UNIVERSITY OF CALIFORNIA
SANTA CRUZ

PREDICTION AND EXPANSION OF BIOLOGICAL PATHWAYS FROM PERTURBATION EFFECTS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSPHY

in

BIOINFORMATICS

by

Charles J. Vaske

September 2009

The Dissertation of Charles J. Vaske
is approved:

Professor Joshua Stuart, Chair

Professor Kevin Karplus

Professor Fitnat Yildiz

Lisa C. Sloan
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures **vi**

List of Tables **viii**

Abstract **ix**

Dedication **x**

I Background **1**

1 Biological Networks and Processes **2**

1.1 Multi-molecule bonding and complexes 3

1.2 Gene regulation 4

1.3 Signal transduction 8

1.4 Transitivity and scope in biological networks 10

1.5 Traditional biomolecular methods for network discovery 13

2 Measuring Biological Networks **15**

2.1 Perturbation Methods 16

2.1.1 Genetic perturbation 16

2.1.2 Impulse perturbation 17

2.2 High-throughput gene expression measurement 19

2.3 High-throughput network structural measurement 21

2.3.1 Protein-Protein Interaction 22

2.3.2 Protein-DNA Interaction 25

3 Probabilistic and Other Graphical Models **27**

3.1 Graph formulations 28

3.1.1 Markov random fields 30

3.1.2 Bayesian networks 32

3.1.3 Factor graphs 33

3.2 Inference methods 35
3.2.1 Exact Inference 36
3.2.2 Approximation with message passing 38
3.3 Structure learning 40
3.4 Causality in graphical models 40

4 Network Prediction Methods 42
4.1 *De novo* methods and network refinement methods 42
4.2 Computational methods 43
4.2.1 Protein-protein based methods 44
4.2.2 Perturbed gene-expression methods 44

II Methods and Results 46

5 Joint Intervention Networks 47
5.1 Relationship to published work 48
5.2 Modeling epistasis with Bayesian Networks 49
5.3 Regulatory Model 54
5.4 Regulatory Model Score 55
5.5 Network Search 56
5.6 Prediction of *V. cholerae* biofilm network 58
5.7 Expansion of *V. cholerae* biofilm network 63
5.8 Conclusions 64

6 Factor graph Nested Effect Model 66
6.1 Nested Effect Model Averaging 67
6.1.1 Linear models recover *D. melanogaster* immune response 68
6.2 Interaction Modes 70
6.3 Pairwise Network Formulation 77
6.4 Transitivity Constraints and Model Priors 80
6.5 Model Inference Using a Factor Graph 82
6.6 Pathway Expansion 87

7 FG-NEM Performance on Artificial Networks 90
7.1 Network and Data Generation 91
7.1.1 Calculating Network Structure Recovery Accuracy 95
7.1.2 Calculating Network Expansion Accuracy in Artificial Data . . . 96
7.2 Comparison to Unsigned FG-NEM variants and NEM 97
7.3 FG-NEM Structure Prediction Performance 99
7.3.1 Varying Amount of Inhibition 100
7.3.2 Varying Number of Data Replicates 103
7.3.3 Varying Fraction of Known Network 105
7.4 FG-NEM Pathway Expansion Performance 106

8 Pathway Prediction and Expansion in *S. cerevisiae* 109

8.1 Expression Data and Gene Sets 109
8.2 Pathway Expansion Performance 110
8.3 Physical Structure Priors 114
8.4 Pathway Structure Prediction 116
8.5 Predicted inhibition in ion homeostasis pathway 118

9 Cancer Invasion Pathway Prediction and Expansion 121

9.1 Previous Data and Methods 122
9.2 Initial Network and Frontier 123
9.3 Frontier Verification 129
9.4 Expanded Cancer Invasion Network 133
9.4.1 Membrane/protease domain 135
9.4.2 Calcium domain 136
9.4.3 Secretory/trafficking domain 136
9.5 Expanded Cancer Invasion Frontier 137

10 Discussion 141

10.1 Conclusions 141
10.2 Future Directions 142

Bibliography 146

List of Figures

1.1 Gene regulation in the *E. coli* lactose metabolism pathway. Biological entities are labeled with text, and arrows indicate interactions between entities. 5
1.2 Extended *lac* operon network. This network shows the same core as in Figure 1.1 but with more of the regulators. In addition, different contexts are modeled, namely an intracellular context above “Cell Wall” and an extracellular context below. 9
1.3 A reduced scope *lac* network. 11

3.1 Example Markov network. 31
3.2 Correspondence between Markov random fields, Bayesian networks and factor graphs 34
3.3 Transformation of graph under variable elimination 36

5.1 Switch regulatory epistasis analysis 49
5.2 Overview of the experimental setup 51
5.3 Example Joint Intervention Network 53
5.4 Predicted Biofilm Network 59
5.5 Significance of biofilm network. 61
5.6 Biofilm network expansion 62

6.1 *D. melanogaster* model averaging predictions 68
6.2 Observed inhibitory effects and signaling in a yeast compendium. 71
6.3 Hypothetical example network and expression 73
6.4 Knock-out expression for a known inhibitory interaction 75
6.5 Interaction Modes 76
6.6 Factor graph structure 85

7.1 Estimating the difference between up- and down- distributions using proteasome- and ribosome- related genes as contrast sets 94
7.2 Comparison of uFG-NEM and exhaustive NEM model search for structure recovery 98

7.3 Influence of inhibition on network recovery 100
7.4 Network recovery as a function of increasing inhibition in E-gene attachment 102
7.5 Influence of replicates in network recovery 103
7.6 Influence of incomplete pathway knowledge on network recovery 105
7.7 Accuracy of FG-NEM network expansion compared to Template Matching107

8.1 Yeast Pathway Expansion Precision/Recall Comparison 111
8.2 Yeast Expansion Performance AUC Comparison 112
8.3 Compatibility of high-throughput physical evidence and predicted S-gene interactions 116
8.4 Predicted S-gene networks for the Ion Homeostasis pathway 118

9.1 Expression changes of selected E-genes following targeted S-gene knockdowns in HT29 colon cancer cells 124
9.2 Cancer invasion network predicted by FG-NEM 126
9.3 Matrigel Invasion Assay in HT29 Colon Cancer Cells 132
9.4 Inferred S-gene network and Frontier 134
9.5 S-gene interaction confidence 137
9.6 Frontier enrichment 138

List of Tables

5.1 Top new *vps* predictions 63

8.1 Pathway expansion AUC on a yeast gene-deletion compendium 113
8.2 FG-NEM Ion Homeostasis pathway expansion frontier 120
8.3 uFG-NEM Ion Homeostasis pathway expansion frontier 120

9.1 Initial cancer invasion frontier 131

Abstract

Prediction and expansion of biological pathways from perturbation effects

by

Charles J. Vaske

Complex phenotypes, such as cancer invasion, result from the actions and interactions of many genes and gene products. Though pathway-based analysis can offer improved predictions over single-gene or set-of-gene analyses, few pathways have been characterized. New high-throughput technology offers the opportunity for individual investigators to learn entire pathways from a small number of gene perturbation and gene expression experiments. I present two pathway inference methods using data from downstream perturbation effects and use the inferred structures to predict novel pathway members in cancer invasion and *Vibrio cholerae* biofilm.

In a *V. cholerae* system, microarray gene expression data under gene perturbations from deletion knockouts was analyzed using a new method called a Joint Intervention Network. This analysis resulted in an inferred regulatory network of the perturbed genes, and prediction of biofilm-associated genes that was more accurate than a correlation-based method.

I next developed a signed version of the Nested Effects Model and an associated efficient structure inference method, named Factor Graph-Nested Effects Model (FG-NEM). On synthetic data I show improved performance of FG-NEM over unsigned versions of the algorithm. In yeast, FG-NEM predicts Gene Ontology categories more accurately than a correlation-based method. And finally, in a cancer cell line I predicted an invasion network and identified fourteen new genes necessary for cancer invasion.

I dedicate this dissertation to my brilliant wife Yvette,
whose support has been critical to all my work here.

Part I

Background

Chapter 1

Biological Networks and Processes

Biological networks aim to describe the large scale activity of cells in terms of the interactions of a cell's components. Discoveries in molecular biology have told us that these activities are performed by various classes of biological entities: *e.g.* DNA, RNA, proteins, membranes, and small molecules. Molecular biology has also informed us of many common types interactions that happen between biological entities. Thus, "biological network" is a rather broad term, that can refer to a description of how a complex mulitigenic phenotype arises from genotype, to a simple biochemical reaction involving only an enzyme, substrate molecule, and a product molecule.

This network representation of biological processes, specifically a graph representation, pervades system biology [112]. The dominant machine-readable formats, SBML [55, 32], BioPAX [81, 126], PSI MI [51] and CellML [23] all describe biological processes as graphs, but each format places different emphases on structure, dynamics, and supporting evidence.

In this dissertation, I will focus on three primary interaction classes: complex formation, gene regulation, and signal transduction. A fourth class of interaction, biochemical reaction, is important and represented in nearly all systems biology ontologies, but there is currently no high-throughput assay that allows for computational predictions of such reactions, so I will not treat them explicitly. In the sense that

1.1 Multi-molecule bonding and complexes

Often, multiple molecules of RNA, DNA, or protein act together as a single entity in a biological process. When these molecules stably bond, it is referred to as a complex. Networks model both the individual components and the complex.

As an example, the ribosome is a complex that consists of a stable core of two major RNA units and many proteins. Humans have approximately 80 ribosomal proteins [125] and *E. coli* have more than 50 [3]. This core complex is stable enough to exist while the ribosome is not performing translation, and to be isolated from cell extracts. During translation, additional proteins and RNAs associate with the ribosome in a more transient manner. Biological networks attempt to capture and describe not only the stable parts of the complex, but also the transient associations.

The spliceosome, like the ribosome, consists of both RNA and protein, but instead of being a stable complex, assembles as needed. This complex removes introns from precursor mRNAs as part of mRNA processing. The spliceosome can be modeled as a complex of complexes, each referred to in general as small nuclear ribonucleoproteins (snRNPs). Full complex assembly is performed in sequential steps, delimited by ATP-dependent energy wells [98].

1.2 Gene regulation

As the first step in the physical link between genotype and phenotype, the determining factor between morphologically distinct cells/tissues in the same organism, and perhaps the major evolutionary difference between humans and their closest neighbors [72], gene regulation features prominently in the structure of biological networks. Gene regulation refers to the production of an active product from a single genomic locus, where the product is a protein or RNA. There are known examples of gene regulation occurs at nearly all instantiations of a gene: controlling of transcription of the gene, modulation of the steps in between transcription and translation, post-translational modifications of protein, and protein degradation. Biological networks can model the state of the regulatory elements, the presence or abundance of active gene product, and the state of the gene product intermediates, though the simplest networks model only the presence of the final gene product. Often, models of gene regulation will not state the exact mechanism of regulation. A gene regulation network may elide some elements in a chain of multiple steps of regulation. For example, if a protein A activates chromatin machinery to silence a gene b, preventing the protein B from activating transcription of gene c, then it may be said that a regulates c.

Both activation and inhibition are important aspects of regulation, and both

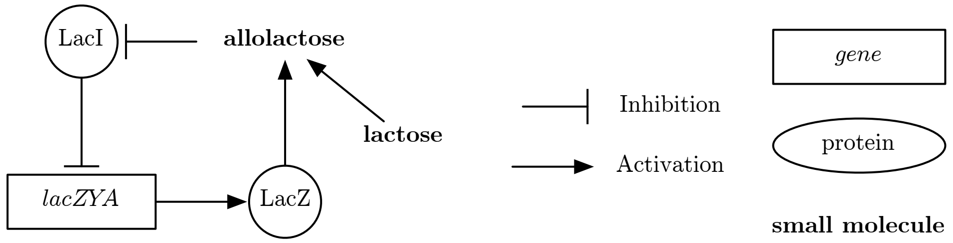


Figure 1.1: Gene regulation in the *E. coli* lactose metabolism pathway. Biological entities are labeled with text, and arrows indicate interactions between entities.

are found extensively in gene regulation networks. As an example of gene regulation, I will describe the regulation of the *E. coli lac* operon. Prior to discovery of the *lac* operon, it was known that some enzymes in bacteria would only be produced when needed. The work of Jacob and Monod [64] on the *lac* operon was the first characterization of gene regulation. *E. coli* activates this network only when its preferred sugar, glucose, is not present. In the presence of lactose and absence of glucose, *E. coli* uses the *lac* operon to convert lactose to glucose. The network that includes the *lac* operon exhibits activation, inhibition, multiple types of components, extra-cellular signaling, and cycles. This network has all the characteristics that I aim to predict in this dissertation.

Figure 1.1 shows a cartoon of the *lac* operon network. The two proteins involved in this network, LacI and LacZ, are an inhibitory regulator and an enzyme respectively. The nodes **allolactose** and **lactose** are both small molecules. The node *lacZYA* is an operon, a gene with multiple protein products. The tee-arrow (⊣) indicates that the presence of LacI inhibits *lacZYA*. Similarly, the tee-arrow from node **allolactose** to node LacI indicates that the biochemical observation that allolactose inhibits LacI, preventing it from acting. The arrow (→) from node *lacZYA* to node LacZ indicates "activation," which mean that the target of the link, LacZ, is also activated, here by transcription and translation. There are two arrows into the **allolactose** node, one from LacZ and one from **lactose**, both indicating activation. In my representation of networks, the presence of more than one link into a node indicates that their effects combine "multiplicatively." This means that when the incoming links are activation, all of the regulators must be present in order for the regulatee to be activated. In this case, it has been experimentally observed that LacZ converts lactose into allolactose, and that the presence of both enzyme and substrate are necessary for the product.

Recall that the enzyme for conversion of lactose, LacZ, is only active when the substrate is present, meaning the LacZ is regulated by lactose. The cycle in the network explains how LacZ is regulated and produced in only the proper conditions. The cycle in the network consisting of nodes LacI, *lacZYA*, LacZ, and **allolactose**, is regulated by the upstream node **lactose**. Ignoring this upstream node, the sequence of links in the cycle dictates two consistent solutions for the presence/activation and absence/inactivation of the entities: { LacI = *inactive*, *lacZYA* = *active*, LacZ = *present*, **allolactose** = *present*} and { LacI = *active*, *lacZYA* = *inactive*, LacZ = *absent*, **allolactose** = *absent*}. When we consider the link from the node **lactose** to node **allolactose** we see that in the absence of **lactose** then **allolactose** must also be absent, and therefore node LacI must be active and node *lacZYA* inactive. When the node **lactose** is present, the other solution is consistent, as long as LacZ is also present. If there is absolutely no LacZ, then the addition of lactose will not change the system.

The usage of all these terms is very loose, and such looseness is necessary in order to have a generalized way of talking about these systems. The concept of "presence" and "absence" is different for different entities in the system. When referring to node LacI, the active and inactive refer to its ability to inhibit node *lacZYA*. Biologically, LacI binds the promoter of the *lacZ* gene, preventing transcription. Biochemically it was observed that allolactose is an allosteric effector of LacI, and when bound LacI can no longer bind the promoter, making LacI "inactive." However, when referring to LacZ, **allolactose**, and **lactose**, the terms "presence" and "absence" mean that the cell has a larger or smaller quantity of the entity. And though I refer to node LacZ as "absent," it is never entirely absent biologically, even in the absence of lactose. This is because *lacZYA* is expressed whenever not bound by LacI, and LacI binds loosely enough for a few errant transcripts to occur, resulting in a small amount of endogenous LacZ. This small amount allows the cycle to switch to the solution with **allolactose** present when lactose is introduced to the network.

Transcriptional repressors such as LacI are just one type of a larger class called transcription factors. There are also transcriptional activators, that increase the rate of transcription. Transcription factors act on *cis*-regulatory elements, DNA sequences that are on the same DNA molecule that contains the gene.

Eukaryotes have a very rich toolbox for regulating gene expression beyond *cis*-regulatory elements. The methods of regulation in eukaryotes include chromatin structure/domains [68], DNA methylation and imprinting [100], microRNAs [50] and RNA mediated interference [33], nonsense-mediated decay [78], and even the threedimensional organization of chromosomes in the nucleus [34]. Many of these regulatory methods have only been recently discovered, and advances in molecular biology may discover yet more.

1.3 Signal transduction

Signal transduction is the process of a cell turning the perception of something external, usually a molecule, into a response inside the cell. This response will almost always be energy dependent, and often involve protein phosphorylation. In this section I will first describe how the *E. coli lac* network performs a function like signal transduction. Then, I will explain some of what is known about signal transduction in the datasets that I use later in this dissertation.

Figure 1.2 shows an elaborated network for the *lac* operon. The core components are LacI, *lacZYA*, LacZ, **cellular lactose**. What was labeled **lactose** in Figure 1.1 has been more specifically relabeled as **cellular lactose** here. The new network includes LacY, another product of the *lacZYA* operon. LacY is a transmembrane protein that uses ion flux to move lactose into the cell. In the absence of extracellular lactose, there are very low levels of LacY and LacZ, since LacI almost fully represses *lacZYA*. When extracellular lactose is present, it will interact with the small amount of endogenous LacY, resulting in a small amount of cellular lactose, which results in a small amount of allolactose, releasing LacI and permitting the *lac* operon to be transcribed.

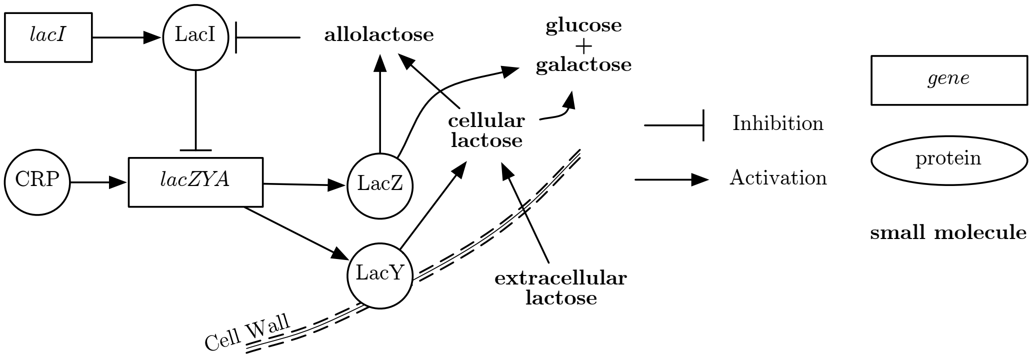


Figure 1.2: Extended *lac* operon network. This network shows the same core as in Figure 1.1 but with more of the regulators. In addition, different contexts are modeled, namely an intracellular context above “Cell Wall” and an extracellular context below.

The eukaryotic signal transduction repertoire includes ion-flux transporters such as in the *E. coli lac* network but alse many more kinase-based systems. These are referred to as mitogen-activated protein (MAP) kinases, since they are activated as the result of external small molecules. Though there are many families of signal transduction, here I will only discuss two as representatives: the G-protein couple receptor (GPCR) family and the receptor tyrosine kinase (RTKs).

GPCR-based signal transduction is responsible for a large class of the studied cellular responses in humans, including processes as broad as vision, neurotransmission, and histidine response. GPCRs are a class of proteins that integrate into the plasma membrane via seven-transmembrane alpha helices. The portion of the GPCR on the exterior of the membrane binds to specific ligand or a specific class of ligands. When the ligand is bound, the conformation of the GPCR on the inner side of the membrane changes. On the inner side of the membrane, GPCRs are coupled to G proteins, which are named for their guanosine-binding properties. G proteins are localized on the cellular membrane, next to a GPCR. Upon conformational change of the GPCR, the G protein exchanges bound guanosine diphosphate (GDP) for guanosine triphosphate (GTP), and is no longer localized to the cell membrane. This G protein is now activated to continue a MAP kinase response elsewhere in the cell.

Similar to GPCRs, RTKs are located in the cellular membrane and contain a ligand-binding receptor domain on the exterior of the membrane. Upon binding a ligand in the receptor domain, the RTK is activated to phosphorylate a tyrosine target on a phosphotyrosine binding (PTB) domain on another protein.

1.4 Transitivity and scope in biological networks

An important aspect of biological networks for my proposal is that they can consistently and accurately be viewed at multiple scope and scales. By scope, I mean the portion of the network that is under examination or investigation. In Figure 1.2, the network shows some indication of how the *lac* network connects to other parts of the entire cellular network. The protein CRP is another transcriptional regulator that effects not just the *lac* operon, but almost 200 other transcriptional units in *E. coli* according to the database EcoCyc [71]. Also shown are the primary products of LacZ, glucose and galactose. Using data from EcoCyc, we could connect **glucose** to four other biochemical pathways in *E. coli* as a substrate.

In principle, the entire cell and all its functions could be modeled with such

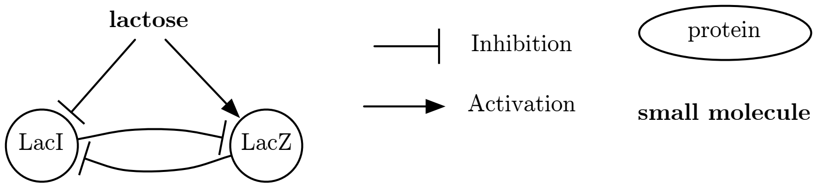


Figure 1.3: A reduced scope *lac* network.

a network of a very large size. Despite the size and interconnectedness of the cellular network, we are able to narrow the scope of investigation to a small number of genes.

In addition to narrowing our investigation to a small number of entities, we can also narrow scope to just a few types of entities and retain a consistent and informative view of the biological process. Figure 1.3 shows a network of just two proteins and the external lactose input. If we were only able to measure these two proteins, and control the external availability of lactose, this would still be an accurate description of the network. In this reduced network, the link from LacI to LacZ is the the result of the transitive closure of LacI⊣*lacZYA*→LacZ from Figure 1.1 and Figure 1.2.

This consistency under reduction of scope is essential to our ability to investigate and discover biological networks. We currently have no physical techniques for the simultaneous measurement of all entities in a cellular network, and the size of such measurements would outstrip our mathematical and computational tools for inference. However, we are able to explore small pieces of the network, and even without knowing the full context of the scope, we are able to accurately infer interactions.

When reducing scope in this way, our activation and inhibition links may no longer correspond to direct physical interaction or immediate causes. Figure 1.3 shows direct links from **lactose** node to the protein nodes. However, there is a distinction between the model and the known biology: we know that the physical interactions are mediated by allolactose in the case of the link from the **lactose** node to LacI. The chain of physical interactions is longer for the link from node **lactose** to LacZ. The entities **allolactose**, LacI, and *lacZYA* are all involved in describing the activation link from node **lactose** to node LacZ. These links are therefore also the result of transitive physical links.

Even the extended network in Figure 1.2 omits much of our knowledge of the network. For example, we know more of the structure of the promoter of *lacZYA*, and we ignore such essential components such as the transcription and translational machinery.

This consistency under change of scope is particularly essential to the methods in this dissertation. My methods focus on determining these networks not by measuring any of the entities directly, but only by looking for downstream changes in gene expression. In addition, I am only placing a single node for a gene in the network, mixing the mRNA and protein species of a gene while ignoring small molecules. Under these conditions, I expect to predict networks that are consistent with the true biological network, but amenable to further study in two ways. First, there may be additional intervening proteins on the links that I predict. Second, by adding more types of entities using other investigation techniques, my predictions could be filled with more direct physical causes.

1.5 Traditional biomolecular methods for network discovery

Recall the *lac* operon network from Figure 1.2. This network describes the activation of the *E. coli* lactase enzyme, LacZ, in the presence of lactose. The protein LacI serves as a general inhibitor of LacZ and the operon *lacZYA*, preventing the expression of both the lactase and the transporter which moves lactose across the cell wall. However, if there is a small amount of LacZ and a small amount of cellular lactose, then LacZ will produce some amount of allolactose, which inhibits LacI, freeing up the promoter of *lacZYA* and activating the metabolic pathway in response to an external signal.

All of this was discovered in piecemeal fashion through the use of structural perturbations to the cellular network. This network was the culmination of over a decade's worth of work, and resulted in the discovery of the nature of *cis*- and *trans*-gene regulatory elements. Inferences in all steps of the network were aided by perturbations: analogs of the small molecules allolactose and lactose were used to perturb the corresponding nodes in the network and mutants were discovered that perturbed the function of the proteins LacI, LacZ, and LacY. Critically, an *E. coli* strain was used which had a functioning LacI protein, but a mutation in the binding site of LacI near the promoter of *lacZYA*, establishing the existence of *cis*-regulatory elements.

The framework I propose for biological network discovery follows much the same path. To begin network inference, we propose a set of models based on a limited amount of perturbation data. From this set of models, we determine which further effects are most likely to disambiguate our existing models and expand upon the search. The search procedure can then iterate until the network is completely elaborated or the limits of gene-expression in the network prevent further research.

Chapter 2

Measuring Biological Networks

The discovery of the *lac* operon network was the culmination of many studies using a wide array of techniques to measure and perturb biological entities. Today, the biological techniques for network inference follow the same lines. The particular techniques for measurement and perturbation have become easier; perturbation is more directed and easier, and measurement covers mary more entities at once and more types of biological entities. However, the general principles are the same in that network discovery relies both on the directed perturbation of a system and measurements of responses to perturbation.

This chapter discusses the biological methods that are useful for investigating networks and that are relevant to my methods. I will first discuss the methods for perturbing elements in a biological network. I will then discuss the measurement technique that my method uses. Finally, I will discuss high-throughput methods for detecting the presence of direct links in the biological network, which I can use either for verifying predictions or as prior knowledge when predicting networks.

2.1 Perturbation Methods

Perturbation is an extremely powerful tool for establishing causal relationships, as discussed in §3.4. The essence of the increased causal power of perturbation is that perturbation disconnects the perturbed element from its other causes, and therefore causes a structural change in the network. There exist several techniques for directing perturbation of biological systems. These fall into two classes: genetic perturbation and impulse perturbation. With genetic perturbation, measurements can only be taken at least one cell cycle after the perturbation has occurred, allowing responses to fully propagate throughout the cellular network. Impulse perturbation allows measurements to be taken within the same cell cycle, when not all responses to perturbation have yet traveled through the cellular network. Both techniques are used in the datasets I will analyze in my aims.

2.1.1 Genetic perturbation

Genetic perturbation usually involves reducing the functionality of a gene. This can be done either by gene deletion in which case there is complete loss of gene function, or by merely reducing the activity, resulting in a hypomorph.

Gene deletions are of such utility that the Saccharomyces Genome Deletions Consortium has deleted over 90% of *S. cerevisiae* open reading frames [124, 47]. These deletions proceed in several steps [9]. First, a DNA construct is created to replace the target gene. This construct contains a selectable gene, so that cells with successful replacements can be differentiated from those that have not been successfully replaced. The other ends of the construct are homologous to the 5' and 3' end of the target gene, allowing for recombination once the construct is inserted into cells. These constructs are inserted into a yeast culture using a lithium acetate treatment [48], which is called transformation in microbes. Importantly for the method and results in Chapter 5, techniques have also been developed for *V. cholerae* that allow gene deletion [39, 80].

Since such deletions last for the entire lineage of the cell, the effects on the rest of the cellular network are very broad. There has been some concern that using observations from broad changes will make confound network inference. However, recent experiments with synthetic data [108] give preliminary indications that long term perturbations like these might in fact allow better predictions.

2.1.2 Impulse perturbation

Impulse perturbation, unlike genetic perturbation, allows for the evaluation of the change in a network over time. The cellular network will respond to such perturbations dynamically, and depending on when measurements are taken after the perturbation, different effects will be observed. Though I do not use such temporal effects in my proposed methods, it is important to note the effect as it can have consequences for the interpretation of observations of the network and is a concern when using data of this sort.

In the *E. coli lac* network we saw an example of lactose, an external signal, providing a perturbation in the network, and more general systems have since been developed. There are several techniques for affecting specific genes in the cell at a controllable time: genetic mutations to produce temperature-sensitive versions of protein products, promoters to genes that respond to a drug treatment, and the RNAi system for degradation of mRNA.

Signal transduction, as discussed in §1.3, induces changes in the state of the network in the cell. These external stimuli activate, or deactivate, different parts of the cellular network. These perturbations are extremely useful for discovering which biological entities are related to each other, by identifying a connected component of the cellular network. However, these perturbations are not structural, in that they do not disconnect any elements from their causes, since by definition external variables do not have any causes within the cell.[1]

An example of impulse perturbation of internal elements of the network is the tetracycline-regulatable promoter in *S. cerevisiae* [41]. In this system, a gene is placed behind a special promoter such that whenever the cell culture is exposed to tetracycline the gene is effectively transcriptionally silent. When tetracycline is not present, the promoter is promiscuous and transcription of the gene may be increased 1000-fold. Here, it is clear how perturbation removes other causes: the promoter is changed such that the normal inputs into the gene are entirely missing. The normal gene regulatory program has been replaced with one that is easy to manipulate externally.

