how or why they are activated. By modeling the upstream pathways, the set of TFs that DREM identifies improves substantially from the initial application (when the signaling network is not yet utilized) to the final iteration. In the short model there were 17 TFs selected by DREM in the first iteration that were dropped in subsequent iterations due to lack of support in the oriented PPI network. Of these only one, Mcm1, is present in the HOG gold standard, and even Mcm1 is considered a HOG pathway member in only one of the seven gold standard sources. On the other hand, there were eight active TFs in the final model that were missed in the first iteration. These eight TFs include Cin5 and Rox1, TFs for which our experimental results and prior literature strongly support their role as active regulators in the osmotic stress response. Thus, the use of signaling data leads to more accurate regulatory models, which in turn provide new targets allowing for better reconstruction of the signaling pathways.

Likewise, while the network orientation algorithm performs very well when given a set of sources and targets, its applicability and utility are greatly reduced if it is limited to conditions in which the target TFs are completely known. In the osmotic stress response, DREM detected active TFs such as Cin5, Gcn4, Nrg1, Rox1, and Yap6 that play a role in the response and recovery but are not included in canonical HOG pathway representations and thus would not be included in the target set for the network orientation algorithm.

Physical Network Models [217, 218] and ResponseNet [119, 219] are existing algorithms for connecting upstream sources (deleted genes and genetic screen hits, respectively) to downstream transcriptional effects via physical interaction networks. Although not expressly designed to infer directed pathways from the initial nodes in signaling networks to the TFs active in the downstream transcriptional response, we assessed whether they can successfully address this task as well (see [64] for the comparison methodology). As we demonstrate in Tables 4.2 and 4.3, SDREM outperforms both PNM and ResponseNet in modeling the osmotic stress response when given the same upstream proteins and expression data, which we quantify via the HOG gold standard. Using the short expression data, for which both PNM and ResponseNet perform best, the overlaps between the predicted TFs and gold standard are insignificant (p-values 0.770 and 0.162, respectively). Only SDREM correctly recovers all four core TFs, demonstrating that modeling

the dynamic transcriptional response enhances identification of active TFs.

PNM predicts a very large set of proteins, 445 for the short expression data and 309 for the long expression data. Even though it predicts over 6 times more proteins than SDREM in both cases, it recovers only 2 or 3 more gold standard proteins than SDREM, resulting in less significant overlaps. The noteworthy omission of Hog1 in the PNM network constructed with the long expression data and aforementioned insignificant overlap between the predicted TFs and gold standard TFs indicate the HOG pathway is not well-represented even in the large predicted network.

When run with the default settings, ResponseNet's capping parameter is set to 0.7, which controls the maximum edge weight in the network. The majority of the edge weights in our interaction network are ≥ 0.7 so this leads to a network where most of the edges have the same maximum weight of 0.7. Consequently, ResponseNet struggles to predict internal signaling nodes in this setting and includes only 4 such proteins in its network. Hog1 is not among those 4 proteins, and the large set of predicted TFs does not significantly overlap the HOG gold standard TFs. Therefore, we also ran ResponseNet with a capping parameter of 0.9, which allows the varying confidence in our edge weights to be represented. In this case, 4 of the 8 predicted signaling proteins are in the gold standard, and the significance of the overlap is comparable to that obtained by SDREM. However, the TF predictions are once again insignificant, as ResponseNet only identifies a single gold standard TF.

Table 4.2: Overlap significance for PNM predictions and HOG gold standard. The sizes of the networks predicted by PNM are shown alongside the SDREM models for comparison. PNM predicts very large networks and does not recover all of the core HOG TFs. The five sources are not included in the counts.

Algorithm	SDREM	PNM	SDREM	PNM
Expression data	Short	Short	Long	Long
Total predictions	58	445	51	309
Predicts Hog1	Y	Y	Y	N
Predicted internal	30	374	17	248
Gold standard internal	30	30	30	30
Internal overlap	6	9	5	7
Internal significance	1.11E-8	1.61E-4	2.55E-8	3.88E-4
Predicted TFs	28	71	34	61
Gold standard TFs	7	7	7	7
TF overlap	4	2	4	1
TF significance	0.008	0.770	0.016	0.922

Table 4.3: Overlap significance for ResponseNet (RN) predictions and HOG gold standard. The sizes of the networks predicted by ResponseNet are shown alongside the SDREM models for comparison. ResponseNet was run twice on each of the osmotic stress expression datasets, once with the default parameters and again with the capping parameter set to 0.9. In all cases, ResponseNet is unable to recover the primary HOG TFs as well as SDREM. The five sources are not included in the counts.

Algorithm	SDREM	RN	RN	SDREM	RN	RN
Expression data	Short	Short	Short	Long	Long	Long
Settings	Default	Default	Cap 0.9	Default	Default	Cap 0.9
Total predictions	58	61	13	51	61	13
Predicts Hog1	Y	N	Y	Y	N	Y
Predicted internal	30	4	8	17	4	8
Gold standard internal	30	30	30	30	30	30
Internal overlap	6	1	4	5	1	4
Internal significance	1.11E-8	0.023	5.99E-8	2.55E-8	0.023	5.99E-8
Predicted TFs	28	57	5	34	57	5
Gold standard TFs	7	7	7	7	7	7
TF overlap	4	2	1	4	2	1
TF significance	0.008	0.632	0.162	0.016	0.632	0.162

4.3.8 Parameter selection and robustness

Whenever possible, SDREM's parameters were selected in accordance with existing biological data or computational approaches. We used condition-specific osmotic stress data to obtain an estimate for the active TF influence parameter, which represents the portion of bound genes that are expected to be affected by an active TF. The TFs Hot1 and Sko1 are the two HOG pathway TFs for which we have condition-specific binding data [30], and both are known to be active in the osmotic stress response. 79% of the genes bound by Hot1 are differentially expressed in both the short osmotic stress expression data and the long expression data. Likewise, 79% of genes bound by Sko1 are affected in the short expression data and 68% in the long expression dataset. Therefore, we set this parameter's default value to 80%.

Several parameters such as the path length, PPI edge weight threshold, and number of top paths used for scoring were selected based on our analysis we performed on yeast PPI networks (Section 3.3). In this analysis the number of targets in the network was fixed, which suggested using a fixed threshold for the number of top paths. The number of paths considered by SDREM is equal to 5 times the number of targets instead of the fixed value of 100 to account for the fluctuating number of targets over all iterations. In the short model, this flexibility results in using between 95 and 140 top paths.

The protein-DNA edge weights are motivated by the ResponseNet weighting scheme [219]. Most existing approaches for weighting protein-DNA interactions are unable to simultaneously account for experimental p-value, motif presence, and experimental condition, which all influence edge weight in our network. The Physical Network Models strategy, for instance, uses p-values alone. ReponseNet fixes an arbitrary weight of 0.7 for interactions with a conserved binding motif in multiple species, which we do as well except with a weight of 0.95 (for consistency with our PPI weights). ResponseNet assigns the remaining edge weights based on motif presence and conservation, whereas we incorporated experimental condition and p-value in addition to these features.

For those parameters that could not be directly estimated from biological data, we made an initial choice of value based on our intuition of the algorithm's behavior. We then tested the robustness of this selection to small fluctuations in the parameter value (where robustness is measured in terms of the overlap in the outcomes between different parameter values), following the approach of Kim *et al.* [103]. These parameters were consistent across all SDREM runs and are suitable for analysis of other conditions or organisms.

Table 4.4 describes the eight parameters that were varied during the robustness testing, all of which was performed using the short osmotic stress expression data. In addition to

the two runs per parameter (using a lower/higher value than the default), we ran SDREM with an unweighted version of our protein-DNA interaction network to observe whether our weighting scheme enhanced SDREM's predictions. The topology of this unweighted network was identical to the original protein-DNA interaction network, but the weights were uniformly set to 1. The PPI edge weights were not varied because they have been justified previously [65].

Although varying these parameters does have an effect on the SDREM output, the core of the predicted network remains the same. Nearly all of the new runs generate fewer predictions than the baseline run, but in the majority of the runs over 90% of the new predictions are also found in the baseline prediction (Table 4.5). The notable exception is the set of predictions from the unweighted protein-DNA interaction network, which has a greater effect than varying the algorithm's parameters. Only 25 of the 58 baseline short model predictions also appear in this run, lower than any of the overlaps obtained when only the parameters are varied. Figure 4.8 shows that out of the 58 proteins in the baseline short model, 31% are still predicted in all 16 runs where a parameter is varied and 79% are predicted in at least half of the runs. In contrast, the majority (56%) of the proteins that are predicted only when the parameters are varied appear in the output of a single run.

When varying the parameters, the overlap between SDREM's predictions and the HOG gold standard is significant in all cases and comparable to the overlap obtained when using the original parameters (Tables 4.6 and 4.7). However, once again we observe that the run that uses the unweighted network is an outlier and performs markedly worse than the baseline prediction. Only six signaling proteins are predicted, and Hog1 is not among them, confirming that the protein-DNA edge weights we assigned improve predictive capabilities.

4.3.9 Limitations of the learned models

While SDREM identified the majority of the gold standard proteins, it missed two important HOG pathway proteins, Ssk1 and Ssk2, that are present in all seven gold standard sources. The most likely explanation for their absence is that both proteins have a low degree in the protein interaction network. Consequently, it is unlikely that these proteins will have a large number of source-target paths through them in the directed network, which means that they will have low connectivity scores and not be recognized as important HOG members. This suggests a possible bias in our technique against low-degree proteins, which we address in Section 5.3.

SDREM is designed to discover directed cascades between the upstream sources and

Table 4.4: Parameters perturbed for robustness testing. In addition to the protein-DNA network weight, the eight parameters below were varied for robustness testing. The baseline run uses the default value for all parameters.

Run name	Parameter being varied	Default value	New value
baseline	None		
active.tf.influence.0.7	Percent of bound genes that are influenced by a TF that is active in the stress response	80%	70%
active.tf.influence.0.9	Percent of bound genes that are influenced by a TF that is active in the stress response	80%	90%
dist.tfs.25	Number of TFs used to build random activity score distribution	50	25
dist.tfs.100	Number of TFs used to build random activity score distribution	50	100
dist.thresh.0.4	Percentile in the random activity score distribution that real TF scores must exceed	50th	40th
dist.thresh.0.6	Percentile in the random activity score distribution that real TF scores must exceed	50th	60th
min.prior.0.005	Minimum activity prior allowed	0.01	0.005
min.prior.0.05	Minimum activity prior allowed	0.01	0.05
node.thresh.0.005	Node score threshold	0.01	0.005
node.thresh.0.05	Node score threshold	0.01	0.05
random.target.ratio.0.5	Number of random targets added to network during target scoring per real target	1	0.5
random.target.ratio.2	Number of random targets added to network during target scoring per real target	1	2
target.thresh.0.7	Target score distribution threshold	0.8	0.7
target.thresh.0.9	Target score distribution threshold	0.8	0.9
top.paths.100	Number of top-ranked paths used to calculate target and node scores in the network	5 times number of targets	100
top.paths.1000	Number of top-ranked paths used to calculate target and node scores in the network	5 times number of targets	1000
pdi.no.weight	Protein-DNA interaction network edge weights	See text	See text

Table 4.5: Baseline overlap during perturbation testing. The number of proteins predicted by the baseline model and the runs in which a single parameter was varied. The five sources are present in all models and are not included in the counts. Overlap percentages are calculated with respect to the baseline ('Baseline overlap') and the robustness testing run ('Run overlap').

Run name	**Baseline predictions**	**Run predictions**	**Overlap**	**Baseline overlap**	**Run overlap**
active.tf.influence.0.7	58	46	41	71%	89%
active.tf.influence.0.9	58	28	28	48%	100%
dist.tfs.25	58	41	36	62%	88%
dist.tfs.100	58	52	41	71%	79%
dist.thresh.0.4	58	53	49	84%	92%
dist.thresh.0.6	58	36	36	62%	100%
min.prior.0.005	58	51	47	81%	92%
min.prior.0.05	58	50	47	81%	94%
node.thresh.0.005	58	57	55	95%	96%
node.thresh.0.05	58	39	32	55%	82%
random.target.ratio.0.5	58	58	55	95%	95%
random.target.ratio.2	58	52	49	84%	94%
target.thresh.0.7	58	58	55	95%	95%
target.thresh.0.9	58	42	41	71%	98%
top.paths.100	58	53	49	84%	92%
top.paths.1000	58	68	45	78%	66%
pdi.no.weight	58	35	25	43%	71%

the inferred active TFs. Therefore, it is unable to recover pathway members that are further upstream of the given sources, which explains the absence of Opy2 and Hkr1 in SDREM's predictions. Furthermore, any proteins that do not lie between the sources and TFs are missed. For example, the HOG gold standard diagrams show that Ptc2, Ptc3, Ptp2, and Ptp3 are outside of the source-TF paths, and are consequently omitted in both models. However, it is possible that linking SDREM with our edge prediction algorithm (Section 3.5) would enable us to identify such nodes. In addition, similar to all other modeling methods, SDREM is dependent on the input data it uses. For example, three gold standard proteins — Ctt1, Glo1, and Gpd1 — were not identified by SDREM because these proteins were not present in the high-confidence PPI input network. Similarly, the gold standard TF Msn1 is missing from both the TF binding data [133] and the PPI network.

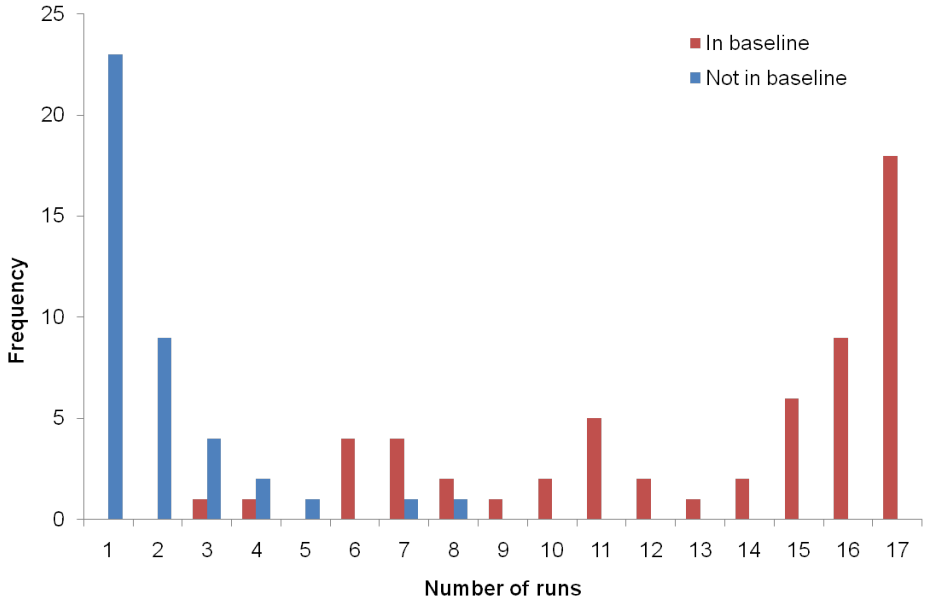

Figure 4.8: Histogram of the number of occurrences of each protein across all perturbation testing. The number of runs specifies how many models include a particular protein. The frequency provides the number of proteins that fall into each bin. For example, 23 proteins are predicted in only a single run. The 'pdi.no.weight' run is not included in the counts. The five sources appear in all models and are not counted.

Table 4.6: Robustness testing signaling protein overlap significance. The significance of the overlap between the HOG gold standard signaling proteins (those that are not sources or TFs) and signaling proteins in the SDREM models. The total predictions include the signaling proteins and TFs, but not the five sources.

Run name	**Total predictions**	**Predicts Hog1**	**Predicted signaling**	**Gold standard signaling**	**Signaling overlap**	**Signaling significance**
baseline	58	Y	30	30	6	1.11E-8
active.tf.influence.0.7	46	Y	18	30	6	3.65E-10
active.tf.influence.0.9	28	Y	11	30	4	2.79E-7
dist.tfs.25	41	Y	22	30	6	1.44E-9
dist.tfs.100	52	Y	16	30	5	1.80E-8
dist.thresh.0.4	53	Y	21	30	6	1.05E-9
dist.thresh.0.6	36	Y	17	30	5	2.55E-8
min.prior.0.005	51	Y	21	30	6	1.05E-9
min.prior.0.05	50	Y	22	30	6	1.44E-9
node.thresh.0.005	57	Y	30	30	6	1.11E-8
node.thresh.0.05	39	Y	11	30	4	2.79E-7
random.target.ratio.0.5	58	Y	31	30	6	1.37E-8
random.target.ratio.2	52	Y	21	30	6	1.05E-9
target.thresh.0.7	58	Y	32	30	6	1.68E-8
target.thresh.0.9	42	Y	20	30	5	6.30E-8
top.paths.100	53	Y	29	30	6	8.94E-9
top.paths.1000	68	Y	28	30	8	1.19E-12
pdi.no.weight	35	N	6	30	2	4.68E-4

Table 4.7: Robustness testing TF overlap significance. The significance of the overlap between the HOG gold standard TFs and the SDREM model TFs.

Run name	**Total predictions**	**Predicted TFs**	**Gold standard TFs**	**TF overlap**	**TF significance**
baseline	58	28	7	4	0.008
active.tf.influence.0.7	46	28	7	4	0.008
active.tf.influence.0.9	28	17	7	4	0.001
dist.tfs.25	41	19	7	4	0.002
dist.tfs.100	52	36	7	4	0.020
dist.thresh.0.4	53	32	7	5	0.001
dist.thresh.0.6	36	19	7	4	0.002
min.prior.0.005	51	30	7	4	0.010
min.prior.0.05	50	28	7	4	0.008
node.thresh.0.005	57	27	7	4	0.007
node.thresh.0.05	39	28	7	4	0.008
random.target.ratio.0.5	58	27	7	4	0.007
random.target.ratio.2	52	31	7	5	0.001
target.thresh.0.7	58	26	7	4	0.006
target.thresh.0.9	42	22	7	4	0.003
top.paths.100	53	24	7	4	0.004
top.paths.1000	68	40	7	5	0.004
pdi.no.weight	35	29	7	5	0.001

Chapter 5

Enhancing SDREM

The analysis of well-studied yeast stress response pathways in Section 4.3 demonstrates that SDREM is adept at recovering the majority of the contributing proteins (both signaling proteins and TFs) and suggesting novel responders. Here we address several limitations of SDREM that were revealed in the yeast study. A primary goal in developing SDREM is to provide insight into poorly understood response pathways and transcriptional dynamics, particularly those with clinical relevance. However, as we discuss in Section 5.1, scaling to human datasets is a nontrivial challenge, and we developed several algorithmic and data-driven approaches to enable the analysis of human data. These extensions have enabled us to study the human response to influenza infection, yielding several exciting predictions (Section 5.2). Moreover, as discussed in Section 4.2.1, the two components of SDREM involve very different underlying computational approaches (inference in a probabilistic graphical model and combinatorial optimization). In Section 5.3, we demonstrate that both the network orientation and dynamic gene expression analysis can be represented with a single, unified probabilistic graphical model.

5.1 Scaling to human datasets

Conceptually, SDREM as presented in Section 4.2.1 is a general algorithm and can be readily applied to any organism. In practice, however, the relatively small size of the yeast proteome allows for algorithmic approaches that are unacceptably slow when moving more complex organisms, in particular human. As the number of proteins and edges in the interaction network grows, so does the number of potential source-target paths. A larger set of paths typically leads to additional contention over the direction of a particular

edge, and both factors complicate the edge orientation. In addition, there are more potential transcriptional regulators, slowing the analysis of the gene expression data and making the identification of well-connected TFs even more important.

We can address these challenges by incorporating new types of biological data (Sections 5.1.1 and 5.1.2) and making algorithmic improvements (Sections 5.1.3 and 5.1.4). Note that the emphasis here is on human datasets, but these extensions are useful (or necessary) for analysis of many other complex organisms (e.g. mouse, *Arabidopsis thaliana*, *Drosophila melanogaster*).

5.1.1 Incorporating RNAi screens

Due to the larger PPI and regulatory networks, there are many more possible connections from sources to targets and disagreements about the orientation of individual PPI in human models. In order to increase our ability to distinguish true signaling pathways that activate relevant TFs from other potential connections, we integrate genome-wide RNA interference (RNAi) screens into the SDREM framework. RNAi screens knock down a gene and report whether the loss of that particular gene impacts a phenotype of interest such as viral load in an infected cell line. If a gene is associated with a phenotypic change in an RNAi screen, we are more likely to believe that pathways containing that gene are controlling the stress response and should be preferred during the network orientation. Note that due to redundancy, the converse is not necessarily true. Genes that are negative screen hits may still be highly relevant to signaling pathways.