[1] Note that so far we have only discussed cellular networks, and not discussed the interaction of a cell with its environment. In some situations, such as quorum sensing signals in *V. cholerae* growth, the cellular network affects the extracellular environment and other cellular networks.

Microbial organisms have long had systems for genetic transformation, but a quick and inexpensive system for general perturbation of metazoans has only recently been discovered. First reported in *C. elegans*, cytoplasmic double-stranded RNA induces a pathway that digests messenger RNAs complementary to the double-stranded RNA. This response is called RNAi, and it has quickly gained popularity as a method performing gene knockdowns in many eukaryotic systems. Since its initial discovery, it has been found that regulation via microRNAs shares many of the same components as the RNAi pathways [73]. The two primary entities involved in the process are called Dicer and RISC [59]. If the double-stranded RNA is long, then the Dicer enzyme cuts the dsRNA into 21-25 nucleotide double-stranded fragments.

Any such small 21-25 nt dsRNA, referred to as small interfering RNA (siRNA), is incorporated into the RNA-induced silencing complex (RISC). The siRNA strands are unwound to single strands, and the RISC complex is remodeled in an ATP-dependent process, which results in an activated RISC capable of recognizing and degrading complementary RNA. This process also works with synthetic siRNA, so by inserting the siRNA into the cell, nearly any transcript can be targeted with specificity.

2.2 High-throughput gene expression measurement

In recent years, gene expression profiling has emerged as a powerful tool for quickly and easily assaying thousands of phenotypes in a cell. These phenotypes are particularly valuable, as they elementally correspond to a particular gene sequence, and also correlate well with a protein. Therefore, a gene-expression phenotype can be targeted directly by the methods above, and the semantics of the expression phenotype correlate with the activity of the protein that is encoded by the gene. Gene expression profiling also provides a wider scope on the cellular network than any other single technique available, in that they capture an essential stage for a very large proportion of the entire cellular network. Whole-genome gene expression microarrays also offer a relatively unbiased way to search for activity, and next-generation RNA sequencing methods promise the ability to assess RNA quantities even for sequences that are unknown *a priori*.

First used to report on 45 gene expression profiles in *Arabidopsis thaliana* [103], DNA microarrays are somewhat analogous to a large-scale Southern or Northern blot. Gene expression microarrays aim to quantify the amounts of many different types of RNA species in a cell. This is done by using the tendency of complementary nucleotide sequences to base-pair, or hybridize. An individual run of a microarray is often called a hybridization.

For each RNA assayed by the microarray, there are one or more single stranded DNA probes of length 25-1000 nt which complement that RNA. Most probes are designed to match only one sequence of RNA or a single gene. When a probe complements more than one sequence (perhaps with a few mismatches), that probe will suffer from cross-hybridization, which complicates the interpretation of that probe's signal. Additional probe issues originate from variation in melting temperature, density of probe DNA, and homogeneity of the probes. Various wet-lab and mathematical techniques have been developed to normalize different probes to each other to allow comparison of responses between probes.

There are many methods for manufacturing microarrays, but the types can be grouped into two broad categories: cDNA and oligo arrays. In cDNA arrays the probes are made from reverse-transcribing full length mRNAs, and can be of varying length and quality. Second, Oligo arrays use short, synthetic DNA probes, with uniform lengths between 25-70 nt. Though they are a consistent size, oligo probes can vary in their binding affinity due to differences in GC-content or because of secondary structure effects.

2.3 High-throughput network structural measurement

At the most detailed scope of the cellular network, links are defined by physical interactions. Ultimately we hope to find explanations of biological processes that identify all the components, and specifically notice how they interact. Lab techniques used for detecting protein-protein and protein-DNA interactions are now being scaled to the degree that thousands or tens of thousands of such assays can be performed in a single study. This permits investigation of the structure of the cellular network without bias towards previously investigated genes, and provides a dataset which is useful to investigations that later use any of the assayed genes.

2.3.1 Protein-Protein Interaction

The network of protein-protein interactions determines much of the skeleton of allowed protein-kinase pathways. Nearly all known cellular processes, from transcription and translation to signal transduction, depend on the binding of proteins to each other in a highly specific manner. These protein-protein interactions can refer to transient binding, more lasting binding in a complex, and the "self" binding of homo-multimers.

Ascertaining protein-protein interactions is complicated by the context specificity of protein interaction. Many protein-protein interactions only make sense in a specific context. For example, many mitogen-activated protein kinase signaling pathways are highly localized, and this localization is essential the signal specificity [20]. Some protein-protein interactions must be mediated by chaperons. The PDZ family of domains [58], present in both prokaryotes and eukaryotes, including approximately 350 human proteins [76], bind specific peptide sequences to assist in the assembly of complexes and general protein targeting. Presently both *in vitro* and *in vivo* methods suffer this specificity problem. Using a localization database in conjunction with protein-protein interaction can help resolve this context specificity.

The two-hybrid system is a technique for reporting when two proteins bind well enough to activate a transcript. It requires fusing a domain to each queried protein. The TAP-Mass spectrometry system requires fusing a sequence to each queried protein. Most high-throughput studies have been performed in yeast, which has wellestablished and favorable genetic systems and culturing conditions for performing large-scale experiments.

2.3.1.1 Two-hybrid systems

Initially reported in 1989 by Fields and Song [30] in *S. cerevisiae*, the two-hybrid system is inspired by the activity of the GAL4 gene. GAL4 contains two domains, one of which binds upstream of its genomic location, and own which activates transcription of itself. Both domains are required: if the activation domain is not localized to the promoter the activator will not work and if the binding domain is not attached to the activating domain, GAL4 is not activated. To test if two proteins X and Y interact, two hybrid proteins are created. The first hybrid protein is the fusion of the binding domain of GAL4 and X, the second hybrid protein is the fusion of Y and the GAL4 activating domain. These hybrids are introduced into strains without GAL4 and with a β-galactosidase reporter gene downstream of the sequence bound by GAL4. Significant amounts of β-galactosidase activity then indicate binding of proteins X and Y.

By 1994, the two-hybrid system was in widespread use [31], using a variety of reporter, binding, and activating constructs. The system has also been expanded to DNA-protein, RNA-protein, and protein-small molecule interaction [120]. The first genome-wide study of protein linkage was conducted in 1996 in *E. coli* bacteriophage T7 [7].

Two independent, large-scale investigations in *S. cerevisiae* used the two-hybrid system to determine interactions. Both studies created libraries of strains for both binding and activating hybrids, and then crossed all strains to create an array of double-hybrids. Uetz *et al.* [117] compared two different methods of detecting positives in a 192 by 6000 (binding and activating hybrids, respectively) screen, one with higher accuracy and one with greater throughput. Ito *et al.* [63] completed a thorough scan (3,278 proteins in interactions) subsequently, with somewhat non-overlapping results.

Due to the necessary gene fusion steps, two-hybrid systems inherently have high false negative rates [65]. The fused binding or activating domain has the potential of obstructing both normal protein folding and the interacting sites of the proteins. The Uetz *et al.* higher accuracy screen found an interaction for a binding hybrid only 45% of the time.

2.3.1.2 TAP-Mass spectrometry

Increasingly precise mass spectrometric methods now allow the identification of proteins and protein complexes from cell extracts. A single species, isolated *e.g.* by gel electrophoresis or centrifugation, may be digested by trypsin, resulting in small fragments whose amino acid constituents can be identified via mass spectroscopy. Searching against a database of potential peptide sequences can quite often identify unique proteins that match the observed weights.

In order to improve isolation of single species of compounds, studies often use a single "bait" protein, which has been modified with an easily immunoprecipitated tag [18]. The FLAG tag is particularly popular due its ability to precipitate without denaturing complexes [28]. Interactions found this way may not necessarily be direct since a whole compound may be pulled down and not all members may share full contact.

Again in *S. cerevisiae*, two genome wide screens have been performed to find binding ability [52, 45]. Rates of positive interactions were much higher than in the yeast two-hybrid experiments, with 78%–82% of the baits finding partners, compared to a best case of 45% for yeast two-hybrid. Positives were also much more repeatable than in the two-hybrid studies, approximately 75% vs. 20%.

2.3.1.3 Databases

There are several databases of protein-protein interaction with both literature-curated interactions, high-throughput interactions, and computationally predicted studies. Mathivanan *et al.* [86] compare the human-specific parts of seven such databases (BIND, DIP, IntAct, Mint, MIPS, PDZBase, and Reactome) to their own database of protein-protein interactions [97]. Several of these databases include additional information beyond protein-protein interaction.

2.3.2 Protein-DNA Interaction

Chromatin immunoprecipitation paired with DNA microarray analysis (ChIP-chip)[53] or DNA sequencing (ChIP-seq)[61] promises to give high-throughput results of protein-DNA interaction. Previous wet-lab methods for determining protein-DNA interaction included DNase footprinting, primer extension, and gel shift assays, and were generally limited to a very small number of queries. Alternatively, a protein's binding site could characterized computationally (often by a weight matrix), and then genomic binding sites could be predicted, usually with a high false positive rate.

ChIP-chip and ChIP-seq experiments first cross link transcription factors to bound DNA with formaldehyde *in vivo*. DNA-protein complexes are extracted and cut randomly via sonication or other method. The protein, and any cross-linked DNA, is selected via immunoprecipitation of the target factors or epitope tags. Finally, the DNA is amplified via PCR, and the sequences that were bound are queried with either microarrays or with DNA sequencing. Both methods produce similar results among their top-ranking predictions [29], but require appropriate controls for identifying entirely novel binding motifs [66].

Large-scale ChIP-chip studies have been published in human [137] and *S. cerevisiae* [75]. In *C. elegans*, protein-DNA interactions have been investigated using a yeast one-hybrid system [26].

Chapter 3

Probabilistic and Other Graphical Models

Computational tools for dealing with networks of variables under probabilistic constraints have been developed and widely used in statistical physics, machine learning, computer vision, coding theory, and bioinformatics. These computational tools are now being applied in biological networks, both for modeling and discovery. This chapter contains the relevant computational background for my work in Part II. I discuss previously and commonly used computational graphical models, algorithms for inferring values from a given model, algorithms for learning a model from data, and the implications of causality and perturbation in these models.

3.1 Graph formulations

Traditionally, probabilistic graphical models are presented in two categories: undirected and directed models. In both of these types of models, each node of the graph is a random variable. The edges in both types of models encode the probabilistic dependencies between random variables, though both models encode the dependencies in different ways. Decoding the probabilistic dependencies requires examination of the local structure around a variable.

These two categories, also known by the names Markov random fields and Bayesian networks, are capable of representing different probability spaces, but share some overlap. Sometimes, a third type of graphical modeling containing both directed and undirected edges is presented to generalize the two and unify the set of representable probability spaces. However, a different formulation of this generalization, called factor graphs, has seen growing popularity. I find the factor graph representation far preferable to the mixed directed/undirected model formulation, and in many cases preferable to both Markov random fields and Bayesian networks due to the explicit representation of the characteristic function of the network.

Before proceeding with the definitions, I will define some notation conventions. First, capital letters such as A or X refer to variables over a domain. In general, the domain of a variable can be any set, countable or uncountable, but in my proposal I use only finite sets or the real numbers, $\Re$.

A *factor* is a function whose domain is a set of variables and whose range is the real numbers. Many functions can be interpreted as factors, so their notation varies. For example, $\Pr(A, B)$, f_{AB}, or $\Pr(A|B)$ could all be considered factors in the rest of this proposal. Factors are sometimes referred to as *potentials.* Both joint and conditional probability distributions are quintessential examples of factors, though factors need not be restricted in the ways that probability distributions are defined. For example, the identity $F = MA$ over three variables with a real-valued domain could be defined as a factor N with f, m, and a all real numbers:

$$N(f, m, a) = \begin{cases} 0 & \text{if } f = ma \\ -\infty & \text{otherwise} \end{cases}$$

As is often done with probability distributions, I will use abbreviated notation for factors where an argument in parentheses to a factor, such as X and Y in $f(X, Y)$, is a set and a component of the domain of the factor. The total domain of this factor is the cross product of all the sets that were used as arguments to the factor. Thus the expression $f(X, Y)$ simultaneously names a factor and defines the domain of the factor.

There are two primary factor operations that are used in graphical models: factor product and factor marginalization. The product of two factors $f(X)$ and $g(X, Y)$, denoted $f(X)g(X, Y)$ or fg, is another factor. The resulting factor's domain is the cross product of the union of the domains of each operand, $X \times Y$ in this example. The value for each element in the result is the value of operand evaluated at that point, *i.e.* in this example $(x, y) \mapsto f(x)g(x, y)$ for all $x \in X$ and $y \in Y$. The other operation used on factors is marginalization, sometimes called summarization. Marginalization results in a factor with one less variable in the domain. There are two variants of marginalization which are commonly used; one variant uses addition and the other the max function. If X in the above examples is discrete, then marginalization of X out of g is denoted and defined as:

$$\sum_X g(X,Y) \equiv y \mapsto \sum_{x \in X} g(x,y) \text{ for all } y \in Y$$

If X is a continuous variable, then

$$\int_X g(X,Y) \equiv y \mapsto \int_{x \in X} g(x,y) \text{ for all } y \in Y$$

It is possible to marginalize a factor down to zero variables, in which case the result is a factor with no variables and a single real number in the range. This full marginalization has sensible interpretations in some contexts: marginalizing probabilistic factors by addition results in a probability mass, marginalization by max of probabilistic factors results in the most probable assignment, and marginalization of an arbitrary factor results in the partition function.

3.1.1 Markov random fields

Markov random fields were originally developed in statistical physics to describe systems of small particles, where the state of one particle interacts with the state of nearby particles. Such interactions are represented by lines between variables. A

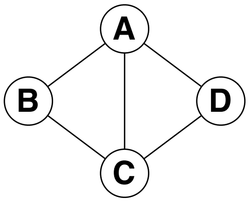


Figure 3.1: Example Markov network.

Markov network describes a probability distribution over its variables. For every maximal clique C in the graph, a factor ϕ_C describes the interactions of those variables. If the set of all such factors is denoted Φ, then the probability distribution over all variables $\bar{X}$ is defined to be:

$$\Pr\left(\bar{X}\right) \equiv \frac{1}{Z} \prod_{\phi_c \in \Phi} \phi_c \tag{3.1}$$

For the example in Fig. 3.1, $\Pr\left(A, B, C, D\right) = \frac{1}{Z}\phi_{ABC}(A, B, C)\phi_{ACD}(A, C, D)$. The constant Z is known as the partition function. It serves to normalize the function to a proper probability function and can be calculated by $Z = \sum_{\bar{x}} \prod_{\phi_c \in \Phi} \phi_c$. Though the expression is simple, such a calculation is not usually trivial, and is in effect similar to calculating the probability of the data in a Bayesian statistics model. Note that there is flexibility in this parameterization, and that many different pairs of ϕ_{ABC} or ϕ_{ACD} will result in identical probability distributions with the same normalizing constant Z. This is because information about the joint dependence between A and C, $\Pr\left(A, C\right)$ can be "shifted" between ϕ_{ABC} and ϕ_{ACD}. For any factorization $\phi_{ABC} = \phi'_{ABC}\phi'_{AC}$, where ϕ'_{AC} does not have zero elements, let $\phi'_{ACD} = \phi_{ACD}/\phi'_{AC}$ where factor division is defined similarly to factor multiplication. Then $\Phi = \{\phi_{ABC}, \phi_{ACD}\}$ results in the same probability distribution and normalization constant as $\Phi' = \{\phi'_{ABC}, \phi'_{ACD}\}$.

3.1.2 Bayesian networks

Bayesian networks are a representation of a probability distribution based on the conditional probabilities. Conditional probabilities offer the advantage of often being able to characterize a known real-world system, and in conjunction with a Bayesian network the meaning of each conditional probability has a fairly clear interpretation. This is an advantage over clique potentials in Markov random fields, which may have an unclear meaning. However, specifying a Bayesian network from a probability distribution is done through the conditional independences, which can be difficult to assess.

In a Bayesian network, there is a conditional probability distribution for each node. The directed graph structure is determined by these conditional probability tables: there is an arrow into each node from every variable on which it is conditioned. Additionally, this directed graph must be acyclic. Let $\bar{X}$ be the set of variables in the network, and $\text{Parents}\left(X\right)$ for $X \in \bar{X}$ be the set of variables on which X is conditioned.

$$\Pr\left(\bar{X}\right) = \prod_{X \in \bar{X}} \Pr\left(X \middle| \text{Parents}\left(X\right)\right) \tag{3.2}$$

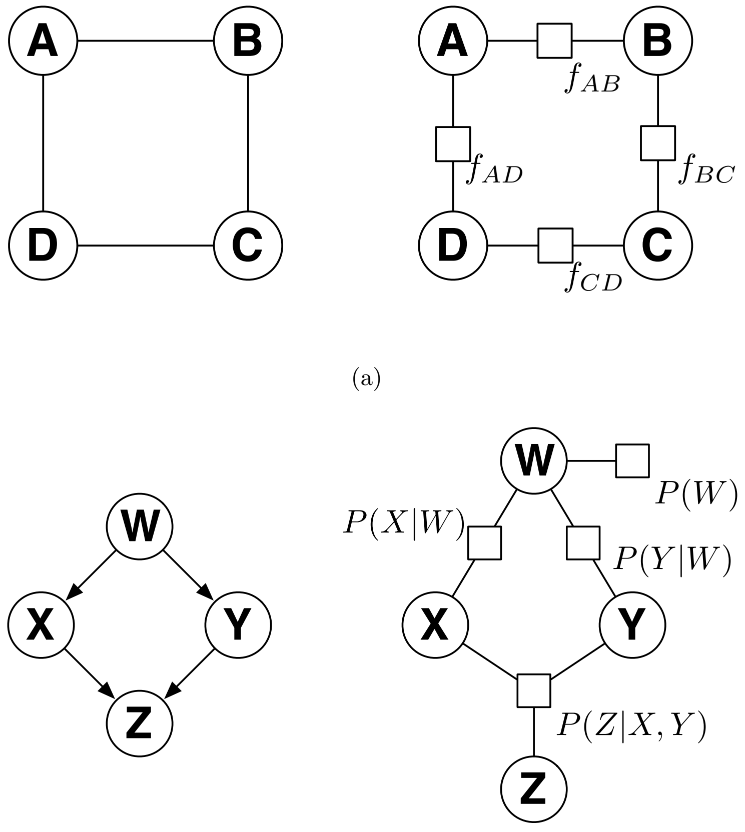
There is great flexibility in the parameterization here, as in the Markov Random Field case, since any probability distribution can be expanded into conditional probabilities in any order. There are many cases where the probability distribution can be conditioned in a different order, but still encode precisely the same conditional independences, resulting in a network with flipped arrows but the same probability distribution as the original Bayesian network. For this reason Bayesian networks can be slightly deceptive, as the directionality is sometimes assumed to encode causality, but this may not be the case.

Many graph-based bioinformatics problems are easily formulated as Bayesian Networks of a certain form. Algorithms on phylogenetic trees, for example, are special cases of the algorithms used on Bayesian networks. Hidden Markov models can be represented as a chain with one variable for each hidden state and one variable for each observed symbol.

3.1.3 Factor graphs

Both Markov random fields and Bayesian networks are mathematically specified by an objective function, namely their probability distribution. Factor graphs are a representation of any such objective function over a set of variables, and thus generalize both Markov random fields and Bayesian networks in that sense.

Figure 3.2 shows the factor graph representations of both a Markov random field and a Bayesian network. Factor graphs represent both the variables as nodes and the factors as nodes, with edges from each factor to the variables in that factor's domain, resulting in a bipartite graph. A factor graph is then a very general representation of constraints on variables, and has even been used to represent problems such as n-SAT [89] and fast Fourier transforms [1, 2].



(a)

(b)

Figure 3.2: A Markov random field and a Bayesian network next to their corresponding factor graphs. (a) A Markov random field (left) and the corresponding factor graph representation (right). The decomposition of the joint probability, $\Pr(A, B, C, D) = \frac{1}{Z} f_{AB} f_{BC} f_{CD} f_{AD}$, corresponds directly to the factors in the factor graph. (b) A Bayesian network (left) and the corresponding factor graph (right). As with the Markov random field in (a), the decomposition of the joint probability, $\Pr(W, X, Y, Z) = \Pr(Z|X, Y) \Pr(X|W) \Pr(Y|W) \Pr(W)$ is made explicit in the factor graph. Though the Markov random field and the Bayesian network share the same edges when directionality is discarded, their probabilistic formulations are quite different. The factor graph representation of each structure makes this difference explicit visually.

3.2 Inference methods

Inference in graphical models aims to find out something about the distribution of variables. For example, common goals are to calculate the maximally likely assignment of values to each variable or the distribution of some set of variables given observations of the value of some disjoint set of variables. In a probabilistic setting, any conditional query reduces to two probability functions:

$$\Pr\left(\bar{X}|\bar{Y}\right) = \frac{\Pr\left(\bar{X},\bar{Y}\right)}{\Pr\left(\bar{Y}\right)} \tag{3.3}$$

Such computations are in general bounded by the size of the largest domain during execution, since the size of a discrete domain factor grows exponentially with the number of variables in the domain. Therefore the main aim of inference algorithms is to minimize the largest such factor that is created during the calculation of the final answer.

Inference algorithms for Markov random fields, Bayesian networks, and factor graphs are common to all representations [74]. In general exact inference on a graphical model is an NP-hard problem. However, inference can be performed in linear time if the factor graph representation of a graphical model has no cycles, and has inspired the highly successful method of approximation via message passing on graphical models with cycles. I have used both exact inference and approximation via message passing in this thesis.

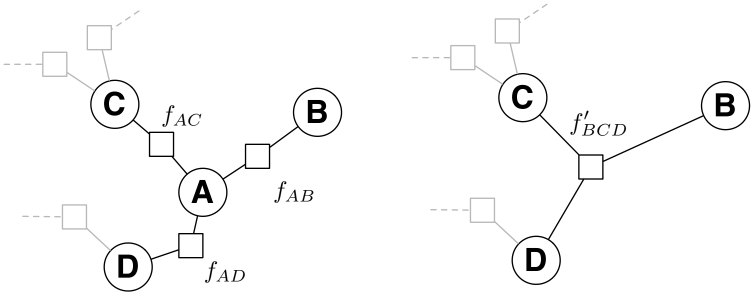


Figure 3.3: Variable elimination step for A. The new factor f'_{BCD} is equal to $\sum_A f_{AC} f_{AD} f_{AB}$.

3.2.1 Exact Inference

There are two basic methods for exact inference in graphical models: variable elimination and message passing on acyclic structures. Both of these algorithms take as input an ordering over all the variable nodes in the graph, and the largest factor used in algorithm depends on both the ordering and the structure of the graphical model. The general solution to finding the most efficient such ordering is NP-hard, but there are certain cases where an ordering is known to be optimally efficient. For example, in factor graph trees, any order that respects the connectivity of the tree is optimal.

Variable elimination transforms the factor graph, removing one variable per step, until only the variables of interest are left. The elimination of a variable node removes that variable node and all adjacent factor nodes, replacing them with a factor node. If the neighbors of variable X are the set of factors F, then the factor f' that results from eliminating X is

$$f' = \sum_{X} \prod_{f \in F} f \tag{3.4}$$

Figure 3.3 shows the graph transformation under elimination of A. The memory and time costs of a variable elimination step depend on the maximum size of product in Equation 3.4. Heuristic variable elimination algorithms try to choose an ordering of variable elimination that results in the smallest such product.

In the special case of a tree-shaped factor graph, it can be shown that the most efficient ordering always takes a variable node with the fewest number of neighbors. Using transformations of the graph, namely by combining two variables into a single variable with a larger domain, any cycle in the graph can be eliminated, resulting in a tree structure. Such a structure is called a *junction tree*.

Variable elimination on a junction tree collapses in a predictable manner, and does not create a factor larger than the largest node in the junction tree. This predictable collapse of the junction tree suggests an algorithm where messages are passed between all nodes, each message a factor itself which represents the current local belief in the setting of a variable or a set of variable. If, during variable elimination no nodes are removed from the graph, but instead the newly created factors are stored as "messages", the results can be remembered and reused such that all variables are solved for in turn. I will discuss this algorithm in greater detail in the next section, where the message passing occurs on a cyclic structure, and the algorithm generates an approximate inference of variable posteriors.

3.2.2 Approximation with message passing

The message passing algorithm, also known as belief propagation or affinity propagation, has been invented many times, at least twice in the coding community with low-density parity check codes [40] and turbo codes [11], and once in Bayesian network community [96]. It has gained great popularity, as the message passing has been found to approximate exact results with very little computation in difficult problems. Approximate message passing has seen a surge of recent interest, with the creation of generalized message passing algorithms [133, 132] and surprisingly successful applications on problems such as clustering [35].

There are two types of messages passed in the algorithm. Variable nodes pass messages to neighboring factors, where each message is a valuation over the possible states of the variable. Similarly, factor nodes pass a message to each neighbor variable node v summarizing that factor's "belief" that v is in each possible state. Every message is a valuation over the settings of a single variable, a normalized or unnormalized probability distribution, and is itself a factor. Let $m_{n_1 \to n_2}$ denote the message from node n_1 to node n_2, and $\mathrm{Neighbors}(n)$ the set of all nodes adjacent to n. Then the message sent from a variable node v to a factor node f, where $f \in \mathrm{Neighbors}(v)$ is simply:

$$m_{v \to f} = \prod_{f' \in \mathrm{Neighbors}(v) \setminus \{f\}} m_{f' \to v} \tag{3.5}$$

All the incoming messages are factors over just the variable v, so therefore all the outgoing messages have the same domain. The message from a factor node f to a variable node v is calculated by:

$$m_{f \to v} = \sum_{\text{Neighbors}(f) \setminus v} f \cdot \prod_{v' \in \text{Neighbors}(f) \setminus v} m_{v' \to f} \tag{3.6}$$

Note that the product of the factor and the incoming messages is marginalized by every variable except for v, and there the message is a factor over just the domain of v. In this scheme messages are passed iteratively. At any given moment, the belief in a variable X is approximated by multiplying all the incoming messages.

Scheduling of message passing is an area of active research, with few general results. In the case of a tree factor graph, waiting until all incoming messages are ready, and passing each message at most once results in the exact result. When there are cycles in the graph, passing messages according to a schedule that causes data to be counted more than once can lead to poor approximations.

Generally, message passing is performed until the messages "converge," usually detected by measuring successive changes in the messages. Some difficulties can be encountered when there are long-range correlations, meaning when the value of one variable is highly correlated to the value of a variable a large number of nodes away. Additionally, it is known that message passing can be prevented from detecting convergence when there is a multi-modal cycle to the pattern of the messages. Message damping by averaging successive messages can help prevent such oscillations.

3.3 Structure learning

Though an important aspect of graphical models, there are few general methods for structure learning as it is a hard problem. One challenge is computational, in that number of possible structures is super-exponential in the number of variables. Further, identification of conditional independence between variables can require very large amounts of data, and statistical power to detect dependence and independence depends on the particular probability distribution being estimated. Therefore, most structure learning is highly dependent upon the nature of the data and problem space. Typical applications in biology for modeling gene expression have used between five hundred and thousands of joint samples of variables in the network. In the next chapter, Chapter 4, I describe some previous methods for structure search for finding biological networks.

3.4 Causality in graphical models

The arrows in Bayesian networks are highly suggestive of causal influences, particularly when combined with the common introductory Bayesian network examples. However, a Bayesian network need not represent any causal structure at all, and for any Bayesian network which coincidentally does represent a causal structure, there are many other Bayesian networks which represent the exact same probability space but have entirely different structures. Therefore, Bayesian networks are not causal networks, and learning a Bayesian network on a dataset will not learn causality on that dataset.

The notion of causality has often been avoided by statisticians, as there are many philosophical pitfalls to avoid in addition to a general lack of theory. However, recent years have seen some theoretical work on rigorously defining causality in probabilistic and modeling terms. In particular, a framework of causality has been developed with directed and undirected graphical models [95].

As it concerns this dissertation, causation provides a framework for predicting effects under the perturbation, or intervention, of a variable. In particular, perturbing a variable disconnects that variable from its causes, while leaving the effects of that variable intact. Thus, a full causal model predicts not only a standard distribution over the variables, but also predicts the distribution of the variables under all perturbations.

The methods in this dissertation use causal models in this form. These methods use graphical models that specify how biological elements interact together, but also predict how they behave under perturbation. In this sense, these methods are causal, and links in the network represent cause and effect. These causes may be from direct physical interaction, or they may represent the transitive chain of several direct physical interactions.

Chapter 4

Network Prediction Methods

Computational construction of biological networks from high-throughput data is an active area of research. Gene expression data, protein-protein interaction data, and protein-DNA interaction data have all been used in various combinations to predict networks. In recent years, availability of gene-expression under knockdown has increased the number of methods dealing with such data. In this chapter, I will first review the methods used to infer the connectivity of the *lac* operon in *E. coli.* I will then describe the methods used for modeling biological networks, followed by some methods used to predict pathways from data *de novo*.

4.1 *De novo* methods and network refinement methods

Being able to accurately model a biological network is a necessary step towards being able to learn them *de novo.* Gat-Viks *et al.* [44, 42, 43] sought to use interaction and regulatory links found in literature to create a network. This network modeled protein interactions, gene regulation, mRNA quantities, and protein quantities. Accurately modeling a network also requires being able to model any cycles that may appear, so they used a factor graph formulation of the regulatory network. Combined with expression data, they were able to identify regions where the regulatory network was poorly characterized, and then refine it. Gat-Viks *et al.* used several *S. cerevisiae* systems for biological verification, including the osmotic stress response network.

The growing abundance of protein-protein interaction and protein-DNA interaction data, described in §2.3.1 and §2.3.2, is particularly amenable to refinement due to uncertainty of the accuracy of these new methods. In addition, working on the scaffolding of a protein-protein and protein-DNA net permit the exploration of direct causal links in a network. Yeang *et al.* [130] sought to explain knockout gene expression data in conjunction with protein-protein and protein-DNA interactions. Their inference algorithm assigned direction to protein-protein edges, and a sign (activation or inhibition) to protein-protein and protein-DNA edges. Yeang *et al.*, explained the *S. cerevisiae* pheromone response network, among others.

4.2 Computational methods

Building upon their experience with Bayesian networks in other contexts, Friedman *et al.* [37] were the first to build Bayesian networks from gene expression data to explain a pathway. Their methods are somewhat different from the efforts of their successors in that they built their networks entirely from observational gene-expression microarrays. Additionally, they were able to learn causal Bayesian networks by searching over equivalency classes of partially directed acyclic graphs. Friedman *et al.*, were able to predict the *S. cerevisiae* cell cycle network over the Spellman [109] data set, which consists of an unusually large number of microarrays over a time-course on a synchronized culture. Observational studies with such large numbers of microarrays are very rare, limiting the application of this technique.

4.2.1 Protein-protein based methods

Interpretation of high-throughput protein-protein interaction data is often confounded by two factors: the inaccuracy of the assay, and false positives from proteins that bind but are never expressed such that they can co-localize. Several separate groups have been able to learn accurate networks from noisy protein-protein interaction data by combining it with co-expression data, with increasingly computationally-efficient methods [111, 106, 56]. This methods relies on searching the protein-protein interaction network for regions with high co-expression. It has recently been extended to include not only protein-protein interaction data, but also protein-DNA interaction data and the results from knockdown experiments, allowing the inference of some causal relationships [93].

4.2.2 Perturbed gene-expression methods

Markowetz *et al.* [83, 84] developed the Nested Effects Model (NEM) for predicting networks in a common and practical experimental setup: gene expression profiling under perturbation of genes known to be involved in the phenotype of interest. Their method starts with genes known to be related due to a common loss-of-function phenotype when each gene is knocked down. Next, microarrays are used to profile expression of the genome under each perturbation. Finally, they build a model based on the nesting of effects under each knockdown.

Since NEM predictions are based on secondary effects, these models can predict networks over biological entities which are difficult to directly assay. For example, Markowetz *et al.* predict a signaling network in *D. melanogaster* which involves no change in gene expression among the signaling genes and is therefore invisible to current high-throughput techniques.

I believe their data and model setup to be the most appropriate for the prediction of signaling networks. I have therefore based my methods on the same fundamental idea: learning networks among perturbed genes from downstream effects. In the following chapter, I present such a method based on Bayesian networks. In subsequent chapters, I present extensions of the Nested Effects Model both in terms of the biological model and an inference method, along with new results.

Part II

Methods and Results

Chapter 5

Joint Intervention Networks

If multiple genes contribute to the same phenotype, it is often the case that these genes interact. With quantifiable phenotypes, measurements of the phenotype under both single- and multiple-gene perturbations can be collected. When genes contribute to the phenotype completely independently, we can construct an estimate of the phenotype under the multiple-gene perturbation by using an arithmetic combination of the single-gene phenotypes, typically addition or multiplication. When the multiple-gene phenotype does not match the value predicted from single-gene perturbation phenotypes, then there is an *epistatic* interaction between the genes in the double-gene perturbation.

The concept of *epistasis* has been in the genetics literature for a century [8], and has allowed geneticists to predict interactions between genes without requiring the ability to measure the activity of the genes themselves. Predicting an epistatic interaction requires the ability to perturb genes and measure a downstream phenotype.

This chapter presents an extension of epistasis analysis that allows the inference of complex interaction between more than two genes from high-throughput data.

An overview of the experimental setup for an application of this methodology is shown in Figure 5.2. Data is collected from competitive expression measurements contrasting two different genetic strains. In §5.2 through §5.4 I show how to model epistasis analysis on a complex regulatory network with a Bayesian network. In §5.5 I describe how to search for a regulatory network that best describes the data. Finally, in §5.6 and §5.7 I describe an application of JIN to a *V. cholerae* biofilm network.

5.1 Relationship to published work

The work in this chapter has been published previously in the proceedings of the 2009 Pacific Symposium on Biocomputing [70]. The computational model was developed jointly with Dr. Chen-Hsiang Yeang and Dr. Joshua Stuart. I developed and conceived the computational model as a single Bayesian network under the advisorship of by Dr. Stuart and me. I implemented the general Bayesian network library that performs inference on a Bayesian network. Mrs. Pinal Kanabar and I jointly planned the implementation of a conditional probability table (CPT) that follows the model in §5.3, and Mrs. Kanabar coded it to our specification. Similarly, Mrs. Kanabar and I jointly planned the implementation of network scoring and network search, and Mrs. Kanabar coded the implementation. Mrs. Kanabar also ran the implementation on the *V. cholerae* biofilm data provided by Dr. Fitnat Yildiz's lab. Mrs. Kanabar and

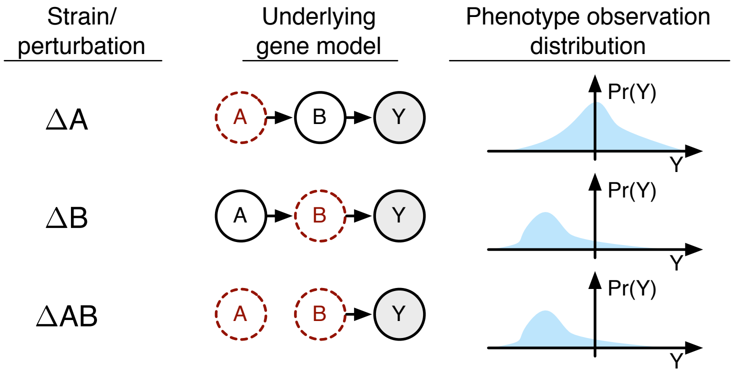


Figure 5.1: Switch regulatory epistasis. Three perturbations of a cellular network are presented. Each cellular network consists of three nodes, A, B, and Y. Hidden gene states are shown by nodes with solid borders, perturbed gene states are denote with dashed, red borders, and observed gene states are shaded gray. Switch regulatory epistasis analysis assumes that the observed phenotype, Y, is downstream of both of the perturbed genes, A and B.

Dr. Yildiz analyzed the network and expansion for biological relevance.

5.2 Modeling epistasis with Bayesian Networks

The Joint Intervention Network automates switch regulatory epistasis analysis with complex gene regulation. In general, the term *epistasis* refers to any interaction between gene perturbations that is non-additive. Using definitions from Huang and Sternberg (2006) [54], *switch regulatory* epistasis analysis assumes the measured effect is downstream of the perturbed genes, whereas *substrate dependent* epistasis analysis assumes the observed effect is on the path between the two perturbed genes. Switch regulatory epistasis analysis is generally used in gene regulatory models, while substrate dependent epistasis analysis is often used for metabolic pathways. Recently, Van Driessche *et al.* [118] used switch regulatory analysis to manually build a regulation pathway by comparing gene expression data under single and double knockouts.

Figure 5.1 illustrates the logic of epistasis analysis. Node Y represents an observable phenotype downstream of the genes A and B. In the top network, A is deleted; in the middle network, B is deleted; in the bottom network, both A and B are deleted. Under the deletions of ΔB and ΔAB, observations of the phenotype Y will be similar, since B has identical states in both molecular networks. However, under ΔA Y may have a different value as B is less constrained.

This epistatic reasoning matches the results of a Bayesian network modeling the same perturbations. In this Bayesian network representation, let genes A and B be unobserved, hidden variables. Let the phenotype, Y, be an observed variable. Interactions between genes are represented by conditional probability tables (CPTs). Gene deletions in a strain are modeled as causal perturbations of the gene's variables in the Bayesian network, both setting the value of the perturbed gene and removing the influence from parent genes. The phenotype observation distribution in Figure 5.1 shows hypothetical probability distributions for one possible setting of CPTs, where ΔB and ΔAB perturbations have identical distributions for $\Pr(Y)$, and the ΔA network has a different distribution for $\Pr(Y)$. By the conditional independence assumptions of this network structure, the identity of phenotype distributions under ΔB and ΔAB will hold for any possible CPTs in addition to those in the example. Only in degenerate cases, i.e. deterministic CPTs, will ΔA have the same phenotype distribution as ΔB or ΔAB. Thus, Bayesian networks provide a natural generalization of epistatic reasoning.

The Joint Intervention Network is a method for evaluating competitive gene

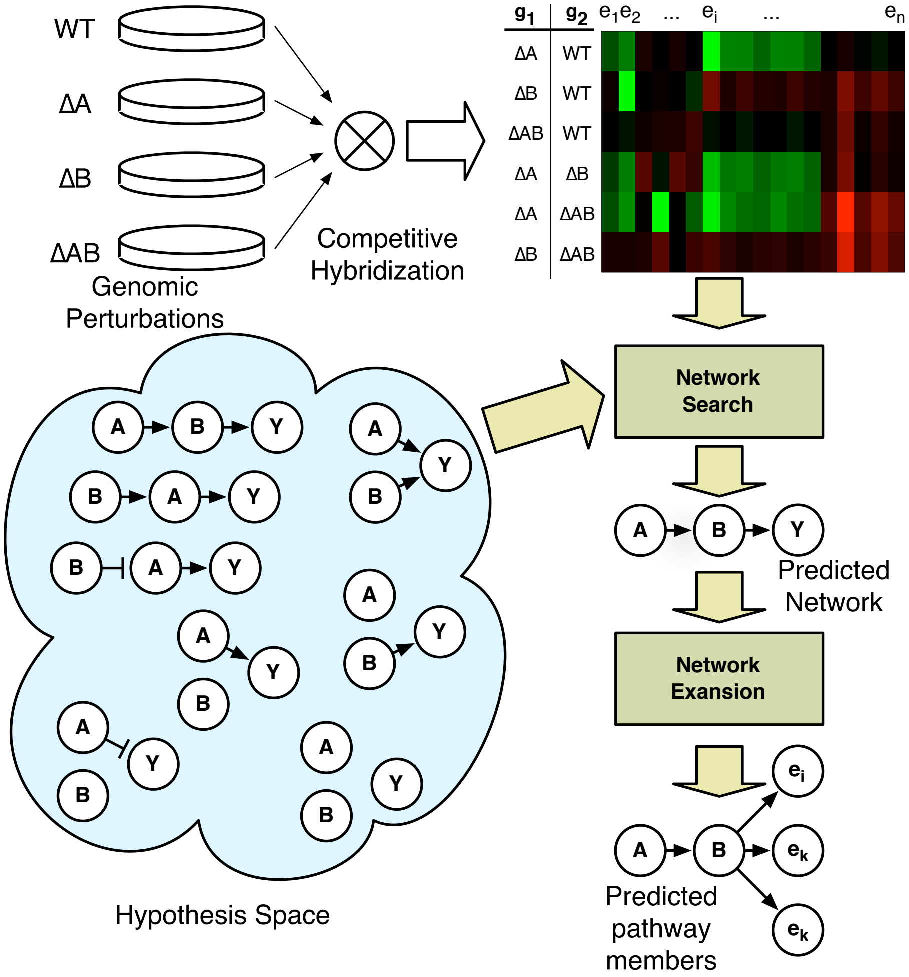


Figure 5.2: Overview of the experimental setup for a Joint Intervention Network application. For one biological system, many perturbations are gathered (upper left). Competitive expression hybridizations are performed for all combinations, resulting in all possible comparisons (upper right). The hypothesis space (lower left) consists of all possible regulatory networks among the perturbed genes. Network search finds which regulatory model best fits the data, resulting in a single predicted regulatory network. This network is then used to find other genes that may be under the same regulation model, resulting in a new network with additional genes taken from the original data matrix.

expression data with epistatic reasoning on a complex regulatory model. Given a regulatory model M, a data matrix D, and a genotype comparison matrix G, the Joint Intervention Network $JIN(M, D, G)$ is a Bayesian network for evaluating the likelihood of competitive expression observations. Figure 5.2 shows an example data matrix and genotype matrix along with many potential regulatory models from the hypothesis space. The regulatory model M is a boolean Bayesian network over some unobservable and perturbed genes, and a replicated phenotype variable labeled Y. Except the phenotype node Y, every node in M is a *regulatory node*. A regulatory node is a binary variable that describes the hidden activity state of the corresponding regulatory gene. Interactions between genes are modeled with CPTs between nodes in the Bayesian network. In the following section I describe the constraints on CPTs used aid in biological relevance and interpretability. The $m \times n$ matrix D consists of competitive observations between perturbed strains, and is a discretized log ratio of two genotypes, where there are m comparisons and n different observations under each comparison. The genotype matrix G, has dimension $m \times 2$, and identifies the strains used for the competitive observations in D. The ith row in D is the log-ratio gene expression in genotype G_{i1} over genotype G_{i2}.

The hidden activity state of every regulatory gene in every strain is modeled in $JIN(M, D, G)$. For a given genotype deletion strain ΔX, where X is some set of regulatory nodes in M, let $M(X)$ be a copy of the Bayesian network that has been causally perturbed for each gene in the set X. Specifically, for each variable V in M, add a variable V_X to the subnetwork $M(X)$. Next, for each variable V_X in $M(X)$, if

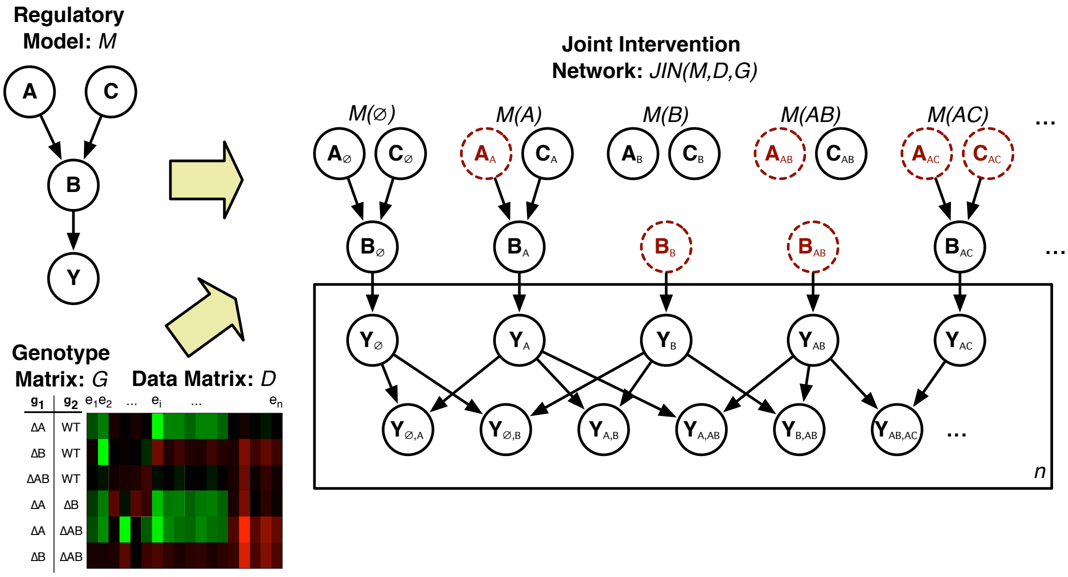


Figure 5.3: Structure of an example Joint Intervention Network. The regulatory model on the left, M, is a Bayesian network. The genotype matrix, G and data matrix, D, contain the conditions and competitive expression observations, respectively. Part of the corresponding Joint Intervention Network, $JIN(M, D, G)$, is shown on the right. For each perturbation genotype, M is copied and then causally perturbed according to the genotype. For example $M(B)$ has a copy of each variable in M, but is causally perturbed for deletion of B. Below the genotype networks are observation nodes that compare the phenotypes of various genotype networks. Boxes in Bayesian networks are plate notation–each variable in a box is replicated the number of times shown in the lower right.

$V \in X$ then clamp Vs state to 0. Otherwise, if $V \notin X$, set the CPT of V_X to be the same as the CPT of V in M.

To construct $JIN(M, D, G)$, we first create variables to model the hidden states of gene activity in each genotype, and then connect them to observation nodes that correspond to elements of D. For every unique genotype g that is in the genotype matrix G, add $M(g)$ to $JIN(M, D, G)$. Next, for every row i in the data matrix D, construct an observation node $Y_{G_{i1},G_{i2}}$ to model the competitive expression observation, and connect this node as the child of $Y_{G_{i1}}$ and $Y_{G_{i2}}$. The competitive observation node has a ternary domain of $\{-1, 0, 1\}$, while the parent nodes $Y_{G_{i1}}$ and $Y_{G_{i2}}$ are boolean with domain $\{0, 1\}$. Construct a CPT for $Y_{G_{i1},G_{i2}}$ such that:

$$\Pr\left(Y_{G_{i1},G_{i2}} \mid Y_{G_{i1}}, Y_{G_{i2}}\right) = \begin{cases} 1 & \text{if } Y_{G_{i1},G_{i2}} = Y_{G_{i1}} - Y_{G_{i2}} \\ 0 & \text{otherwise} \end{cases}$$

Finally, for every column 1 to n in the data matrix D, replicate every Y node and associated CPTs. In Figure 5.3 these replicated variables are shown with plate notation.

5.3 Regulatory Model

In the Joint Intervention Network we assign regulatory roles to each interaction in the model. We allow three types of interaction: repression, additive activation, and multiplicative activation. Each edge in a regulatory model M is of one of these three types of interactions, and each CPT in M is restricted to a parameter space that matches the interaction types on the incoming edges to a node. By M we denote just the structure and regulatory class of each edge in the regulatory model. For a vector of parameters p for all CPTs, let M_p refer to the full Bayesian network with all parameters specified.

The CPT for each node specifies the probability distribution for that node under each variable setting of its parents. For a child node C in M, the parent set $\text{Parents}\,(C)$ can be partitioned into three sets according to the type of interaction on each edge: the set $R(C)$ for all repressive parents, the set $A^+(C)$ for all additive activating parents, and the set $A^{\times}(C)$ for multiplicative parents. This definition allows a regulatory gene to have different interaction types with each of its child genes. Given a setting of the variables $\text{Parents}\,(C)$, we define a function that maps the repression and activation to an expected state for C:

$$f(R(C), A^{+}(C), A^{\times}(C)) = \left(\prod_{r \in R(C)} (1-r)\right)\left(\prod_{a \in A^{\times}(C)} a\right)\left(\max_{a \in A^{+}(C)} a\right)$$

With this helper function, the CPT for C in M is restricted to:

$$\Pr\left(c \mid \text{Parents}\,(C)\right) = \begin{cases} p_{\text{Parents}(C)} & \text{if } c = f(R(C), A^{+}(C), A^{\times}(C)) \\ \\ 1 - p_{\text{Parents}(C)} & \text{otherwise} \end{cases}$$

where $p_{\text{Parents}(C)}$ is restricted to be in the range $(0.5, 1.0]$. Note that $p_{\text{Parents}(C)}$ is unique to each possible setting of $\text{Parents}\,(C)$. For a node N with no parents, the prior is $\Pr\left(N = 1\right) = 0.5$.

5.4 Regulatory Model Score

To score a model M, we evaluate the likelihood of the data D given the full JIN model:

$$Score(M) \quad = \quad \max_{p} \Pr\left(D \mid JIN(M_p, D, G)\right) - \frac{|p|}{2} \ln n$$

where the $|p|$ denotes the number of free parameters for all CPTs. This score corresponds to the Bayesian Information Criterion [105].

Due to time and implementation constraints, this maximization was implemented by Mrs. Kanabar as an approximation rather than as an exact maximization.

The maximum was taken over 200 independent random parameter samples p, rather than the full parameter space.

5.5 Network Search

To infer a regulatory model for our data, we must search through the hypothesis space of regulatory models in order to find a network that has a high score. We used a stochastic hill-climbing search method to find a M with a high score.

This stochastic hill-climb starts at a particular model, M^i, and then performs modifications to M^i, looking for a small change that results in a higher score, or with a small probability, a small change that decreases the score. This procedure resulted in identical results to an exhaustive greedy hill climb, but was much faster due to fewer models being scored at each step of the hill climb. Once such an improvement is found, the current model is changed to this slightly modified model we is now labeled M^{i+1}, and the process iterates using M^{i+1} in place of M^i. The procedure stops when no modifications pass the selection criteria.

Given a regulatory model M, we define the set $Succ(M)$ to consist of all regulatory models that are modifications of M in one and only one of these ways:

- **Modification of interaction.** For one edge e in M, the regulatory class of the interaction is changed. See §5.3 for all regulatory classes.

- **Deletion of interaction.** One of the edges e in M is removed from M.

- **Addition of interaction.** One new edge is added to M, with any possible regulatory class, such that M remains acyclic.

- **Reversal of interaction.** One edge in M is reversed, keeping the same regulatory class, such that M remains acyclic.

During one iteration of the hill climb, elements of $Succ(M)$ are tested to see if any pass the selection criteria, and the first element that does is used to start the next iteration. To prevent bias towards any particular modification, elements are drawn from $Succ(M)$ in a random order. The selection criteria for selecting the successor of model M is stochastic, and has two sufficient criteria. First, any model M' such that $Score(M') >= Score(M)$ will pass. Second, if $Score(M') < Score(M)$, then M' will pass with with probability $\frac{Score(M')}{Score(M)}$. This non-determinism is intended to escape a weak local maximum during the search .

Network search is accomplished by performing hill-climbing from 1000 random starting points, and saving the final model and score from each of these 1000 random climbs. The final predicted network is the best-scoring model from any hill climb. Confidence in a particular edge of this final model is assessed by averaging over the results of all hill climbs. Let $Contains(e, M)$ have value 1 if M contains the edge e with any regulation class, and 0 if M does not contain e. Then the edge score is defined as

$$EdgeScore(e) = \frac{\sum_M Contains(e, M) Score(M)}{\sum_M Score(M)}.$$

The edge score can also be calculated for potential edges that were not in the highest scoring network, so that confidence in non-interaction is also assessed.

5.6 Prediction of *V. cholerae* biofilm network

V. cholerae like many bacteria, form colony biofilms. These biofilms help to protect colonies in various environments, and are thought to be highly involved in changes in pathogenicity. The Yildiz lab has created deletion strains of three regulators of *V. cholerae* biofilm, HapR, VpsT, and VpsR, and of all possible combinations of these deletions [12]. Each of these deletion strains change the visual phenotype of biofilm. In addition, deletions of each of these strains affect the mRNA expression of the *Vibrio* polysaccharide (VPS) gene clusters *vps*-I (VC0917-VC0927) and *vps*-II (VC0934-VC0939). The Yildiz lab expression profiled each deletion strain through competitive hybridization on a microarray with the wildtype strain, resulting in a data matrix D of log ratios of expression between deletion strain and wildtype strain. This original data resulted in a genotype matrix G with seven rows, with all seven possible deletion strains in the first column and the wildtype strain in every entry in the second column.

We augmented the genotype matrix G and data matrix D with additional *virtual* competitive hybridizations by comparing all pairs of deletion strains. These virtual observations alleviate the loss of information from discretization, as the differences of values in discretized space is far less accurate than the discretization of differences. For every unique pair (i, j) such that $1 \leq i \leq j$, $2 \leq j \leq 7$, and $G_{i2} = G_{j2}$

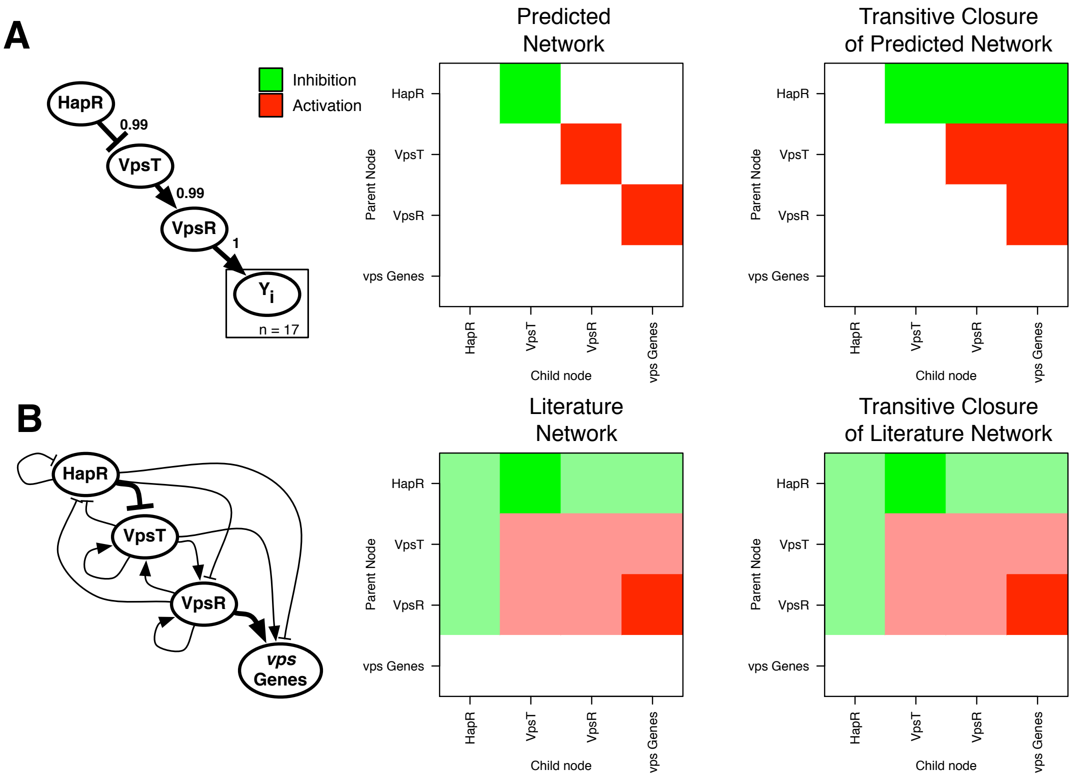


Figure 5.4: Predicted Biofilm Network. A. The predicted *V. cholerae* biofilm regulatory model from JIN in graph (left) and matrix (middle) form. The transitive closure of the predicted model is shown in matrix form on the right. Each edge in the graph is labeled by its edge score. In the matrix representation, red entries represent activating interactions, green entries represent inhibiting interactions, and blank entries indicate the lack of direct interaction. B. Potential regulatory interactions supported by literature. As in A., the network is shown in graph and matrix form. The data for the literature network is adapted from Kanabar *et al.*[70]. Less saturated colors indicate potential interactions with less support.

we add to G an additional row with index k. The values of this row are $G_{k1} = G_{i1}$ and $G_{k2} = G_{j1}$. This virtual hybridization is therefore a competitive hybridization between the genotypes G_{k2} and G_{j1}. We also add a corresponding data row to D with index k, such that $D_{kl} = D_{il} - D_{jl}$. Since elements of D are log ratios between strains, and the denominator strain is identical for i and j, these augmented data rows are log-ratios between deletion strains if we assume that the background strain contains approximately the same value in i and j. Finally, the data matrix D was discretized into the states $\{-1, 0, 1\}$ at boundaries of -1.3 and 1.3, corresponding to fold-change cutoffs used in a previous analysis of this data [12].

Network search was performed as described in §5.5 and resulted in the network shown in Figure 5.4A. All predicted edges had extremely high edge scores, indicating that all models without those edges had extremely small data likelihood compared to models with the edges. The edges scores for potential edges that were not predicted did not exceed 0.01, indicating that there is a strong separation between predicted edges and predicted non-edges.