Although related approaches (e.g. ResponseNet [219]) use the screen hits directly as sources in the network, we place less trust in the RNAi data. Independent RNAi screens can exhibit very low overlap [189] in part due to the impact of differences in methodology [14] or cell population context [184]. Table 5.1 demonstrates this disagreement for RNAi screens relevant to H1N1 influenza infection. No genes are hits in all five screens, and only a single gene is detected in four of the five screens (note that two of the screens [25, 178] are targeted, not genome-wide).

To incorporate this notion into our models, we convert the RNAi screen data into vertex weights following our approach for PPI weighting (Section 3.2.4).

$$w(v) = \begin{cases} 1-(1-c)^n, & \text{if } n > 0 \\ 0.5, & \text{otherwise} \end{cases}$$

where $w(v)$ is the weight assigned to a vertex (gene), c is the confidence in the screen in the range $[0, 1]$, and n is the number of screens that report v as a hit. We set $c = 0.75$

Table 5.1: Overlap among five H1N1 influenza infection RNAi screens [25, 28, 100, 107, 178]. The vast majority of the 1009 genes are hits in only a single screen.

n	**Genes detected in n screens**
1	940
2	62
3	6
4	1
5	0

in all analyses here but could incorporate biological knowledge to set different confidence levels for different screens if it was available (as we did for the PPI). These node weights can be used directly in the formula for path weights (Section 4.2.3) such that paths that contain many screen hits have higher weights. We show in Section 5.2.2 that integrating RNAi screen data in this manner improves SDREM models.

5.1.2 Fixing edge directions

During the network orientation phase of SDREM, each greedy step of the local search must consider the change in objective function value obtained when flipping each edge's direction. In addition, the number of possible source-target paths that are enumerated during the depth first search is larger when there is a greater fraction of unoriented edges in the network. Consequently, reducing the number of unoriented edges (i.e. fixing edge orientations using prior information whenever possible) can offer substantial reduction in runtime.

Although PPI are generally reported as undirected edges, post-translational modifications (PTMs) have also been collected in databases. These interactions have a definitive directionality — one of the member proteins acts upon the other. In particular, the Human Protein Reference Database (HPRD) [141] houses thousands of PTMs, which can be used to assign a fixed edge direction to the corresponding PPI. Therefore, to reduce uncertainty in the human interaction network, we integrate the HPRD PTMs with HPRD PPI and BioGRID PPI [187]. Whenever the two proteins involved in a PTM are also reported in the set of PPI, we remove the undirected PPI from the network and associate its supporting experimental evidence with the corresponding PTM, thereby increasing the weight for the directed PTM edge. In very rare cases, for instance the human proteins CDK2 and CDK7, PTMs exist in both directions and we include both directed edges in the network. Protein-

DNA binding interactions are also directed and included in the human interaction network, but because these interactions are between a protein and gene instead of two proteins we do not use them to resolve PPI orientation. Note that PTM have been reported at a large scale in yeast as well [166], but it was less critical to integrate this data during the SDREM yeast analysis because the smaller network allows for accurate and efficient inference of the edge directions.

5.1.3 Algorithm parallelization

Integrating additional data types improves accuracy in the more complex human networks and does reduce runtime, but scaling to human data requires several algorithmic improvements as well. SDREM was originally written as a single-threaded application, but we extended it to run on a cluster to better handle the human datasets. We studied the execution times of the various phases of SDREM and parallelized the code in a targeted manner so as to maximally speedup execution.

In order to generate a distribution of random TF activity scores (Section 4.2.2), each iteration of SDREM analyzes the gene expression data many times (typically 10 or more) using randomized protein-DNA interactions. We isolated these SDREM calls, making requisite sections of the code thread-safe, such that each run involving a new (randomized) protein-DNA dataset is independent from the other runs and can be executed in parallel. This (approximately) reduces runtime of the gene expression component of SDREM by the factor $\min(c, r)$, where c is the number of cores available and r is the number of randomizations.

In the network orientation phase, enumerating all possible source-target paths in the large human interaction network composed a substantial portion of the runtime. Therefore we parallelized the depth first search using a synchronized priority queue to track the highest confidence paths found across all parallel threads. The enumeration tasks are divided based on the source node such that each core independently initiates a depth first search from a single source. If the number of sources is greater than the number of cores, which was the case in our analysis of influenza infection (Section 5.2.2), the sources are placed in a work queue and cores can dequeue a source upon terminating the depth first search from the current source.

5.1.4 Source-target pathway approximations

In addition to the aforementioned parallelization, a significant speedup to the network orientation component of SDREM can be obtained by precomputing and writing all possible source-target paths to disk. In each iteration of the original version of SDREM, paths were enumerated many times because the target connectivity scores are computed by orienting a network that includes random targets, which changes the set of paths. However, it is reasonable to limit the set of potential random targets to be only TFs, or even only those TFs that are present in the protein-DNA dataset (i.e. those TFs that could be identified as active regulators during the gene expression analysis). We now search for all paths from a source to any TF, write these paths to file, and read the appropriate stored paths for each new set of putative active targets and random TFs. Enumerating paths once instead of many times at each iteration offers immense savings computationally.

However, precomputing paths does not reduce the time it takes to orient the network. Depending on the network size, maximum path length, and number of sources and targets it is possible to obtains millions or billions of paths when running SDREM on human data versus a couple hundred thousand paths in the yeast analysis (Section 4.3). Even if the number of edges to orient remains fixed, the orientation runtime is highly dependent on the number of paths. Evaluating the objective function requires summing the weights of all satisfied paths, and for every potential edge flip that is considered at a greedy local search step we must determine which paths are still satisfied.

Therefore, we explored whether it would be possible to restrict how many paths are considered and still obtain equally good results in practice. The Maximum Edge Orientation problem is NP-hard and we already solve it using a heuristic approach (based on the results in Section 3.3) suggesting that a further approximation may not negatively impact our results. We modified the parallel path enumeration algorithm to only store the top m paths to each TF, ranking the paths by path weight. When given a set of targets, we merge the top m paths for each target and keep only the top m paths from any source to any of these targets. Considering only the top m paths also enables us to include early termination in the depth first search's branch traversal. By tracking the lowest weight path in the top m found so far, we can determine whether a path has the capability to be in the top m before traversing the entire path in some cases. For example, if the current lowest weight path in the top m has a weight of 0.8 and the first edge we traverse has weight of 0.1, we do not need to traverse the second or subsequent edges. Any paths using that edge will have weight less than 0.8 because all edge and vertex weights are in $[0, 1]$ and the target weights are in $[0, k]$, where k, the path length bound, is assumed to be 5 in this example.

To test the impact of this approximation, we used the H1N1 influenza infection data

described in Section 5.2.2 but only considered a high-confidence subset of the source proteins so that it was possible to enumerate all paths repeatedly in a reasonable amount of time. After enumerating all ~ 3 million paths, we oriented the network 25 times and calculated node scores and the total path weight of the top 1000 paths for each orientation. We similarly calculated node scores and cumulative top path weights for 100 orientations in which only the 100000 or 200000 highest-confidence paths were enumerated.

Figure 5.1 shows that the actual node scores, which are used to identify which proteins participate in the signaling pathways of interest, are highly comparable to the approximated node scores. In addition, increasing m from 100000 to 200000 does little to improve the approximation. The correlation between the actual node scores and the approximated node scores is greater than 0.999 in both cases. Similar results were obtained when using the top 100, 10000, or 50000 paths to calculate the node score (instead of the top 1000).

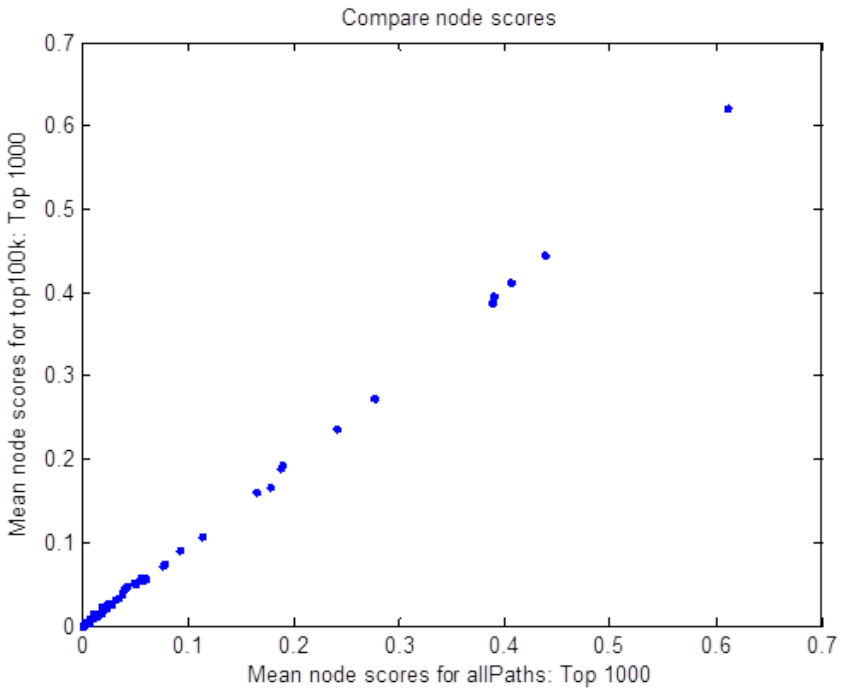

Figure 5.1: Node scores, the fraction of the top 1000 paths that pass through a particular protein, are very similar when enumerating all paths or only the top 100000 paths. The node score obtained when using all paths is shown along the x-axis. The y-axis provides the approximated node score.

Figure 5.2 shows that the top-ranked paths obtained when enumerating only m paths are not identical to those recovered when enumerating all paths. The sum of the path weights are similar, however, indicating that the sets of paths are of similar confidence. Interestingly, enumerating fewer paths results in top-ranked paths with greater cumulative weight. The low-confidence paths (that are not enumerated) no longer affect the orientation, which means that there are fewer conflicts preventing the high-confidence paths from being satisfied. This effect becomes more pronounced as a larger number of top-ranked paths are considered (e.g. 10000 and 50000), suggesting that it is preferable to consider only the top 1000 paths when limiting the number of paths that are enumerated.

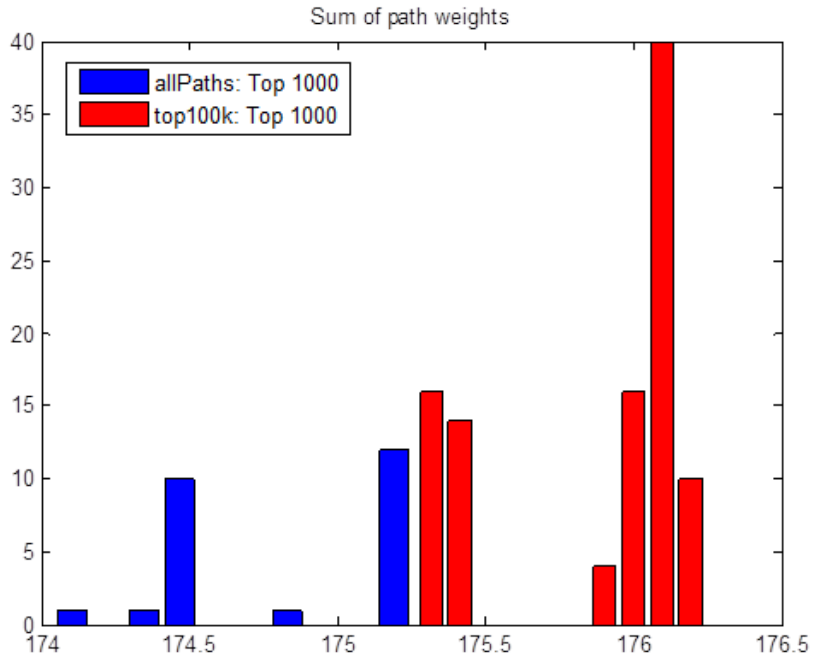

Figure 5.2: Histograms of sum of the path weights for the top 1000 paths (the number of paths used to calculate node scores). The blue histogram shows the distribution of the cumulative top path weights when all paths are enumerated. The red histogram corresponds to the approximation where only 100000 paths are used. Note that only 25 runs were used to generate blue histogram versus 100 for the red histogram, accounting for the taller peaks in the red histogram.

5.2 Human immune response to influenza infection

When moving to models of human stress response, the immune response to pathogen infection provides an excellent opportunity for SDREM. Perturbations caused by pathogen infection are ideal for SDREM analysis because there are clearly defined sources that initiate the subsequent signaling and transcriptional response. In particular, many viruses encode only a small number of proteins allowing us to generate specific models that assume the host response was triggered by host proteins that detect or interact with the viral proteins or RNA.

We initially target influenza A viruses because of the rich datasets available and, more importantly, their importance to global health. The 2009 swine-origin H1N1 virus outbreak received great public attention and was declared a global pandemic in June of that year [227]. More recently, research concerning mutations of avian H5N1 influenza that could allow it to be transmitted among mammals via aerosols have sparked immense controversy [21, 54, 92], highlighting the threat influenza A viruses pose and the need to better understand their interaction with the human immune system.

SDREM's role in comprehending the effects of influenza infection is extensive. In Section 5.2.2 we develop models of H1N1 infection and show them to have relevant functional enrichments. Section 5.2.3 contrasts H1N1 infection with the immune response to other influenza and respiratory viruses. We conclude by showing how to identify specific proteins with potential clinical reference based on SDREM models (Sections 5.2.4 and 5.2.5).

5.2.1 Related work

A comprehensive biological study of the immune response to H1N1 infection [178] is representative of current approaches used to map the affected signaling pathways. The authors identified curated pathways that are enriched for human proteins that interact with viral proteins and genes that are differentially expressed following stimulation (pathway enrichment has been used to explore other viral infections as well [41]). They focused on the four pathways that are significantly represented in both datasets, annotating the pathways with their experimental results to construct an integrated model of the response to influenza infection. Although their approach does expand the set of biological hits by including the neighboring proteins in the human PPI network (neighbors of the viral-interacting human proteins) and other members of the significant curated pathways, it does so indiscriminately. That is, all of these proteins that are putatively involved in the response were treated equally, and further experiments (targeted RNAi screens) were required to

determine which are truly relevant. In contrast, when given this same data, SDREM can predict which candidate proteins are involved in the signaling pathways and which are not. Furthermore, our yeast analysis demonstrated that even the best curated pathway databases are incomplete. Thus, such modeling approaches (even those rooted in extensive and high-quality biological experimentation) are limited in that they cannot recover hidden pathway members that are not detected experimentally (or neighboring proteins of experimental hits in this case).

Topological analysis of the human proteins that directly interact with viral proteins revealed several interesting biological principles [41, 147]. The subnetwork of the human PPI network composed of only these human proteins has higher degree, shorter shortest path length, and higher betweenness than the entire human PPI network. Even more striking is that human proteins targeted by multiple viruses have even higher degree than those that interact with a single virus [147]. Host-pathogen interactions in other species [142] similarly show that pathogen effectors preferentially interact with host hub proteins. However, none of these studies generated end-to-end models of infection response that connect the pathogens to the downstream host transcriptional response.

5.2.2 H1N1 influenza model

H1N1 influenza is the best-profiled strain in terms of proteomics and transcriptomics, with much of the data coming from a study by Shapira *et al.* [178] in which the authors conducted yeast two-hybrid experiments to map PPI between the 10 major viral proteins and roughly 12000 human proteins. Follow up work by Tafforeau *et al.* [191] similarly used yeast two-hybrid and literature mining to expand the H1N1-human interactome, and we combine these datasets along with interactions reported in the VirHostNet database [148] to obtain a comprehensive set of host-pathogen PPI. These human proteins that directly interact with viral proteins along with TLR3, TLR7, TLR8, RIG-I, and NLRP3 — proteins known to be involved in the innate immune response to influenza [91, 109, 206] — compose the set of source nodes from which the immune response originates.

Shapira *et al.* also collected time series gene expression data to identify differentially expressed genes following treatment with the wild type virus, viral RNA, interferon beta, and a NS1-deficient viral mutant. This diverse set of experiments was designed such that each would reveal unique aspects of the complex signaling and transcriptional response. The detailed time series dataset includes 10 measurements from 15 min to 18 hr. However, our preliminary analysis showed that most of the significant transcriptional changes do not appear until the 2 hr time point so we leave out the earlier measurements. We complement the expression data with condition-independent predicted protein-DNA binding

interactions [49] and the protein interaction network described in Section 5.1.2.

H1N1 infection has also been studied extensively through a collection of genome-wide [28, 100, 107] and targeted [25, 178] RNAi screens. We use these screen hits to calculate node scores for all of the proteins in the PPI network as described in Section 5.1.1. The H1N1 screen data affirms our assertion that screen hits are not a suitable choice for the signaling pathway source nodes because they may not capture the most upstream proteins involved. Of the 204 sources, only 42 (21%) are screen hits.

We ran SDREM on the wild type virus expression data to identify the TFs that control the immune response and the signaling pathways that activate them. We identified 33 target TFs and 36 proteins that connect these targets to the upstream nodes in the signaling network. These include several proteins known to be involved in immune response (e.g. the IRF family of TFs and NFKB1) and interestingly several cancer-related proteins (e.g. BRCA1, MYC, and SMAD7). Although this model of H1N1 infection is useful on its own, SDREM's flexibility allows us to pursue more in-depth analysis. Specifically, the H1N1 NS1 protein is known to suppress the signaling pathways induced by viral RNA or IFNβ, and infection with the NS1 mutant virus can reveal a broadened immune response [178]. To analyze the NS1 mutant expression data, we simply removed the 21 human proteins that interact only with NS1 from the set of sources and reran SDREM. The resulting model does contain an expanded set of active TFs (42) and a similar number of internal nodes along the signaling pathways (32). Figure 5.3 overlays the wild type and NS1 mutant immune response models. Approximately 2/3 of the predicted proteins are present in both models. The lower left corner of the network figure shows proteins that are predicted uniquely in the NS1 deletion model, proteins that are likely inhibited by the viral protein in the wild type infection.

Using the screen hits to place priors (vertex weights) on the nodes in the interaction network successfully leads SDREM to prefer pathways that contain these proteins. Of the 69 signaling proteins and TFs predicted in the wild type model, 38 (55%) are hits in one or more screens. Furthermore, running SDREM without the screen data does generate different signaling pathways. Only 39 of the 69 original predictions (57%) are still predicted when the screen data is left out, and this alternate SDREM model contains 22 new predictions among its 61 total predicted proteins (36%). Notably, NFKB1 (which was identified in one of the five screens) is omitted from the screen-independent SDREM model. Because the screen hits were provided to SDREM as input, their recovery alone does not signify that SDREM can accurately reconstruct the H1N1-activated immune response pathways. Therefore, we used DAVID [85, 86] to compute the GO [8] and curated pathway (KEGG [99] and Biocarta [151]) enrichment to assess the predictions made by the SDREM models. Source proteins were excluded in our enrichment analysis because these

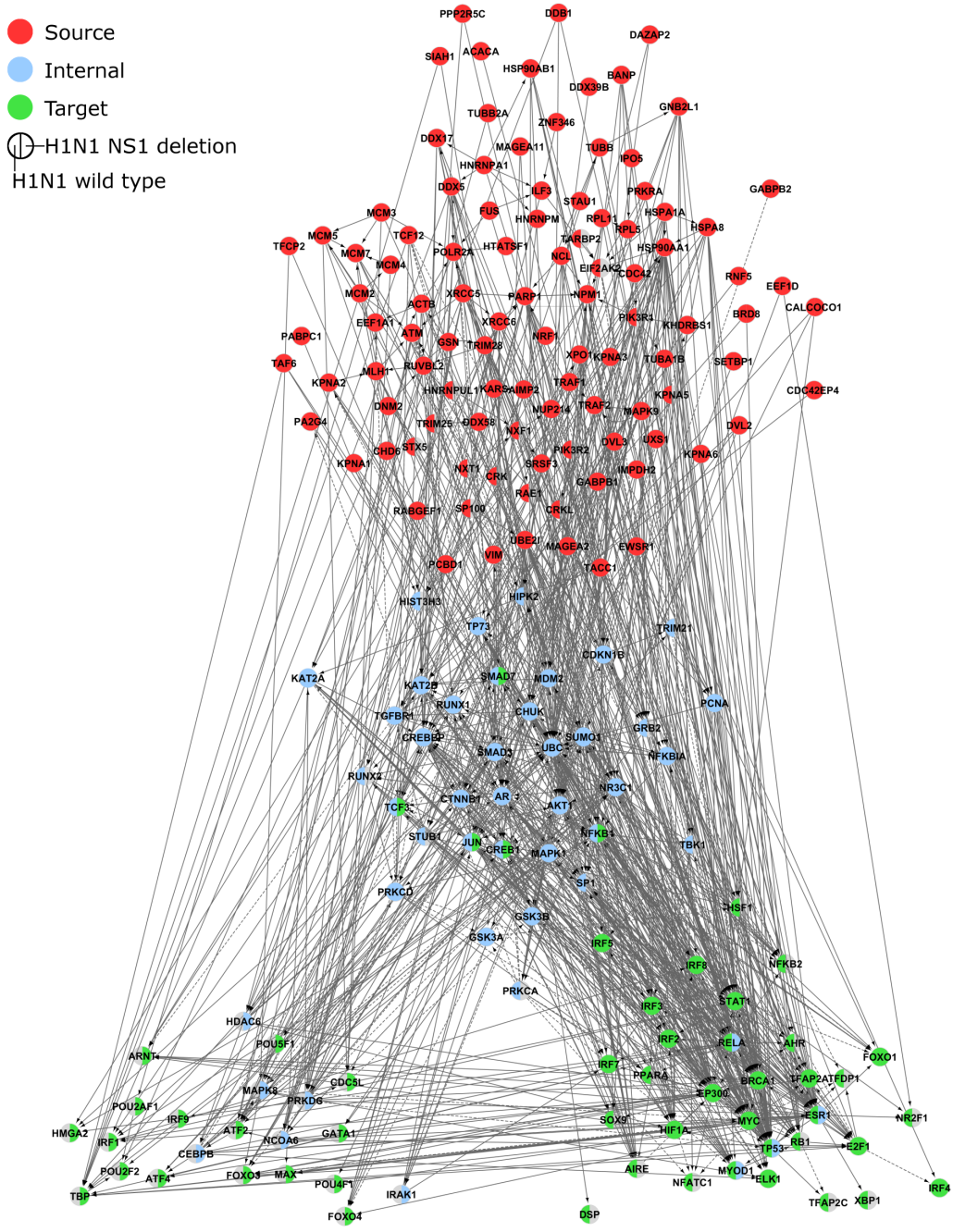

Figure 5.3: A comparison of the wild type and NS1 deletion H1N1 SDREM models. Each node in the network is present in one of both of the models. The left side of the node is colored according to that protein's role in the wild type model and the right side is for the deletion model.

are known to be virus- and infection-related and would therefore favorably bias the enrichment. Table 5.2 compares most significant GO biological process terms in the H1N1 wild type model, NS1 mutant model, and the alternate version of the H1N1 model in which the screen hits were not used to weight the nodes. Similarly, Tables 5.3 and 5.4 compare the most highly enriched pathways. All entries in these tables were significant with p-values less than 0.05 after Benjamini-Hochberg correction [20].

The enriched GO terms and KEGG pathways are rather similar across the three models, making Biocarta enrichment the most distinguishing form of evaluation. The most significant GO terms are dominated by terms related to transcriptional regulation due to the prevalence of target TFs in our predictions. However, beyond the top 10 there are many other significantly enriched terms that support SDREM's model including 'immune system development', 'immune response', 'response to virus', and 'virus-host interaction'. Many of the most significant KEGG pathways are cancer-related, consistent with previous enrichment analysis of influenza infection response [178]. Even though the TLR source nodes were left out of the enrichment analysis, the TLR signaling pathway is enriched signifying that the downstream members of this pathway were recovered successfully. The RIG-I signaling pathway only appears in the top 10 pathways when the screens are omitted, but the enrichment is actually just as strong for the other two SDREM models. The same number of predicted proteins belong to the pathway in all three models and the p-values are comparable. The Biocarta enrichment best demonstrates the advantages of including the RNAi screen hits in the input. Several virus- and immune-related pathways such as 'Human Cytomegalovirus and Map Kinase Pathways', 'The information-processing pathway at the IFN-beta enhancer', and 'T Cell Receptor Signaling Pathway' are in the top 10 pathways for the wild type H1N1 model but are not when the screens are omitted. As expected, for the NS1 deletion mutant response an additional immune pathway 'The 4-1BB-dependent immune response' is in the top 10 but is not among the top 10 for the wild type virus and not significantly enriched when screens are not used. Because the NS1 protein suppresses the immune response, SDREM is able to recover additional immune pathway proteins when NS1 is absent.

5.2.3 Comparing responses to different respiratory viruses

Our comparison of SDREM's H1N1 wild type and NS1 mutant infection models demonstrated the insights that can be gained by examining variants of a particular immune response. However, because both expression datasets were generated in the same experimental conditions by the same researchers, it would have been feasible to directly compare the expression datasets [178]. This approach does not reveal the diversity of the signaling

Table 5.2: Most significantly enriched GO biological process terms in the SDREM H1N1 models

Wild type	**NS1 deletion**	**No screens**
regulation of transcription, DNA-dependent	regulation of transcription	regulation of transcription from RNA polymerase II promoter
regulation of RNA metabolic process	regulation of transcription, DNA-dependent	regulation of transcription, DNA-dependent
regulation of transcription from RNA polymerase II promoter	regulation of RNA metabolic process	regulation of RNA metabolic process
positive regulation of macromolecule metabolic process	positive regulation of gene expression	positive regulation of gene expression
regulation of transcription	positive regulation of macromolecule biosynthetic process	regulation of transcription
positive regulation of gene expression	regulation of transcription from RNA polymerase II promoter	positive regulation of transcription
positive regulation of transcription	positive regulation of transcription	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
positive regulation of nitrogen compound metabolic process	positive regulation of macromolecule metabolic process	positive regulation of macromolecule metabolic process
positive regulation of macromolecule biosynthetic process	positive regulation of biosynthetic process	positive regulation of nitrogen compound metabolic process
positive regulation of cellular biosynthetic process	positive regulation of nitrogen compound metabolic process	positive regulation of macromolecule biosynthetic process

Table 5.3: Most significantly enriched KEGG pathways in the SDREM H1N1 models

Wild type	**NS1 deletion**	**No screens**
Prostate cancer	Prostate cancer	Chronic myeloid leukemia
Pathways in cancer	Pathways in cancer	Pathways in cancer
Chronic myeloid leukemia	Chronic myeloid leukemia	Prostate cancer
Pancreatic cancer	Neurotrophin signaling pathway	Cell cycle
Toll-like receptor signaling pathway	Toll-like receptor signaling pathway	Toll-like receptor signaling pathway
Cell cycle	Pancreatic cancer	Pancreatic cancer
B cell receptor signaling pathway	Colorectal cancer	Cytosolic DNA-sensing pathway
Small cell lung cancer	Small cell lung cancer	TGF-beta signaling pathway
Colorectal cancer	Cell cycle	RIG-I-like receptor signaling pathway
Neurotrophin signaling pathway	Endometrial cancer	Bladder cancer

pathways involved like SDREM, but provides a basic understanding of the different transcriptional activity. One of SDREM's major strengths is that it allows for the comparison of different stress responses for which only incompatible data is available.

To highlight this feature, we applied SDREM to two additional influenza A strains as well as the severe acute respiratory syndrome (SARS) virus. H3N2 influenza is a common seasonal (as opposed to pandemic) influenza. H5N1 influenza (aka "bird flu") can occur in highly pathogenic forms and recent work to mutate the virus sparked fears of a potential pandemic (Section 5.2). Although it is also a single-stranded respiratory virus, the SARS virus and its interactions with the immune system are less similar than those among the three influenza subtypes. The 2003 SARS epidemic generated interested in the mechanisms of the SARS virus's interactions with the immune system [55], and we include it here to determine what similarities may exist between the host responses to diverse respiratory viruses. This task is quite challenging due to the vastly different datasets available for these four viruses (Table 5.5). H1N1 is by far the best-characterized of the four and is the only one for which genome-wide RNAi screens are available. H3N2-human PPI have been thoroughly studied, but for both H5N1 and SARS large-scale PPI screens have not yet been conducted, likely because of the pathogenicity of these viruses. Moreover, the expression datasets come from very disparate settings affecting the magnitudes of differential expression and the significant genes. The H3N2 expression data is difficult to

Table 5.4: Most significantly enriched Biocarta pathways in the SDREM H1N1 models

Wild type	**NS1 deletion**	**No screens**
Influence of Ras and Rho proteins on G1 to S Transition	NFkB activation by Nontypeable Hemophilus influenzae	Acetylation and Deacetylation of RelA in The Nucleus
NFkB activation by Nontypeable Hemophilus influenzae	ATM Signaling Pathway	Pelp1 Modulation of Estrogen Receptor Activity
Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa(alpha)	The information-processing pathway at the IFN-beta enhancer	Influence of Ras and Rho proteins on G1 to S Transition
ATM Signaling Pathway	Influence of Ras and Rho proteins on G1 to S Transition	NFkB activation by Nontypeable Hemophilus influenzae
Acetylation and Deacetylation of RelA in The Nucleus	MAPKinase Signaling Pathway	Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa(alpha)
Human Cytomegalovirus and Map Kinase Pathways	Acetylation and Deacetylation of RelA in The Nucleus	Cell Cycle: G1/S Check Point
The information-processing pathway at the IFN-beta enhancer	AKT Signaling Pathway	Role of ERBB2 in Signal Transduction and Oncology
MAPKinase Signaling Pathway	The 4-1BB-dependent immune response	Cell Cycle: G2/M Checkpoint
Cell Cycle: G1/S Check Point	Erythropoietin mediated neuroprotection through NF-kB	Sumoylation by RanBP2 Regulates Transcriptional Repression
T Cell Receptor Signaling Pathway	Hypoxia and p53 in the Cardiovascular system	CARM1 and Regulation of the Estrogen Receptor

compare to the other expression datasets because it was collected from human volunteers in a clinical setting and the expression changes are affected by the heterogeneity of these individuals. For our analysis we only consider expression data from patients who exhibited symptoms and truncate the last time points where the expression changes are more moderate.

Table 5.5: The conditions in which the expression data was collected are quite diverse and there are varying degrees of network information available for the respiratory viruses. The data sources used for H3N2, H5N1, and SARS are also indicated (see Section 5.2.2 for H1N1).

Feature	**H1N1**	**H3N2**	**H5N1**	**SARS**
Expression data collected from	Cell lines	Volunteers [90]	Cell lines [125]	Cell lines [223]
Time scale	18 hr	108 hr	24 hr	48 hr
Expression data time points	10	14	5	3
Time points used in analysis	6	6	5	3
Source proteins	204	153 [91, 109, 148, 178, 191, 206]	41 [35, 88, 91, 109, 123, 130, 148, 179, 191, 206, 207]	27 [55, 148]
RNAi screen hits	1009	0	32 [25]	0

Despite these differences, we successfully analyzed all four viruses with SDREM and obtained models that can be directly compared with one another. Table 5.6 shows that even among the various influenza A viruses SDREM identifies unique signaling proteins and TFs that are active in the immune response. Given the similarity of the flu viruses, it is surprising how little overlap there is among the human proteins that interact with the viral proteins or RNA, although some of this discrepancy could be due to the limited coverage. Figure 5.4 shows the complete signaling pathways predicted for all four viruses. Each quadrant of the nodes is colored to indicate what role, if any, the proteins play in each SDREM model.

Some of the more interesting proteins in the SDREM models are those that are included in all four models or all but one model. Proteins that are common to the immune response as well as proteins that are only omitted in a single model are worthwhile predictions for further follow up. Table 5.7 lists these important proteins for the respiratory virus models.

Table 5.6: The overlaps among the curated source proteins and the SDREM predictions (preds) for each virus. The percentages are calculated by taking the overlap relative to the number of proteins for the virus named in the row.

Compare to	**H1N1 sources**	**H3N2 sources**	**H5N1 sources**	**SARS sources**	**H1N1 preds**	**H3N2 preds**	**H5N1 preds**	**SARS preds**
H1N1	-	37%	8%	3%	-	38%	41%	55%
H3N2	50%	-	11%	4%	20%	-	19%	21%
H5N1	42%	39%	-	15%	31%	28%	-	39%
SARS	22%	22%	22%	-	34%	25%	31%	-

5.2.4 Predicting RNAi screen hits

SREM produces very specific, complete models of human immune response to viral infection at the signaling and transcriptional levels, but it is not always clear which predicted proteins are most likely to have clinical relevance or most appropriate for experimental follow up. Therefore, we developed a technique for ranking the genes predicted by SDREM to determine which are most likely to produce a phenotypic effect (e.g. change in viral load) when silenced by RNAi experiments. The ability to predict RNAi targets is important even for viruses for which genome-wide screens are available because when tens of thousands of genes are screened, it is impossible to globally verify that each silencing RNA is targeting the correct gene and only that specific gene. The disagreement among genome-wide screens (Table 5.1) implies that there is still work to be done in generating high-quality sets of H1N1 screen hits. Furthermore, viruses like H5N1 are challenging to work with because they require a biosaftey level 3 lab, and it is unlikely that genome-wide RNAi screens will be produced in the near future, if ever, for such pathogens. An approach that used screen hits from H1N1 to guide small-scale H5N1 RNAi screens successfully identified H5N1-relevant genes [25], but this strategy is fundamentally unable to identify which H5N1-affected genes contribute to the oftentimes severe response to H5N1 but not the immune response to milder forms of H1N1 or H3N2. Our goal is to first hone a method that can accurately predict RNAi screen hits and then apply it to H1N1 and H5N1 to search for candidate H5N1 screen targets that have not been implicated in H1N1. Recovering such genes would have significant clinical importance if they could be validated experimentally.

Our strategy for identifying RNAi screen hits is to estimate the *in silico* effects of removing a protein from the signaling network. Specifically, we compute how the connectivity to the targets is affected when a node is removed. Because the simulated phenotype

Table 5.7: The genes that encode the proteins common to all or all but one of the SDREM respiratory virus models. The table also provides the number of H1N1 screens that the gene was identified in, if any.

Gene	SDREM models	H1N1 screens
AKT1	All but H5N1	1
AR	All	0
BRCA1	All	0
CREB1	All but H3N2	1
CREBBP	All	0
CTNNB1	All but H5N1	1
DDX58	All	0
DSP	All but H3N2	1
E2F1	All	1
EP300	All	1
ESR1	All	0
GABPB1	All but SARS	0
HIF1A	All but SARS	0
HSF1	All but H3N2	0
JUN	All	2
KAT2B	All but H5N1	0
KPNA1	All but SARS	1
KPNA2	All	1
KPNA6	All but SARS	0
MAPK1	All	1
MDM2	All	2
MYC	All but H5N1	2
NFKBIA	All but H3N2	1
NLRP3	All but SARS	0
NR3C1	All	0
PPIA	All but H1N1	0
RB1	All	0
RELA	All but H3N2	0
SMAD3	All but SARS	0
SMARCA4	All but H1N1	0
SP1	All	0
SRC	All but H1N1	0
STAT1	All but H3N2	0
STAU1	All but SARS	0
SUMO1	All	1
TCF3	All but H5N1	1
TLR3	All	0
TLR7	All	0
TLR8	All but SARS	0
TP53	All	0
UBC	All	1
UBE2I	All but H5N1	0

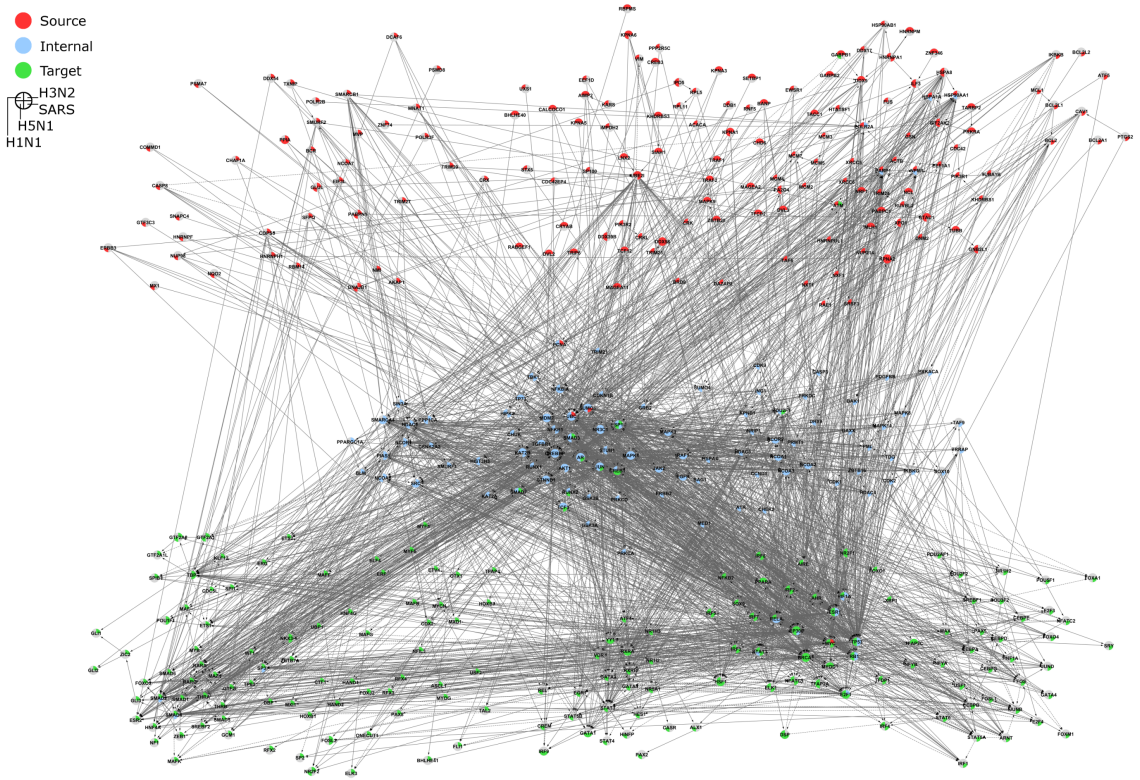

Figure 5.4: A comparison of the four respiratory virus SDREM models. The quadrants correspond to the four viruses, and the colors indicate the role of the protein. Grey quadrants indicate that the protein is not included in the SDREM model for that particular virus.

is computed with respect to the connectivity of a target or targets, it is applicable only to the source and internal nodes. Removing a target node from the network would trivially cause that target to be completely disconnected from the sources, but this is not informative about the likelihood that the target's deletion would have a phenotypic effect. We devise several scoring metrics to quantify the effect of a node deletion upon the targets that vary three factors. *All* versus *Top* denotes whether all satisfied paths or only the top-ranked paths are used to calculate change in connectivity. Source-target *Pairs* versus *Targets* determines whether a target's connectivity is evaluated separately for every source (i.e. each source activates a target differently) or if a target is considered to be disconnected only when it is no longer reachable from any source (i.e. any source can activate the target). *Weighted* versus *Unweighted* specifies if connectivity is treated in a binary (connected or

disconnected) or continuous (how much connectivity is lost) fashion. The score for the weighted variant is

$$score_w(n) = \frac{\sum_{t \in T} \frac{\sum_{p \in P(t)} w(p) I(n \notin N(p))}{\sum_{p \in P(t)} w(p)}}{|T|}$$

where n is the deleted node, T is the set of all targets, $P(t)$ is the set of paths to the target t to be considered (depending on the choice of All vs. Top and Pairs vs. Targets), $w(p)$ is the path weight, $I(*)$ is an indicator function, and $N(p)$ is the set of nodes on the path. Intuitively, this score is the fraction of path weight that exists along paths that can still activate t after n is deleted averaged across all targets. The unweighted score is

$$score_u(n) = \frac{\sum_{t \in T} \left(1 - \prod_{p \in P(t)} I(n \in N(p))\right)}{|T|}$$

This score is the fraction of targets that are still reachable after removing n.

To determine which scoring metric is most predictive of RNAi screen hits, we ran SDREM on the H1N1 data but excluded the screen hits from the input. We used the metrics to rank all 252 non-target proteins in the model using node connectivity score (Section 4.2.3) and interaction network degree to break ties in that order. Given the rankings for each metric, we calculated the area under the curve (AUC) [129] using the H1N1 screen hits as the positive set. 57 of the 252 nodes in the SDREM model are screen hits, and Table 5.8 shows the AUC for each metric and how many of these 57 hits are identified at various thresholds. The best-performing metric considers only the top paths, allows targets to be activated by any source, and uses the weighted score. Not only does this metric yield the best AUC, but it also finds the most (or ties for the most) screen hits at the 10, 20, and 100 prediction thresholds. This metric's predictions significantly overlap with the known screen hits at all thresholds, which is also the case for most of the other metrics. Because the overlaps are significant even for the worst-performing metric when 20 or more genes are predicted, we conclude that the SDREM model itself is a powerful filter for predicting screen hits.

Having used the H1N1 data to select a ranking metric, we then turned to the (independent) H5N1 datasets to predict RNAi screen hits for this influenza strain. Using the scores from the best metric, we ranked all of the sources and internal nodes in the H5N1 SDREM model and compared the ranks to those we obtained using the same scoring metric on the H1N1 data. We also examined the degree of the top-ranked predictions because we expected that high-scoring nodes would be of high degree since such nodes are likely to affect a large number of targets when deleted. Table 5.9 shows the top 30 H5N1 predictions. The first few predictions have substantially lower scores than the subsequent

Table 5.8: The scoring metrics that were used to predict known H1N1 screen hits. The metrics are sorted by AUC. The number of screen hits recovered at various thresholds is shown along with the significance (in parentheses) calculated using Fisher's exact test.