The predicted network is highly concordant with evidence from the literature. Figure 5.4B shows the transitive closure of the predicted model and potential interactions from previous data. The transitive closure is relevant for comparison to previous data, as the previous measurements of interaction may be the result of transitive interactions. Since potential regulatory models in JIN must be acyclic, we can not predict all possible potential interactions shown in the literature network. However, the predicted network is consistent with the highest confidence potential interactions.

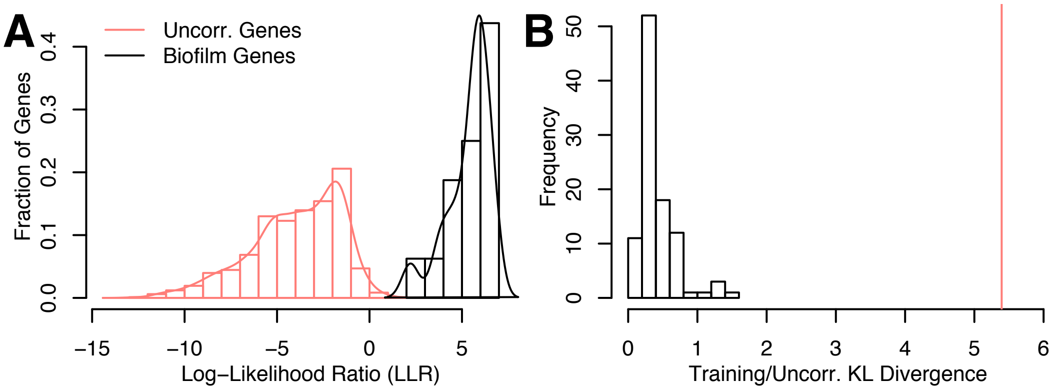


Figure 5.5: Significance of biofilm network. A. A histogram and density estimation of the LLR of *vps* genes and uncorrelated genes under the inferred biofilm regulation network. B. Histogram of Kullback-Liebler divergences between training and uncorrelated genes for 100 null sets of training genes. The vertical red line indicates the divergence for the biofilm genes shown in A.

We assessed statistical significance of the network by the separation in model fit between the training genes and a background made up of uncorrelated genes. A single gene's fit to the model can be assessed by a log-likelihood ratio (LLR), where the ratio is between the likelihood of the gene's data under the predicted model over the likelihood under a null model. For a gene with column index j, this is

$$LLR(j \mid M_p) = \prod_i \frac{\Pr\left(D_{ij} \mid J(M_p, D, G)\right)}{\Pr\left(D_{ij} \mid J(M^{\varnothing}, D, G)\right)}.$$

Here, M_p is the regulatory model and parameters found through network search, and $M^{\varnothing}$ is the regulatory model with no interactions. Figure 5.5A shows a histogram of the log-likelihood ratios for each gene in the *vps* gene cluster and a histogram of the LLRs for genes uncorrelated to the *vps* gene cluster. We defined *uncorrelated* genes as those with an absolute Pearson correlation to the median of training genes that is less than 0.2.

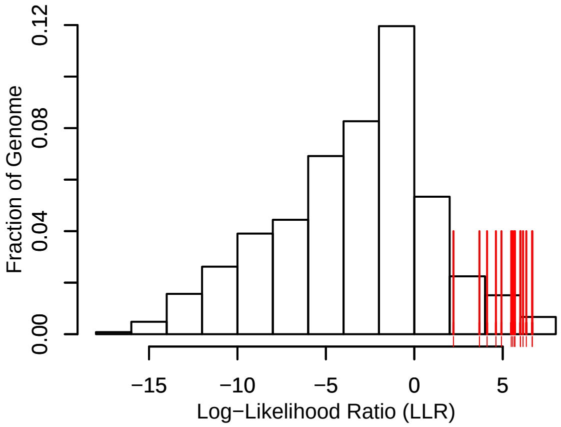


Figure 5.6: Biofilm network expansion. A histogram of log-likelihood ratios for all genes on the microarray. The LLR of genes used to infer the network are shown with red lines.

We found a significant difference between the fit of biofilm genes and uncorrelated genes. The Kullback-Liebler (KL) divergence between the fit of biofilm genes and uncorrelated genes was 5.4 bits. To establish a null distribution of KL divergences between query genes and uncorrelated genes, we simulated 100 random query sets to be the same size as the *vps* gene cluster, learned a network, and calculated the KL divergence between query genes and uncorrelated genes. This null distribution is shown in Figure 5.5B, along with the KL divergence of the *vps* training set at 5.4 bits. Fitting a gamma distribution to the null KL divergences with maximum likelihood results in a p-value of 3×10^{-12}. Thus, the *vps* cluster's fit to the data is indeed significant with respect to this background model.

LLR	Corr. Rank	Locus	Name	Description
7.14	8	VC2445	*exeA*	general secretion pathway protein A
7.03	34	VC1888	*bap1*	biofilm-associated protein
6.83	116	VC2732	*epsE*	general secretion pathway protein E
6.77	99	VC2730	*epsG*	general secretion pathway protein G
6.77	270	VCA0570		Sui1 family protein
6.74	69	VC0483		hypothetical protein
6.73	287	VC1064		lipoprotein-related protein
6.69	114	VCA0612	*mscL*	large-conductance mechanosensitive channel
6.67	86	VC0930	*rbmC*	rugosity-biofilm modulator
6.67	30	VC0931	*rbmD*	rugosity-biofilm modulator
6.67	12	VC1701		hypothetical protein
6.62	67	VC1320	carR	DNA-binding response regulator
6.51	62	VC1935		CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase-related protein
6.48	133	VC1195		lipoprotein, putative
6.48	9	VC0933	*rbmF*	rugosity-biofilm modulator

Table 5.1: Top predictions for new *vps* pathway members. Genes are sorted by decreasing LLR. The correlation rank column is calculated by sorting genes by decreasing Pearson correlation to the median expression profile of the *vps* gene clusters.

5.7 Expansion of *V. cholerae* biofilm network

Though a number genes in *V. cholerae* are known to function in biofilm formation, it is likely that there are many more yet to be discovered. Many genes in *V. cholerae* have no characterized function. Some only have a homolog in another species with a known function, and those genes with known functions may have fulfill additional roles in the cell. Therefore, finding new biofilm-associated genes is of great interest. We can use the predicted JIN model to find additional biofilm genes, instead of a simpler gene expression correlation method. Indeed, the top predictions from JIN had more evidence for biofilm-association than those predicted by correlation to known biofilm genes (Table 5.1).

By scoring the LLR of every gene in the microarray, we found genes other than the training genes that have high LLR. Figure 5.6 shows a histogram of LLR for every gene measured on the microarray, with marks for known biofilm response genes in the *vps*-I and *vps*-II gene clusters. There are several other genes with similar or higher LLRs under this regulation model. Table 5.1 shows the 15 genes with highest LLR which were not used for training.

There is independent evidence that several of these genes are related to biofilm formation in *V. cholerae*. The 5' upstream regions of VC2445, VC0483, and VC0930 contain sequences with high similarity to the VpsR consensus binding site. Several of the genes, VC0930, VC0931, and VC0933, are located between *vps*-I and *vps*-II. The gene *carR* (VC1320), has been verified as a DNA-binding biofilm regulator [13].

5.8 Conclusions

In this chapter I have presented the Joint Intervention Network, a method for learning a gene network from downstream expression phenotypes under gene knockouts. This method captures epistatic reasoning. The network predicted by JIN for *V. cholerae* biofilm closely matches interactions found in the literature. In addition, the regulatory network predicted by JIN is useful for finding other genes under the same regulation program. The biofilm network learned by JIN predicted many novel biofilm-associated genes. This structured expansion of the biofilm network was more accurate at finding biofilm genes than using a simple correlation method.

Chapter 6

Factor graph Nested Effect Model

The availability of RNAi for knockdown and microarrays for profiling gene expression make for a natural combination of measuring many phenotypes under perturbation. Markowetz *et al.* [83] describe a method for inferring a signaling network from such data. They defined a signaling network as a graph with two parts. The genes subject to RNAi knockdowns, referred to as S-genes, are arranged in a directed graph describing the transference of signal from the upstream source to downstream effects. Some genes from the microarrays, referred to as E-genes, were each attached to a single signaling gene. The signaling network describes the response of the E-genes under perturbation.

Given such a network, Markowetz's method is able to score the likelihood of the observed microarray data. When the network space is tractably enumerable, then the best network can be found simply by scoring all networks. However, if the number of signaling genes is n then number of possible networks scales as $\mathrm{O}\left(2^{(n^2)}\right)$.

Thus, finding the optimal network using exhaustive search is tractable only for a small handful of genes. I wished to use the method to place eight human genes in a pathway, and so have developed faster methods for inferring networks. In this chapter I introduce an intuitive extension of the original NEM method, model averaging. I then introduce the Factor Graph Nested Effects Model, which is used in the rest of this dissertation

6.1 Nested Effect Model Averaging

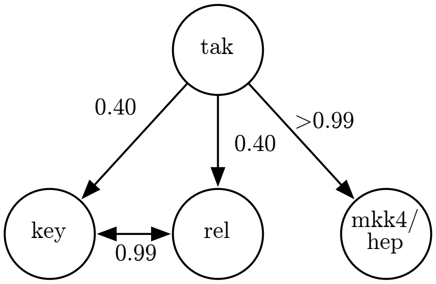
I evaluated the likelihood of a subset of models and then assigned a posterior probability to each potential link. By "link," I mean either the "forwards" and the "backwards" directions between a pair of genes. For example, $A \rightarrow B$ and $B \rightarrow A$ are two separate links, present or absent in the network independently. Denote the likelihood of the data as D, and the subset of possible networks as $\mathcal{M}$. Then I assign the probability of any edge between A and B as approximately:

$$\Pr(A \rightarrow B \mid D) \approx \frac{\sum_{M \in \mathcal{M}} \Pr(D \mid M) \Pr(A \rightarrow B \mid M) \Pr(M)}{\sum_{M \in \mathcal{M}} \Pr(D \mid M) \Pr(M)} \Pr(A \rightarrow B) \tag{6.1}$$

I assemble a network from posterior probabilities by choosing a threshold probability, and then taking only those links that pass that threshold. Since there are $\mathrm{O}\left(n^2\right)$ links, each with its own posterior probability, there are at most $\mathrm{O}\left(n^2\right)$ thresholds. Each threshold results in a unique network, and I find the most likely network by evaluating the likelihood of each.

	key	rel	tak	mkk4hep
key		$-2.5 \cdot 10^{-7}$	-130	-94
rel	$-5.7 \cdot 10^{-7}$		-130	-94
tak	$-9.2 \cdot 10^{-1}$	$-9.2 \cdot 10^{-1}$		$> -1 \cdot 10^{-99}$
mkk4hep	$-7.8 \cdot 10^{1}$	$-7.0 \cdot 10^{1}$	-69	

(a) Log posterior probabilities for all links



(b) Correct links and their estimated posterior probabilities

Figure 6.1: Estimated posterior probabilities of each link. The likelihood of the data for every linear network was calculated, and the posterior probability was calculated with equation 6.1 and a prior link probability of 0.25. In table 6.1(a) the source gene of the link is on the left and the target is along the top. Highlighted links were chosen to be in the network.

6.1.1 Linear models recover *D. melanogaster* immune response

In order to test model averaging, I evaluated a *D. melanogaster* LPS signaling dataset [14] that had previously been analyzed via exhaustive search [83] successfully. In order to ensure a sample that considers all possible orientations of edges, I used the subset $\mathcal{M}$ of all linear permutations of the signaling genes. This subset, though still super-exponential, *i.e.* $\mathrm{O}\left(2^{n \ln n}\right)$, can still be evaluated up to approximately 10 genes on a desktop computer in less than a day using my current implementation. In order for the computation to be numerically stable, I performed all calculations in log space using the usual identity for addition of log space numbers [27].

Linear model averaging recovers the signaling network found by both the original biological investigation [14] and the subsequent computational investigation [83]. I chose a prior on network features of 0.25, on the assumption that genes are at least minimally connected, *i.e.* that there are at least three of the possible twelve directed edges. The gene *tak* is at the top of the signaling hierarchy, signaling to all other genes. The genes *rel* and *key* are linked equivalently, indicating that their phenotypes under knockdown are indistinguishable. Linear networks without the link *tak* → *mkk4/hep* have data likelihoods so much worse than those with the link that the posterior probability of the link is greater than $1 - 2^{-2048}$

The links from *tak* to *key* and *rel* have the lowest posterior probability of the true links, but are still orders of magnitude stronger than the posterior probabilities of any of the false links. Further, with a prior probability on the links that is based on the true model, 5/12, these links would have a posterior probability of 2/3. One possible reason for the lower probability of these links is due to the equivalence of *key* and *rel* in that they have links in both directions between them. Therefore removing just one of the links from *tak* will still result in an identical signaling network, since the signal is transitive in this model

An advantage of the model averaging approach over exhaustive search is that it provides posteriors on individual network links. Using Markowetz's method assigns a single likelihood to the entire network. Ranking individual features can guide further investigations, and help assess where to trust the most likely network.

Attempting to predict cancer networks failed to produce any confident networks. Discretization of the data proved quite challenging, as there are few replicates from which to estimate variance. Without a reliable estimate of variance, placing a discretization boundary at an arbitrary level only emphasized platform effects, and a proper normalization was difficult.

6.2 Interaction Modes

My goal is to automatically identify genetic interactions among a set of signaling genes from gene expression changes observed under their knock-down. The signaling genes represent a set of genes that prior experimental evidence suggests participate in a common pathway. To infer a network, I use an extension of the Nested Effect Model (NEM) introduced by Markowetz *et al.* in 2005[83]. The set of silenced genes are denoted as the set S (or S genes). An NEM is a probabilistic formulation that measures how well a directed graph of the S-genes is consistent with expression changes collected under the separate silencing of each S-gene (i.e. only single knock-downs are considered in NEM). While the method can make use of either complete deletion mutants or genes that may be partially silenced, here I use the term knock-down to refer to either case. I denote the knock-down of S-gene A as ΔA. I also refer to a set of effect genes as the set E (or E genes), for which gene expression data is available. The expression of an E-gene e is assumed to be influenced by at most one S-gene. The key assumption of NEMs is the expression changes observed under ΔA are an approximate superset of the changes observed under ΔB if gene A acts upstream of gene B in a pathway. I use

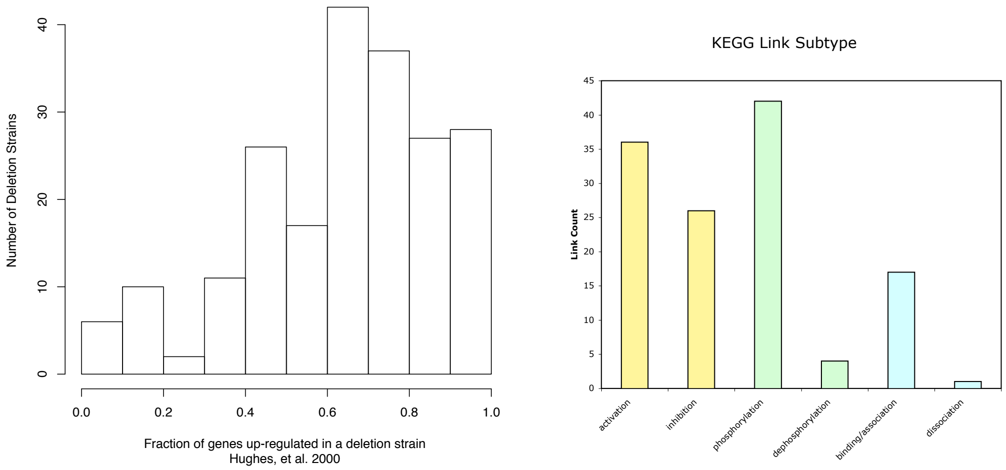


Figure 6.2: Observed inhibitory effects and signaling in a yeast compendium. 6.2(a) Histogram of percentage of up-regulated gene expression from gene knockdown in the Hughes *et al.* (2000) compendium. In each deletion strain, gene expression changes with a p-value better than 0.05 were selected, and then assigned to up-regulated or down-regulated according to their expression log-ratio. The percentage of each strain's up-regulated expression change is plotted in the histogram. Presence of up-regulated expression under gene deletion is evidence of an inhibitory interaction close to the deleted gene. 6.2(b) Histogram of interaction types in *S. cerevisiae* in the KEGG pathway database. In the KEGG ontology, each interaction may be categorized with more than one label. Activation and inhibition are antonyms, as are expression and repression.

the shorthand $A{\rightarrow}B$ to represent this generic directed interaction.

In addition to identifying $A{\rightarrow}B$, the E-gene expression changes on the microarray can be used to infer the "sign" of the interaction, either activating or inhibiting. In this framework, I extend the interactions so that an upstream gene can have either an inhibitory or stimulatory effect on downstream genes. Biological networks exhibit a large degree of inhibitory interactions. Figure 6.2 shows two estimates of the degree of inhibitory interactions from expression and pathway compendiums. An increase in a gene's expression under gene knockout indicates that there is some inhibitory path between the gene knockout and the gene whose expression increased. Figure 6.2(a) indicates that for the Hughes compendium of deletions, that a gene deletion chosen at random will most likely result in more genes up-regulated than down-regulated. This provides evidence that a significant number of transitive paths are inhibitory in the yeast biological network. Figure 6.2(b) shows the number of genes annotated as either activating or inhibiting in hand-curated pathways. A significant fraction of these direct interactions are annotated as inhibitory. Taken together, these figures show that inhibition plays a large role in the yeast cellular network, and therefore is likely to play a large role in the cellular networks of most other organisms. Modeling and distinguishing between activation and inhibition will therefore be quite informative. Additionally, I show in the following chapter that distinguishing between activated and inhibitory interactions and effects allows for better performance in both network inference and network expansion.

Figure 6.3 presents an example, similar to the work of Fröhlich *et al.* in 2008[38] that motivates the use of signed interactions. E-genes E_1 through E_{13} are listed from top to bottom according to where they are attached to the network. Depending on the connections of the S-genes to one another and to the E-genes, a disruption in an S-gene will cause E-genes to either increase or decrease in expression relative to wild-type. For example, E-gene E_7 decreases under ΔB relative to wild-type because the wild-type activation by B is absent in the deletion. On the other hand, the expression of E_{10} also decreases under ΔB relative to wild-type but as a result

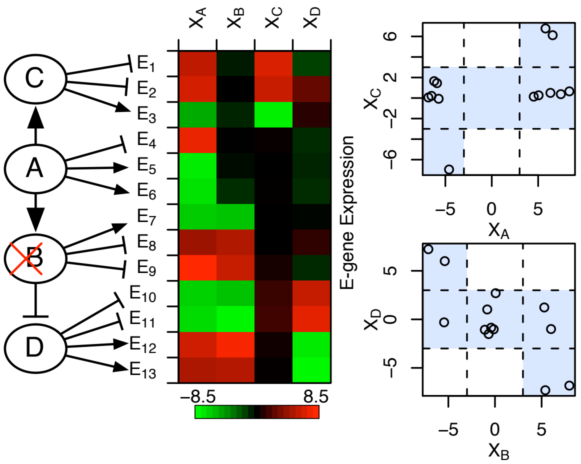


Figure 6.3: Hypothetical example with four S-genes, A, B, C, and D. The graph contains one inhibitory link, $B \dashv D$ (left). A heatmap of E-gene expression under knockdown of each S-gene shows both inhibitory and stimulatory effects (middle). Scatter plots of the C, A, B, and D knockouts show that expression fits in the shaded preferred regions of each interaction (right). The inhibitory link explains some of the "observed" data: expression changes under ΔD (bright red or bright green entries in the heatmap) occur in a subset of the E-genes for which the opposite changes occur in ΔB.

of a different mechanism. In wild-type, E_{10} is expressed at a baseline level because its repressor, the product of gene D, is inhibited by B's product. However, in the B deletion, D is de-repressed, leading to inhibition of E_{10}. This toy example illustrates the disambiguation of inhibition and activation both for S-gene interactions and E-gene attachments making it possible to account for an expanded set of mechanisms leading to the observed expression changes.

The E-gene expression changes are available in a data matrix X where each column gives the difference in expression of each E-gene under the deletion of a single S-gene relative to wild-type. X may also contain replicates in the form of repeated S-gene knock-downs. The entry X_{eAr} represents e's expression change under the rth replicate of ΔA. Furthermore, I assume that an unknown expression "state" for each E-gene under each knock-down, determines its set of expression changes observed across the $\{X_{eAr}\}$ replicates in the microarray data. The matrix Y records a hidden state for each E-gene under each knock-down, where entry Y_{Ba} is the state of E-gene e under ΔA. I allow the states to be ternary-valued $+1, -1, 0$ representing whether e is up-regulated, down-regulated, or unchanged under ΔA relative to wild-type respectively.

Nested effects models include two sets of parameters. The parameter set Φ records all pair-wise interactions among the S-genes and the parameter set Θ describes how each E-gene is attached to the network of S-genes. In the original NEM formulations[83, 84, 38] Φ is a binary matrix with entry ϕ_{AB} set to one if S-gene A acts above S-gene B and zero otherwise. If $\phi_{AB} = \phi_{BA} = 1$ then the S-genes are assumed to operate at an equivalent position in the pathway. Note that indirect interactions are also represented in Φ so that if $\phi_{AB} = 1$ and $\phi_{BC} = 1$ it implies that $\phi_{AC} = 1$. A parsimonious network among the S-genes is solved for by computing the transitive reduction of Φ.

To allow for both stimulatory and inhibitory interactions in this formulation, ϕ_{AB} can assume six possible values for each unique unordered S-gene pair A, B. I refer to these values as interaction modes. The possible values are: i) A activates B, $A{\rightarrow}B$; ii) A inhibits B, $A{\dashv}B$; iii) A is equivalent to B, $A = B$; iv) A does not interact with B, $A{\neq}B$; v) B activates A, $B{\rightarrow}A$; and vi) B inhibits A, $B{\dashv}A$. Additional interaction modes, such as $A{\longmapsto}B$, $A{\longleftarrow\!\!\!\dashv}B$, and $A{\vdash\!\!\!\dashv}B$ are possible, but have not been considered in this dissertation and are saved for future work.

Plotting the response of E-genes under ΔA and ΔB yields a scatter-plot that

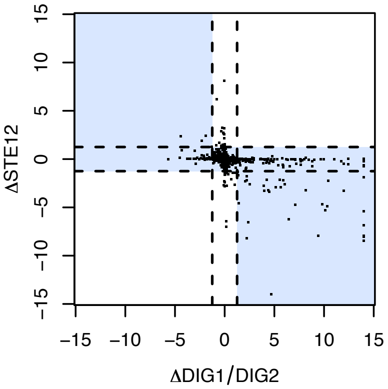


Figure 6.4: Knock-out expression for a known inhibitory interaction.Expression levels of effect genes under the DIG1/DIG2 knock-out (x-axis) plotted against their levels under the STE12 knock-out (y-axis) as detected in an expression compendium[57]. The $\alpha = 0.05$ significance of expression change is indicated in dashed lines. DIG1/DIG2 is known to inhibit the activity of STE12

may provide a signature for the type of interaction between A and B. For example, Figure 6.4 shows a scatter-plot of gene expression changes from the Hughes *et al.* yeast knock-out compendium[57] for a pair of knock-outs of the well-known pheromone-response genes: Δ*STE12* and the Δ*DIG1/DIG2* double knockout. Comparing the scatter-plot for these pheromone-response genes to the patterns in Figure 6.5, it can be seen to match the inhibitory interaction mode more closely than the other modes, which is consistent with *DIG1/DIG2*'s known inhibition of *STE12*. Figure 6.5 shows an example of the first four modes from the previous paragraph. Shaded regions denote consistent E-gene responses for each mode.

An interaction mode determines a constraint on the observed E-gene expression changes. For example, plotting the expression changes of E-genes that act downstream of either A or B for the generic $A \rightarrow B$ interaction mode produces points in one of

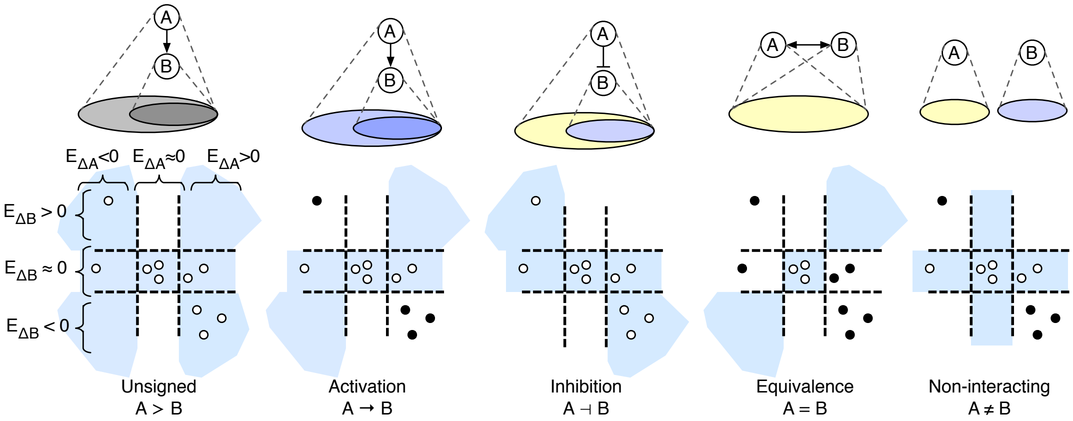


Figure 6.5: Interaction Modes. Observed E-gene expression changes are compared to five possible types of interactions between two S-genes, A and B (i-v). The top row illustrates the expected nested effects relationship for each type of interaction mode: circles represent sets of E-genes with expression changes consistent with either activation (blue circles) or inhibition (yellow circles). Scatter-plots for each interaction mode show the hypothetical expression changes under ΔA (x-axis) and ΔB (y-axis) for all E-genes (circles). E-gene levels are either consistent (open) or inconsistent (filled) with the mode. Shaded regions demark expression levels consistent with each interaction model. The example shows expression changes that most closely match the inhibition mode (indicated by the greatest number of closed circles).

the five shaded regions shown in Figure 6.5 above the label Activation. Figure 6.5 shows an example where the inhibitory interaction mode is the best match to the data because a higher number of E-gene changes fall within consistent regions (open points) than in inconsistent regions (filled points). Genome-wide expression changes detected on the microarrays can be used as quantitative phenotypes to identify a variety of interactions between pairs of S-genes. Note that two genes are equivalent if their knock-downs lead to significantly similar expression changes, which may predict, for example, that they form a complex.

Figure 6.5 also illustrates the generic interaction mode $A{>}B$ equivalent to the interaction of genes in previous NEM methods. In §7.2 I compare FG-NEM results to two unsigned variants to estimate the change in predictive power as a function of the introduction of sign.

6.3 Pairwise Network Formulation

My goal is to find a structure among the S-genes that provides a compact description of X. To find a network that best fits the data, I take a maximum a posteriori approach as in [84, 38] to identify the Φ that maximizes the posterior:

$$\begin{aligned} J(X) &= \arg\max_{\Phi} \{\Pr(\Phi \mid X)\} && (6.2) \\ &= \arg\max_{\Phi} \left\{ \sum_{\Theta} \sum_{Y} \Pr(\Phi, \Theta, Y \mid X) \right\} && (6.3) \end{aligned}$$

where Θ refers to the attachment point of each E-gene into the network and Y refers to the hidden E-gene states. Attachment points for effects, as shown previously in

Figure 6.3, determine the predicted response for an effect from a knockdown. Later, Θ will be parameterized slightly differently than in previous NEM formulations. Applying Bayes' Rule and dropping $\Pr(X)$, which is constant with respect to the maximization obtains:

$$\begin{aligned} J(X) &= \arg\max_{\Phi} \left\{ \Pr(\Phi) \sum_{\Theta} \Pr(\Theta \mid \Phi) \sum_{Y} \Pr(Y \mid \Phi, \Theta) \Pr(X \mid Y) \right\} && (6.4) \\ &= \arg\max_{\Phi} \left\{ \Pr(\Phi) \sum_{\Theta} \sum_{Y} \Pr(Y \mid \Phi, \Theta) \Pr(X \mid Y) \right\} && (6.5) \end{aligned}$$

The approximation in the last step uses the assumption that any E-gene attachments are equally likely given a network structure; i.e. $\Pr(\Theta \mid \Phi)$ is assumed to be uniformly distributed and is ignored in this approach. $\Pr(\Phi)$ represents a prior over S-gene networks.