Paths used	**Connectivity**	**Score**	**AUC**	**Hits in top 10**	**Hits in top 20**	**Hits in top 50**	**Hits in top 100**	**Hits in top 150**
Top	Targets	Weighted	0.722	6 (1.97 E-5)	8 (3.44 E-5)	18 (3.24 E-9)	42 (9.42 E-23)	47 (4.85 E-19)
Top	Pairs	Weighted	0.717	3 (2.87 E-2)	7 (2.88 E-4)	20 (4.59 E-11)	40 (8.68 E-21)	47 (4.85 E-19)
Top	Pairs	Unweighted	0.716	3 (2.87 E-2)	8 (3.44 E-5)	20 (4.59 E-11)	37 (5.59 E-18)	47 (4.85 E-19)
All	Targets	Unweighted	0.711	2 (0.153)	5 (1.09 E-2)	20 (4.59 E-11)	39 (7.82 E-20)	47 (4.85 E-19)
All	Targets	Weighted	0.706	3 (2.87 E-2)	6 (1.97 E-3)	18 (3.24 E-9)	39 (7.82 E-20)	49 (1.13 E-20)
Top	Targets	Unweighted	0.704	2 (0.153)	6 (1.97 E-3)	19 (4.02 E-10)	39 (7.82 E-20)	47 (4.85 E-19)
All	Pairs	Weighted	0.702	3 (2.87 E-2)	6 (1.97 E-3)	18 (3.24 E-9)	36 (4.43 E-17)	49 (1.13 E-20)
All	Pairs	Unweighted	0.676	2 (0.153)	6 (1.97 E-3)	18 (3.24 E-9)	36 (4.43 E-17)	45 (1.83 E-17)

genes. Recall from the weighted scoring equation above that the scores denote the fraction of target connectivity remaining after the *in silico* deletion so lower scores translate to a greater predicted effect. Ten of the H5N1 predictions — STAT3, CASP8, HSF1, ERBB3, NRIP1, PRMT1, YY1, RXRA, STUB1, and NR3C1 — are particularly interesting because they are neither known H1N1 RNAi screen hits nor in the top 100 H1N1 predicted hits. These genes are therefore the most appropriate for experimental follow up in order to determine whether their removal truly does affect H1N1 and H5N1 differently. Two of these genes, CASP8 and ERBB3, have been reported to directly interact with H5N1 viral proteins but not H1N1 even though the H1N1-human PPI have much greater coverage, suggesting that they may indeed play distinct roles. Many of the top-ranked H5N1 genes are sources (which directly interact with the viral proteins) or high-degree nodes, and in both cases it is perhaps unsurprising that deletion of such genes would have substantial impact on the downstream targets. In contrast, NRIP1 and PRMT1 are neither sources nor of high-degree. Their inclusion in the top predictions is noteworthy because the paths through these nodes affect targets to a greater extent than expected.

Table 5.9: The top-ranked H5N1 RNAi screen hit predictions alongside the known and top-ranked predicted H1N1 screen hits. N/A indicates that the gene was not included in the SDREM H1N1 model and was therefore not ranked.

Gene	**H1N1 source**	**H5N1 source**	**Degree**	**H1N1 screens**	**H5N1 screens**	**H5N1 score**	**H1N1 rank**	**H5N1 rank**
HSPA8	Y	Y	95	1	1	0.765	78	1
PA2G4	Y	Y	26	1	1	0.815	66	2
AR	N	N	452	0	0	0.836	12	3
ILF3	Y	Y	39	1	1	0.901	75	4
ESR1	N	N	502	0	0	0.908	11	5
KPNA2	Y	Y	50	1	1	0.915	93	6
TP53	N	N	655	0	0	0.918	2	7
STAT3	N	N	419	0	0	0.924	151	8
CREBBP	N	N	265	0	0	0.928	53	9
SP1	N	N	365	0	0	0.931	92	10
RB1	N	N	257	0	0	0.934	5	11
GNB2L1	Y	Y	68	0	0	0.937	69	12
CASP8	N	Y	104	0	0	0.940	262	13
UBC	N	N	485	1	0	0.948	4	14
EIF2AK2	Y	Y	40	1	0	0.948	7	15
HSF1	N	N	217	0	0	0.950	N/A	16
EP300	N	N	377	1	0	0.951	3	17
BRCA1	N	N	301	0	0	0.954	49	18
NUP98	N	Y	36	2	0	0.955	N/A	19
ERBB3	N	Y	37	0	0	0.963	N/A	20
NRIP1	N	N	48	0	0	0.964	N/A	21
STAT1	N	N	642	0	0	0.964	22	22
PRMT1	N	N	70	0	0	0.964	147	23
KPNA1	Y	Y	26	1	1	0.967	216	24
HSP90AA1	Y	N	144	2	1	0.968	9	25
YY1	N	N	380	0	0	0.969	215	26
XPO1	Y	Y	64	1	0	0.970	55	27
RXRA	N	N	1020	0	0	0.970	N/A	28
STUB1	N	N	118	0	0	0.971	124	29
NR3C1	N	N	561	0	0	0.971	184	30

5.2.5 Predicting genetic interactions

Moving beyond predicting the effects of single gene deletions, we can also use the above approach to identify genetic interactions. Genetic interactions are functional interactions between pairs of genes where simultaneous double deletion has a smaller or greater than expected effect. Humans have tens of thousands of genes, making it impossible to comprehensively screen for all possible genetic interactions in a condition of interest as is done to identify the phenotypic effects of single gene loss. Therefore, computational analysis that can suggest pairs for experimental testing is very valuable.

Experimental studies of genetic interactions in yeast [38, 39, 94, 197] guide our strategy for prediction genetic interactions that are relevant to influenza infection. Initially pairs of genes were classified as having or not having a genetic interaction with less emphasis on the strength of the interaction [197]. More recent work has focused on quantifying genetic interactions on a continuous scale as

$$\epsilon_{AB} = P_{AB}^{\text{observed}} - P_{AB}^{\text{expected}}$$

where ϵ_{AB} is the interaction between genes A and B and P_{AB} is the phenotype when both A and B are deleted [38]. Typically the expected phenotype is defined as the product of the phenotypes observed in the individual single deletions such that

$$\epsilon_{AB} = P_{AB}^{\text{observed}} - P_{A}^{\text{observed}} P_{B}^{\text{observed}}$$

Using this definition, negative interactions occur when the double knockout has a stronger effect than expected because stronger effects correspond to lower values of P_{AB}^{observed}. Positive interaction may suggest that the pairs of genes are members of the same pathway because the observed double deletion effect is weaker than expected. If the pathway is disabled by the deletion of one of the two genes, loss of the second gene has little additional effect.

In yeast experimental work, colony size is a typical phenotype [38, 39, 197] because it approximates growth rate, but other possibilities exist [94]. In our simulations, the score defined in Section 5.2.4 is the *in silico* phenotype. Like colony size, in our score more significant deletions result in lower values, and it is meaningful to take the product of the scores from two individual deletions. We generalized the score defined for predicting RNAi screen hits such that

$$P_{AB}^{\text{observed}} = score_w(A, B) = \frac{\sum_{t \in T} \frac{\sum_{p \in P(t)} w(p) I(A \notin N(p)) I(B \notin N(p))}{\sum_{p \in P(t)} w(p)}}{|T|}$$

Based on our previous results for the known H1N1 screens (Table 5.8) we again consider only the top-ranked paths and allow targets to be activated by any single source. Although it is possible to calculate the degree of genetic interaction for all human genes using our approach, we only predict such interactions for pairs of nodes that are present in the SDREM model. This choice is motivated by experimental work that claims focusing on a targeted set of genes improves the signal-to-noise ratio of the results [38]. We predicted genetic interactions that affect H1N1 (Table 5.10) and H5N1 (Table 5.11) infection. Some of these predicted pairs such as ILF3-PA2G4 are especially interesting because the two proteins are not high degree nodes in the PPI network.

Table 5.10: The top 20 predicted H1N1 genetic interactions.

Gene A	**Gene B**	ϵ_{AB}	P_{AB}^{ob}	P_{AB}^{ex}	P_{A}^{ob}	P_{B}^{ob}
EP300	TP53	-0.0077	0.8152	0.8229	0.9158	0.8986
TRAF2	UBE2I	-0.0070	0.8275	0.8345	0.9348	0.8927
UBC	UBE2I	-0.0070	0.8256	0.8326	0.9327	0.8927
RB1	TP53	-0.0068	0.8316	0.8384	0.9330	0.8986
TP53	TRAF2	-0.0066	0.8333	0.8400	0.8986	0.9348
RB1	UBE2I	-0.0057	0.8272	0.8329	0.9330	0.8927
EP300	UBC	-0.0057	0.8485	0.8541	0.9158	0.9327
EP300	TRAF2	-0.0055	0.8506	0.8561	0.9158	0.9348
EIF2AK2	UBE2I	-0.0053	0.8432	0.8485	0.9505	0.8927
NPM1	UBE2I	-0.0052	0.8442	0.8494	0.9515	0.8927
HSP90AA1	UBE2I	-0.0052	0.8447	0.8498	0.9520	0.8927
PARP1	UBE2I	-0.0051	0.8456	0.8506	0.9529	0.8927
TP53	UBE2I	-0.0047	0.7974	0.8022	0.8986	0.8927
RB1	UBC	-0.0045	0.8657	0.8702	0.9330	0.9327
POLR2A	UBE2I	-0.0044	0.8512	0.8557	0.9585	0.8927
TCF12	UBE2I	-0.0044	0.8515	0.8559	0.9588	0.8927
AR	TP53	-0.0043	0.8562	0.8605	0.9576	0.8986
EP300	RB1	-0.0042	0.8503	0.8545	0.9158	0.9330
TCF12	TP53	-0.0042	0.8573	0.8615	0.9588	0.8986
EP300	EIF2AK2	-0.0042	0.8663	0.8705	0.9158	0.9505

Table 5.11: The top 20 predicted H5N1 genetic interactions.

Gene A	**Gene B**	ϵ_{AB}	P_{AB}^{ob}	P_{AB}^{ex}	P_{A}^{ob}	P_{B}^{ob}
HSPA8	PA2G4	-0.0435	0.5798	0.6233	0.7647	0.8151
HSPA8	AR	-0.0370	0.6025	0.6396	0.7647	0.8363
HSPA8	ILF3	-0.0234	0.6655	0.6888	0.7647	0.9008
HSPA8	KPNA2	-0.0199	0.6801	0.7000	0.7647	0.9154
ILF3	PA2G4	-0.0184	0.7159	0.7342	0.9008	0.8151
ILF3	AR	-0.0162	0.7371	0.7533	0.9008	0.8363
KPNA2	PA2G4	-0.0156	0.7305	0.7462	0.9154	0.8151
GNB2L1	HSPA8	-0.0148	0.7018	0.7166	0.9371	0.7647
ESR1	PA2G4	-0.0142	0.7261	0.7403	0.9082	0.8151
HSPA8	CASP8	-0.0142	0.7045	0.7187	0.7647	0.9398
CREBBP	HSPA8	-0.0141	0.6952	0.7093	0.9275	0.7647
AR	KPNA2	-0.0138	0.7517	0.7656	0.8363	0.9154
HSPA8	STAT3	-0.0130	0.6934	0.7064	0.7647	0.9238
ESR1	AR	-0.0123	0.7473	0.7596	0.9082	0.8363
HSPA8	EIF2AK2	-0.0122	0.7130	0.7251	0.7647	0.9483
GNB2L1	PA2G4	-0.0116	0.7522	0.7639	0.9371	0.8151
AR	TP53	-0.0114	0.7563	0.7677	0.8363	0.9179
PA2G4	CASP8	-0.0111	0.7549	0.7660	0.8151	0.9398
HSPA8	NUP98	-0.0105	0.7199	0.7305	0.7647	0.9552
PA2G4	TP53	-0.0104	0.7378	0.7482	0.8151	0.9179

5.3 Fully probabilistic model

Despite SDREM's practical successes it lacks a global probabilistic interpretation and unified objective function, which motivates us to explore reframing the approach as a single probabilistic graphical model. Rather than iteratively focusing on transcriptional activity and network connectivity, the unified graphical model presented in this section simultaneously accounts for both forces.

5.3.1 Model definition

Given DREM's success in modeling dynamic gene expression, our goal is to preserve as much of DREM's underlying graphical model (the IOHMM) as possible as we extend it.

As in SDREM, variables for TF activity are introduced. However, in the unified model the TF variables are connected to additional sets of variables that account for the connectivity in the signaling network. Similarly, we still require a small set of known upstream proteins that initiate the stress response (the sources), a partially oriented physical interaction network (which may contain unoriented PPI and directed protein-DNA binding and PTM edges), dynamic gene expression data, and protein-DNA binding interactions for the TFs of interest. The objectives are to infer the directionality of all unoriented edges in the interaction network, determine which TFs control the stress response, model the transcriptional changes of the genes regulated by these TFs, and learn the directed signaling cascades that activate these TFs.

Three sets of variables are used to orient the interaction network and represent TF connectivity in this network: T for TFs, P for potential source-target paths, and D for edge directions. The construction of T is the most straightforward. For each potential regulator in the input protein-DNA interaction dataset we create a $t \in T$. The discrete t can take the values $\{0.1, 0.5, 0.9\}$ reflecting that regulator's activity in the response (larger values denote greater activity). Given a path length bound, we enumerate all possible paths from each source to each t (as in Section 3.2.1) via depth first search in the interaction network. All edges that are used in the same direction in all paths have their orientation deterministically fixed (Figure 5.5). For the other edges, termed conflict edges, we create a binary variable $d \in D$. If all proteins in the interaction network are assigned a numeric index arbitrarily, then let $d = 1$ correspond to orienting the edge toward the protein with the greater index. Lastly, for each source-target path, we create a binary path variable $p \in P$. $p = 1$ means that the path participates in activating the target TF.

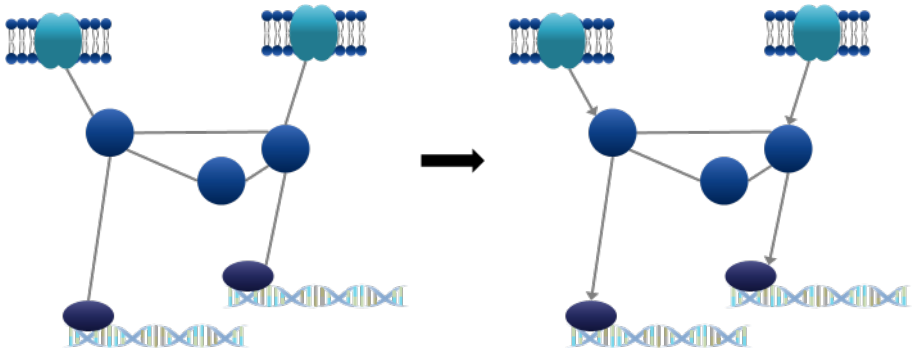

Figure 5.5: Physical interactions that are used in the same manner by all possible source-target paths can be oriented deterministically in a preprocessing step. Edge variables are not required for these edges or other interactions with a previously known orientation.

Given these variables, we now describe the relationships between them and the induced graphical model structure. To make these relationships explicit, we employ a factor graph representation [106, 114]. In a factor graph, circles represent variables and squares represent the factors, or functions, defined over those variables. Figure 5.6 shows how each source-target path in the interaction network is transformed into variables and factors in the factor graph and how these components of the factor graph are combined.

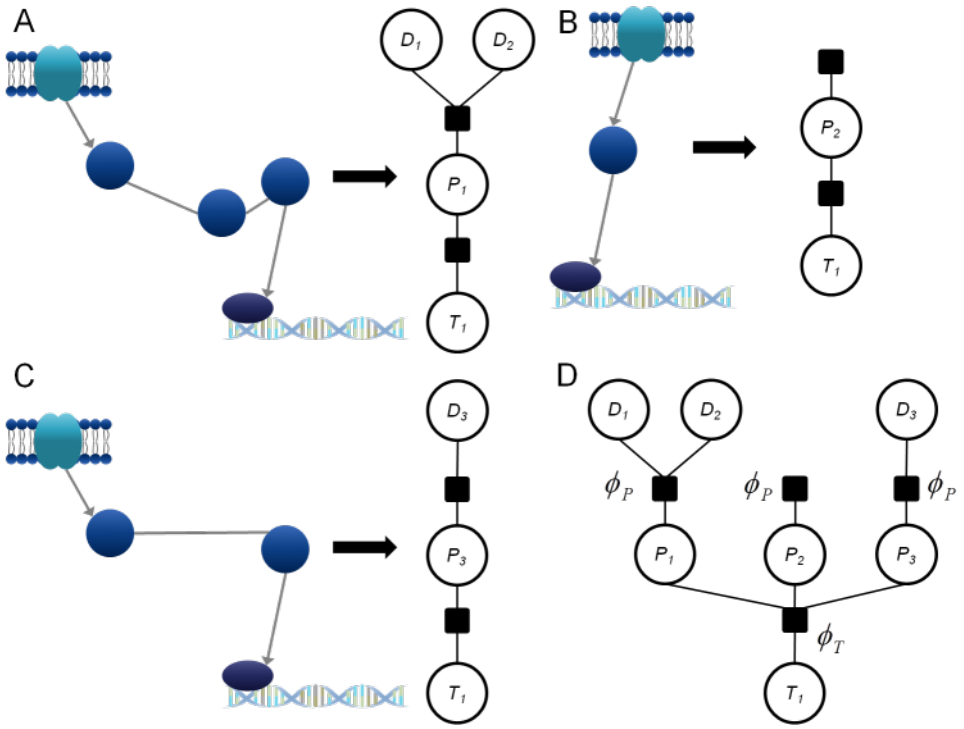

Figure 5.6: The factor graph components created for all paths to the rightmost target in Figure 5.5. A) For a source-target path with conflict edges, an edge direction variable is created for each conflict edge. A unique path variable is created for every path. B) Paths that use only edges with a known orientation do not connect to edge variables in the factor graph. C) The final path to the target. D) The complete factor graph structure. A single target factor is created and connected to t and all p that end at t.

Factors ϕ_P are defined over the edge direction variables and path variable for each

source-target path in the interaction network such that

$$\phi_P(d \in O(p), p) = \begin{cases} w(p), & \text{if } O(p) = \emptyset \\ w(p), & \text{if } \left(I(p=1)\prod_{d_i \in O(p)} I(d_i = \widehat{d}_i)\right) = 1 \\ 1, & \text{if } p = 0 \\ \epsilon, & \text{otherwise} \end{cases}$$

$O(p)$ is the set of edge variables on the path p. $w(p)$ is the path weight precomputed from the interaction network as in Section 4.2.3 divided by the constant C. That is, $w(p) = \frac{w'(p)}{C}$ where $w'(p)$ is the original unscaled path weight. We require that C is set such that $C < \min_{p \in P}(w'(p))$, which ensures $w(p) > 1 \; \forall p \in P$. The path weight normalization is required because some path weights may be less than 1 but we always want a configuration in which a path is satisfied to be more rewarding than the configuration where $p = 0$. $I(*)$ is an indicator function and $\widehat{d}_i$ is the direction that the edge was traversed in the depth first search of the interaction network. Intuitively, ϕ_P returns a large value, the normalized path weight, if the path is satisfied. A path is satisfied when it does not contain any conflict edges or when all conflict edges are oriented toward the target. If the path is not used to activate TFs ($p = 0$), the path is ignored in the objective. However, configurations where the path is not satisfied but $p = 1$ are penalized by taking the very small value ϵ.

Factors ϕ_T are defined over all paths that lead to a particular target in the interaction network and the corresponding target variable t. Table 5.12 provides the function values for each of the nine possible states. $E(t)$ is the set of paths that end at target t. ϕ_T examines the weighted fraction of *satisfied* paths that end at t (as opposed to other targets) relative to expectation. The expectation can be precomputed from the interaction network structure and is the ratio of *all* paths that end at t.

$$expected(t) = \frac{\sum_{p \in E(t)} w(p)}{\sum_{p \in P} w(p)}$$

This simple form of the expectation does not model the actual expectation of each path being satisfied, which varies depending on the number of conflict edges on the path. The definition of ϕ_T encourages targets with more satisfied paths than expected to take on higher activity values. Although ϕ_T could have instead been compared to the ratio of paths to t that are satisfied or to a fixed ratio, e.g. 0.01, in order to mimic the SDREM node scores more closely, we deviate from that design to help remove the bias against low-degree nodes that was observed in the SDREM yeast stress response analysis (Section 4.3.9). Comparing to the expectation helps ensure that high-degree and highly-connected TFs do not dominate. Although these TFs will have a large number of potential pathways to them,

they are required to have a greater number of satisfied paths to them than less-connected TFs in order to be considered as well-connected in the oriented network.

Table 5.12: Values of $\phi_T(p \in E(t), t)$. The rows describe the connectivity of t relative to expectation and the columns denote the activity level of t.