As in previous NEM formulations, I assume that each E-gene is attached to a single S-gene and that each E-gene observation vector across the knock-downs is independent of other E-gene observations. The maximization function can then be written:

$$\begin{aligned} J(X) &= \arg\max_{\Phi} \left\{ \Pr(\Phi) \sum_{\Theta} \sum_{Y} \prod_{e \in E} \Pr(Y_e \mid \Phi, \theta_e) \Pr(X_e \mid Y_e) \right\} && (6.6) \\ &= \arg\max_{\Phi} \left\{ \Pr(\Phi) \prod_{e \in E} \sum_{\Theta} \sum_{Y} \Pr(Y_e \mid \Phi, \theta_e) \Pr(X_e \mid Y_e) \right\} && (6.7) \\ &= \arg\max_{\Phi} \left\{ \Pr(\Phi) \prod_{e \in E} L_e(\Phi) \right\} && (6.8) \end{aligned}$$

where X_e and Y_e are the row vectors of data and hidden states for E-gene e respectively, and θ_e records the attachment point information for E-gene e. After rearranging the products and sums, I introduce the shorthand L_e to represent the likelihood of the data restricted only to E-gene e under a particular model Φ and θ_e.

Previous approaches decompose L_e over the knock-downs, which assume the S-gene observations are independent given the network and attachments (see Fröhlich 2008[38] for an example of such a derivation). To facilitate scoring the expanded set of interaction modes mentioned earlier, I replace L_e with a function proportional to L_e, L'_e. L'_e is defined as a product of pair-wise S-gene terms:

$$L'_e = \prod_{\substack{\{A,B\}\subset S\\ A\prec B}} \sum_{\theta_{eAB}} \sum_{Y_{eA}} \sum_{Y_{eB}} \Pr\left(Y_{eA}, Y_{eB} \mid \phi_{AB}, \theta_{eAB}\right) \Pr\left(X_{eA} \mid Y_{eA}\right) \Pr\left(X_{eB} \mid Y_{eB}\right)$$

where θ_{eAB} represents the *local* attachment of E-gene e, i.e. the attachment of e in the network relative only to the pair of S-genes A and B. This local attachment represents whether there is a path from A or B to e, and the sign of that path. The parameter θ_{eAB} can therefore take on five possible values from the set $\{A, -A, B, -B, 0\}$ representing that e is either up- or down-regulated by A, either up- or down-regulated by B, or not affected by either S-gene, respectively. Note that both θ_{eAB} and ϕ_{AB} are indexed by a pair, A, B, and that an arbitrary total ordering of the variables, $\prec$, has been introduced so that ϕ_{AB} is only counted once. ϕ_{AB} defines the interaction mode between S-genes A and B. Given an interaction mode ϕ_{AB} and the attachment point θ_{eAB}, the expected response of e under each knockdown, Y_{eA} and Y_{eB}, is entirely determined. Therefore the probability $\Pr\left(Y_{eA}, Y_{eB} \mid \phi_{AB}, \theta_{eAB}\right)$ has value 1 if Y_{eA} and Y_{eB} are consistent with the interaction and attachment point, and 0 if inconsistent.

Assuming the replicates are independent given the E-gene states, $\Pr(X_{eA}|Y_{eA})$ can be written as a product over replicate terms: $\prod_{r \in R_A} \Pr(X_{eAr} \mid Y_{eA})$ where R_a is the set of replicates for ΔA and $\Pr(X_{eAr} \mid Y_{eA})$ is modeled with a Gaussian distribution having mean μ and standard deviation σ estimated from the data. While I used hard constraints to model consistent and inconsistent expression changes (corresponding to the rigid boundaries of the regions drawn in Figure 6.5), such constraints could be softened to use factors with belief potentials between zero and one. Also, note that even though the interaction modes in Figure 6.5 show boundaries, observations that fall outside these illustrated boundaries do not have zero probability since $\Pr(X_{eA} \mid Y_{eA})$ is modeled as a Gaussian distribution and therefore assigns non-zero probabilities over all possible expression values.

Substituting L'_e for L_e in Equation (6.8) and approximating summation over attachment points θ_{eAB} with maximization results in the maximizing function used in the FG-NEM approach:

$$J(X) = \arg\max_{\Phi} \Pr(\Phi) \prod_{\substack{e \in E \\ \{A,B\} \subset S \\ A \prec B}} \max_{\substack{\theta_{eAB} \\ Y_{eA} \\ Y_{eB}}} \Pr(Y_{eA}, Y_{eB} | \phi_{AB}, \theta_{eAB}) \Pr(X_{eA}|Y_{eA}) \Pr(X_{eB}|Y_{eB})$$

6.4 Transitivity Constraints and Model Priors

The prior over interactions, $\Pr(\Phi)$, can represent preferences over specific interactions in the S-gene graph, allowing the incorporation of biologically-motivated constraints to guide network search. For example, the interaction priors for genes in a common pathway or genes whose products have been detected to interact in protein-

protein interaction screens could be set higher than the priors for arbitrary pairs of S-genes. In this thesis I have used the prior both with and without external biological information. Without external biological information, the prior encodes a basic property of the S gene graph: that it should exhibit transitivity to force pair-wise interaction modes to be consistent among all triples. Using transitivity, all paths between any two genes, A and B, are guaranteed to have the same overall effect; i.e. the product of the signs of individual links along different paths between A and B are equal.

In order to preserve the transitivity of identified interaction modes, the prior is decomposed over interaction configurations into transitivity constraints on all triples of S-genes:

$$\Pr(\Phi) \propto \left(\prod_{\substack{\{A,B,C\} \subset S \\ A \prec B \prec C}} \tau_{ABC}\left(\phi_{AB}, \phi_{BC}, \phi_{AC}\right) \right) \left(\prod_{\substack{\{A,B\} \subset S \\ A \prec B}} \rho_{AB}\left(\phi_{AB}\right) \right) \tag{6.9}$$

where τ is zero if the triple of interactions are intransitive, and one if the interactions are transitive. The product over ρ factors in Equation (6.9) encodes evidence from high-throughput assays, such as protein-protein binding and protein-DNA binding interactions (see §8.3, "Physical Structure Priors"). The transitivity constraint includes both the direction of interactions and the sign of interactions. As S-gene interactions are signed, the transitivity constraint forces the sign of the product of two edges to equal the sign of the third; e.g. if $A \dashv B$ and $B \dashv C$, then $A \rightarrow C$.

The function τ can be formally defined as follows. For each interaction ϕ_{AB} let $f(\phi_{AB})$ denote the forward signal of A to B, and let $b(\phi_{AB})$ denote the sign of the signal from B to A:

$$f(\phi_{AB}) = \begin{cases} 1 & \text{if } \phi_{AB} \in \{\rightarrow, =\}, \\ 0 & \text{if } \phi_{AB} \in \{\leftarrow, \neq, \vdash\}, \\ -1 & \text{if } \phi_{AB} \in \{\dashv\}. \end{cases} \qquad b(\phi_{AB}) = \begin{cases} 1 & \text{if } \phi_{AB} \in \{\leftarrow, =\}, \\ 0 & \text{if } \phi_{AB} \in \{\rightarrow, \neq, \dashv\}, \\ -1 & \text{if } \phi_{AB} \in \{\vdash\}. \end{cases}$$

Additionally, using Iverson notation where let [*predicate*] has value one if *predicate* is true, and zero if false, let

$$t(A, B, C) = 1 - [f(\phi_{AB}) \neq 0][f(\phi_{BC}) \neq 0][f(\phi_{AB})f(\phi_{BC}) = b(\phi_{AC})]$$

The value of a transitivity constraint $\tau_{ABC}\left(\phi_{AB}, \phi_{BC}, \phi_{AC}\right)$ can now be defined as:

$$\tau_{ABC}\left(\phi_{AB}, \phi_{BC}, \phi_{AC}\right) = t(A, B, C)t(A, C, B)t(B, A, C)t(B, C, A)t(C, A, B)t(C, B, A).$$

A result of modeling transitivity is that a directed cycle of stimulatory interactions will also imply activation between any pair of S-genes in the cycle, in both directions. Therefore, the method clusters such S-genes into equivalence interactions.

While network structures are constrained to reflect more intuitive models, the decomposition introduces interdependencies among the interactions, adding complexity to the search for high-scoring networks. Importantly, max-sum message passing in a factor graph[74] provides an efficient means for estimating highly probable S-gene configurations. We next describe how the problem is recoded into message-passing on a factor graph.

6.5 Model Inference Using a Factor Graph

The formulation above provides a definition of the objective function to be maximized but says nothing about how to search for a good network. The search space of networks is very large making exhaustive search intractable for networks larger than five S-genes[83]. To apply the method to larger networks, I required a fast, heuristic approach. Markowetz *et al.* [84] introduced a bottom-up technique to infer an S-gene graph. They identify sub-graphs of S-genes (pairs and triples) and then merge the sub-graphs together into a final parsimonious graph. Fröhlich *et al.* (2008) [38] use hierarchical clustering to first identify modules, subsets of S-genes with correlated expression changes. Networks among the modules are exhaustively searched and a final network is identified by greedily introducing interactions across modules that increase the likelihood.

Here, I introduce the use of a graphical model called a factor graph to represent all possible NEM structures simultaneously. The parameters that determine the S-gene interactions, Φ, are explicitly represented as variables in the factor graph. Identifying a high-scoring S-gene network is therefore converted to the task of identifying likely assignments of the Φ variables in the factor graph. A factor graph is a probabilistic graphical model whose likelihood function can be factorized into smaller terms (factors) representing local constraints or valuations on a set of random variables. Other graphical models, such as Bayesian networks and Markov random fields, have straightforward factor graph analogs. A factor graph can be represented as an undirected, bi-partite graph with two types of nodes: variables and factors. A variable is adjacent to a factor if the variable appears as an argument of the factor. Factor graphs generalize probability mass functions as the joint likelihood function requires no normalization and the factors need not be conditional probabilities. Each factor encodes a local constraint pertaining to a few variables.

Figure 6.6 shows the factor graph representing the NEM for the example S-gene network from Figure 6.3. Each random variable is represented by a circle and each conditional probability term in Equation (6.3) and Equation (6.9) is represented by a square. The factor graph contains three types of variable nodes. First, every unique unordered pair of S genes $\{A, B\}$ has a corresponding variable, ϕ_{AB}, that takes on values equal to one of the previously mentioned interaction modes (Figure 6.6, "S-Gene Interactions" level). Second, every E-gene-S-gene pair is associated with a variable, Y_{eA} for the hidden expression state of effect gene e under knock-down A, (Figure 6.6, "E-gene Expression State" level). Third, every observed expression value is associated with a continuous variable, X_{eAr}, where r indexes over replications of ΔA (Figure 6.6, "E-gene Expression Observation" level). Figure 6.6 also shows the expression factors, interaction factors, and transitivity factors of Equation (6.3) and Equation (6.9).

A Φ that maximizes the posterior is found using max-sum message passing using all terms from Equation (6.3) and Equation (6.9) in log space. For acyclic factor graphs, the marginal, max-marginal and conditional probabilities of single or multiple variables can be calculated exactly with the max-sum algorithms[74]. Message-passing algorithms demonstrate excellent empirical results in various practical problems even on graphs containing cycles such as feed-forward and feed-back loops[36, 35, 82, 132].

Here, I use a message passing schedule that performs inference in two phases. In the first phase, messages from observations nodes X_{eAr} are passed through the ex-

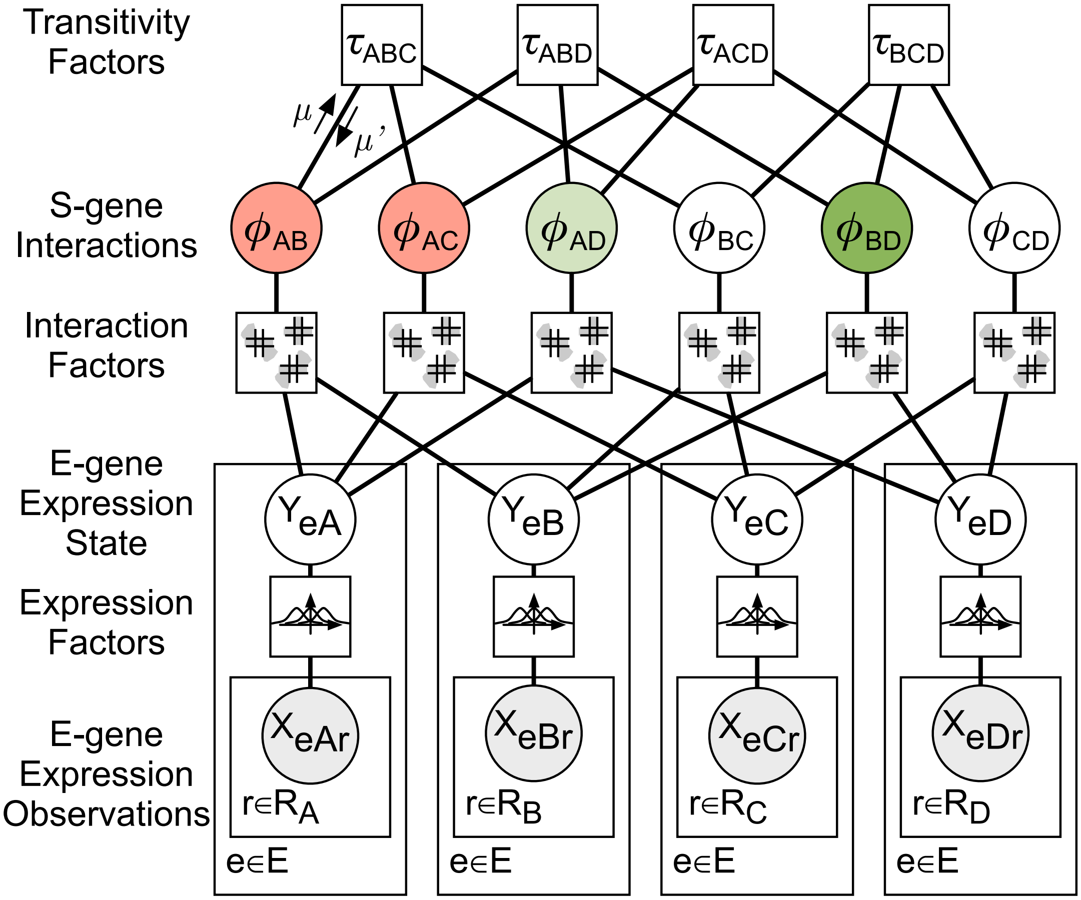


Figure 6.6: Structure of the factor graph for network inference. The factor graph consists of three classes of variables (circles) and three classes of factors (squares). X_{eAr} is a continuous observation of E-gene e's expression under ΔA and replicate r. Y_{eA} is the hidden state of E-gene e under ΔA, and is a discrete variable with domain $\{-1, 0, 1\}$. ϕ_{AB} is the interaction between two S-genes A and B. Expression Factors model expression as a mixture of Gaussian distributions. Interaction Factors constrain E-gene states to the allowed regions shown in the interaction modes of Figure 6.5. Transitivity Factors constrain pair-wise interactions to form consistent triangles. The arrows labeled μ and μ' are messages encoding local belief potentials on ϕ_{AB} and are propagated during factor graph inference.

pression factors and hidden E-gene state variables, to calculate all messages $\mu(Y_A \rightarrow \phi_{AB})$in a single upward pass. In the second phase, messages are passed between only the interaction variables and transitivity factors until convergence, as described below. In the example shown in Figure 6.6, running inference results in assignments of activation for ϕ_{AB} and ϕ_{BC} (shaded red), inhibition for ϕ_{BD} and ϕ_{AD} (shaded green), and non-interaction for ϕ_{AB} and ϕ_{BC} (unshaded), which match the NEM structure from Figure 6.3. For display of inferred S-gene networks, I compute the transitive reduction of Φ by removing all links for which there is a longer redundant path[121].

Each message $\mu(\phi_{AB} \rightarrow \tau_{ABC})$ or $\mu(\tau_{ABC} \rightarrow \phi_{AB})$, (denoted by μ and μ' in Figure 6.6, respectively), consists of a length six vector, representing the source's log-space valuation of six possible interaction types for ϕ_{AB}. The message passing schedule consists of two-step iterations. $\mu(\phi_{AB} \rightarrow \tau_{ABC})$ $\mu(\tau_{ABC} \rightarrow \phi_{AB})$ The first step of an iteration calculates every message from an interaction variable to a transitivity constraint, $\mu(\phi_{AB} \rightarrow \tau_{ABC})$. The second step of an iteration calculates all messages from transitivity constraints to interaction variables, $\mu(\tau_{ABC} \rightarrow \phi_{AB})$. In order to calculate the first step of the first iteration, every $\mu(\tau_{ABC} \rightarrow \phi_{AB})$ is set to $[0, 0, 0, 0, 0, 0]$. To prevent non-convergence from cyclic activities, each new message is dampened by arithmetically averaging it (in log-space) with the message from the previous iteration as well as by normalizing each factor-to-variable message to sum to one. At the end of each iteration, the change in $\mu(\tau_{ABC} \rightarrow \phi_{AB})$ messages is calculated by forming a matrix of all messages, and calculating the Frobenius norm of the new message matrix and the previous message matrix. Message passing is terminated when the Frobenius norm is less than 0.01, or after 50 iterations.

During message passing the summation over Y_{eA} is approximated with a maximization step. Max-sum approximation of a marginalization step has been shown to produce accurate results and converge efficiently for several factor graph applications to large problems including SAT and clustering[35, 89]. The maximization chooses the highest probability configuration of attaching an effect at some point relative to a pair of S genes as well as the sign of the attachment.

6.6 Pathway Expansion

Once a signaling network is identified using the message passing inference procedure above, the network can be used to search for new genes that may be part of the pathway. The NEM and FG-NEM framework predict new members that act in the pathway by "attaching" E-genes to S-genes in the network, or leaving them detached if their expression data does not fit the model. Attaching E-gene e to S-gene A asserts that the expression changes of e over all knock-downs are best explained by a network in which e is directly downstream of A. The E-genes attached to the network are collectively referred to as the frontier. Frontier genes may be good candidates for further characterization (e.g. knock-down and expression profiling) in subsequent experiments.

To gain a global picture for where e is connected, I use a modified NEM scoring from Markowetz *et al.* (2005)[83]. The pair-wise attachments for a single E-gene connection variable θ_{eAB}, provide local "best guesses" for e's attachment. Rather than aggregate e's collection of local attachments, I use NEM scoring, modified to incorporate both stimulatory and inhibitory attachments, to estimate the attachment point using the full network learned in the previous step.

Using the notation of Markowetz *et al.* (2005)[83], NEM scoring assigns a likelihood to a hidden attachment variable, θ_e, for each E-gene e. The attachment variable can take on possible values $\{1, 2, \ldots, |S|\}$, where each number corresponds to a single S-gene attachment point. To allow both inhibitory and stimulatory effects, I allow θ_e to take on possible values $-|S|, \ldots, 1, 0, 1, \ldots, |S|$, where a positive index means e is activated by S gene A while a negative index means it is inhibited by A. The zero attachment state represents detachment from the entire signaling network, allowing e's expression levels to be independent of any single S gene. Note that this possibility was not included in the original NEM framework, but a generalization has been described in Tresch *et al.* (2008) [115] in which effects can be attached to "null actions" that are analogous to including a detachment possibility introduced here. NEM model scoring assigns a likelihood to each possible value of θ_e, and I choose the value that maximizes the likelihood as the attachment point.

I calculate a log-likelihood ratio that measures the degree to which e's expression data is explained by the network if it is attached to one of the S-genes, compared to being disconnected from the network, i.e. its likelihood was generated entirely by the background Gaussian distribution. For E-gene e, the the log-likelihood of attachment ratio (LAR) is:

$$LAR(e) = \log \left(\frac{\max_{i \neq 0} \Pr\left(X_e \mid \Phi, \theta_e = i\right)}{\Pr\left(X_e \mid \Phi, \theta_e = 0\right)} \right).$$

We rank all of the E-genes according to their LAR scores. Top-scoring genes have data that is more likely to have arisen from the model than a null background. Any E-gene that has a positive LAR score is considered for inclusion as a frontier gene. In the applications in subsequent chapters, significance of attachment is tested by several methods.

Chapter 7

FG-NEM Performance on Artificial Networks

One difficulty with current pathway prediction algorithms is that there are very few examples where all the members of a signaling pathway are known, or all of the interactions between genes. It can therefore be of interest to examine a prediction algorithm's performance on artificial networks and data, as performance can be more fully assessed than on the incomplete knowledge of biological networks.

In this chapter I first compare FG-NEM methods to the prior NEM method[83] to show that the factor graph has similar network structure recovery capability with similar data. Second, I examine the performance of FG-NEM under different data generation conditions, including the amount of inhibition in the network, the amount of the network which is included in the S-gene set, the amount of noise in the effect measurements, and the number of data replicates. These experiments can help guide the experimental design of biological experiments, for example in deciding how many data replicates to measure.

7.1 Network and Data Generation

To create a synthetic signaling network, I first generated the network among S-genes containing both inhibition and cycles, and then attached E-genes to this S-gene network. I created a full underlying network for a set of S-genes, T, by first generating a random connected acyclic graph, and then adding additional links by randomly sampling pairs of S-genes. Each interaction was associated with a strength chosen uniformly between 0.75 and 1. The strength of an interaction determines how well a signal passes from one S-gene to another, with 1 equal to perfect transmission and 0.75 equal to 25% transmission loss. Any interaction was also chosen to be inhibitory with probability λ. The parameter λ controls the proportion of inhibitory interactions between S-genes as well as inhibitory E-gene attachments. For example, $\lambda = 0$ produces a network containing only activating S-gene connections and only activated E-gene attachments. After generating the S-gene network, a fixed amount of E-genes were added by choosing a parent S-gene uniformly from all S-genes. This S-gene to E-gene connection was inhibitory with probability λ.

Given a synthetic signaling network, expression data was generated as follows. To generate a single data replicate for A, quantized expression responses under ΔA were simulated by propagating a signal from A to the connected effect genes, keeping track of the sign and the strength of the signal. The final signal arriving at E-gene e, s_e, was calculated as the product of the strengths and signs of all of the interactions on the path between A and e. E-gene e was then assigned to the activated (or inhibited) distribution with probability $|s_e|$ for $s_e > 0$ (or $s_e < 0$) and was assigned to the unaffected distribution with probability $1 - |s_e|$. E-genes thus chosen to be activated by A in a replicate were given expression levels that were less than their expression levels in wild-type, as the deletion of the gene results in the loss of activation. Similarly, E-genes inhibited in a replicate were given expression values from a expression distribution with a higher mean than wild-type. I refer to the distribution of expression changes of E-genes normally activated by A as the "activated distribution" and to the distribution of expression changes for repressed genes as the "inhibited distribution." An E-gene e's expression value under the ΔA replicate was generated by sampling from the activated, inhibited, or unaffected distributions depending on whether the path from A to e was inhibitory, stimulatory, or not connected (i.e. either e was not downstream of A or e was disconnected from the network completely). Replicated hybridizations were simulated by repeating the above procedure R times.

To model incomplete knowledge of the network, a subset of S-genes was selected from T such that the resulting set of S-genes had size $k \cdot |T|$ for $0 < k \leq 1$. Only data from this subset was used to infer networks.

The parameters λ, k, and R were varied for generating artificial networks. The number of E-genes per S-gene was set to 20 and cycles were always present in the network to model feedback loops that are often present in real biological networks.

Unless otherwise noted, the number of S-genes in the network was sampled uniformly in the range of five to 15 and all S-genes were used in network recovery.

The E-gene expression distributions were modeled as Gaussians with unit variance and means equal to zero for the unaffected distribution, 1.75 for the inhibited distribution, and 1.75 for the activated distribution. Note that the activated distribution has a negative expression change on average since E-genes normally activated by a signal have lower expression levels relative to wild-type in knock-outs that block the activating signal. By the same reasoning, inhibited E-genes have positive expression changes relative to wild-type if a deletion blocks a normally inhibiting signal.

To estimate realistic levels of separation between the activated and inhibited distributions, I calculated the difference in mean expression of genes residing in pathways with somewhat opposing operations in the cell, such as subunits of the proteasome or ribosome. While the cell may turn on genes in both pathways, some conditions might favor protein expression over degradation (or vice versa). I reasoned that these conditions provide an estimate of the average difference in expression of an affected (up- or down-regulated) gene compared to an unaffected gene. I collected expression data for all of the genes in the proteasome and ribosome GO categories from a compendium of 3883 *H. sapiens*, 1049 *S. cerevisiae*, and 334 *D. melanogaster* microarray samples (see Figure 7.1). For each sample, a standardized absolute difference was calculated by computing the difference between the mean ribosome and proteasome levels and dividing the difference by the geometric mean of the respective standard deviations. The 95% quantile of the standardized absolute differences was found to be 1.75 and was used

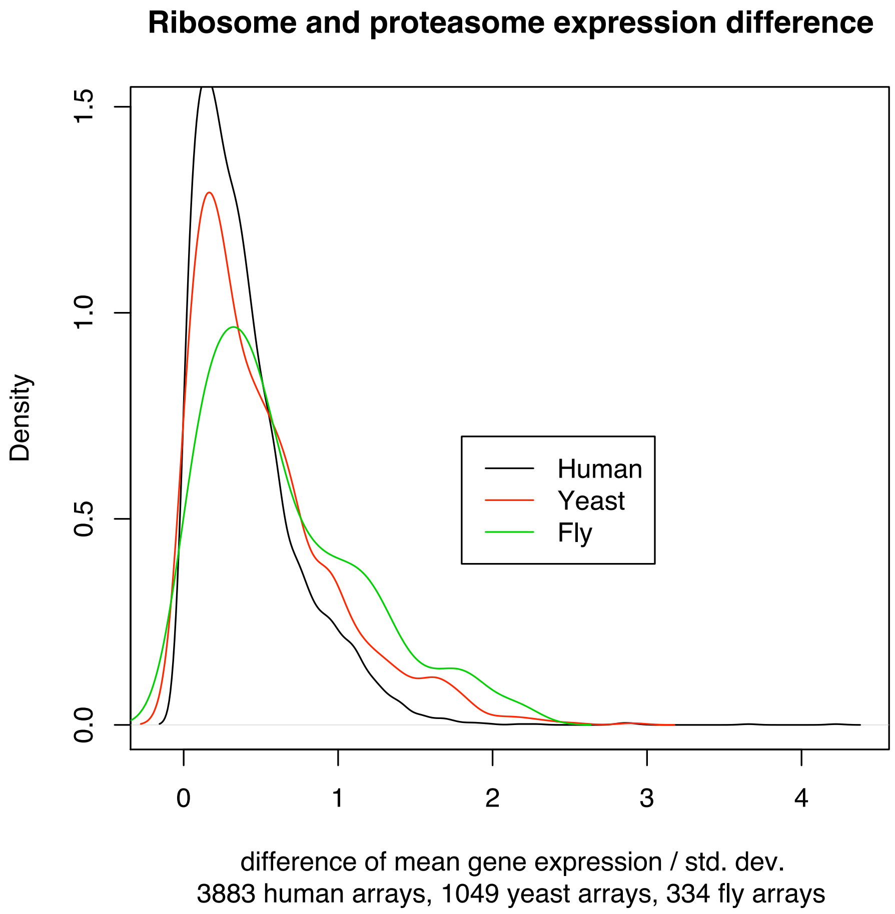


Figure 7.1: Estimating the difference between up- and down- distributions using proteasome- and ribosome- related genes as contrast sets. In *C. elegans*, *S. cerevisiae*, and *H. sapiens*, I determined members of the ribosome by collecting all genes with a GO annotation term that begins with "ribosome." Similarly, I constructed a proteasome gene set in each species from any gene annotations with a GO term beginning with "proteasome." I collected two-channel microarrays from databases and supplementary material for a total of 334 *C. elegans*, 1049 *S. cerevisiae*, and 3883 *H. sapiens* microarrays. Within each microarray I calculated the mean and standard deviation of the set of ribosome genes and the set of proteasome genes. For each microarray I then calculated the absolute value of the difference of the mean of ribosome and proteasome expression, and divided by the geometric means of the standard deviations. Finally, I plotted density estimates for each species.

as the separation between the activated (or inhibited) distribution and the unaffected distribution for simulating E-gene expression changes. To avoid detecting differences in method performance due to fitting Gaussian parameters to simulated E-gene responses, the parameters for the expression factors in the FG-NEM and uFG-NEM were set to the same values used during simulation.

7.1.1 Calculating Network Structure Recovery Accuracy

I calculated the ability of FG-NEM and variants to recover artificially generated S-gene networks using simulated E-gene data was measured. I simulated and predicted 500 networks, calculated the area under the precision-recall curve (AUC) for each predicted network, and recorded the mean and standard deviation of these AUCs.