	$t = 0.1$	$t = 0.5$	$t = 0.9$
$\frac{\sum_{p \in E(t)} w(p) I(p=1)}{\sum_{p \in P} w(p) I(p=1)} < 0.75 expected(t)$	0.9	0.5	ϵ
$0.75 expected(t) \leq \frac{\sum_{p \in E(t)} w(p) I(p=1)}{\sum_{p \in P} w(p) I(p=1)} < 1.25 expected(t)$	0.5	0.9	0.5
$1.25 expected(t) \leq \frac{\sum_{p \in E(t)} w(p) I(p=1)}{\sum_{p \in P} w(p) I(p=1)}$	ϵ	0.5	0.9

Because the regulatory path component of SDREM is already a probabilistic graphical model, it requires no fundamental adaptations. In SDREM, the IOHMM took the protein-DNA binding interactions and TF activity priors as input for the transition probability functions. Here, the activity priors are replaced with the T variables, which serve as the interface between the network and regulatory components. As a graphical model, the IOHMM can be represented as a factor graph with factors for each transition and emission probability. Figure 5.7 shows an example of the regulatory portion of the unified model for two time points, but this structure generalizes for any number of time points.

The factor $\phi_{H_t, H_{t+1}, H'_{t+1}}$ corresponds to the IOHMM's transition probability, which uses logistic regression to map from protein-DNA interactions and TF activity to state transitions. An alternate version $\phi_{H_t, H_{t+1}}$ is used when there is a deterministic state transition because there is only one possible subsequent state. Figure 5.7 depicts both variants, and as in [50] the sets of hidden states H branch in a tree structure. Note that when used as a subscript t refers to a time point in the gene expression data not a target.

$$\phi_{H_t, H_{t+1}, H'_{t+1}}(g, B, T) = \frac{1}{1 + e^{\left(-\psi_0 - \sum_{i=1}^{|T|} \psi_i B_{i,g} t_i\right)}}$$

ψ is the weight vector learned by the logistic regression classifier, as described in Section 5.3.2. $B_{i,g}$ is the binding value for TF t_i and gene g in the protein-DNA interaction dataset. The equation is shown for a 2-way split, but generalizes to multi-way splits as in [50, 112]. When the graphical model's structure search determines there is no split

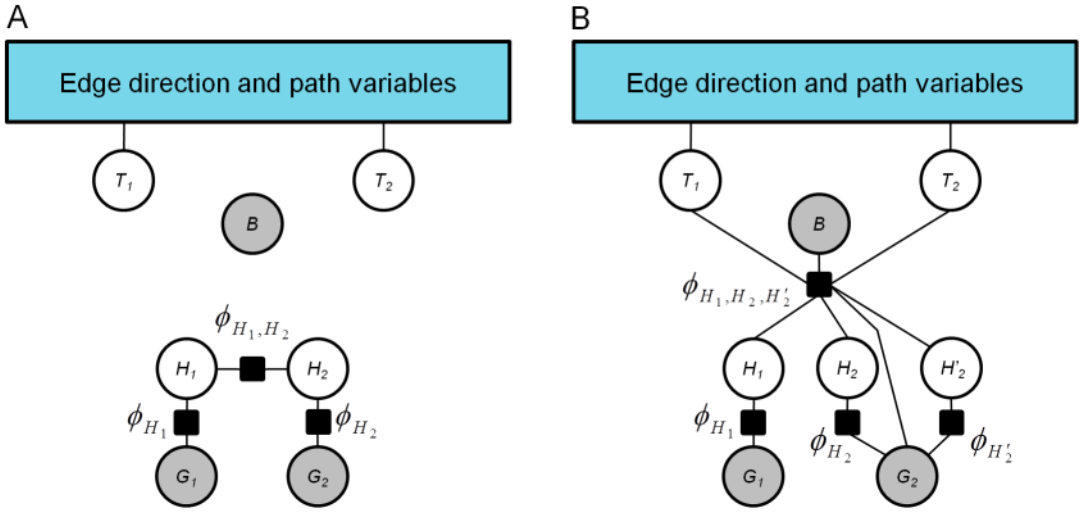

Figure 5.7: The remainder of the unified graphical model. The D and P variables connected to the T variables are still present but are not shown in order to emphasize the IOHMM representation. Shaded nodes are observed data; namely, the gene expression (G_1 and G_2) and protein-DNA binding interactions (B). A) The case corresponding to an IOHMM transition where there is no split node. B) At an IOHMM split node, there are two possible next states at time point 2, H_2 and H'_2.

for some time point, $\phi_{H_t,H_{t+1}} = 1$. Figure 5.7 reflects the fact that this factor's value is independent of the TF activity, protein-DNA binding, and gene expression.

The final factor corresponds to the emission probability from the IOHMM, which is modeled as a Gaussian distribution

$$\phi_{H_t}(g) = N(\mu_{H_t}, \sigma_{H_t})$$

Learning the Gaussian distribution parameters μ_{H_t} and σ_{H_t} is described in the next section.

The likelihood function follows from [50] except it also includes the factors ϕ_P and ϕ_T that were introduced to model the signaling paths.

$$L(D, P, T, \theta \mid B, G) \propto$$
$$\prod_{p \in P} \phi_P(d \in O(p), p) \prod_{t \in T} \phi_T(p \in E(t), t) \sum_{g \in G} \left(\prod_{H_t} \phi_{H_t,H_{t+1},H'_{t+1}}(g, B, T) \phi_{H_t}(g) \right)$$

where θ is the set of variables from the factors derived from the IOHMM (the ψ, μ_{H_t} and σ_{H_t} variables). The normalization constant (aka partition function) [106] that is required to form a probability distribution is omitted.

5.3.2 Inference

Parameter estimation and inference are substantially more challenging in the unified graphical model than they are in related work that analyzes either the network [217] or the expression dynamics [50] but not both. PNM [217] uses belief propagation to infer edge directions along source-target paths. However, as we describe in [65], belief propagation in this model converges too slowly even when applied to yeast datasets. Furthermore, if using an expectation-maximization strategy for parameter estimation, the inference procedure will be run many times and must be simple and/or very efficient. DREM [50] leverages the highly efficient learning and inference algorithms available for HMMs, whose independence assumptions are violated in the unified model due to the influence of the T variables. However, we exploit the fact that if we are given the T variables, the standard HMM algorithms are applicable for the IOHMM-derived parameters and factors.

Therefore, we adopt a sampling-based approach for inference [106]. Because the unified model is a generative model, we can calculate the likelihood of the binding and gene expression data given a configuration of the variables using the likelihood function above. Furthermore, by sampling a set of edge directions it is possible to generate the locally optimal states of the other variables as follows.

Given a randomly sampled assignment to all D variables, there is a deterministic assignment of the P variables that maximizes ϕ_P. Namely, if for some p all $d_i \in O(p)$ are set to $\widehat{d_i}$ or $O(p) = \emptyset$, then $p = 1$. Otherwise $p = 0$. Given these assignments to P variables, an assignment to all T follows similarly from ϕ_T. The assignment of P determines the row of Table 5.12 to use, and there is a unique assignment to t that causes ϕ_T to be 0.9 once the row is fixed.

Once all T have been assigned, structure search, parameter estimation, and inference over the remaining variables proceeds as in DREM [50]. The fixed T behave like SDREM's TF activity priors and are incorporated into the logistic regression but do not affect any of the learning or inference algorithms. Specifically, the Baum-Welch algorithm [17, 47] is used for parameter estimation. Like DREM, during the logistic regression training we place an L_1 penalty on the logistic regression weight vector ψ and weight the training instances (genes). A gene's weight is determined by the probability of transitioning to each child state (e.g. H_2 and H_2' in Figure 5.7) from the current state given the current

model parameters. Using the above procedure thus provides a manner for calculating the likelihood of a configuration given an assignment to the D variables.

The question that remains is how to search for the optimal configuration of D. The search space includes $2^{|D|}$ possible assignments making it impractical to randomly sample D until a suitable configuration is found. Instead, we randomly sample a limited number of starting points and perform a greedy search from each. At each step in the greedy search, we calculate for each $d \in D$ the new likelihood obtained after flipping the assignment to d (maintaining all other edge directions). We flip the d that yields the maximum increase in likelihood and repeat, stopping when no d can improve the likelihood. Note that changing a single d is unlikely to affect the assignment to the T variables, which will only change when the target connectivity crosses the expectation thresholds. By caching the values of the θ parameters and likelihood function terms from the IOHMM-derived portion of the factor graph for each configuration of T that we observe in the search, we can attain substantial computational savings. In practice, the change in likelihood when a d is flipped will primarily come from the updated values of ϕ_P, which can be calculated very quickly and are sensitive to a single edge flip.

5.3.3 Relation to SDREM and PNM

Conceptually, SDREM and the unified model target the same problem and share the same motivations. Beyond the obvious distinction that lead to the development of the unified model — that SDREM cannot simultaneously account for network connectivity and gene expression changes when searching for active TFs — there are several other notable differences. SDREM relies extensively on randomization of the protein-DNA binding interactions and interaction network target nodes to determine whether various scores are significant, which is not required in the unified model. Furthermore, the unified model directly optimizes the likelihood of the model and observed expression data. The objective function for SDREM's network orientation is biologically motivated, but only indirectly ties into the gene expression changes. SDREM does allow a more diverse set of TF activities as the unified model is constrained to three discrete values (this constraint could be relaxed but would reduce the benefits of caching during the greedy search). However, the unified model explicitly controls for high-degree and highly-connected TFs when assessing TF activity based on the signaling network. In addition, SDREM sums path weights for the network orientation objective function, whereas the likelihood function in the unified model multiplies the normalized path weights (the output of ϕ_P).

In practice, the scalability of the two approaches may prove to be a substantial difference. The network orientation phase of SDREM has a reasonable runtime, especially when

portions have been parallelized as described in Section 5.1.3, and the IOHMM phase is run a fixed number of times. For example, in the yeast analysis in Chapter 4, the IOHMM was run 110 times using the real TF-binding interactions and ten randomized sets of interactions at each of the ten SDREM iterations. In the unified graphical model, the greedy search over D is roughly as complex as SDREM's network orientation (assuming for the moment that the cached likelihood of the IOHMM portion can be used), and the SDREM approximation in which only the top m paths are enumerated so as to reduce the runtime (Section 5.1.4) is applicable as well. The critical difference is that the total number of times the IOHMM inference algorithm is run is instance-dependent for the unified model. Using the caching strategy described above, in the worst case this inference will be performed $3^{|T|}$ times (once for every combination of TF activities), which is essentially intractable. In the best case, any t whose final value is 0.1 or 0.9 at the local optimum will be changed from 0.5 to that value at some point in the search and then maintain that value for the rest of the search. In this case, the IOHMM inference will only be performed at most $|T|$ times, which is roughly as often as in SDREM given that current TF-gene binding interaction datasets typically cover a few hundred TFs. The caveat is that the greedy search described above is unlikely to achieve this best case performance, suggesting that more advanced search strategies may be required (see Section 6.2.5 for one possibility).

The unified graphical model search and inference algorithms can be partially parallelized to help achieve scalability. The greedy searches from the different starting points can be run in parallel trivially. For instances where $|D|$ is large the calculation of the change in likelihood for each d that is flipped can be parallelized as well. However, any such flips that change a TF activity variable and require running IOHMM inference will form a bottleneck. The best d to flip cannot be determined until all of the IOHMM runs terminate at a given greedy step. To avoid this bottleneck, the next search step could proceed by assuming a particular outcome from the IOHMM inference and backtracking if the assumption proves to be wrong (like speculative execution in modern pipelined computer processors). Some of the IOHMM inference could potentially be parallelized as well, although doing so is nontrivial. Even with such parallelization, the number of times the IOHMM inference is run may make the unified model orders of magnitude slower than SDREM. Substantially reducing the unified model's runtime may require approximations in the greedy search that only rerun IOHMM inference when multiple TF activity values have changed, thereby reducing how often the IOHMM inference is run overall.

The PNM factor graph [217] is similar to the network portion of the unified model here. Both seek to infer source-target pathways through a network of physical interactions and use variables for edge direction and path activity. PNM also includes edge sign (activation or repression) variables , which could be incorporated in the unified model.

However edge sign variables are less appropriate in this setting where the targets are TFs as opposed to differentially expressed genes affected by a knockout, which have a clear sign. Whereas PNM requires edge presence variables to reflect uncertainty in the interaction data, the unified model eliminates the need for these variables by precomputing path weights and incorporating them into ϕ_P. In addition, PNM infers whether a source-target pair is explained or not, but does not infer the target activity level like the unified model. The greatest difference between the two is in the treatment of parallel paths. PNM does encourage parallel paths in that it contains potential functions whose values increase as additional paths are used to explain the source-target pair. However, the emphasis is on finding at least one explanatory path, whereas the unified model explicitly rewards active TFs that are connected to the sources with many satisfied paths.

Chapter 6

Conclusions and future work

Across all organisms the ability to thrive and adapt to dynamic, sometimes hostile, environmental conditions is a fundamental necessity, yet the manner in which signaling and transcriptional regulatory networks mediate the response to environmental stresses is quite complex and far from being completely understood. Even in the most comprehensively studied signaling pathways in model organisms, canonical pathway representations are incomplete and inherently static. Mapping these pathways is challenging at both the biological and computational levels due to pathway and regulator redundancy. In this thesis we have developed computational approaches to elucidate detailed models of stress response. Our data-integrative analysis has provided insights into the use of knockout data in such modeling as well as specific yeast and human stress responses.

6.1 Conclusions

In Chapter 2 we addressed the puzzling disagreement between genome-wide TF knockouts and protein-DNA binding datasets in yeast, and better understanding this apparent discrepancy motivated our subsequent computational strategies. By incorporating sequence-level and PPI features for the TFs, we were able to predict which TFs had a putative redundant partner that could compensate for their loss in a single deletion strain. The effects of such transcriptional backup mechanisms were confirmed when we observed that genes bound by TFs that are mostly likely to have a redundant partner are not affected at all when that TF is deleted. Complementary network analysis reaffirmed that PPI networks can explain indirect effects (i.e. why genes that are not directly bound by a TF are affected by its knockout). These biological principles helped shape SDREM, which leverages phys-

ical interaction networks to link source proteins to downstream TFs and does not rely on knockouts to define the sources due to the confounding effects of redundancy. More generally, our contributions to understanding fault tolerance in transcriptional regulatory networks has potential relevance to constructing robust computational systems [144].

The PPI network orientation algorithms presented in Chapter 3 excel not only at reconstructing signaling pathways when the all endpoints are known, but also at directing biological networks for independent algorithms that require or benefit from such directionality. Specifically, our oriented networks have been used to predict missing edges in signaling pathways (Section 3.5) and analyze the topological redundancy of yeast interaction networks [1]. The Maximum Edge Orientation problem itself is formulated based on the biological principles studied in Chapter 2 (redundancy and explanatory paths in interaction networks) and others observed in signaling databases (prevalence of high-confidence interactions and short paths). Surprisingly, we showed that a simple algorithm composed of random restarts and local search outperforms more theoretically complex approaches derived from Boolean satisfiability.

Chapter 4 builds on this work to address the more common scenario in which the targets (TFs) that drive the stress response are unknown. We developed a widely-applicable algorithm, SDREM, that requires a minimal amount of data that is specific to the stress response, namely a partial set of source proteins and time series gene expression data. Integrating this condition-specific data with condition-independent interactome data that is prevalent for many species enables SDREM to infer end-to-end models that specify which TFs control the response, how they are influenced by the upstream sources, and when they exert regulatory influence on their target genes. When applying SDREM to very well-studied yeast MAPK pathways it not only recovered the core transcriptional unit but also extended the pathways with new predictions, some of which we validated experimentally.

Similarly, Chapter 5 demonstrates SDREM's relevance to the clinical setting through our study of respiratory virus infection in humans with particular emphasis on H1N1 and H5N1 influenza. Scaling to human datasets required further algorithmic improvements and benefited from the integration of additional data sources, RNAi screens and PTMs. Using our influenza models, we developed techniques for prioritizing the proteins that we predict to participate in the immune response. These algorithms can be used to guide experimental follow up, making it easier to identify genes that directly impact a disease-relevant phenotype (in this case viral load post-infection) when silenced with RNAi. Furthermore, these same ideas enable the prediction of condition-relevant genetic interactions, which cannot be identified via exhaustive experimental screening due to sheer number of human gene pairs. Finally, we explore theoretical improvements to the SDREM framework by formulating it as a single, unified probabilistic graphical model.

Throughout the thesis, we aspired to ensure that our computational work had direct biological relevance. This was accomplished in part with extensive literature searches to place our predictions into biological context (Sections 3.3.2, 3.5.2, 3.5.3, and 4.3.5) and also with targeted experimental validation. Despite our finding that redundancy can make gene knockouts ill-suited for defining the sources in a network analysis problem, we leverage the valuable causal information provided by gene deletion to verify many of our predictions (Sections 2.3.2, 3.5.3, and 4.3.3). Backup mechanisms make it difficult to interpret the absence of knockout effects, complicating the modeling of expression data in knockout strains. However, in the network setting the effects that are observed can be clearly attributed to protein that was removed from the network, a useful strategy for validating the predicted order of proteins along a signaling pathway. When investigating the SDREM osmotic stress models, the knockouts were complemented with microscopy and FACS analysis to determine whether the predicted active TFs exhibit differential nuclear localization or protein expression.

Data integration is a major theme in our work, and combining different types of data can "elevate" (infer additional properties of) more readily-available data. This idea is rooted in SDREM's predecessor [50] in which dynamic gene expression was used to annotate static protein-DNA binding interactions with the time at which the binding took place or was active. We carry this principle further in our network analysis. SDREM's integration of condition-specific and condition-independent datasets allows it to annotate the general PPI network with condition-specific properties. Specifically, a subset of the undirected interactions are identified as being responsible for the stress response of interest, and the direction in which these edges are used in the pathways can be determined. As we discuss below, there are opportunities to leverage additional types of data, such as genetic and functional associations, that will further improve SDREM's stress response modeling.

6.2 Future work

6.2.1 Applications to new stress responses

Now that SDREM has been improved to scale to larger (e.g. mammalian) datasets, one next step is to apply it broadly to different stress responses to increase our understanding of the biological mechanisms and principles at work. SDREM is a robust, accurate stress response modeling algorithm, but is nevertheless limited by the availability, coverage, and quality of the biological data it requires. In our experience, acquiring the PPI networks

and condition-specific time series gene expression data will not be the bottleneck in analyzing further responses and species. BioGRID, just one of many PPI databases, contains several thousand physical interactions for each of seven species (as of version 3.1.86) and is steadily growing [186]. The Gene Expression Omnibus [13] has exhibited exponential growth in the number of time series expression datasets that have been deposited [66]. Rather, the challenge will be the protein-DNA interactions and condition-specific source nodes (e.g. host-pathogen interactions). Fortunately, recent trends suggest that these data types too will become richer and more prevalent in the near future.

The ENCODE and modENCODE projects [61, 168, 195, 196] are rapidly generating experimental data detailing TF binding (among many other things) in human, *Caenorhabditis elegans*, and *Drosophila melanogaster*. These efforts will ultimately provide the ideal resource for SDREM. However, current coverage of TF binding is far from complete (e.g. less than 10% of human TFs have been profiled so far). In the meantime, computational advances can provide TF-gene interactions across a larger number of TFs. State of the art algorithms [27, 40, 49, 161, 212] combine TF binding motifs with epigenetic data and other other genome-wide features (such as sequence conservation or proximity to a transcription start site) to predict protein-DNA interactions for hundreds of TFs. Furthermore, methods that leverage DNase I sensitivity [27, 40, 161] can make cell type- or condition-specific binding predictions. These are especially relevant to SDREM, as we expect SDREM will be able to more accurately identify the TFs active in a stress response when their binding interactions coincide with the condition in which the gene expression data was collected. Protein binding microarray data [11, 170] is another TF-DNA binding resource that SDREM can utilize until experimental TF binding datasets are available on a larger scale. Drawing upon one or more of these approaches will enable SDREM to be applied to a wide range of species, even those in which TF binding is poorly characterized by ChIP-chip and ChIP-Seq.

Detailed mappings of host-pathogen interactions have garnered the interest of the biological community in recent years resulting in an increasing number of genome-wide host-pathogen PPI and RNAi screens for a multitude of pathogens. In addition to host-pathogen PPI databases [33, 46, 56, 148], which contain many literature-curated interactions for viruses and other pathogens, large-scale host-pathogen PPI experiments have been performed for HIV [93], hepatitis C [41], dengue [102], Epstein-Barr virus [29], herpesviruses [201], human T-cell leukemia virus [182], and even non-human hosts such as *Arabidopsis thaliana* [142]. Although SDREM can be run using these pathogen interaction partners alone, we showed in Section 5.2.2 that including RNAi screen hits improves our models of the immune response despite the incredibly low overlap among these screens. Fortunately, genome-wide RNAi screens for host factors related to pathogen infection have

also been conducted for HIV [28, 108, 220, 229], hepatitis C [127, 192], West Nile virus [111], and *Mycobacterium tuberculosis* [116]. Targeted screens are even more abundant. Our analysis of four respiratory viruses revealed common pathways and host proteins involved in the immune response and provided a blueprint for future comparative analysis. Three of the viruses we studied, the influenza A strains, were very closely related, and it will be quite interesting to see what conclusions can be drawn from the analysis of more distant viruses or even different types of pathogens (e.g. bacteria) or hosts.