The FG-NEM and uFG-NEM (see §7.2) methods associate a likelihood score to each interaction mode for each pair of S-genes. Predicted networks for each method were produced by sorting the list of interactions by their mode preference scores and then keeping interactions with scores above a threshold. A prediction of an interaction between S-genes A and B was considered a true-positive if there was a direct or indirect path between A and B in the generated network having the same direction. For example, the prediction $A \rightarrow B$ would be considered correct if the path $A \rightarrow C \rightarrow B$ was present in the generated network and would not be considered correct if instead $A \rightarrow C \leftarrow B$ was present. Note that because the top-scoring interaction mode for any pair was chosen, it is possible a method cannot reach a recall of 100% as an incorrect interaction will never be considered correct even for the most relaxed threshold. Precision was calculated as the fraction of true-positives out of all predicted positives, while recall was calculated as the fraction of true-positives out of all of gene pairs having a path in the generated network.

The area under the precision-recall curve (AUC) was calculated for each predicted network learned from artificial data. Predicted interactions were associated with their mode preference score, which was calculated by taking the log of the ratio of the likelihood of the top-scoring mode to the likelihood of the second best scoring mode. A precision-recall curve was then generated for a predicted network by ranking its interactions by their mode preference score and calculating the precision and recall for a series of score cutoffs.

7.1.2 Calculating Network Expansion Accuracy in Artificial Data

To evaluate the ability of a method to predict new genes involved in a network, I performed a type of leave-one-out cross-validation. Each S-gene was held out in turn by removing the simulated expression data associated with its knock-down and then obtaining a LAR score for the S-gene in the recovered network. Each S-gene has simulated expression changes under the other S-gene knock-downs and thus the held-out S-gene can be scored like an E-gene. For a given pathway with a set of S-genes, each gene A was iteratively removed from the list of S-genes and included as one of the E-genes. In each iteration, the knockout data for A was deleted, resulting in the data set X_{-A} representing the full matrix of expression for the pathway in which the columns (i.e. replicates) corresponding to the A knockout were removed. I used the percentiles of the LAR scores when comparing methods. The LAR percentile reflects how likely a held-out S-gene is attached to the network relative to all of the E-genes and is therefore more comparable across methods than using the LAR score directly.

The expression changes for the remaining S-genes were recorded in a reduced dataset X_{-A}. I used an artificial network of 32 genes from which eight S-genes were randomly selected for simulated knock-down under which expression changes were simulated. I then ran FG-NEM and unsigned variants (see next section) on X_{-A}, sorted all effects by their LAR, and recorded rank of effect A divided by the total number of effects. I also compared expansion performance to an unstructured method, Pavlidis's Template Matching (TM) [94]. This was performed on an augmented matrix $[X_{-A} : W]$ where W was a matrix of the same size as X_{-A} drawn from the unaffected distribution. TM compares a gene's expression vector to an "idealized" vector, in this case a vector with a value of 1 in entries that correspond to expression from a pathway knockdown, X_{-A}, and a 0 in entries that correspond to expression values from the control arrays, W. To perform TM expansion, the Pearson correlation of each row in $[X_{-A} : W]$ with the idealized is calculated, each row is sorted by descending correlation, and the percentile of the rank calculated.

7.2 Comparison to Unsigned FG-NEM variants and NEM

One of the key differences of the FG-NEM method is the addition of both signed interactions between S-genes and the modeling of signed responses. To estimate

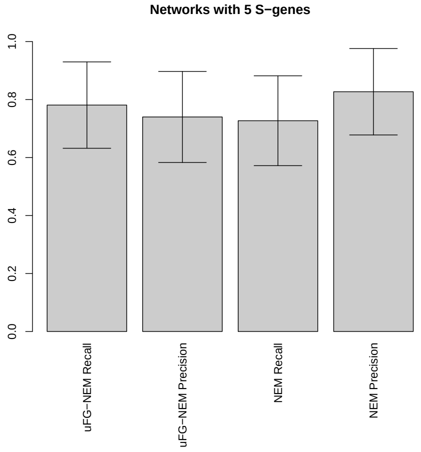


Figure 7.2: Comparison of uFG-NEM and exhaustive NEM model search for structure recovery. Each bar shows the mean and standard deviation of precision or recall for S-gene links over 100 artificial networks with five S-genes. Networks were generated with no inhibition and an average of 20 E-genes per S-gene. Expression was generated for one replicate using standard deviation of 1 and a mean for the down distribution of -1.75. Precision and recall for both uFG-NEM and exhaustive NEM search do not differ significantly.

the importance of these parts of the model, I compare FG-NEM results to two unsigned variants, FG-NEM AVT and uFG-NEM, to estimate the change in predictive power as a function of the introduction of sign. In effect, both variants consider four interaction modes: i) $A > B$; ii) $B > A$; iii) $A \neq B$; and $A = B$. For comparison purposes, a predicted unsigned interaction was treated as activation. In the FG-NEM AVT variant, FG-NEM is run on the absolute value of the data. This is a change of the data that simulates what happens in the original implementation of NEM. In the uFG-NEM method, I remove the component of FG-NEM which models repressive links between S-genes and E-genes. In effect, this makes both the top row and the right column of each interaction mode a disallowed region (see Figure 6.5).

In addition to comparison of FG-NEM to uFG-NEM, I also compared uFG-NEM to the original NEM algorithm on networks with five S-genes, the maximum tractable network size. Figure 7.2 shows similar performance between uFG-NEM and NEM on a set of 100 artificial networks. I conclude that the ability to operate on networks larger than five S-genes does not hamper FG-NEM's ability to predict networks of size five.

7.3 FG-NEM Structure Prediction Performance

I evaluated FG-NEM's ability to recover artificial networks from simulated data. Data was generated by propagating signals in networks containing simulated knock-downs and then sampling expression data from activated, inhibited, or unaffected expression change distributions (see §7.1). I focused on how the FG-NEM approach increased recovery of networks that contain both activation and inhibition. Because FG-NEMs explicitly incorporate inhibition, I hypothesized that they would recover networks containing an appreciable amount of inhibition more accurately than an approach lacking separate modes for inhibition and activation. The uFG-NEM method, as described in §7.2, implements this unsigned approach.

To make the comparison of FG-NEM to uFG-NEM fair, I measured network recovery in two ways. First, I calculated a measure of structure recovery: a predicted interaction was called correct if it matched an interaction (of either sign) in the simulated network. In this case, whether the interaction was inhibitory or stimulatory was

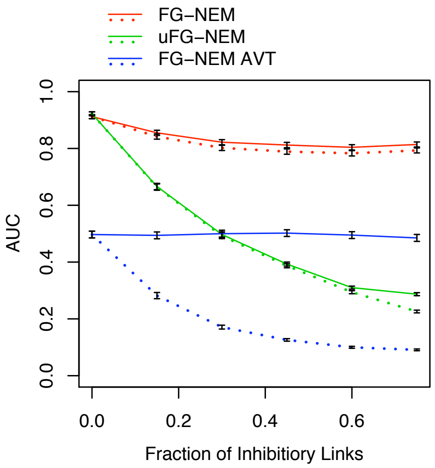


Figure 7.3: Influence of inhibition on network recovery. AUC (y-axis) plotted as a function of the percent of inhibitory links (x-axis). Four replicate hybridizations were used in all simulations. Points and error bars represent means and standard deviations computed across 500 synthetically generated networks respectively. Lines in each plot represent the performance of FG-NEM (red) and uFG-NEM run on the original data (green) or on absolute-value-transformed (AVT) data (blue) for both structure recovery (solid lines) and sign recovery (dotted lines).

ignored. Second, I measured sign recovery: a predicted interaction was recorded as correct if it matched an interaction in the simulated network and had the matching sign.

7.3.1 Varying Amount of Inhibition

I tested the ability of FG-NEM and uFG-NEM to recover the structure of networks simulated with varying fractions of inhibition, $0 \leq \lambda \leq 0.75$, for both the amount of inhibitory connections between S-genes and inhibitory attachments of E-genes. I simulated and predicted 500 networks, calculated the area under the precision-recall curve (AUC) for each predicted network (see §7.1.1), and recorded the mean and standard deviation of these AUCs. As expected, when no inhibition was present, FG-NEM and uFG-NEM were equivalent in terms of AUC when run on non-transformed data (Figure 7.3). Surprisingly, FG-NEM run on the AVT data performs much worse than FG-NEM even with no inhibition. This may be due to its interpretation of unaffected E-gene changes as affected changes which adds noise to its estimates of hierarchical nesting. As increasing amounts of inhibition is added into simulated networks, the performance of uFG-NEM degrades precipitously for structure recovery, under performing FG NEM by a margin of more than 0.20 units of AUC at the highest levels of simulated inhibition (Figure 7.3). Even at moderate levels of inhibition, for example at the 15% inhibition level, FG-NEM's AUC is already significantly higher than uFG-NEM's AUC. I also calculated the AUC for recovering the correct sign of the interactions for the unsigned models. In this case, unsigned interactions were interpreted to be activating interactions. As expected, the AUC decreases quadratically since both the precision and recall decrease linearly with increasing fraction of inhibition. Given these results, I expect FG-NEMs to have significantly better performance on real genetic networks where appreciable amounts of inhibition exist (see Figures 6.2).

I repeated the experiment of varying inhibition to match our expectations for application to the cancer invasion network discussed subsequently in Chapter 9. In the cancer invasion network the known S-genes were recovered in such a way that only activating S-gene connections were identified, however, there is still a large degree of inhibitory signaling to downstream effect genes. To simulate this situation, I created networks containing only activating S-gene interactions but varied the proportion of

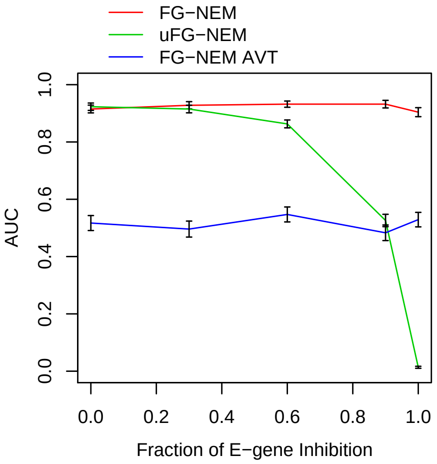


Figure 7.4: Network recovery as a function of increasing inhibition in E-gene attachment. In this experiment, only activating S-gene interactions were simulated while varying the proportion of inhibited E-genes. AUC (y-axis) for each method's ability to recover synthetic networks is plotted as a function of the fraction of E-genes having inhibitory attachments (x-axis). Points and error bars represent means and standard deviations, respectively, computed across 100 synthetic networks. Lines in each plot represent the performance of FG-NEM (red) and uFG-NEM run on the original data (green) or on AVT data (blue) for structure recovery. Four replicates were used. Network sizes were varied uniformly between 5 and 15 S-genes.

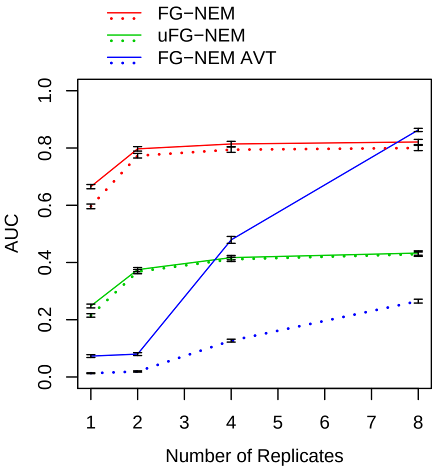


Figure 7.5: Influence of replicates in network recovery. AUC (y-axis) plotted as a function of varying numbers of microarray hybridization replicates (x-axis). Artificial networks contained 40% inhibitory interactions.

inhibiting E-gene attachments. Even in this situation where all of the known S-genes have activating interactions, FG-NEM's performance begins to significantly surpass uFG-NEM's performance when 40-60% of the E-genes are connected with inhibitory attachments (Figure 7.4). Thus, according to these simulations, even in cases where activation predominates the S-gene interactions, incorporating sign in the model for E-gene changes can lead to higher network recovery accuracies. I expect the signed FG-NEM to also perform well for the invasion network where 40-60% of the expression changes are consistent with inhibited E-gene attachments, as shown below.

7.3.2 Varying Number of Data Replicates

Uncertainty in microarray measurements is lessened by repeated hybridization either using biological or technical replicates. However, replication is costly or not possible in many situations. Because the datasets analyzed in this paper have at most two replicates, I was interested in testing the methods' abilities to recover networks from data containing few replicates. I compared the performance of FG-NEM and uFG-NEM for varying numbers of replicates (Figure 7.5). My results are consistent with Markowetz *et al.*'s findings[84] that the unsigned approach has a positive predictive value (PPV) about 0.1–0.15 higher when comparing four replicates to a single replicate. In contrast, the new FG-NEM methods performs well even if a single hybridization is available for each S-gene knock-down. With one to four replicates, the performance of FG-NEM was significantly higher than uFG-NEM using either the original or the AVT data. As the number of replicates increased to eight, the two methods achieved comparable performance for FG-NEM run on the AVT data. In a noisy data model, such as the one used in the synthetic data to model microarray expression, the distribution for effects that has a positive mean still has a tail with some degree of mass to the left of zero. The absolute value transformation confounds this tail with the no-effect distribution. However, replicate samples from the distribution can quickly correct the confounding of the distributions, explaining the increased performance of AVT with more replicates. To match the typical number of replicates included in microarray studies, four replicates were used for the rest of the artificial network experiments. At this number of replicates, FG-NEM significantly outperforms both uFG-NEM and FG-NEM AVT variants.

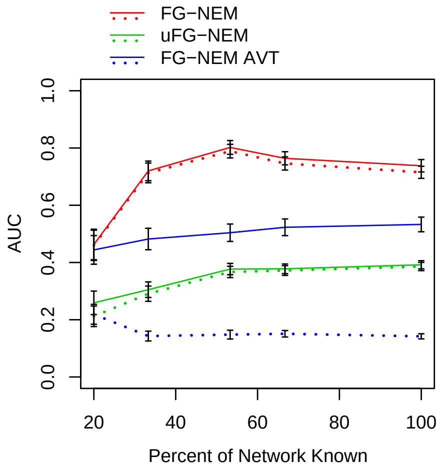


Figure 7.6: Influence of incomplete pathway knowledge on network recovery. Expression data for a subset of S-genes was made available for network recovery. AUC (y-axis) plotted as a function of the percent of S-genes having available data (x-axis). 40% inhibition and four replicates were used. Solid lines indicate AUC of S-graph connectivity. Dashed lines indicate AUC using both S-graph connectivity and sign.

7.3.3 Varying Fraction of Known Network

In practice, I expect studies of well-characterized pathways to include many of the S-genes while studies of poorly characterized pathways will lack many of the pathway's S-genes. A full network of fifteen genes was generated and then a subset of $(k \times 100)\%$ of the pathway genes was randomly selected from it for use as the set of S-genes. I tested the ability of both FG-NEMs and uFG-NEMs to recover the structure of the hidden network from S-gene subsets of different sizes by varying k.

As expected, the performance of both methods increased as the proportion of known S-genes was increased (Figure 7.6). Methods achieve their best performance when at least half of the S-genes are known. Performance decreases slightly as more of the network is known, as the prediction problem has more S-genes and therefore the hypothesis space is larger and the search problem becomes more difficult. This is promising as I can expect the best possible performance of the methods even if I use only half of the known underlying network. Conversely, all methods' performances were low when smaller proportions of known S-genes were used, especially at 20–33% (3–5 S-genes out of fifteen). Compared to more complete S-gene subsets, using fewer S-genes introduces longer expected distances between any two S-genes, forcing the models to infer a greater proportion of indirect interactions. In the simulated networks, longer paths between S-genes will be associated with weaker signals compared to shorter paths. Therefore, the expression changes have more chances to diverge from an idealized nesting relationship compared to shorter paths. The FG-NEM remains more accurate than the unsigned counterpart for known levels of 33% and higher. For example, when 33% of the network is known, the FG-NEM method achieves an average AUC of 0.72, which is 50% higher than the AUC achieved by its unsigned counterpart.

7.4 FG-NEM Pathway Expansion Performance

As pathway expansion is one of the goals of my pathway inference methods, I measured the ability of the FG-NEM method to expand the network to new genes involved in the pathway compared to a correlation-based method I refer to as Template Matching (TM) used by Irby *et al.* (2005)[62]. Briefly, Template Matching[94] ranks genes based on the correlation of their expression profiles to an idealized profile/template that reflects a phenotype of interest. TM has been used in several studies

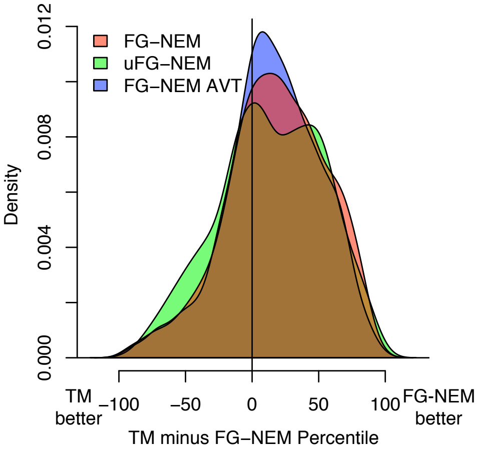


Figure 7.7: Accuracy of FG-NEM network expansion compared to Template Matching. The percentile of an S-gene obtained from Template Matching was subtracted from the percentile of the LAR score (see Methods) assigned by FG-NEM and uFG-NEM obtained from the leave-one-out expansion test. A smoothed histogram for FG-NEM (red), uFG-NEM run on the original data (green) and the AVT data (blue) was plotted and shows the proportion of S-genes (y-axis) with a particular difference in method percentile (x-axis). The underlying simulated network had 32 S-genes, eight S-genes were used for network recovery, and twenty E-genes were attached to each S-gene.

to identify genes with expression patterns that follow a series of phenotypes[5, 77]. I found that FG-NEMs significantly outperform TM when used to expand artificial networks (Figure 7.7). TM was compared to FG-NEM using a leave-one-out test in which knock-down data from one S-gene was removed from the dataset (see §7.1). I found that both FG-NEM and uFG-NEM rank a held-out signaling gene higher than TM on average. All three distributions of LAR percentile differences are shifted to the right of zero. The uFG-NEM exhibits a bi-modal performance improvement over TM, but I was unable to identify the source of this bi-modality. On average, FG-NEM predicts a held-out S-gene 25.3 (+/-15) percentile units higher than TM.

Chapter 8

Pathway Prediction and Expansion in *S. cerevisiae*

Due to easy laboratory manipulation, *S. cerevisiae* is perhaps the best understood eukaryote in terms of gene function and pathways. In particular, there is a large compendium expression profiles of gene knockout strains from Hughes *et al.*[57]. This compendium contains whole-genome expression profiles of 276 yeast gene-deletion mutants and p-values for differential gene expression.

8.1 Expression Data and Gene Sets

In each deletion strain, gene expression changes with a p-value smaller than 0.05 were selected, and then labeled as activated or inhibited according to the sign of their expression log-ratio. The p-values were converted to continuous expression values using the method of Yeang *et al.* (2004)[129]. This method replaces a p-value with a value obtained by inverting a Chi-square distribution. The value can be interpreted as a log-likelihood ratio reflecting the probability that an E-gene is expressed in the affected distribution compared to a background distribution. Gene sets, as proxies for pathways, were taken from Gene Ontology (GO)[4], KEGG[92] and Reactome[67] information. There were 25 non-redundant pathways selected that had at least five genes included as knock-outs in the knockout compendium. The largest pathway, chromosome organization and biogenesis, contained 45 S-genes. On a 2.83 GHz processor, factor graph inference using 5046 E-genes took a total of 1828 seconds. A pathway with 12 genes, such as nitrogen compound metabolism, took 38 seconds for network inference.

8.2 Pathway Expansion Performance

The accuracy of FG-NEMs for expanding each pathway to include new genes was measured. The likelihood of attachment ratio (LAR) score for each gene in the genome was calculated and the area under the precision-recall curve (AUC) was computed (see §7.1.2 for details). For each pathway, an AUC ratio was then calculated by dividing each method's AUC by the AUC calculated from randomly guessing E-genes for attachment to the network. Pathways sharing 25% or more of their genes with another pathway of higher AUC were ignored. Five non-redundant pathways were found that had AUCs significantly better than random guessing for at least one of the methods. Also included in the comparison is a sFG-NEM method, that includes priors

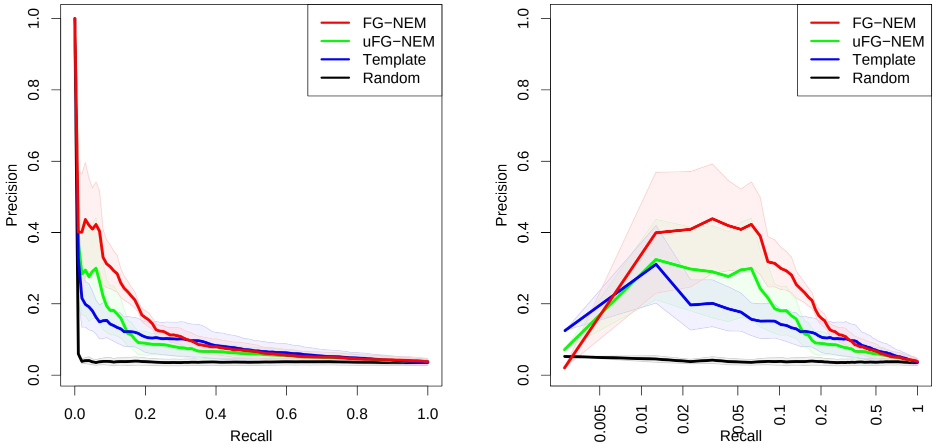


Figure 8.1: Yeast Pathway Expansion Precision/Recall Comparison. Each method's ability to expand a pathway was compared. Thick lines indicate mean precision and shaded regions represent standard error of mean calculated over the networks with the five highest AUCS from any of the tested methods.

on FG-NEM structures from high-throughput data. The details of this method are provided in the next section, §8.3.

While the precision of FG-NEM over uFG-NEM was not significant at any specific recall range, its overall higher precision across a broad range of recalls reflects a systematic improvement. Figure 8.1 shows the precision-recall curves averaged across these five pathways. The AUC ratios for the selected pathways are shown in Figure 8.2 and are sorted by the AUC achieved under the best-performing method.

Except for ribosome biogenesis, FG-NEM performed comparably or better than uFG-NEMs and TM (Figure 8.2 and Table 8.1). For sexual reproduction, ion homeostasis, and cell wall, FG-NEM outperformed the other methods by the largest margins, outperforming TM by a ratio of 4.17, 3.98, and 2.64 respectively. The signaling networks of both sexual reproduction and ion homeostasis consist of several inhibitory

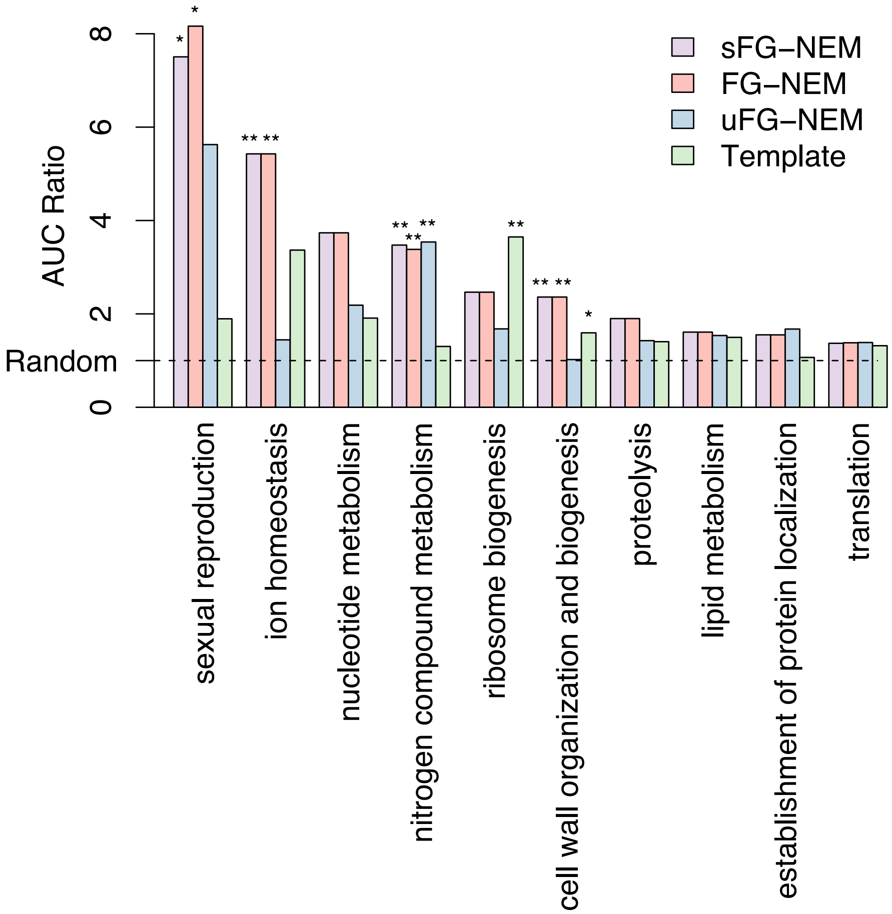


Figure 8.2: Yeast Expansion Performance AUC Comparison. Networks were predicted for a non-redundant set of GO categories containing four or more S-genes in the Hughes *et al.* (2000) compendium and used to predict held-out genes from the same category. The area under the curve (AUC) for each pathway was calculated for each method. AUC ratios (y-axis) were calculated for each method relative to the lowest AUC. Prediction methods that are significantly better than the lowest performing method, excluding random, at the 0.05 level (*) and 0.01 level (**) were determined by a proportions test on the top 30 predictions from each method.

Pathway	sFG-NEM	FG-NEM	uFG-NEM	Template
ribosome biogenesis	0.131	0.131	0.089	0.194
sexual reproduction	0.113	0.123	0.085	0.029
nitrogen compound metabolism	0.127	0.124	0.130	0.048
ion homeostasis	0.102	0.102	0.027	0.063
translation	0.087	0.087	0.088	0.084
cell wall organization and biogenesis	0.092	0.092	0.040	0.062
chromosome organization and biogenesis	0.069	0.069	0.081	0.075
nucleotide metabolism	0.056	0.056	0.033	0.029
vesicle-mediated transport	0.047	0.047	0.055	0.059
lipid metabolism	0.055	0.055	0.052	0.051
carbohydrate metabolism	0.049	0.049	0.038	0.037
establishment of protein localization	0.072	0.072	0.077	0.049
phosphorus metabolism	0.046	0.046	0.035	0.033
proteolysis	0.053	0.053	0.040	0.039
cytoskeleton organization and biogenesis	0.034	0.034	0.040	0.043
cellular respiration	0.017	0.017	0.013	0.016
response to osmotic stress	0.015	0.015	0.012	0.014
protein complex assembly	0.018	0.018	0.020	0.018
filamentous growth	0.015	0.015	0.015	0.015
positive regulation of transcription	0.017	0.017	0.018	0.016
interphase	0.019	0.018	0.020	0.018
glycoprotein metabolism	0.014	0.013	0.014	0.014
protein folding	0.014	0.014	0.015	0.013
regulation of enzyme activity	0.014	0.014	0.016	0.010
response to temperature stimulus	0.006	0.006	0.007	0.005

Table 8.1: Pathway expansion AUC on a yeast gene-deletion compendium. Gene Ontology categories are sorted in descending order according to their maximum expansion performance in any of the methods.

interactions[22, 25], consistent with FG-NEM's ability to capture negative as well as positive regulatory interactions. TM may perform the best on ribosome biogenesis because the proteins involved in ribosome assembly are all tightly co-regulated and their knock-outs lead to severe (and uninformative) effects. The signatures of expression changes for the ribosome biogenesis genes are not distinct from arbitrary genes because knocking out any of the ribosome biogenesis genes leads to drastic fitness defects in yeast and a concomitant alteration in gene expression to many genes in the genome.

8.3 Physical Structure Priors

The factor graph approach allows prior information to be incorporated. I tested a supervised variant of FG-NEMs (sFG-NEM) in which additional factors were incorporated to reward models that included known interactions. Three classes of physical data were downloaded for use as interaction priors: protein-DNA interactions, phosphorylation target data, and protein-protein interactions (PPI). Protein-DNA interactions with a p-value less than 0.001 were selected from the study of Lee, *et al.* (2002)[75]. Data describing kinase targets was taken from the study of Ptacek, *et al.* (2005)[99]. PPI data was downloaded from the BioGRID database[15] on July 30, 2008. For each gene set under study, I selected any interaction between S-genes in that category, resulting in 27 Protein-DNA interactions, 4 phosphorylation interactions, and 64 PPIs for the gene sets discussed in this chapter. For each unique physical interaction, I added an additional factor to the corresponding interaction variable to increase the likelihood of consistent interaction modes and decrease the likelihood of inconsistent modes.