6.2.2 Integrated model of multiple conditions

Instead of building separate SDREM models for related stress responses and comparing the resulting networks (as we did for osmotic stress in Section 4.3.1 and viral response in Section 5.2.3), SDREM could be extended to jointly model similar gene expression datasets that share a common signaling network structure. Each of the transcriptional response patterns will provide a different view of the active TFs, which are presumed to be similar across the individual conditions in this setting. Therefore, combining this information should allow more accurate reconstruction of the signaling pathways. We discuss this extension of the context of the viral responses studied for illustrative purposes. When calculating vertex weights as described in Section 5.1.1, the RNAi screen hits from all of the viruses would be used instead of deriving separate node weights for each condition. Similarly, source proteins that interact with proteins from multiple viruses could be given higher vertex weights such that paths starting from them are regarded with higher confidence. This would change the semantics of the network and the TF activity priors derived from it, which would no longer be specific to one individual response.

The expression profiles in the different conditions would still be modeled individually as these are likely to be dissimilar even for highly related condition due to the effects of the experimental methodology. Therefore, it would be necessary to merge the TF activity scores from each IOHMM into a single target weight for the subsequent network orientation. In general, TF activity scores are not directly comparable across different models because their scale can depend on the number of genes in the model (see the equations in Section 4.2.2), which suggests we should not take the maximum or average TF activity score across all conditions. To merge the TF activity scores, we could instead create a combined activity score background distribution. This distribution would be formed by randomizing the TF-gene binding data, building IOHMMs for each expression dataset using the same randomized interactions, and then storing the product of the activity scores across all conditions for each TF. The combined activity score (product of the individual activity scores) would likewise be computed using the real binding data and compared to

the combined background distribution to generate target weights (as SDREM does currently).

There are several advantages to jointly modeling the signaling network of related stress responses. It is difficult to build accurate models for conditions where limited data is available. For example, SARS lacks extensive host-virus PPI studies and genome-wide RNAi screens, and the data used in our analysis comes from small scale experiments reported in the literature. Leveraging the abundant influenza source proteins and RNAi screens would allow SDREM to reconstruct more trustworthy (but less SARS-specific) signaling pathways that still take advantage of the SARS time series gene expression data available. Even for responses where coverage is not a problem, we can place greater confidence in pathways that involve source proteins and screen hits that are implicated in multiple responses as opposed to a single condition. Jointly modeling the signaling network also allows SDREM to integrate multiple expression datasets from the same condition (e.g. the yeast osmotic stress response), which should help reduce the effects of noise in the expression data in terms of the predicted active TFs. This approach also provides a more direct way to identify proteins that are common to multiple responses instead of examining the intersection of multiple SDREM models as we did previously.

On the other hand, creating a single signaling network for conditions that are not sufficiently similar would lead to inferred pathways that are inconsistent with the actual biological mechanisms. Because SDREM is most useful when applied to responses that are not already well-understood, it is difficult to know in advance how similar the involved pathways are. Another potential disadvantage is that conditions with lower coverage in the source nodes and screen hits could be dominated by those with higher coverage such that the network is not representative of all conditions. This could also occur if differences in the expression profiles caused one condition to have many more TFs with high activity scores than the others such that the targets for the network orientation were primarily from a single condition. The combined background distribution is expected to control for but not entirely eliminate this effect.

6.2.3 Leveraging additional types of data

RNAi screens are well-suited for the study of pathogen infection and the vertex weights that we derived from the H1N1 screen hits improved SDREM's predictive capability. One future direction will be to incorporate other types of data that can be used to generate node weights for stress responses where RNAi screens are either unavailable or inappropriate due to the lack of a clearly defined phenotype. Many disease phenotypes have been linked to genetic variations (e.g. single nucleotide polymorphisms) through genome-wide asso-

ciation and other types of studies. Databases such as the Online Mendelian Inheritance in Man [4] catalog these relationships, which could be used to create gene priors when using SDREM to study any of the thousands of diseases with known associations. Phosphoproteomic data [71], which can quantify differential phosphorylation in response to a stimulus, could also provide a very direct measure of which proteins are most likely to be involved in a stress response pathway. This type of data has already been successfully used to define the source nodes in a different signaling network inference algorithm [89] and is especially useful because it provides quantitative information that could be transformed into continuous-valued node weights.

We have already shown how SDREM models can be used to predict genetic interactions, but exploring the converse question of how to integrate genetic interactions into SDREM's model building is another interesting potential extension. Genetic interaction data alone is informative enough to infer pathways and the order of the pathway members [15, 159]. However, the edges learned by such algorithms may not reflect direct physical interactions because genetic interactions themselves represent functional, oftentimes indirect, relationships [15]. Integrating genetic interaction and PPI data revealed that $\sim 40\%$ of genetic interactions could be "explained" as being between or within physical interaction pathways using predefined pathways (densely connected groups of proteins) [101].

Rather than rely on a fixed set of pathways or infer paths from genetic interactions alone, SDREM could incorporate genetic interaction datasets to influence the network orientation. The genetic interactions could help constrain the set of biologically reasonable orientations. For instance, assume we observe a positive genetic interaction between A and B, which suggests that they act within the same linear pathway. The network orientation therefore should be consistent with this interaction and prefer the orientations $A \rightarrow C \rightarrow B$ or $A \leftarrow C \leftarrow B$ over the orientations $A \leftarrow C \rightarrow B$ or $A \rightarrow C \leftarrow B$. Similarly, negative interactions would reward the orientation algorithm for placing the interacting genes in parallel pathways instead of a linear chain. The challenge would come in integrating individual, sometimes conflicting, constraints imposed by the genetic interactions with each other and the original objective of optimally connecting the signaling network sources and targets. However, the major advantage of exploiting genetic interactions in this manner is that the inferred pathways would have physical interpretations. A caveat is that such an extension would only be advantageous in organisms where large-scale genetic interactions have been performed (e.g. yeast). As discussed in Section 5.2.5, the number of possible gene pairs in higher-order organisms like humans make it unlikely that global genetic interaction screens will be feasible (given current experimental technology), but even small sets of targeted, condition-relevant interactions would be useful.

6.2.4 Feedback loops

Feedback loops are an important component of biological networks that have been shown to buffer noise [82] and affect robustness to perturbations [118]. However, SDREM currently considers only simple (cycle-free) paths when searching for optimal source-target connections. In practice SDREM models can contain cycles if two independent source-target paths happen to contain the same two vertices and disagree about which of the two vertices is upstream of the other (e.g. paths $A \rightarrow B \rightarrow C \rightarrow D$ and $E \rightarrow C \rightarrow F \rightarrow B \rightarrow G$). Such cycles do not carry any special meaning and do not model feedback in the signaling or regulatory network. The signaling network in SDREM represents directed connections along which messages can reach the TFs from the sources, but does not explain how or when the signaling proteins are (de)activated. Because the timing of the signal transduction is not modeled, SDREM implicitly assumes that the pathways operate instantaneously allowing the targets to be activated at any time after the stimulus is detected. In this context, there is no benefit of including feedback loops in the source-target paths. Feedback loops would not provide new connections among the sources and targets and would not affect the activity levels or timing of the internal nodes since these aspects of the pathways are not modeled. It is difficult to assess the degree to which omitting feedback impacts SDREM because the KEGG and *Science Signaling* Database of Cell Signaling representations of the HOG pathway (the well-studied pathway we used to quantify SDREM's accuracy) are also cycle-free. Nevertheless, modeling feedback has the potential to generate more realistic pathways.

To extend SDREM to account for feedback we must transition from a high-level view of the network that depicts pathway members and interactions but not the precise control mechanisms to a more detailed model. Feedback is an inherently dynamic process [82, 118, 164] so the first step will be to model the timing of the signaling pathways and the activity of the signaling proteins. The time series gene expression data and inferred TF activity dynamics are insufficient to determine the temporal behavior of the proteins upstream of the TFs because protein activity does not necessarily correspond to differential gene expression, signaling events can take place much more quickly than transcriptional regulation, TFs are oftentimes active at multiple time points, and many paths lead to the same active TFs. Phosphoproteomic analysis, which was discussed above as an alternative data type for generating node weights, can also be used to track the differential changes in phosphorylation over time in response to a stimulus [153]. Such data has been used to validate inferred edge directions in a study of EGF/ERK signaling [203] but was not incorporated during model construction. In SDREM, the times at which proteins are (de)phosphorylated could be used to determine the dynamics of some of the internal nodes and constrain the network orientation. In addition, we would no longer use the maximum

TF activity over all time points to determine the target weight in the signaling network, but rather track the predicted activity changes over time with the goal of explaining them via the network dynamics.

Because positive and negative feedback loops have different biological functions [82], we could also extend SDREM to model edge signs (activating or inhibiting) like PNM [217]. The network orientation phase would then search for the edge directions, edge signs, and dynamic protein activities that are most consistent with the phosphoproteomic data, sources, and targets. Paths would only be satisfied if all edges were oriented toward the target and if the edge signs and protein activities agree with the phosphoproteomic data. This is clearly a much more difficult optimization problem and greedy search may no longer be an appropriate strategy.

6.2.5 Theoretical and algorithmic improvements

Further future work concerns the theoretical aspects of SDREM and the unified graphical model. Our theoretical analysis of Maximum Edge Orientation provided approximation algorithms and an inapproximability bound, but the gap between the two is large. This suggests that there are superior approximation algorithms for MEO, which could potentially also improve SDREM's accuracy in practice. It is also feasible that better approximation algorithms would not outperform the simple random orientation with local search in a biological evaluation based on our MAX-CSP results. In addition, we have shown that the Shortcuts and Shortcuts-X problems are NP-hard, but there is much progress to be made on their approximability.

Lastly, the sampling-based algorithm for the graphical model presented in Section 5.3.2 is advantageous because it fixes the TF activity variables, allowing the efficient HMM parameter estimation and inference algorithms to be applied, but the complexity of the graphical model warrants additional exploration of inference strategies. The current inference algorithm does not directly use the gene expression data to change the values of the TF activity variables. The expression data only indirectly influences these variables in that states where the TF activity better agrees with the gene expression will have higher likelihood in the search over edge direction configurations. There is likely room for improvement in the current sampling-based inference. Rather than randomly sample new starting points from a uniform distribution after the greedy search achieves a local optimum, we could adopt a weighted sampling strategy that attempts to sample reasonable edge directions. Instead of uniformly weighting the probability of sampling each edge direction, we could derive a weight for each edge based on the pathways that use that edge and the targets they connect. The activity scores that SDREM generates (not to be confused with the TF activity vari-

ables in the graphical model) could also be computed in the IOHMM-derived portion of the graphical model in a post-processing step. These scores tell which TFs are most active in the stress response based on the gene expression data. We can sum over all paths that wish to use an edge in one direction, assigning each path the value of the activity score of the TF at its end. This sum would be compared to the sum over all paths that use the edge in the opposing direction and when normalized would provide a non-uniform distribution for the edge direction sampling that encourages paths that lead to highly active TFs to be satisfied.

Bibliography

[1] R. Albert, B. DasGupta, R. Hegde, G. S. Sivanathan, A. Gitter, G. Gürsoy, P. Paul, and E. Sontag. Computationally efficient measure of topological redundancy of biological and social networks. *Physical Review E*, 84(3): 036117, 2011. 1, 6.1

[2] R. Albert and H. G. Othmer. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. *Journal of Theoretical Biology*, 223(1): 1–18, 2003. 4.1

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17): 3389–3402, 1997. 2.3.1

[4] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh. McKusick's online mendelian inheritance in man (OMIM). *Nucleic Acids Research*, 37(suppl 1): D793–D796, 2009. 6.2.3

[5] A. Amon, M. Tyers, B. Futcher, and K. Nasmyth. Mechanisms that help the yeast cell cycle clock tick: G2 cyclins transcriptionally activate G2 cyclins and repress G1 cyclins. *Cell*, 74(6): 993–1007, 1993. 3.3.2

[6] B. Anchang, M. J. Sadeh, J. Jacob, A. Tresch, M. O. Vlad, P. J. Oefner, and R. Spang. Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. *Proceedings of the National Academy of Sciences*, 106(16): 6447–6452, 2009. 4.1

[7] V. Archambault, E. J. Chang, B. J. Drapkin, F. R. Cross, B. T. Chait, and M. P. Rout. Targeted proteomic study of the cyclin-Cdk module. *Molecular Cell*, 14(6): 699–711, 2004. 3.3.2

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1): 25–29, 2000. 4.3.6, 5.2.2

[9] S. M. Assmann and R. Albert. Discrete dynamic modeling with asynchronous update, or how to model complex systems in the absence of quantitative information. In D. A. Belostotsky and J. M. Walker (eds.), *Plant Systems Biology*, vol. 553 of *Methods in Molecular Biology*, 207–225. Humana Press, 2009. 4.1

[10] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1): 78–85, 2004. 3.5.2

[11] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935): 1720–1723, 2009. 6.2.1

[12] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J. Franois, and R. Zecchina. Finding undetected protein associations in cell signaling by belief propagation. *Proceedings of the National Academy of Sciences*, 108(2): 882 –887, 2011. 4.1

[13] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, *et al.* NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Research*, 39(suppl 1): D1005–D1010, 2011. 6.2.1

[14] N. J. Barrows, C. L. Sommer, M. A. Garcia-Blanco, and J. L. Pearson. Factors affecting reproducibility between genome-scale siRNA-based screens. *Journal of Biomolecular Screening*, 15(7): 735 –747, 2010. 5.1.1

[15] A. Battle, M. C. Jonikas, P. Walter, J. S. Weissman, and D. Koller. Automated identification of pathways from quantitative genetic interaction data. *Molecular Systems Biology*, 6(379), 2010. 6.2.3

[16] F. Bauer, M. Urdaci, M. Aigle, and M. Crouzet. Alteration of a yeast SH3 protein leads to conditional viability with defects in cytoskeletal and budding patterns. *Molecular and Cellular Biology*, 13(8): 5070–5084, 1993. 4.3.5

[17] L. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3: 1–8, 1972. 5.3.2

[18] G. Bebek and J. Yang. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics*, 8: 335, 2007. 1, 3.1, 3.2.4, 3.3

[19] Y. Bengio and P. Frasconi. An input output HMM architecture. In *Advances in Neural Information Processing Systems*, vol. 7, 427–434, 1995. 4.2.1

[20] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289300, 1995. 5.2.2

[21] K. I. Berns, A. Casadevall, M. L. Cohen, S. A. Ehrlich, L. W. Enquist, J. P. Fitch, D. R. Franz, C. M. Fraser-Liggett, C. M. Grant, M. J. Imperiale, *et al.* Adaptations of avian flu virus are a cause for concern. *Science*, 335(6069): 660–661, 2012. 5.2

[22] D. Bertsimas, C. Teo, and R. Vohra. On dependent randomized rounding algorithms. *Operations Research Letters*, 24(3): 105–114, 1999. 3.2.3, 3.2.3

[23] P. L. Blaiseau, A. D. Isnard, Y. Surdin-Kerjan, and D. Thomas. Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Molecular and Cellular Biology*, 17(7): 3640–3648, 1997. 2.3.3

[24] M. Blondel, J. M. Galan, Y. Chi, C. Lafourcade, C. Longaretti, R. J. Deshaies, and M. Peter. Nuclear-specific degradation of Far1 is controlled by the localization of the F-box protein Cdc4. *The EMBO Journal*, 19(22): 6085–6097, 2000. 3.3.2

[25] E. Bortz, L. Westera, J. Maamary, J. Steel, R. A. Albrecht, B. Manicassamy, G. Chase, L. Martínez-Sobrido, M. Schwemmle, and A. García-Sastre. Host- and strain-specific regulation of influenza virus polymerase activity by interacting cellular proteins. *mBio*, 2(4): e00151–11, 2011. 5.1.1, 5.1, 5.2.2, 5.5, 5.2.4

[26] A. Bossi and B. Lehner. Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, 5(260), 2009. 3.5

[27] A. P. Boyle, L. Song, B. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford, and T. S. Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3): 456–464, 2011. 6.2.1

[28] A. L. Brass, D. M. Dykxhoorn, Y. Benita, N. Yan, A. Engelman, R. J. Xavier, J. Lieberman, and S. J. Elledge. Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, 319(5865): 921–926, 2008. 1, 5.1, 5.2.2, 6.2.1

[29] M. A. Calderwood, K. Venkatesan, L. Xing, M. R. Chase, A. Vazquez, A. M. Holthaus, A. E. Ewence, N. Li, T. Hirozane-Kishikawa, D. E. Hill, *et al.* Epstein-Barr virus and virus human protein interaction maps. *Proceedings of the National Academy of Sciences*, 104(18): 7606–7611, 2007. 6.2.1

[30] A. P. Capaldi, T. Kaplan, Y. Liu, N. Habib, A. Regev, N. Friedman, and E. K. O'Shea. Structure and function of a transcriptional network activated by the MAPK Hog1. *Nature Genetics*, 40(11): 1300–1306, 2008. 3.5.2, 3.3, 3.5.3, 4.3, 4.3.1, 4.3.4, 4.3.8

[31] T. Chan, J. Carvalho, L. Riles, and X. F. S. Zheng. A chemical genomics approach toward understanding the global functions of the target of rapamycin protein (TOR). *Proceedings of the National Academy of Sciences*, 97(24): 13227 –13232, 2000. 4.3.6

[32] M. Charikar, K. Makarychev, and Y. Makarychev. Near-optimal algorithms for maximum constraint satisfaction problems. *ACM Transactions on Algorithms*, 5(3): 1–14, 2009. 3.2.3

[33] A. Chatr-aryamontri, A. Ceol, D. Peluso, A. Nardozza, S. Panni, F. Sacco, M. Tinti, A. Smolyar, L. Castagnoli, M. Vidal, *et al.* VirusMINT: a viral protein interaction database. *Nucleic Acids Research*, 37(suppl 1): D669–673, 2009. 1, 4.2, 6.2.1

[34] M. Chaves, R. Albert, and E. D. Sontag. Robustness and fragility of boolean models for genetic regulatory networks. *Journal of Theoretical Biology*, 235(3): 431–449, 2005. 4.1

[35] J. Chen, S. Huang, and Z. Chen. Human cellular protein nucleoporin hNup98 interacts with influenza A virus NS2/nuclear export protein and overexpression of its GLFG repeat domain can inhibit virus propagation. *Journal of General Virology*, 91(10): 2474–2484, 2010. 5.5

[36] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191): 98–101, 2008. 3.5

[37] M. H. Cobb and E. J. Goldsmith. How MAP kinases are regulated. *The Journal of Biological Chemistry*, 270(25): 14843–14846, 1995. 1

[38] S. R. Collins, A. Roguev, and N. J. Krogan. Quantitative genetic interaction mapping using the E-MAP approach. *Methods in Enzymology*, 470: 205–231, 2010. 5.2.5

[39] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, *et al.* The genetic landscape of a cell. *Science*, 327(5964): 425–431, 2010. 5.2.5

[40] G. Cuellar-Partida, F. A. Buske, R. C. McLeay, T. Whitington, W. S. Noble, and T. L. Bailey. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, 28(1): 56–62, 2012. 6.2.1

[41] B. de Chassey, V. Navratil, L. Tafforeau, M. S. Hiet, A. Aublin-Gex, S. Agaugué, G. Meiffren, F. Pradezynski, B. F. Faria, T. Chantier, *et al.* Hepatitis C virus infection protein network. *Molecular Systems Biology*, 4(230), 2008. 5.2.1, 6.2.1

[42] E. de Nadal and F. Posas. Multilayered control of gene expression by stress-activated protein kinases. *The EMBO Journal*, 29(1): 4–13, 2010. 3.5.2, 3.5.2, 3.5.2, 4.3.1

[43] E. Demaine and M. Zadimoghaddam. Minimizing the diameter of a network using shortcut edges. In *Algorithm Theory - SWAT 2010*, vol. 6139 of *Lecture Notes in Computer Science*, 420–431. Springer Berlin / Heidelberg, 2010. 3.5

[44] Q. Diao and W. J. Van Der Linden. Automated test assembly using lp_Solve version 5.5 in R. *Applied Psychological Measurement*, 35(5): 398–409, 2011. 3.2.3

[45] L. Dirick and K. Nasmyth. Positive feedback in the activation of G1 cyclins in yeast. *Nature*, 351(6329): 754–757, 1991. 3.3.2

[46] T. Driscoll, M. D. Dyer, T. M. Murali, and B. W. Sobral. PIG–the pathogen interaction gateway. *Nucleic Acids Research*, 37(suppl 1): D647–D650, 2009. 1, 4.2, 6.2.1