I incorporated physical data as a prior on the interaction modes of an interaction variable in the factor graph. Some possible interaction modes are compatible with the evidence and some are not. I denote the set of interaction modes compatible with the evidence I^+ and the set that is incompatible I^-. For protein-DNA and phosphorylation data, where the protein of gene A binds the DNA of gene B or phosphorylates the product of gene B, $I^+_{AB} = \{\rightarrow, \dashv\}$ and $I^-_{AB} = \{\vdash, \neq, =, \leftarrow\}$. For PPI data, $I^+_{AB} = \{\rightarrow, \dashv, \vdash, =, \leftarrow\}$ and $I^-_{AB} = \{\neq\}$. If there is no interaction data for A and B, then all interaction modes are in I^+_{AB}. I then define the evidence priors as $\rho_{AB}(\phi_{AB}) = 1$ if $\phi_{AB} \in I^+_{AB}$ and $p = 10^{-7}$ otherwise. Varying the parameter p did not produce significantly different results.

Incorporating physical interaction priors showed little effect on network expansion performance. For most of the pathways, the performance of sFG-NEMs was indistinguishable from its unsupervised counterpart. A slight improvement was seen for the nitrogen metabolism pathway. Incorporation of structural priors adds activation from GLN3 to YEA4, and from ARG80 to ARG5,6, and slightly boosts the predictive power of the network. Thus, FG-NEM can usually identify new pathway genes in the unsupervised setting as well as when known interactions are provided.

Interestingly, the largest change in performance resulting from the use of prior information was a small drop observed for predicting genes involved in the sexual reproduction pathway. We investigated this decrease and found that using protein-DNA priors forced the placement of a transcription factor STE12 to the top of the pathway, whereas placement toward the bottom seemed to better fit the expression changes. Consequently, FG-NEM ranks the sexual reproduction E-genes higher than sFG-NEM.

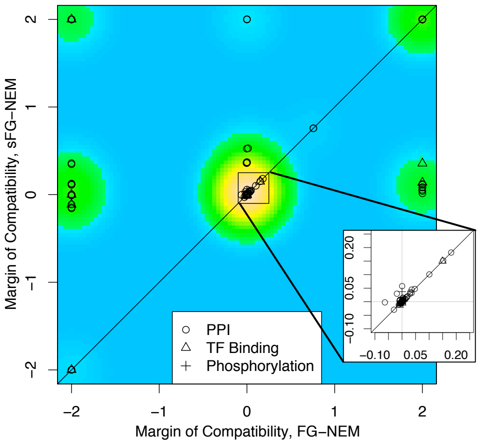


Figure 8.3: Compatibility of high-throughput physical evidence and predicted S-gene interactions. Each point is the margin of compatibility (MOC, see Eq. (8.1)) of a predicted genetic interaction to high-throughput physical interaction data when physical interaction evidence was used (y-axis) and when it was not used (x-axis). Coloring indicates two-dimensional density estimation of points. Inset shows detail of the highest density region.

8.4 Pathway Structure Prediction

On average, physical interaction priors increase the compatibility of FG-NEM predictions with high-throughput physical data. A leave-one-out analysis was used to test the ability of physical interaction data to improve pair-wise interaction predictions. To compare improvement in network structure prediction, I calculated the margin of compatibility (MOC) to reflect how well predicted interactions match held-out physical evidence.

Given the inferred log-likelihoods of each interaction mode $LL_{AB}(\phi_{AB})$, I define the compatible log-likelihood as $LL^+_{AB} = \max_{I \in I^+_{AB}} LL_{AB}(I)$ and the incompatible log-likelihood as $LL^-_{AB} = \max_{I \in I^-_{AB}} LL_{AB}(I)$. To compare pair-wise interaction predictions with physical evidence I defined the *margin of compatibility* (MOC) to be the difference in the log-likelihood of the most-likely interaction mode compatible with the evidence, LL^{+}_{AB}, relative to the most-likely mode incompatible with the evidence, LL^{-}_{AB}:

$$MOC_{AB} = \frac{LL^{+}_{AB} - LL^{-}_{AB}}{0.5\left(LL^{+}_{AB} + LL^{-}_{AB}\right)}. \tag{8.1}$$

Negative MOCs are assigned to predicted interactions that are incompatible with the physical evidence, while positive MOCs assigned to compatible predictions. For each held-out physical interaction, a network was computed using all other physical interaction data. Figure 8.3 shows the MOC of using priors plotted against the MOC without priors.

Of the 163 physical interactions, 104 (63%) have higher while 43 (26%) have lower MOC in sFG-NEM than FG-NEM. Of these 43, 33 have positive MOCs for both approaches (i.e. both agree with the physical evidence). Notably, of the 93 that achieved higher compatibilities in sFG-NEM, 38 (23%) became compatible only when the physical evidence was included. One example is the interaction between CDC42 and FAR1 in the sexual reproduction pathway. FAR1 acts downstream of CDC42 in the pheromone response signal cascade. The FAR1 gene deletion shows little expression change and is not placed downstream of CDC42 even though CDC42 is placed at the top of the signaling cascade by FG-NEM. With the inclusion of other structural priors, FAR1 is correctly placed downstream of CDC42. Thus, incorporating known interactions, even from possibly noisy high-throughput sources, can increase the likelihood of finding other interactions. However, the caveat is that such information

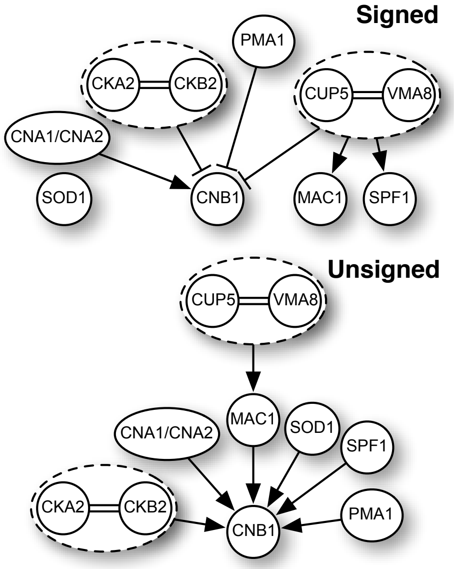


Figure 8.4: Predicted S-gene networks for the Ion Homeostasis pathway. Shown are predicted networks from the FG-NEM method (Signed) and the uFG-NEM method (Unsigned). Arrows indicate activating interactions and tees indicate inhibiting interactions. The absence of a link between a pair of S-genes indicates the most likely mode for the pair was the non-interaction mode. Equivalence interactions are indicated with double lines and S-genes connected by equivalence are grouped into dashed ovals.

may force a poorer fit to the observed expression data which could decrease the accuracy of frontier expansion.

8.5 Predicted inhibition in ion homeostasis pathway

FG-NEMs achieved significant improvement over the unsigned variant on the ion homeostasis pathway. To gain insights into the structural predictions underlying the difference in performance of the methods, I compared the predicted S-gene networks of the FG-NEM and uFG-NEM methods for this pathway (Figure 8.4). In budding yeast, calcineurin regulates gene expression and ion transport in response to calcium signals by dephosphorylating the transcription factor Crz1p, thus allowing Crz1p to rapidly translocate from the cytosol to the nucleus[110]. Conversely, the casein kinase homolog Hrr25p binds to and phosphorylates Crz1p to functionally antagonize calcineurin signaling in yeast[69]. FG-NEMs predicted an ion homeostasis gene network that is comprised of a number of biologically relevant links where CNA1 stimulates CNB1, the casein kinase 2 subunit genes CKA2 and CKB2 are equivalent and repress CNB1, and the vacuolar proton pump subunits CUP5 and VMA8 are likewise equivalent and repress CNB1.

Both the FG-NEM and uFG-NEM correctly predicted the equivalence of CKA2 and CKB2 which together form a complex. Of the top fifteen frontier genes predicted by FG-NEM, eight are annotated by GO as involved in ion homeostasis (Table 8.2), FRE2 is involved in ion transport, YGL039W is an oxidoreductase, and ARO9 is involved in amino acid catabolism. In contrast, only one of the top uFG-NEM frontier genes (Table 8.3), GRX4, is annotated by GO as involved in ion homeostasis. Examining the top 20 true positives predicted to be attached by FG-NEM, 19 were found to be predicted to be repressed by their S-gene. These true positives were not predicted to be attached to the network by uFG-NEM. Thus, the inability to make use of the explicit depression of E-genes may contribute to the poorer performance of the unsigned method.

ORF	GO Ann.	LAR	NAME	Truncated description
YKL220C	-	17.66	FRE2	Ferric reductase and cupric reductase
YGL039W	-	14.85		Oxidoreductase
YGR043C	-	14.17	NQM1	Protein of unknown function
YDR270W	+	13.84	CCC2	Cu(+2)-transporting P-type ATPase
YOR381W	+	13.80	FRE3	Ferric reductase
YHL040C	+	13.74	ARN1	Transporter, responsible for uptake of iron
YER145C	+	12.12	FTR1	High affinity iron permease
YLR205C	+	12.01	HMX1	ER localized, heme-binding peroxidase
YMR090W	-	11.44		Putative protein of unknown function
YHR137W	-	11.17	ARO9	Aromatic aminotransferase II
YEL065W	+	11.14	SIT1	Ferrioxamine B transporter
YOR384W	-	10.94	FRE5	Putative ferric reductase
YOR338W	-	10.62		Putative protein of unknown function
YLR136C	+	10.55	TIS11	mRNA-binding protein
YOL158C	+	10.39	ENB1	Endosomal ferric enterobactin transporter

Table 8.2: FG-NEM Ion Homeostasis pathway expansion frontier. The table is sorted by decreasing natural log of likelihood attachment ratio. The symbol + in the GO Ann. column indicates that the E-gene is annotated as a member of ion homeostasis by Gene Ontology.

ORF	GO Ann.		NAME	Truncated description
YGL009C	-	12.08	LEU1	Isopropylmalate isomerase
YER174C	+	9.66	GRX4	Glutathione-dependent oxidoreductase
YMR120C	-	8.55	ADE17	Enzyme of 'de novo' purine biosynthesis
YGR286C	-	7.52	BIO2	Biotin synthase
YGR234W	-	7.27	YHB1	Nitric oxide oxidoreductase
YNR074C	-	6.87	AIF1	Mitochondrial cell death effector
YOR356W	-	6.37		Mitochondrial protein
YJR048W	-	5.93	CYC1	Cytochrome c, isoform 1
YIL015W	-	5.73	BAR1	Aspartyl protease
YDL171C	-	5.46	GLT1	NAD(+)-dependent glutamate synthase
YER001W	-	4.96	MNN1	Alpha-1,3-mannosyltransferase
YPL274W	-	4.78	SAM3	High-affinity S-adenosylmethionine permease
YEL071W	-	4.35	DLD3	D-lactate dehydrogenase
YLR130C	-	4.20	ZRT2	Low-affinity zinc transporter
YHR216W	-	4.04	IMD2	Inosine monophosphate dehydrogenase

Table 8.3: uFG-NEM Ion Homeostasis pathway expansion frontier. The table is sorted by decreasing natural log of the likelihood attachment ratio. The symbol + in the GO Ann. column indicates that the E-gene is annotated as a member of ion homeostasis by Gene Ontology.

Chapter 9

Cancer Invasion Pathway Prediction and Expansion

Carcinogenesis involves a host of cell-cell communication breakdowns that include the loss of contact inhibition, an increased potential to proliferate, and the ability to invade and spread into foreign tissue. The molecular events involved in this transformation are still poorly understood. New systematic methods are needed to infer the key events responsible for these disease processes. The ability to measure gene expression changes for the entire genome in the presence of molecular perturbations, such as specific gene knock-downs, provides a new opportunity to infer gene networks in a data-driven manner.

Towards this end I applied the FG-NEM approach to a human colon cancer invasiveness network genes by Irby *et al.* (2005) [62]. In this work, the authors identified several "tiers" of genes implicated in the invasion process under the control of SRC kinase. Genes were included in a tier if their knock-downs were found to produce a significant drop in the invasive potential of HT29 colon cancer cells as defined by invasion through Matrigel. To identify additional genes involved in the invasion process, the authors measured gene expression under an RNA interference knock-down of each gene in the tier. Genes whose expression was lower in the knock-downs producing loss-of-invasiveness, and higher in knock-downs that did not produce loss-of-invasiveness, were considered candidates for inclusion in the next tier. In this fashion, each tier was formed by knocking-down each candidate gene and assaying for loss-of-invasion in Matrigel. In this chapter I describe using FG-NEM on the second tier genes to discover the third tier, and the network resulting from the third tier.

9.1 Previous Data and Methods

I applied the FG-NEM to the five S-genes from the second tier of Irby *et al.* (2005) [62]. These five human genes are cytokeratin 20 (KRT20), transcription factor Dp-1 (TFDP1), DEAH (Asp-Glu-Ala-His) box polypeptide 32 (DHX32), ribosomal protein L32 (RPL32), and glutaminase (GLS). Knock-down of each second-tier S-gene has been demonstrated to significantly reduce the invasion phenotype of HT29 colon cancer cells. KRT20 has historically served as a diagnostic marker for colorectal carcinoma [90], whereas high expression of ribosomal protein L32, glutaminase, and DEAD/H box polypeptides has been associated with various cancers and metastatic lesions [16, 135]. For this study, S-genes from the first tier were excluded as the expression profiles from the knock-down experiments were collected on a different microarray platform and therefore cross-platform normalization issues could potentially impact the results. The Expression Factor parameters were estimated from genes found to be up- or down-regulated by running the Statistical Analysis of Microarrays algorithm (SAM) [116], with a False Discovery Rate of 1%, on gene expression data collected on a panel of knock-downs. Using the differentially expressed genes yielded an estimate of 1.75 for the mean $\log_2$ ratio of the inhibited E-gene distribution and -1.75 for the activated E-gene distribution and standard deviations of 0.5. A mixture of Gaussians with these parameters was used for $\Pr(X_e|Y_e)$ (see §6.3). Several of these knock-downs led to loss-of-invasiveness while others produced invasive growth in the Matrigel assay as reported by Irby *et al.* (2005). The hybridization data and associated normalization information can be accessed from the Gene Expression Omnibus (GEO) database [6] under the series accession number GSE11848 and associated platform accession number GPL6978. A subset of this data containing the SAM-selected E-genes can be obtained from Dataset S1 in the supplementary materials from Vaske *et al.* (2009) [119].

9.2 Initial Network and Frontier

I selected E-genes that demonstrate a robust and significant effect under at least two of the knock-downs. Specifically, I choose only genes whose $\log_2$ ratios differ by less than 0.5 in replicate arrays and had an absolute $\log_2$ expression change at least equal to the mean absolute level of the activated distribution (1.75) in at least two

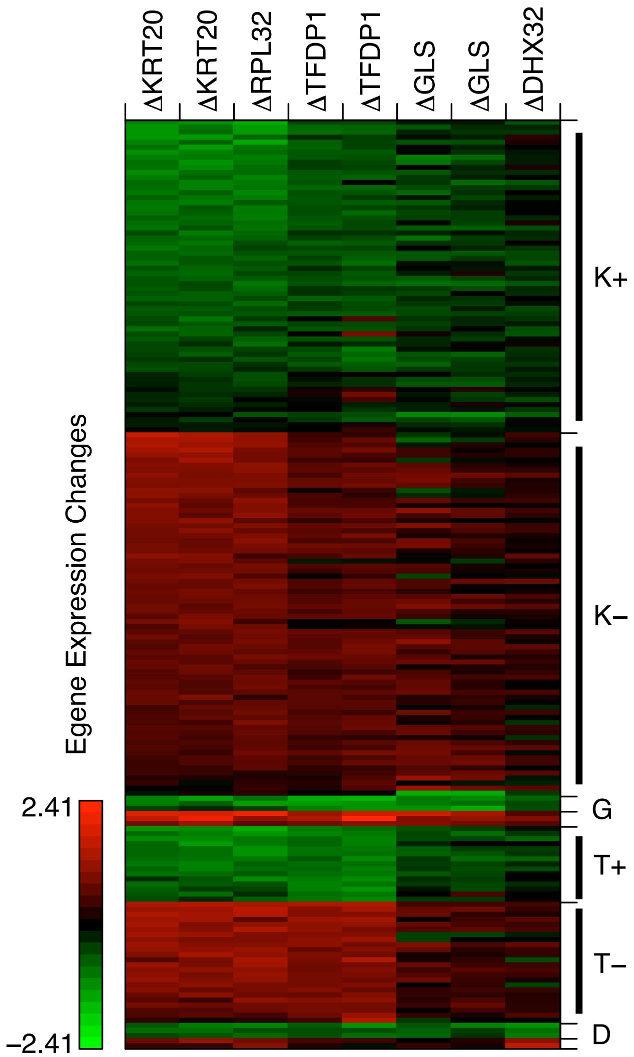


Figure 9.1: Expression changes of selected E-genes following targeted S-gene knockdowns in HT29 colon cancer cells. Gene expression was measured in HT29 cells treated with a shRNA specifically targeting an S-gene relative to cells treated with a scrambled control shRNA (Irby *et al.* 2005)[62]. Colors indicate putatively inhibited E-genes with up-regulated levels relative to control (red), activated E-genes with down-regulated levels relative to control (green), and unaffected E-genes with expression levels not significantly different from control (black). Genes were sorted by their attachment point and then by their LAR scores.

arrays. Using these criteria, I identified 185 E-genes to use for model inference. Figure 9.1 shows the expression data of these E-genes plotted in order of their predicted attachment points as identified by the highest scoring network model. For the most part, E-gene expression changes moved in the same direction following knock-down across the panel of five S-genes, indicating the presence of mostly stimulatory links among the S-genes. This is in contrast to Figure 6.3, where expression changes of a single E-gene move in the opposite direction following knock-down of S-genes connected by an inhibitory link. The absence of inhibitory links among S-genes is expected since, according to the selection criteria of previous studies, all of the S-genes were found previously to act in the same direction (invasion promotion). The method does find many inhibitory links to E-genes, which dramatically increases the fit of the model on the data points. These predicted attachment signs provide information about how an E-gene's involvement in the invasion process can be tested in follow-up experiments. The model predicts that invasion can be suppressed by knocking down genes connected by stimulatory attachments or by over-expressing genes connected by inhibitory attachments.

To calculate the significance of S-gene interactions predicted for the invasiveness network, I permuted the second tier data 1000 times by shuffling the data within microarray hybridizations and then calculating the maximum likelihood interaction mode for each S-gene pair under every data permutation. An empirical P value was calculated independently for each S-gene pair as the fraction of likelihoods from permuted runs that exceed the likelihood of the non-permuted run. To calculate the

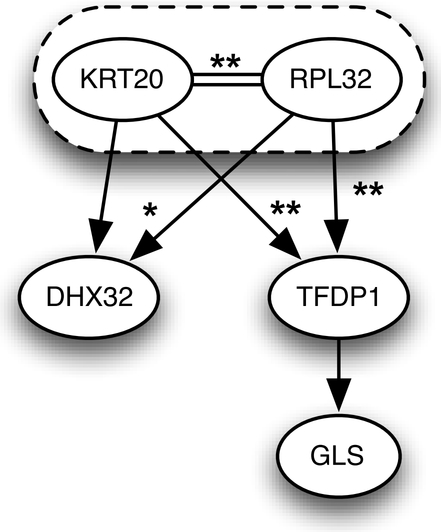


Figure 9.2: Cancer invasion network predicted by FG-NEM. For each pair of S-genes, the most likely interaction mode is shown. The same conventions used for illustrating interactions predicted for the yeast networks were used here. Some interactions were found to be significant at the 0.05 level (*) or 0.01 level (**) using a permutation test (see Methods). KRT20 and RPL32 were predicted to be equivalent and are therefore grouped together in a dashed oval.

significance of E-gene attachments, I permuted the data 1000 times by shuffling within gene rows, then predicting a network and calculating each E-gene's LAR. An empirical P value for E-gene attachment was calculated independently for each E-gene as the fraction of LARs for that gene that exceed the LAR on the non-permuted data.

FG-NEM recovered the network shown in Figure 9.2. KRT20 and RPL32 are predicted to be equivalent. Also, the model predicts TFDP1 and DHX32 are downstream of KRT20 and RPL32. The equivalent interaction of KRT20 and RPL32 received significantly high likelihoods ($p < 0.001$) as well as a strong excitatory downstream connection to TFDP1 ($p < 0.001$). There is a significant excitatory connection between KRT20/RPL32 and DHX32 based on one series of knock-down experiments specifically targeting KRT20 ($p = 0.006$), although a second knock-down experiment (using a silencing RNA differing from the first series that targets a different region of the KRT20 mRNA) resulted in a weaker connection ($p = 0.534$). Consequently, one could designate this link as deserving of follow-up functional studies (e.g. promoter analysis or chromatin immunoprecipitation). Though GLS is connected to the network, the likelihood of interaction was not strong enough to be significant. Hence, the GLS connection may require future knock-downs of additional S-genes coupled with gene expression profiling in order to resolve its tentative connection.

The FG-NEM model predicts that TFDP1 is at the bottom of the signaling cascade, which may reflect its role as part of the E2F transcriptional complex in targeting the expression of downstream genes that promote cell proliferation and invasion [60, 136]. The ribosomal subunit, RPL32 is curiously placed upstream of the DP1 transcription factor and at an equivalent level with the structural molecule KRT20. Aberrant expression of ribosomal proteins has been noted in a variety of cancers, although the molecular consequence of these expression changes is unknown [91]. It has been postulated that ribosomal proteins may play an important extraribosomal role (i.e. beyond translation) in the oncogenic transformation process [91].

Because the number of S-genes in the second tier is small, I compared the heuristic pair-wise search employed by FG-NEM to a random model search. If the heuristic approach is reasonable, it should identify network models that are among the highest scoring models identified by random sampling in less time. I generated 1000 random networks among the five second-tier genes. For each network, I calculated the data likelihood using message passing. Out of the 1000 randomly enumerated networks, the recovered network for the second-tier genes had a likelihood higher than 997 of the random networks. Interestingly, all three of the random networks with higher scores had identical structures to the network recovered by FG-NEM except that all three networks differed in their attachment of DHX32 and GLS. However, FG-NEM ran in only 1.96s whereas it took 232s to score the random models. This result demonstrates that the pair-wise heuristic search employed by FG NEM successfully identifies high-scoring networks in the space of all networks. While I need to test the trend for increasing network sizes, these results are promising for scaling up to larger networks in which random sampling will not be feasible.

9.3 Frontier Verification

I used the highest-scoring model recovered by FG-NEM to search for additional genes involved in colon cancer invasiveness by sorting each gene by its LAR score (see §6.6). I found 19 positive and 31 negative attachments with significant probabilities. Significance of the attachments was assessed by permuting each E-gene's observations, relearning a FG-NEM network, and computing its LAR score to construct an empirical null distribution of LARs. The E-genes with the highest attachment probabilities and positive LAR scores found to be significant via permutation testing are shown in Table 9.1.

Many of the genes in Table 9.1 have roles consistent with cancer cell invasion. For example, three E-genes encode proteases, including the metalloproteases ADAM9 and ADAM19. The metalloproteases represent a class of transmembrane proteins that are known facilitators of cell migration and invasion by proteolytic cleavage of extracellular matrix components [10]. Interestingly, ADAM21 is included among the first tier genes of Irby *et al.* (2005). This demonstrates that FG-NEM is able to identify two additional family members of this first tier gene even though it was not included in the S-gene set used in network learning. Glial fibrillary acid protein (GFAP) and Testes-specific protease 50 (TSP50) are also included in Table 9.1. GFAP is known to interact with the oncogenic tyrosine kinase SRC [102] and involved in astrocyte tumor invasiveness [19], while TSP50 has been shown to be differentially regulated in both breast and testicular cancer [127, 134]. Thus, FG NEMs predict that an expanded set of proteases may play a role in the colon cancer invasion process. Also included among the set of genes in our expanded invasion network is a second keratin family member, keratin 13 (KRT13), which is consistent with the previous identification of KRT20 in the second tier and may reflect a structural underpinning needed for invasion. Several of the genes in Table 9.1 represent novel connections of genes to the colon cancer invasiveness pathway. For example STK24, is a highly conserved protein whose homolog in *S. cerevisiae*, STE20, is involved in signal transduction of pseudo-hyphal growth [24]. It is intriguing to consider the possibility that part of the invasiveness pathway could be due in part to the aberrant regulation of an ancient cell migration process that dates back to single-cellular organisms.

The E-genes with positive LAR scores constitute the network "frontier" of the cancer invasiveness pathway in that they are predicted to directly interact with the second-tier genes. From among the 38 genes with positive and significant LAR scores, two were arbitrary selected to test for a loss-of-invasiveness phenotype in HT29 cells as defined by invasion in Matrigel. In collaboration with Norm Lee's laboratory at George Washington University we selected CAPN12 and expressed sequence tag AA099748 from Table 9.1 for gene knock-down experiments. CAPN12 is a member of the calpain gene family, which has been shown to have fibrillin activity. Genbank EST accession AA099748 aligns to the genome 3' to the gene CHMP4C, along with the EST AW440175, both from cancer tissues. Additionally, the amino acid translations of these ESTs align to the N-terminus of CHMP4C with 48% identity. The C-terminal tail of CHMP4C was recently shown [87] to be bound by the apoptosis inhibitor PDCD6IP,

LAR[a]	E-gene	S-gene	E-gene Description
18.79	CHORDC1[b]	GLS	cysteine and histidine-rich domain (CHORD)-containing 1
11.35	RNF32	GLS	ring finger protein 32
10.93	TSP50	TFDP1	testes-specific protease 50
10.02	HS3ST1[d]	KRT20	heparan sulfate (glucosamine) 3-O-sulfotransferase 1
6.85	CHMP4C[c]	TFDP1	chromatin modifying protein 4C
6.76	ADAM19[b]	KRT20	ADAM metallopeptidase domain 19 (meltrin beta)
6.34	CYP3A43	KRT20	cytochrome P450, family 3, subfamily A, polypeptide 43
5.97	SPTLC3[b]	TFDP1	serine palmitoyltransferase, long chain base subunit 3
5.25	PLEKHM3[b]	KRT20	pleckstrin homology domain containing, family M, member 3
4.92	KRT13	TFDP1	keratin 13
4.28	CAPN12	KRT20	calpain 12
3.87	C1orf34[b]	KRT20	hypothetical protein LOC22996
3.54	ZNF350	KRT20	zinc finger protein 350
3.53	ADAM9	TFDP1	ADAM metallopeptidase domain 9 (meltrin gamma)
2.75	SLC2A1[b]	KRT20	solute carrier fam. 2 (facilitated glucose transporter), member 1
2.38	TCTEX1D1	TFDP1	Tctex1 domain containing 1
2.23	STK24	KRT20	serine/threonine kinase 24 (STE20 homolog, yeast)
2.05	DDX58	KRT20	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58
2.01	GFAP	KRT20	glial fibrillary acidic protein

Table 9.1: Top frontier genes for colon cancer invasiveness ranked by LAR score (see Methods) and filtered for significance as determined by data permutation test. Note that measurements from the microarray expression data are of EST probes, and the E-gene column lists the gene that the EST maps to, or is closest to if the EST has not been associated with a gene model in the UCSC genome browser.

[a] Natural logarithm of likelihood of attachment score. [b] EST is inside an intron of this gene. [c] EST is on the 3' end of this gene. [d] EST is on the 5' end of this gene.