[47] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998. 5.3.2

[48] J. Ernst, Q. K. Beg, K. A. Kay, G. Balázsi, Z. N. Oltvai, and Z. Bar-Joseph. A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. *PLoS Computational Biology*, 4(3): e1000044, 2008. 4.2.1

[49] J. Ernst, H. L. Plasterer, I. Simon, and Z. Bar-Joseph. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Research*, 20(4): 526 –536, 2010. 5.2.2, 6.2.1

[50] J. Ernst, O. Vainas, C. T. Harbison, I. Simon, and Z. Bar-Joseph. Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, 3(74), 2007. 1, 4.1, 4.2.1, 5.3.1, 5.3.2, 6.1

[51] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology*, 3(89), 2007. 1

[52] R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, *et al.* Pfam: clans, web tools and services. *Nucleic Acids Research*, 34(suppl 1): D247–251, 2006. 2.3.1

[53] J. L. Folch-Mallol, L. M. Martinez, S. J. Casas, R. Yang, C. Martinez-Anaya, L. Lopez, A. Hernandez, and J. Nieto-Sotelo. New roles for CDC25 in growth control, galactose regulation and cellular differentiation in saccharomyces cerevisiae. *Microbiology*, 150(9): 2865–2879, 2004. 3.3.2

[54] R. A. M. Fouchier, S. Herfst, and A. D. M. E. Osterhaus. Restricted data on influenza H5N1 virus transmission. *Science*, 335(6069): 662–663, 2012. 5.2

[55] M. Frieman, M. Heise, and R. Baric. SARS coronavirus and innate immunity. *Virus Research*, 133(1): 101–112, 2008. 5.2.3, 5.5

[56] W. Fu, B. E. Sanders-Beer, K. S. Katz, D. R. Maglott, K. D. Pruitt, and R. G. Ptak. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Research*, 37(suppl 1): D417–D422, 2009. 1, 4.2, 6.2.1

[57] A. Fujita, J. R. Sato, H. M. Garay-Malpartida, P. A. Morettin, M. C. Sogayar, and C. E. Ferreira. Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method. *Bioinformatics*, 23(13): 1623–1630, 2007. 4.1

[58] I. Gamzu, D. Segev, and R. Sharan. Improved orientations of physical networks. In V. Moulton and M. Singh (eds.), *Algorithms in Bioinformatics*, vol. 6293 of *Lecture Notes in Computer Science*, 215–225. Springer Berlin / Heidelberg, 2010. 3.1

[59] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11(12): 4241–4257, 2000. 1, 3.5.3, 4.3, 4.3.3

[60] A. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. Michon, C. Cruciat, *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868): 141–147, 2002. 3.3.2

[61] M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, 330(6012): 1775–1787, 2010. 6.2.1

[62] M. Geymonat, A. Spanos, G. P. Wells, S. J. Smerdon, and S. G. Sedgwick. Clb6/Cdc28 and Cdc14 regulate phosphorylation status and cellular localization of swi6. *Molecular and Cellular Biology*, 24(6): 2277–2285, 2004. 3.3.2

[63] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418(6896): 387–391, 2002. 1

[64] A. Gitter, M. Carmi, N. Barkai, and Z. Bar-Joseph. Linking the signaling cascades and dynamic regulatory networks controlling stress responses. 1, 4.3.2, 4.3.7

[65] A. Gitter, J. Klein-Seetharaman, A. Gupta, and Z. Bar-Joseph. Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Research*, 39(4): e22, 2011. 1, 1, 3.1, 3.2.4, 3.3, 3.3.1, 4.3.8, 5.3.2

[66] A. Gitter, Y. Lu, and Z. Bar-Joseph. Computational methods for analyzing dynamic regulatory networks. In I. Ladunga (ed.), *Computational Biology of Transcription Factor Binding*, vol. 674 of *Methods in Molecular Biology*, 419–441, 2010. 4.1, 6.2.1

[67] A. Gitter, Z. Siegfried, M. Klutstein, O. Fornes, B. Oliva, I. Simon, and Z. Bar-Joseph. Backup in gene regulatory networks explains differences between binding and knockout results. *Molecular Systems Biology*, 5(276), 2009. 1, 2.2, 2.3.2

[68] A. Goossens, T. E. Dever, A. Pascual-Ahuir, and R. Serrano. The protein kinase Gcn2p mediates sodium toxicity in yeast. *Journal of Biological Chemistry*, 276(33): 30753 –30760, 2001. 4.3.5

[69] W. Görner, E. Durchschlag, M. T. Martinez-Pastor, F. Estruch, G. Ammerer, B. Hamilton, H. Ruis, and C. Schüller. Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. *Genes & Development*, 12(4): 586–597, 1998. 3.5.3

[70] N. R. Gough. Science's signal transduction knowledge environment: the connections maps database. *Annals of the New York Academy of Sciences*, 971: 585–587, 2002. 1, 3.3, 3.5.2, 4.3, 4.3.1

[71] A. Gruhler, J. V. Olsen, S. Mohammed, P. Mortensen, N. J. Faergeman, M. Mann, and O. N. Jensen. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Molecular & Cellular Proteomics*, 4(3): 310–327, 2005. 6.2.3

[72] F. Gu, H. Hsu, P. Hsu, J. Wu, Y. Ma, J. Parvin, T. Huang, and V. Jin. Inference of hierarchical regulatory network of estrogen-dependent breast cancer through ChIP-based data. *BMC Systems Biology*, 4(1): 170, 2010. 4.2.1

[73] V. Guruswami, D. Lewin, M. Sudan, and L. Trevisan. A tight characterization of NP with 3 query PCPs. In *Proceedings of the 39th IEEE Annual Symposium on Foundations of Computer Science*, 8. IEEE Computer Society, 1998. 3.2.3

[74] E. Halperin and U. Zwick. Combinatorial approximation algorithms for the maximum directed cut problem. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, 1–7. Society for Industrial and Applied Mathematics, 2001. 3.2.2

[75] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. Tagne, D. B. Reynolds, J. Yoo, *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004): 99–104, 2004. 1, 2, 3.5.2, 3.2, 4.3.6

[76] G. T. Hart, A. K. Ramani, and E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11): 120, 2006. 3.5

[77] J. Hastad. Some optimal inapproximability results. *Journal of the ACM*, 48(4): 798–859, 2001. 3.2.2

[78] M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St. Onge, M. Tyers, D. Koller, *et al.* The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, 320(5874): 362–365, 2008. 1, 3.5.3, 4.3.1, 4.3.6

[79] S. Hohmann. Control of high osmolarity signalling in the yeast Saccharomyces cerevisiae. *FEBS Letters*, 583(24): 4025–4029, 2009. 3.5.2, 4.3.1, 4.3.2

[80] S. Hohmann, M. Krantz, and B. Nordlander. Yeast osmoregulation. *Methods in Enzymology*, 428: 29–45, 2007. 4.3.1

[81] P. C. Hollenhorst, G. Pietz, and C. A. Fox. Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation. *Genes & Development*, 15(18): 2445–2456, 2001. 2.2, 2.3.3

[82] G. Hornung and N. Barkai. Noise propagation and signaling sensitivity in biological networks: A role for positive feedback. *PLoS Computational Biology*, 4(1): e8, 2008. 6.2.4

[83] F. Hu, Y. Gan, and O. M. Aparicio. Identification of Clb2 residues required for Swe1 regulation of Clb2-Cdc28 in Saccharomyces cerevisiae. *Genetics*, 179(2): 863–874, 2008. 3.3.2

[84] Z. Hu, P. J. Killion, and V. R. Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nature Genetics*, 39(5): 683–687, 2007. 1, 2, 2.1

[85] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1): 1–13, 2009. 5.2.2

[86] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1): 44–57, 2009. 5.2.2

[87] H. Huang and J. S. Bader. Precision and recall estimates for two-hybrid screens. *Bioinformatics*, 25(3): 372–378, 2009. 3.5

[88] S. Huang, J. Chen, H. Wang, B. Sun, H. Wang, Z. Zhang, X. Zhang, and Z. Chen. Influenza A virus matrix protein 1 interacts with hTFIIIC102-s, a short isoform of the polypeptide 3 subunit of human general transcription factor IIIC. *Archives of Virology*, 154(7): 1101–1110, 2009. 5.5

[89] S. C. Huang and E. Fraenkel. Integration of proteomic, transcriptional, and interactome data reveals hidden signaling components. *Science Signaling*, 2(81): ra40, 2009. 4.1, 6.2.3

[90] Y. Huang, A. K. Zaas, A. Rao, N. Dobigeon, P. J. Woolf, T. Veldman, N. C. Oien, M. T. McClain, J. B. Varkey, B. Nicholson, *et al.* Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza A infection. *PLoS Genetics*, 7(8): e1002234, 2011. 5.5

[91] T. Ichinohe. Respective roles of TLR, RIG-I and NLRP3 in influenza virus infection and immunity: impact on vaccine design. *Expert Review of Vaccines*, 9(11): 1315–1324, 2010. 5.2.2, 5.5

[92] M. Imai, T. Watanabe, M. Hatta, S. C. Das, M. Ozawa, K. Shinya, G. Zhong, A. Hanson, H. Katsura, S. Watanabe, *et al.* Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature*, 2012. 5.2

[93] S. Jäger, P. Cimermancic, N. Gulbahce, J. R. Johnson, K. E. McGovern, S. C. Clarke, M. Shales, G. Mercenne, L. Pache, K. Li, *et al.* Global landscape of HIV-human protein complexes. *Nature*, 481(7381): 365–370, 2012. 6.2.1

[94] M. C. Jonikas, S. R. Collins, V. Denic, E. Oh, E. M. Quan, V. Schmid, J. Weibezahn, B. Schwappach, P. Walter, J. S. Weissman, *et al.* Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science*, 323(5922): 1693–1697, 2009. 5.2.5

[95] R. Kafri, A. Bar-Even, and Y. Pilpel. Transcription control reprogramming in genetic backup circuits. *Nature Genetics*, 37(3): 295–299, 2005. 2.1, 2.3.1, 2.3.3

[96] R. Kafri, O. Dahan, J. Levy, and Y. Pilpel. Preferential protection of protein interaction network hubs in yeast: Evolved functionality of genetic redundancy. *Proceedings of the National Academy of Sciences*, 105(4): 1243–1248, 2008. 2.1

[97] P. Kaiser, V. Moncollin, D. J. Clarke, M. H. Watson, B. L. Bertolaet, S. I. Reed, and E. Bailly. Cyclin-dependent kinase and Cks/Suc1 interact with the proteasome in yeast to control proteolysis of M-phase targets. *Genes & Development*, 13(9): 1190–1202, 1999. 3.3.2

[98] K. Kakiuchi, Y. Yamauchi, M. Taoka, M. Iwago, T. Fujita, T. Ito, S. Song, A. Sakai, T. Isobe, and T. Ichimura. Proteomic analysis of in vivo 14-3-3 interactions in the yeast Saccharomyces cerevisiae. *Biochemistry*, 46(26): 7781–7792, 2007. 3.5.2, 3.2

[99] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1): 27–30, 2000. 1, 3.3, 3.5.2, 3.5.2, 4.2, 4.3.1, 5.2.2

[100] A. Karlas, N. Machuy, Y. Shin, K. Pleissner, A. Artarini, D. Heuer, D. Becker, H. Khalil, L. A. Ogilvie, S. Hess, *et al.* Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature*, 463(7282): 818–822, 2010. 1, 5.1, 5.2.2

[101] R. Kelley and T. Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology*, 23(5): 561–566, 2005. 6.2.3

[102] S. Khadka, A. D. Vangeloff, C. Zhang, P. Siddavatam, N. S. Heaton, L. Wang, R. Sengupta, S. Sahasrabudhe, G. Randall, M. Gribskov, *et al.* A physical interaction network of dengue virus and human proteins. *Molecular & Cellular Proteomics*, 10(12), 2011. 6.2.1

[103] Y. Kim, S. Wuchty, and T. M. Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Computational Biology*, 7(3): e1001095, 2011. 4.3.8

[104] E. Klopf, L. Paskova, C. Solé, G. Mas, A. Petryshyn, F. Posas, U. Wintersberger, G. Ammerer, and C. Schüller. Cooperation between the INO80 complex and histone chaperones determines adaptation of stress gene transcription in the yeast Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 29(18): 4994–5007, 2009. 4.3.5

[105] R. Kohli, R. Krishnamurti, and P. Mirchandani. The minimum satisfiability problem. *SIAM Journal on Discrete Mathematics*, 7(2): 275–283, 1994. 3.2.3

[106] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive Computation and Machine Learning. The MIT Press, 2009. 5.3.1, 5.3.1, 5.3.2

[107] R. König, S. Stertz, Y. Zhou, A. Inoue, H. H. Hoffmann, S. Bhattacharyya, J. G. Alamares, D. M. Tscherne, M. B. Ortigoza, Y. Liang, *et al.* Human host factors required for influenza virus replication. *Nature*, 463(7282): 813–817, 2010. 1, 5.1, 5.2.2

[108] R. König, Y. Zhou, D. Elleder, T. L. Diamond, G. M. Bonamy, J. T. Irelan, C. Chiang, B. P. Tu, P. D. D. Jesus, C. E. Lilley, *et al.* Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, 135(1): 49–60, 2008. 6.2.1

[109] S. Koyama, K. J. Ishii, H. Kumar, T. Tanimoto, C. Coban, S. Uematsu, T. Kawai, and S. Akira. Differential role of TLR- and RLR-signaling in the immune responses to influenza A virus infection and vaccination. *The Journal of Immunology*, 179(7): 4711–4720, 2007. 5.2.2, 5.5

[110] M. Krantz, D. Ahmadpour, L. Ottosson, J. Warringer, C. Waltermann, B. Nordlander, E. Klipp, A. Blomberg, S. Hohmann, and H. Kitano. Robustness and fragility in the yeast high osmolarity glycerol (HOG) signal-transduction pathway. *Molecular Systems Biology*, 5(281), 2009. 4.3.1

[111] M. N. Krishnan, A. Ng, B. Sukumaran, F. D. Gilfoy, P. D. Uchil, H. Sultana, A. L. Brass, R. Adametz, M. Tsui, F. Qian, *et al.* RNA interference screen for human genes associated with West Nile virus infection. *Nature*, 455(7210): 242–245, 2008. 6.2.1

[112] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6): 957–968, 2005. 5.3.1

[113] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, *et al.* Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084): 637–643, 2006. 1, 3.5

[114] F. Kschischang, B. Frey, and H. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2): 498–519, 2001. 5.3.1

[115] O. Kuchaiev, M. Raajski, D. J. Higham, and N. Prulj. Geometric de-noising of protein-protein interaction networks. *PLoS Computational Biology*, 5(8): e1000454, 2009. 3.5

[116] D. Kumar, L. Nath, M. A. Kamal, A. Varshney, A. Jain, S. Singh, and K. V. Rao. Genome-wide analysis of the host intracellular network that regulates survival of Mycobacterium tuberculosis. *Cell*, 140(5): 731–743, 2010. 6.2.1

[117] K. Kuranda, V. Leberre, S. Sokol, G. Palamarczyk, and J. Francois. Investigating the caffeine effects in the yeast saccharomyces cerevisiae brings new insights into the connection between TOR, PKC and Ras/cAMP signalling pathways. *Molecular Microbiology*, 61(5): 1147–1166, 2006. 4.3.6

[118] Y. Kwon and K. Cho. Quantitative analysis of robustness and fragility in biological networks based on feedback dynamics. *Bioinformatics*, 24(7): 987 –994, 2008. 2.1, 6.2.4

[119] A. Lan, I. Y. Smoly, G. Rapaport, S. Lindquist, E. Fraenkel, and E. Yeger-Lotem. ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Research*, 39(suppl 2): W424–W429, 2011. 4.3.7

[120] C. J. Langmead and S. K. Jha. Symbolic approaches for finding control strategies in boolean networks. *Journal of Bioinformatics and Computational Biology*, 7(2): 323–338, 2009. 4.1

[121] M. Lavallée-Adam, B. Coulombe, and M. Blanchette. Detection of locally over-represented GO terms in protein-protein interaction networks. *Journal of Computational Biology*, 17(3): 443–457, 2010. 3.5.1

[122] J. Lee, K. Colwill, V. Aneliunas, C. Tennyson, L. Moore, Y. Ho, and B. Andrews. Interaction of yeast Rvs167 and Pho85 cyclin-dependent kinase complexes may link the cell cycle to the actin cytoskeleton. *Current Biology*, 8(24): 1310–1321, S1, 1998. 4.3.5

[123] J. H. Lee, S. Kim, P. N. Q. Pascua, M. Song, Y. H. Baek, X. Jin, J. Choi, C. Kim, H. Kim, and Y. K. Choi. Direct interaction of cellular hnRNP-F and NS1 of influenza A virus accelerates viral replication by modulation of viral transcriptional activity and host gene expression. *Virology*, 397(1): 89–99, 2010. 5.5

[124] M. Lewin, D. Livnat, and U. Zwick. Improved rounding techniques for the MAX 2-SAT and MAX DI-CUT problems. In *Integer Programming and Combinatorial Optimization*, vol. 2337 of *Lecture Notes in Computer Science*, 67–82, 2002. 3.2.3

[125] C. Li, A. Bankhead III, A. J. Eisfeld, Y. Hatta, S. Jeng, J. H. Chang, L. D. Aicher, S. Proll, A. L. Ellis, G. L. Law, *et al.* Host regulatory network response to infection with highly pathogenic H5N1 avian influenza virus. *Journal of Virology*, 85(21): 10955–10967, 2011. 5.5

[126] C. Li, S. McCormick, and D. Simchi-Levi. On the minimum-cardinality-bounded-diameter and the bounded-cardinality-minimum-diameter edge addition problems. *Operations Research Letters*, 11(5): 303–308, 1992. 3.5

[127] Q. Li, A. L. Brass, A. Ng, Z. Hu, R. J. Xavier, T. J. Liang, and S. J. Elledge. A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proceedings of the National Academy of Sciences*, 106(38): 16410–16415, 2009. 6.2.1

[128] L. Lin, H. Lee, W. Li, and B. Chen. A systematic approach to detecting transcription factors in response to environmental stresses. *BMC Bioinformatics*, 8(1): 473, 2007. 4.1

[129] C. X. Ling, J. Huang, and H. Zhang. AUC: a statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th international joint conference on artificial intelligence*, IJCAI'03, 519–524. Morgan Kaufmann Publishers Inc., 2003. 5.2.4

[130] D. Liu, X. Liu, J. Yan, W. Liu, and G. F. Gao. Interspecies transmission and host restriction of avian H5N1 influenza virus. *Science in China Series C: Life Sciences*, 52(5): 428–438, 2009. 5.5

[131] W. Liu, D. Li, J. Wang, H. Xie, Y. Zhu, and F. He. Proteome-wide prediction of signal flow direction in protein interaction networks based on interacting domains. *Molecular & Cellular Proteomics*, 8(9): 2063–2070, 2009. 3.1

[132] S. Lu, F. Zhang, J. Chen, and S. Sze. Finding pathway structures in protein interaction networks. *Algorithmica*, 48(4): 363–374, 2007. 3.1

[133] K. MacIsaac, T. Wang, D. B. Gordon, D. Gifford, G. D. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, 7(1): 113, 2006. 2.2, 3.5.3, 4.3, 4.3.6, 4.3.9

[134] C. I. Maeder, M. A. Hink, A. Kinkhabwala, R. Mayr, P. I. H. Bastiaens, and M. Knop. Spatial regulation of Fus3 MAP kinase activity through a reaction-diffusion mechanism in yeast pheromone signalling. *Nature Cell Biology*, 9(11): 1319–1326, 2007. 3.3.2

[135] M. Malcher, S. Schladebeck, and H. Mösch. The Yak1 protein kinase lies at the center of a regulatory cascade affecting adhesive growth and stress resistance in Saccharomyces cerevisiae. *Genetics*, 187(3): 717–730, 2011. 3.5.3

[136] F. Markowetz, D. Kostka, O. G. Troyanskaya, and R. Spang. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23(13): i305–i312, 2007. 4.1

[137] A. Medvedovsky, V. Bafna, U. Zwick, and R. Sharan. An algorithm for orienting graphs based on cause-effect pairs and its applications to orienting protein networks. In *Proceedings of the 8th international workshop on Algorithms in Bioinformatics*, 222–232, 2008. 3.1, 3.2.2, 3.3

[138] M. V. Metodiev, D. Matheos, M. D. Rose, and D. E. Stone. Regulation of MAPK function by direct interaction with the mating-specific Galpha in yeast. *Science*, 296(5572): 1483–1486, 2002. 3.3.2

[139] A. Meyerson and B. Tagiku. Minimizing average shortest path distances via shortcut edge addition. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, vol. 5687 of *Lecture Notes in Computer Science*, 272–285. Springer Berlin / Heidelberg, 2009. 3.5

[140] C. Miller, B. Schwalb, K. Maier, D. Schulz, S. Dümcke, B. Zacher, A. Mayer, J. Sydow, L. Marcinowski, L. Dölken, *et al.* Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular Systems Biology*, 7(458), 2011. 4.3.5

[141] G. R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, T. M. Raghavan, *et al.* Human protein reference database–2006 update. *Nucleic Acids Research*, 34(suppl 1): D411–414, 2006. 1, 5.1.2

[142] M. S. Mukhtar, A. Carvunis, M. Dreze, P. Epple, J. Steinbrenner, J. Moore, M. Tasan, M. Galli, T. Hao, M. T. Nishimura, *et al.* Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science*, 333(6042): 596–601, 2011. 5.2.1, 6.2.1

[143] D. Nam, S. Seo, and S. Kim. An efficient top-down search algorithm for learning boolean networks of gene expression. *Machine Learning*, 65(1): 229–245, 2006. 4.1

[144] S. Navlakha and Z. Bar-Joseph. Algorithms in nature: the convergence of systems biology and computational thinking. *Molecular Systems Biology*, 7(546), 2011. 6.1

[145] S. Navlakha, A. Gitter, and Z. Bar-Joseph. A network-based approach for predicting missing pathway interactions. 1, 3.5.1, 3.5.3

[146] S. Navlakha, M. C. Schatz, and C. Kingsford. Revealing biological modules via graph summarization. *Journal of Computational Biology*, 16(2): 253–264, 2009. 3.5, 3.5.2

[147] V. Navratil, B. de Chassey, C. R. Combe, and V. Lotteau. When the human viral infectome and diseasome networks collide: towards a systems biology platform for the aetiology of human diseases. *BMC Systems Biology*, 5(1): 13, 2011. 5.2.1

[148] V. Navratil, B. de Chassey, L. Meyniel, S. Delmotte, C. Gautier, P. André, V. Lotteau, and C. Rabourdin-Combe. VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Research*, 37(suppl 1): D661–D668, 2009. 1, 4.2, 5.2.2, 5.5, 6.2.1

[149] T. Nevitt, J. Pereira, D. Azevedo, P. Guerreiro, and C. Rodrigues-Pousada. Expression of YAP4 in saccharomyces cerevisiae under osmotic stress. *Biochemical Journal*, 379(Pt 2): 367–374, 2004. 4.3.5

[150] L. Ni, C. Bruce, C. Hart, J. Leigh-Bell, D. Gelperin, L. Umansky, M. B. Gerstein, and M. Snyder. Dynamic and complex transcription factor binding during an inducible response in yeast. *Genes & Development*, 23(11): 1351 –1363, 2009. 3.5.2

[151] D. Nishimura. BioCarta. *Biotech Software & Internet Report*, 2(3): 117–120, 2001. 5.2.2

[152] B. Nordlander, M. Krantz, and S. Hohmann. Hog1-mediated metabolic adjustments following hyperosmotic shock in the yeast saccharomyces cerevisiae. In F. Posas and A. R. Nebreda (eds.), *Stress-Activated Protein Kinases*, vol. 20 of *Topics in Current Genetics*, 141–158. Springer Berlin Heidelberg, 2008. 3.5.2, 3.5.3

[153] J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, and M. Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3): 635–648, 2006. 6.2.4

[154] S. M. O'Rourke and I. Herskowitz. Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Molecular Biology of the Cell*, 15(2): 532 –542, 2004. 4.3.4

[155] O. Ourfali, T. Shlomi, T. Ideker, E. Ruppin, and R. Sharan. SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, 23(13): i359–i366, 2007. 2, 3.1, 3.2.4, 4.1

[156] X. Pan and J. Heitman. Sok2 regulates yeast pseudohyphal differentiation via a transcription factor cascade that regulates cell-cell adhesion. *Molecular and Cellular Biology*, 20(22): 8364–8372, 2000. 3.5.3

[157] X. Pan and J. Heitman. Protein kinase a operates a molecular switch that governs yeast pseudohyphal differentiation. *Molecular and Cellular Biology*, 22(12): 3981–3993, 2002. 3.5.3

[158] T. Peleg, N. Yosef, E. Ruppin, and R. Sharan. Network-free inference of knockout effects in yeast. *PLoS Computational Biology*, 6(1): e1000635, 2010. 2, 4.1

[159] H. Phenix, K. Morin, C. Batenchuk, J. Parker, V. Abedi, L. Yang, L. Tepliakova, T. J. Perkins, and M. Kaern. Quantitative epistasis analysis and pathway inference from genetic interaction data. *PLoS Computational Biology*, 7(5): e1002048, 2011. 6.2.3

[160] O. Piloto, M. Wright, P. Brown, K. Kim, M. Levis, and D. Small. Prolonged exposure to FLT3 inhibitors leads to resistance via activation of parallel signaling pathways. *Blood*, 109(4): 1643–1652, 2007. 1

[161] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3): 447–455, 2011. 6.2.1

[162] D. K. Pokholok, J. Zeitlinger, N. M. Hannett, D. B. Reynolds, and R. A. Young. Activated signal transduction kinases frequently occupy target genes. *Science*, 313(5786): 533 –536, 2006. 3.5.2, 3.2, 3.3

[163] T. Pramila, S. Miles, D. GuhaThakurta, D. Jemiolo, and L. L. Breeden. Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. *Genes & Development*, 16(23): 3034–3045, 2002. 2.3.2

[164] R. J. Prill, P. A. Iglesias, and A. Levchenko. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biology*, 3(11): e343, 2005. 6.2.4

[165] M. Proft, A. Pascual-Ahuir, E. de Nadal, J. Ariño, R. Serrano, and F. Posas. Regulation of the Sko1 transcriptional repressor by the Hog1 MAP kinase in response to osmotic stress. *The EMBO Journal*, 20(5): 1123–1133, 2001. 3.5.3

[166] J. Ptacek, G. Devgan, G. Michaud, H. Zhu, X. Zhu, J. Fasolo, H. Guo, G. Jona, A. Breitkreutz, R. Sopko, *et al.* Global analysis of protein phosphorylation in yeast. *Nature*, 438(7068): 679–684, 2005. 3.5.3, 5.1.2

[167] P. Raghavendra and D. Steurer. How to round any CSP. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, 2009. 3.2.3

[168] B. J. Raney, M. S. Cline, K. R. Rosenbloom, T. R. Dreszer, K. Learned, G. P. Barber, L. R. Meyer, C. A. Sloan, V. S. Malladi, K. M. Roskin, *et al.* ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Research*, 39(suppl 1): D871–D875, 2011. 6.2.1

[169] T. Reguly, A. Breitkreutz, L. Boucher, B. Breitkreutz, G. Hon, C. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, *et al.* Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *Journal of Biology*, 5(4): 11, 2006. 2.3.1

[170] K. Robasky and M. L. Bulyk. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 39(suppl 1): D124–D128, 2011. 6.2.1

[171] J. M. Rodríguez-Peña, R. García, C. Nombela, and J. Arroyo. The high osmolarity glycerol (HOG) and cell wall integrity (CWI) signalling pathways interplay: a yeast dialogue between MAPK routes. *Yeast*, 27(8): 495–502, 2010. 4.3.1

[172] L. Romero-Santacreu, J. Moreno, J. E. Pérez-Ortín, and P. Alepuz. Specific and global regulation of mRNA stability during osmotic stress in Saccharomyces cerevisiae. *RNA*, 15(6): 1110 –1120, 2009. 4.3, 4.3.4

[173] P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlić, M. Quesada, G. B. Quinn, J. D. Westbrook, *et al.* The RCSB protein data bank: redesigned web site and web services. *Nucleic Acids Research*, 39(suppl 1): D392–D401, 2011. 1.1

[174] M. Sanchez, S. Bouveret, S. de Givry, F. Heras, P. Jegou, J. Larrosa, S. Ndiaye, E. Rollon, T. Schiex, C. Terrioux, *et al.* Max-CSP competition 2008: toulbar2 solver description. In M. van Dongen, C. Lecoutre, and O. Roussel (eds.), *Proceedings of the Third International CSP Solver Competition*, 63–70, 2008. 3.2.3

[175] D. D. Schlaepfer, K. C. Jones, and T. Hunter. Multiple Grb2-mediated integrin-stimulated signaling pathways to ERK2/mitogen-activated protein kinase: Summation of both c-Src- and focal adhesion kinase-initiated tyrosine phosphorylation events. *Molecular and Cellular Biology*, 18(5): 2571–2585, 1998. 1

[176] J. Scott, T. Ideker, R. M. Karp, and R. Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *Journal of Computational Biology*, 13(2): 133–144, 2006. 3.1, 3.3, 3.3.2

[177] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11): 2498–2504, 2003. 1.2

[178] S. D. Shapira, I. Gat-Viks, B. O. Shum, A. Dricot, M. M. de Grace, L. Wu, P. B. Gupta, T. Hao, S. J. Silver, and D. E. Root. A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*, 139(7): 1255–1267, 2009. 1, 5.1.1, 5.1, 5.2.1, 5.2.2, 5.2.2, 5.2.3, 5.5

[179] K. Sharma, S. Tripathi, P. Ranjan, P. Kumar, R. Garten, V. Deyde, J. M. Katz, N. J. Cox, R. B. Lal, S. Sambhara, *et al.* Influenza A virus nucleoprotein exploits Hsp40 to inhibit PKR activation. *PLoS ONE*, 6(6): e20215, 2011. 5.5

[180] G. Shenhar and Y. Kassir. A positive regulator of mitosis, Sok2, functions as a negative regulator of meiosis in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 21(5): 1603–1612, 2001. 3.5.3

[181] B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. part II computational methods to predict protein and domain interaction partners. *PLoS Computational Biology*, 3(4): e43, 2007. 3.5

[182] N. Simonis, J. Rual, I. Lemmens, M. Boxus, T. Hirozane-Kishikawa, J. Gatot, A. Dricot, T. Hao, D. Vertommen, S. Legros, *et al.* Host-pathogen interactome mapping for HTLV-1 and 2 retroviruses. *Retrovirology*, 9(1): 26, 2012. 6.2.1

[183] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright. Computational prediction of protein-protein interactions. *Molecular Biotechnology*, 38(1): 1–17, 2008. 3.5

[184] B. Snijder, R. Sacher, P. Rämö, P. Liberali, K. Mench, N. Wolfrum, L. Burleigh, C. C. Scott, M. H. Verheije, J. Mercer, *et al.* Single-cell analysis of population context advances RNAi screening at multiple levels. *Molecular Systems Biology*, 8(579), 2012. 5.1.1

[185] L. Song, M. Kolar, and E. P. Xing. KELLER: estimating time-varying interactions between genes. *Bioinformatics*, 25(12): i128–i136, 2009. 4.1

[186] C. Stark, B. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, *et al.* The BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 39(suppl 1): D698–704, 2011. 6.2.1

[187] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1): D535–539, 2006. 3.2.4, 3.5.2, 4.3, 5.1.2

[188] M. Steffen, A. Petti, J. Aach, P. D'haeseleer, and G. Church. Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3(1): 34, 2002. 3.1, 3.3

[189] S. Stertz and M. L. Shaw. Uncovering the global host cell requirements for influenza virus replication via RNAi screening. *Microbes and Infection*, 13(5): 516–525, 2011. 5.1.1

[190] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(suppl 1): D561–D568, 2010. 3.5.2

[191] L. Tafforeau, T. Chantier, F. Pradezynski, J. Pellet, P. E. Mangeot, P. Vidalain, P. Andre, C. Rabourdin-Combe, and V. Lotteau. Generation and comprehensive analysis of an influenza virus polymerase cellular interaction network. *Journal of Virology*, 85(24): 13010–13018, 2011. 5.2.2, 5.5

[192] A. W. Tai, Y. Benita, L. F. Peng, S. Kim, N. Sakamoto, R. J. Xavier, and R. T. Chung. A functional genomic screen identifies cellular cofactors of hepatitis C virus replication. *Cell Host & Microbe*, 5(3): 298–307, 2009. 6.2.1

[193] Y. Tang and S. I. Reed. The Cdk-associated protein Cks1 functions both in G1 and G2 in Saccharomyces cerevisiae. *Genes & Development*, 7(5): 822–832, 1993. 3.3.2

[194] K. Tatebayashi, K. Yamamoto, K. Tanaka, T. Tomida, T. Maruoka, E. Kasukawa, and H. Saito. Adaptor functions of Cdc42, Ste50, and Sho1 in the yeast osmoregulatory HOG MAPK pathway. *The EMBO Journal*, 25(13): 3033–3044, 2006. 3.5.2, 3.3

[195] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146): 799–816, 2007. 6.2.1

[196] The modENCODE Consortium, S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330(6012): 1787–1797, 2010. 4.2.1, 6.2.1

[197] A. H. Y. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, *et al.* Global mapping of the yeast genetic interaction network. *Science*, 303(5659): 808–813, 2004. 5.2.5

[198] H. Tong, C. Faloutsos, and J. Pan. Fast random walk with restart and its applications. In *Proceedings of the 6th International Conference on Data Mining (ICDM)*, 613–622, 2006. 3.5.1

[199] L. Trevisan. Parallel approximation algorithms by positive linear programming. *Algorithmica*, 21(1): 72–88, 1998. 3.2.3

[200] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9): 5116 –5121, 2001. 4.3.3

[201] P. Uetz, Y. Dong, C. Zeretzke, C. Atzler, A. Baiker, B. Berger, S. V. Rajagopala, M. Roupelieva, D. Rose, E. Fossum, *et al.* Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758): 239–242, 2006. 6.2.1

[202] J. Urban, A. Soulard, A. Huber, S. Lippman, D. Mukhopadhyay, O. Deloche, V. Wanke, D. Anrather, G. Ammerer, H. Riezman, *et al.* Sch9 is a major target of TORC1 in Saccharomyces cerevisiae. *Molecular Cell*, 26(5): 663–674, 2007. 4.3.6

[203] A. Vinayagam, U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, J. Timm, H. E. Assmus, M. A. Andrade-Navarro, and E. E. Wanker. A directed protein interaction network for investigating intracellular signal transduction. *Science Signaling*, 4(189): rs8, 2011. 4.1, 6.2.4

[204] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887): 399–403, 2002. 1

[205] W. Voth, Y. Yu, S. Takahata, K. Kretschmann, J. Lieb, R. Parker, B. Milash, and D. Stillman. Forkhead proteins control the outcome of transcription factor binding by antiactivation. *The EMBO Journal*, 26: 4324–4334, 2007. 2.3.2

[206] J. P. Wang, G. N. Bowen, C. Padden, A. Cerny, R. W. Finberg, P. E. Newburger, and E. A. Kurt-Jones. Toll-like receptor-mediated activation of neutrophils by influenza A virus. *Blood*, 112(5): 2028–2034, 2008. 5.2.2, 5.5

[207] P. Wang, W. Song, B. W. Mok, P. Zhao, K. Qin, A. Lai, G. J. D. Smith, J. Zhang, T. Lin, Y. Guan, *et al.* Nuclear factor 90 negatively regulates influenza virus replication by interacting with viral nucleoprotein. *Journal of Virology*, 83(16): 7850–7861, 2009. 5.5

[208] M. P. Ward, C. J. Gimeno, G. R. Fink, and S. Garrett. SOK2 may regulate cyclic AMP-dependent protein kinase-stimulated growth and pseudohyphal development by repressing transcription. *Molecular and Cellular Biology*, 15(12): 6854–6863, 1995. 3.5.3

[209] Z. Wei and H. Li. A hidden spatial-temporal markov random field model for network-based analysis of time course gene expression data. *The Annals of Applied Statistics*, 2(1): 408–429, 2008. 4.1

[210] P. J. Westfall, J. C. Patterson, R. E. Chen, and J. Thorner. Stress resistance and signal fidelity independent of nuclear MAPK function. *Proceedings of the National Academy of Sciences*, 105(34): 12212 –12217, 2008. 4.3.2

[211] M. J. Winters and P. M. Pryciak. Interaction with the SH3 domain protein Bem1 regulates signaling by the Saccharomyces cerevisiae p21-Activated kinase Ste20. *Molecular and Cellular Biology*, 25(6): 2177–2190, 2005. 4.3.5

[212] K. Won, B. Ren, and W. Wang. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology*, 11(1): R7, 2010. 6.2.1

[213] C. T. Workman, H. C. Mak, S. McCuine, J. Tagne, M. Agarwal, O. Ozier, T. J. Begley, L. D. Samson, and T. Ideker. A systems approach to mapping DNA damage response pathways. *Science*, 312(5776): 1054–1059, 2006. 2.1, 4.1

[214] C. Wu, T. Leeuw, E. Leberer, D. Y. Thomas, and M. Whiteway. Cell cycle- and Cln2p-Cdc28p-dependent phosphorylation of the yeast Ste20p protein kinase. *The Journal of Biological Chemistry*, 273(43): 28107–28115, 1998. 3.3.2

[215] Z. Wunderlich and L. A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, 25(10): 434–440, 2009. 2.1

[216] M. W. Xie, F. Jin, H. Hwang, S. Hwang, V. Anand, M. C. Duncan, and J. Huang. Insights into TOR function and rapamycin response: Chemical genomic profiling by using a high-density cell array method. *Proceedings of the National Academy of Sciences*, 102(20): 7215 –7220, 2005. 4.3.6

[217] C. Yeang, T. Ideker, and T. Jaakkola. Physical network models. *Journal of Computational Biology*, 11(2-3): 243–262, 2004. 2, 2.1, 3.1, 4.1, 4.3.7, 5.3.2, 5.3.3, 6.2.4

[218] C. Yeang, H. C. Mak, S. McCuine, C. Workman, T. Jaakkola, and T. Ideker. Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biology*, 6(7): R62, 2005. 4.1, 4.3.7

[219] E. Yeger-Lotem, L. Riva, L. J. Su, A. D. Gitler, A. G. Cashikar, O. D. King, P. K. Auluck, M. L. Geddie, J. S. Valastyan, D. R. Karger, *et al.* Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature Genetics*, 41(3): 316–323, 2009. 4.1, 4.3.7, 4.3.8, 5.1.1

[220] M. L. Yeung, L. Houzet, V. S. R. K. Yedavalli, and K. Jeang. A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. *Journal of Biological Chemistry*, 284(29): 19463–19473, 2009. 6.2.1

[221] N. Yosef, L. Ungar, E. Zalckvar, A. Kimchi, M. Kupiec, E. Ruppin, and R. Sharan. Toward accurate reconstruction of functional protein networks. *Molecular Systems Biology*, 5(248), 2009. 3.1, 4.1

[222] K. Yoshikawa, T. Tanaka, C. Furusawa, K. Nagahisa, T. Hirasawa, and H. Shimizu. Comprehensive phenotypic analysis for identification of genes affecting growth under ethanol stress in Saccharomyces cerevisiae. *FEMS Yeast Research*, 9(1): 32–44, 2009. 3.5.3

[223] T. Yoshikawa, T. E. Hill, N. Yoshikawa, V. L. Popov, C. L. Galindo, H. R. Garner, C. J. Peters, and C. K. Tseng. Dynamic innate immune responses of human bronchial epithelial cells to severe acute respiratory syndrome-associated coronavirus infection. *PLoS One*, 5(1): e8729, 2010. 5.5

[224] H. Yu, A. Paccanaro, V. Trifonov, and M. Gerstein. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7): 823–829, 2006. 3.5, 3.5.2

[225] S. Zaman, S. I. Lippman, X. Zhao, and J. R. Broach. How Saccharomyces responds to nutrients. *Annual Review of Genetics*, 42(1): 27–81, 2008. 3.5.3, 4.3.6

[226] A. Zarrinpar, R. P. Bhattacharyya, M. P. Nittler, and W. A. Lim. Sho1 and Pbs2 act as coscaffolds linking components in the yeast high osmolarity MAP kinase pathway. *Molecular Cell*, 14(6): 825–832, 2004. 3.3.2

[227] L. Zhang, X. Zhang, Q. Ma, F. Ma, and H. Zhou. Transcriptomics and proteomics in the study of H1N1 2009. *Genomics, Proteomics & Bioinformatics*, 8(3): 139–144, 2010. 5.2

[228] X. Zhao, R. Wang, L. Chen, and K. Aihara. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Research*, 36(9): e48, 2008. 3.1, 3.3

[229] H. Zhou, M. Xu, Q. Huang, A. T. Gates, X. D. Zhang, J. C. Castle, E. Stec, M. Ferrer, B. Strulovici, and D. J. Hazuda. Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host & Microbe*, 4(5): 495–504, 2008. 6.2.1

[230] G. Zhu, P. T. Spellman, T. Volpe, P. O. Brown, D. Botstein, T. N. Davis, and B. Futcher. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, 406(6791): 90–94, 2000. 2.3.2

[231] U. Zwick. Approximation algorithms for constraint satisfaction problems involving at most three variables per constraint. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 201–210. Society for Industrial and Applied Mathematics, 1998. 3.2.3