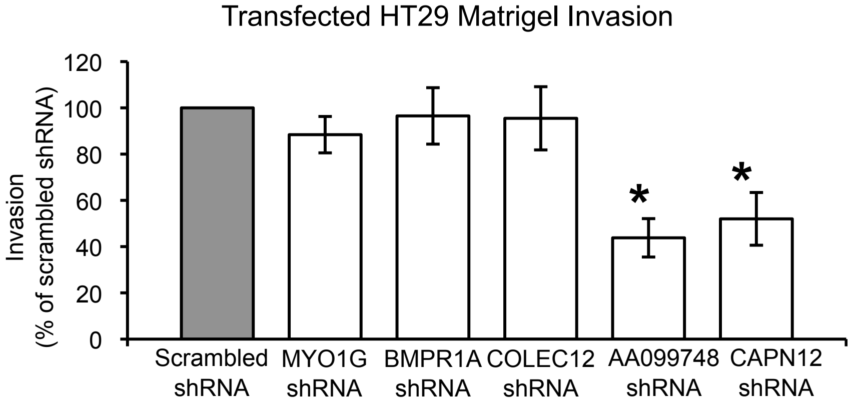


Figure 9.3: Matrigel Invasion Assay in HT29 Colon Cancer Cells. Genes predicted to be significantly attached to the network, CAPN12 and expressed sequence tag AA099748, resulted in a loss of the invasiveness phenotype when knocked-down by RNA interference. Genes not significantly attached to the network, MYO1G, BMPR1A, and COLEC12, did not result in significant loss of the invasive phenotype. A scrambled non-sense sequence also served as a negative control and did not result in a loss of HT29 cell invasiveness. Gene knock-downs in HT29 cells were validated by quantitative real time RT-PCR where mRNA levels of targeted genes were decreased by 70-80% compared to scrambled control shRNA-treated cells (data not shown). Data shown are the mean and S.E. of five independent experiments performed in quadruplicate. Treatments significantly different from scrambled control shRNA-treated cells ($P < 0.05$) by ANOVA and post hoc Tukey test are indicated by (*).

suggesting that the cancer-specific splice form of CHMP4C may have altered binding behavior with PDC6IP. PDC6IP also has been implicated in a broad array of membrane associated processes, including cell adhesion [104]. As negative controls, our collaborators performed knock-down experiments for three E-genes that had low attachment probabilities, namely MYO1G, BMPR1A and COLEC12. As correctly predicted by FG-NEM, both E-genes with high LAR scores produced significant loss of invasion while all three E-genes with low LAR scores did not lead to loss-of-invasion in the Matrigel assay (Figure 9.3).

9.4 Expanded Cancer Invasion Network

After the testing the positive and negative controls shown in Figure 9.3, our collaborators found an additional 12 genes from the cancer invasion frontier to be essential for invasion by a Matrigel assay. Expression profiles were generated from each knockdown, on a microarray platform that consists of a superset of the probes on the previous two platforms. The parameters for E-gene expression distributions were as in the previous tier: means were set to -1.75 and 1.75 for the positive and negative distributions, and all s.d. were 0.5.

In addition to the E-genes selected previously for network inference, E-gene expression spots were chosen from the expression data new to this study. A spot was selected if the means of expression under a knockdown varied significantly more than the expression means of each replicate number. Due to the large number of microarrays, the were necessarily run over a long period of time, and due to experimental artifact where arrays with similar replicate number will co-vary. For each spot on the microarray, the F-statistic $F = MST_{RNAi}/MST_{Replicate}$, where MST_{RNAi} is the mean sum of squares among the RNAi treatment grouping and $MST_{Replicate}$ is the mean sum of squares grouping according to the replicate number. E-gene spots significant at the 0.05 level from an F-distribution with df=(13,2) were included for network inference.

We predicted a signaling network between the invasiveness genes using FG-NEM (see Methods) on 1114 E-genes. The resulting is fully connected, and spans all three tiers of S-gene discovery, using results from three different microarray platforms.

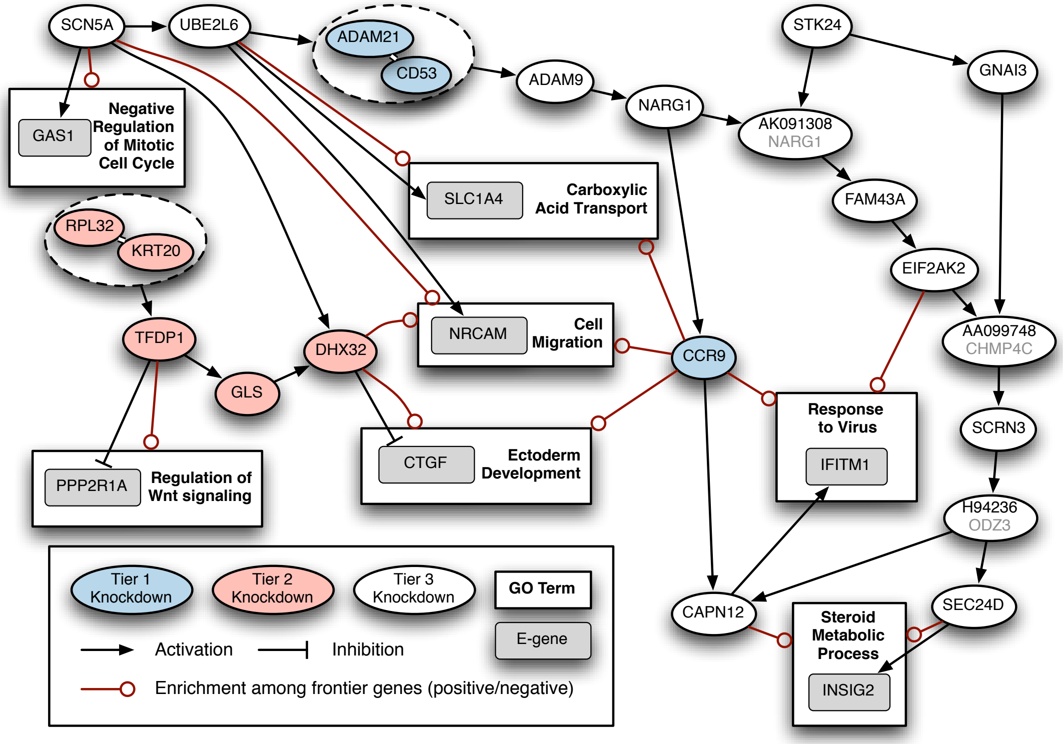


Figure 9.4: Inferred S-gene network and Frontier. Nodes represent S-genes (ovals), E-genes (gray boxes), and Gene Ontology categories (white boxes). Arrows indicate activation, and tees indicate repression. Circular line endings indicate GO set enrichment among both activated and inhibited E-genes.

I assessed the bootstrap confidence of inferred network features. For each of 1000 repetitions, we generated a sampled expression matrix of the same size as the original by selecting rows from the original expression matrix with replacement and inferred a signaling network. For any predicted interaction from the original data, the bootstrap confidence was calculated as the fraction of predictions from sampled matrices that had the same interaction mode as the original predicted interaction.

Each tier of knockdown data was performed on a different microarray platform, at a different time, and by a different technician. There is therefore a great deal of tier-specific signal which confounds attempts to unify the network. Though I was able to eliminate most of this effect for tiers one and three, tier two remains largely disconnected from the rest of the network. I believe that this is due largely to platform specific effects, and that if the tier two knockdowns were replicated on the same platform as the other knockdowns, the tier two genes would be more integrated with the rest of the inferred invasion network.

The predicted signaling network has three entry points: SCN5A, STK24, and KRT20/RPL32. I characterize the network into four topological and functional domains, shown in Figure 9.4. Domain I consists of integral membrane proteins and proteases, and is upstream of all other network domains. Domain II appears to regulate Wnt signaling. Domain III exhibits calcium dependent signaling and protease behavior. Domain IV is downstream of domains I and II, has an independent signaling input from STK24, and contains genes related to secretion.

9.4.1 Membrane/protease domain

The predicted invasion network in the neighborhood of the SCN5A entry point is enriched for membrane proteins, and in particular plasma membrane proteins. On this branch, SCN5A, ADAM21, CD53, and ADAM9 are known to be membrane proteins, and only UBE2L6 is not. The other membrane-associated proteins, ODZ3 and SEC24D are located together at the end of Domain IV and appear to be related to secretion.

9.4.2 Calcium domain

The branch from NARG1 to CCR9 and CAPN12 consists of a putative calcium regulatory channel. Calcium gradients are observed in motile cells, and many essential cell motility proteins in the leading edge require calcium-activation. In addition, cell motility on the leading edge is characterized by flickers of temporarily elevated calcium concentrations [123]. One source of these flickers has been found to be a membrane-tension-gated ion channel. NARG1 is regulated by NMDA receptors, including GRIN1, which admits CA2+ ions into cell and is gated by glutamate. In other cell types, GRIN1 functionally controls lymphocyte activation and are thought to be critical in synaptic plasticity, indicating that GRIN1 may be another neural protein involved in cancer, and which regulates NARG1. We predict NARG1 to signal to the chemokine receptor CCR9, which is known to elevate cytosolic calcium levels when activated by CCL25. The gene product of CAPN12, calpain 12, is a calcium activated protease, which we predict to be regulated by CCR9. Though calpain 12 is not implicated in invasion, other members of the calpain family have been found to regulate integrin-cytoskeletal interactions [PMID: 8999848] and synaptic plasticity [17].

9.4.3 Secretory/trafficking domain

The network branch from STK24 and GNAI3 down to SEC24D consist of signaling proteins and kinases (STK24, GNAI3, EIF2AK2, ODZ), a member of the ESCRT-III complex which is involved in endcytosis and trafficking of multivessicular bodies (CHMP4C), an exocytosis regulator (SCRN3) [122], and a protein involved in

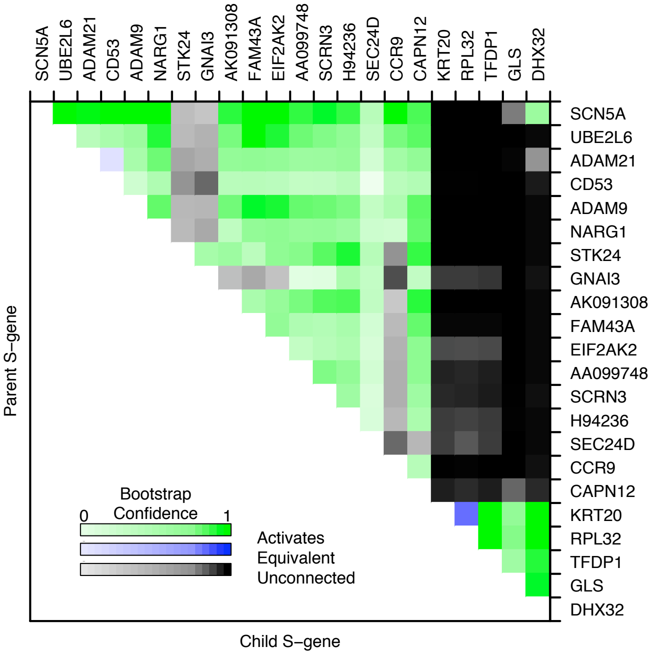


Figure 9.5: S-gene interaction confidence. Each pixel in the heatmap corresponds to an S-gene interaction's bootstrap confidence. For each interaction, the parent S-gene is labeled to the rich, and the child S-gene is labeled to the bottom. Note that though NEM include all transitive interaction, they are not displayed in (B) for simplicity. Therefore, a row shows bootstrap confidence of an S-gene being upstream of other genes, and a column shows bootstrap confidence of a gene being downstream of other genes.

vesicle trafficking and export from the endoplasmic reticulum (SEC24D).

9.5 Expanded Cancer Invasion Frontier

The predicted signaling network was used to predict the specific attachments of each gene in the genome. We found 1752 genes attached to the invasion network at the FDR of 5%, using a permutation test as in the previous tier. We performed a Gene Ontology enrichment analysis for each connection point in the network, and

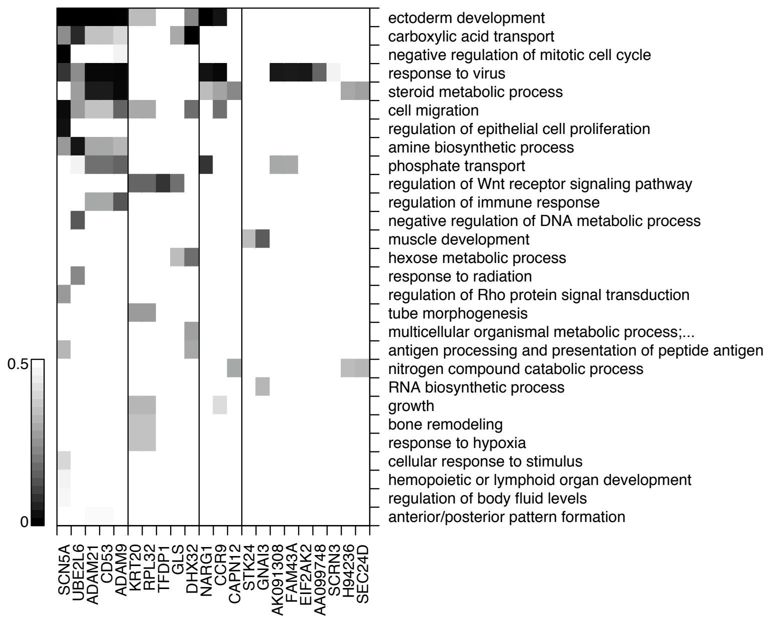


Figure 9.6:

found many GO categories enriched at the FDR of 5%. For Gene Ontology (GO) [4] enrichment analysis of the frontier E-genes, a frontier set was constructed for each S-gene consisting of any attached E-gene that had a positive LAR for inhibitory or activating attachment. Identifiers were mapped to Entrez Gene for both frontier sets and GO sets. Enrichment p-values were calculated for every frontier set-GO set overlap using the hypergeometric distribution as the null distribution, and the GO sets were sorted by their maximum enrichment p-value with any frontier set. The list of GO sets was filtered by 1) removing sets with fewer than 10 or more than 500 members and 2) removing any set whose intersection with a GO set with better enrichment p-value is 10% of the size of either set. Using the p-values for enrichment between all frontier sets and this filtered list of GO sets, I calculated the q-values using the QVALUE package in R with default parameters. A q-value is the minimum false discovery rate at which a test can be considered significant. Figure 9.6 shows the most enriched GO categories and the enrichment at each attachment point.

The gene with the highest Likelihood Attachment Ratio (LAR) was IFITM1. In head and neck squamous cell cancer [49] and in gastric cancer [128], increasing gene product levels of IFITM1 increases invasivity, and suppressing IFITM1 decreases invasivity.

The most significantly enriched GO term was Ectoderm Development. The network predicts that the invasion network represses some members of this set and activates others, mostly collagens, laminins, keratins, and regulators of these structural genes. The gene CTGF, which produces the protein connective tissue growth factor, had the highest LAR in this GO term, and is predicted to be inhibited by the invasion network. CTGF produces an extracellular matrix protein, and in liver has been proposed as a master regulator of the epithelial-mesenchymal transition, and its predicted inhibition in our invasion network is consistent with CTGF's proposed fibrogenic activity.

The significant enrichment of "negative regulation of mitotic cell cycle" represents a novel hypothesis about cancer invasion. The activation of GAS1 and other cell cycle regulators suggests the possibility that cancer invasion requires inactivation of the cell cycle. GAS1 can modulate both cell proliferation and cell differentiation [85]. Some connections in this network indicate that glutamate may play a role in invasion. GAS1 is induced by glutamate/NMDA receptor activation [88] in neurons, and in our network is predicted to be connected to SCN5A, also expressed in neurons. NARG1, another gene in the network, is activated by NMDA receptors. In addition the frontier gene SLC1A4, the top “carboxylic acid transport” hit, transports glutamate, in conjunction with sodium, and is downstream of sodium channel SCN5A.

Cell migration is significantly enriched in the invasion network frontier, with a q-value of 0.0226. Our network predicts several cell migration genes to be activated during colon cancer invasion, and the gene NRCAM has the highest LAR of these. Increased NRCAM expression has previously been shown to enhance cell motility and tumorigenesis [21]. The cell migration gene with the highest LAR, CCDC88A (also known as KIAA1212, GIV, and GIRDIN), is predicted to be inhibited by the invasion network. Previously, CCDC88A has been identified as the binding partner of our S-gene GNAI3 [46], where it has been identified as essential for leading-edge formation. We observe repression of CCDC88A from the probes of three distinct ESTs, covering all known splice forms of the extensively alternatively-spliced CCDC88A. This indicates that we may predict a novel role for CCDC88A in colon cancer invasion.

Genes annotated for the GO term steroid metabolic process were also significantly enriched in the network frontier. The colon cancer marker INSIG2 had the strongest connection to the invasion network and has previously been shown to promote invasion in an over-expression assay [79]. Export of posttranslationally modified peptides serve both as both repellent and attractive cues [101], and enrichment of this GO term in the export/trafficking S-gene domain are suggestive of similar activity for colon cancer invasion.

Chapter 10

Discussion

10.1 Conclusions

In this dissertation I have developed two separate methods for inferring large genetic networks from the downstream effects of perturbations. First, I presented an extension of Bayesian networks that can infer networks according to epistatic reasoning. This method, the Joint Intervention Network (JIN), allows for more refined reasoning than can be accomplished by epistasis analysis alone. However, it is limited to inferences on a small number of genes. Second, I presented a method called Factor Graph-Nested Effects Model (FG-NEM), which can infer much larger networks than can JIN. In addition, it can use effects that are downstream of any network gene, rather than genes that are downstream of all network genes. However, the regulatory logic is simpler, and to infer more complex regulatory logic in the FG-NEM framework would require measurements under multiple simultaneous perturbations.

With the JIN method my collaborators and I were able to predict a regulatory network for *V. cholerae* biofilm that was consistent with known literature interactions. Further, the network was more useful for expanding the biofilm network than correlation-based methods. Of the top fifteen predicted biofilm genes from JIN, eight have independent evidence of involvement in biofilm.

I have applied FG-NEM on synthetic data, *S. cerevisiae* knockout data, and a colon cancer cell line. In each system, modeling the sign of interactions has improved network expansion performance, and where measurable, network structure inference accuracy. In synthetic data, modeling network sign allows reconstruction of the network with fewer data replicates than versions of FG-NEM that do not model sign. With *S. cerevisiae* knockdowns, FG-NEM is better at expanding functional Gene Ontology groups than using correlation. In a colon cancer cell line, FG-NEM predicted a total of fourteen new cancer invasion genes, and was able to distinguish between differentially expressed genes that are necessary for invasion and those that are not.

10.2 Future Directions

The work in this dissertation can be extended in many directions. More sophisticated methods may be used for prioritizing genes for network expansion. Next, the cellular network could be probed at either narrower or broader scopes than single proteins; i.e. the specifics of interactions could be probed by using narrower perturbations or the interconnectivity of gene modules in the cellular network could be probed by treating entire pathways as a single genetic unit. Finally, as large scale genomic efforts are directed at clinical cancer data, the pathway methods from this dissertation may be adapted to treat the genomic alterations of cancer as perturbations, and simultaneously learn both about cancer biology and the underlying human molecular network.

Recent work in active learning suggests another potential approach for choosing new network candidates. In this dissertation, I assumed genes whose expression levels are well-explained by the model are of more interest for subsequent rounds of experimentation. However, it is conceivable to test whether selecting genes based on reducing a measure of uncertainty across models leads to better gene selection as previously performed by Yeang *et al.* [129]. An "active learning" approach prioritizes knock-down experiments based on the reduction of expected entropy of high-scoring models. The "informative" experiments would effectively disambiguate the models which explain the existing data. Fewer experiments might then be needed to narrow down a unique model of the underlying system [131].

Framing FG-NEM expansion in the active learning context requires both a measure of entropy of models and a means of predicting the response to hypothetical interventions. The first requirement, measuring entropy of the inferred model space, is straightforward for graphical models. However, predicting responses of potential knockdown A is more difficult. This requires establishing new probability distributions that correspond to expected data from each possible true location of A in the network.

A different extension to the FG-NEM model could examine either larger or smaller genetic perturbations, to examine biological networks at either a narrower or broader scope. To examine networks at a narrower scope, genetic perturbations must effect a smaller unit than an entire protein. One way this could be pursued is to construct gene mutants that only disturb a single functional domain in a protein or a single active site or binding site. For example, it may be possible to more closely examine the chromatin remodeling machinery by perturbing only small parts of the proteins known to be part of the pathway, and then look at downstream effects from gene expression, histone modification, and nucleosome positioning. The model predicted by FG-NEM from such small scale perturbations may inform the order of or dependence between steps in chromatin remodeling.

Alternatively, it may be possible to establish the connectivity between larger genetic modules of the cellular network. Previous work has shown that gene function can be grouped into modules, such as in the work of Segal *et al.*[107]. Treating an entire module as a single genetic unit, and looking at perturbations of these larger modules rather than perturbations of single genes may reveal the regulatory structure connecting the modules. This may be particularly relevant when combined with clinical cancer data.

New genomic efforts on clinical tumor samples are provide a unique opportunity for pathway-based analysis by methods similar to those in this thesis. Cancer genomes undergo extensive mutation that involves deletion, amplification, and mutation of genes. Many recent studies [114] are measuring these alterations in addition to gene expression. By sampling across many different patients, the assorted genomic alterations could conceivably serve as perturbations in either the JIN or FG-NEM methods. Given a set of S-genes, patients could be clustered into groups with identical genomic alterations on those S-genes. Within each group of identical perturbations, expression that is consistently different compared to other samples can serves as the effects in a JIN or FG-NEM. Many pathways important in cancer have been extensively characterized, offering a better chance for verification of predicted pathways than in other model systems. Additionally, since pathways important to cancer have been so extensively studied, I could estimate the perturbation of each pathway outside of the context of JIN or FG-NEM, for example with the SPIA method [113]. I could then treat each pathway as a single genetic unit, use expression as downstream effects, and learn a network of regulation among the signaling pathways important in cancer. The resulting predicting would be an interpathway map of crosstalk and regulation of high level processes.

Bibliography

[1] S. M. Aji and R. J. McEliece. A general algorithm for distributing information in a graph. In *IEEE International Symposium on Information Theory*, page 6, 1997.

[2] S. M. Aji and R. J. McEliece. The generalized distributive law. *IEEE Transactions on Information Theory*, 46(2):325–343, 2000.

[3] Randy J. Arnold and James P. Reilly. Observation of *Escherichia coli* ribosomal proteins and their posttranslational modifications by mass spectrometry. *Analytical Biochemistry*, 269(1):105–112, 1999.

[4] 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, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

[5] Seetharaman Balasenthil, Anupama E. Gururaj, Amjad H. Talukder, Rozita Bagheri-Yarmand, Ty Arrington, Brian J. Haas, John C. Braisted, Insun Kim, Norman H. Lee, and Rakesh Kumar. Identification of Pax5 as a target of MTA1 in B-Cell lymphomas. *Cancer Research*, 67(15):7132–7138, 8 2007/8/1.

[6] Tanya Barrett and Ron Edgar. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods in Enzymology*, 411:352–369, 2006.

[7] Paul L. Bartel, Jennifer A. Roecklein, Dhruba Sen Gupta, and Stanley Fields. A protein linkage map of *Escherichia coli* bacteriophage T7. *Nature Genetics*, 12(1):72–77, 1996.

[8] W Bateson. *Mendel's Principles of Heredity.* Cambridge University Press, 1909.

[9] A. Baudin, O. Ozier-Kalogeropoulos, A. Denouel, F. Lacroute, and C. Cullin. A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae. Nucleic Acids Research*, 21(14):3329–3330, 1993.

[10] Brigitte Bauvois. Transmembrane proteases in cell growth and invasion: new contributors to angiogenesis? *Oncogene*, 23(2):317–329, Jan 2004.

[11] Claude Berrou, Alain Glavieux, and Punya Thitimajshima. Near Shannon limit error-correcting coding and decoding: Turbo-codes. In *IEEE International Conference on Communications*, volume 2, pages 1064–1070, 1993.

[12] Sinem Beyhan, Kivanc Bilecen, Sofie R Salama, Catharina Casper-Lindley, and Fitnat H Yildiz. Regulation of rugosity and biofilm formation in *Vibrio cholerae*: comparison of VpsT and VpsR regulons and epistasis analysis of vpsT, vpsR, and hapR. *Journal of Bacteriology*, 189(2):388–402, Jan 2007.

[13] Kivanc Bilecen and Fitnat H Yildiz. Identification of a calcium-controlled negative regulatory system affecting *Vibrio cholerae* biofilm formation. *Environmental Microbiology*, April 2009.

[14] Michael Boutros, Herve Agaisse, and Norbert Perrimon. Sequential activation of signaling pathways during innate immune responses in *Drosophila. Developmental Cell*, 3(5):711–722, 2002.

[15] Bobby-Joe Breitkreutz, Chris Stark, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Michael Livstone, Rose Oughtred, Daniel H. Lackner, Jurg Bahler, Valerie Wood, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2008 update. *Nucleic Acids Research*, 36:D637–640, January 2008.

[16] M Causevic, R G Hislop, N M Kernohan, F A Carey, R A Kay, R J Steele, and F V Fuller-Pace. Overexpression and poly-ubiquitylation of the DEAD-box RNA helicase p68 in colorectal tumours. *Oncogene*, 20(53):7734–7743, Nov 2001.

[17] S L Chan and M P Mattson. Caspase and calpain substrates: roles in synaptic plasticity and cell death. *Journal of Neuroscience Research*, 58(1):167–190, October 1999.

[18] Ing-Feng Chang. Mass spectrometry-based proteomic analysis of the epitope-tag affinity purified protein complexes in eukaryotes. *Proteomics*, 6(23):6158–6166, 2006.

[19] M H Chen, W K Yang, J Whang-Peng, L S Lee, and T S Huang. Differential inducibilities of GFAP expression, cytostasis and apoptosis in primary cultures of human astrocytic tumours. *Apoptosis*, 3(3):171–182, 1998.

[20] Melanie H. Cobb. MAP kinase pathways. *Progress in Biophysics & Molecular Biology*, 71:479–500, 1999.

[21] Maralice E. Conacci-Sorrell, Tamar Ben-Yedidia, Michael Shtutman, Elena Feinstein, Paz Einat, and Avri Ben-Ze'ev. Nr-CAM is a target gene of the β-catenin/LEF-1 pathway in melanoma and colon cancer and its expression en-

hances motility and confers tumorigenesis. *Genes & Development*, 16(16):2058–2072, August 2002.

[22] J G Cook, L Bardwell, S J Kron, and J Thorner. Two novel targets of the MAP kinase Kss1 are negative regulators of invasive growth in the yeast *Saccharomyces cerevisiae*. *Genes & Development*, 10(22):2831–2848, Nov 1996.

[23] Autumn A. Cuellar, Catherine M. Lloyd, Poul F. Nielsen, David P. Bullivant, David P. Nickerson, and Peter J. Hunter. An overview of CellML 1.1, a biological model description language. *SIMULATION*, 79(12):740–747, 2003.

[24] I Dan, N M Watanabe, and A Kusumi. The Ste20 group kinases as regulators of MAP kinase cascades. *Trends in Cell Biology*, 11(5):220–230, May 2001.

[25] E de Nadal, J Clotet, F Posas, R Serrano, N Gomez, and J Arino. The yeast halotolerance determinant Hal3p is an inhibitory subunit of the Ppz1p Ser/Thr protein phosphatase. *Proceedings of the National Academy of Sciences*, 95(13):7357–7362, Jun 1998.

[26] Bart Deplancke, Arnab Mukhopadhyay, Wanyuan Ao, Ahmed M Elewa, Christian A Grove, Natalia J Martinez, Reynaldo Sequerra, Lynn Doucette-Stamm, John S Reece-Hoyes, Ian A Hope, Heidi A Tissenbaum, Susan E Mango, and Albertha J M Walhout. A gene-centered *C. elegans* protein-dna interaction network. *Cell*, 125(6):1193–1205, 2006.

[27] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, July 1999.

[28] A Einhauer and A Jungbauer. The flag peptide, a versatile fusion tag for the purification of recombinant proteins. *Journal Biochemical Biophysical Methods*, 49(1-3):455–465, 2001.

[29] Ghia M Euskirchen, Joel S Rozowsky, Chia-Lin Wei, Wah Heng Lee, Zhengdong D Zhang, Stephen Hartman, Olof Emanuelsson, Viktor Stolc, Sherman Weissman, Mark B Gerstein, Yijun Ruan, and Michael Snyder. Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Research*, 17(6):898–909, 2007.

[30] S Fields and O Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, 1989.

[31] Stanley Fields and Rolf Sternglanz. The two-hybrid system: an assay for protein-protein interactions. *Trends in Genetics*, 10(8):286–292, August 1994.

[32] A Finney and M Hucka. Systems biology markup language: Level 2 and beyond. *Biochemical Society Transactions*, 31(Pt 6):1472–1473, 2003.

[33] A Fire, S Xu, M K Montgomery, S A Kostas, S E Driver, and C C Mello. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391(6669):806–811, 1998.

[34] Peter Fraser and Wendy Bickmore. Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447(7143):413–417, 2007.

[35] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[36] Brendan J. Frey and David J. C. MacKay. A revolution: Belief propagation in graphs with cycles. In *Adavnces in Neural Information Processing Systems*, volume 10. MIT Press, 1998.

[37] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

[38] H Fröhlich, M Fellmann, H Sültmann, A Poustka, and T Beissbarth. Estimating large scale signaling networks through nested effect models with intervention effects from microarray data. *Bioinformatics*, 2008 Jan 28.

[39] Karla Jean Fullner and John J. Mekalanos. Genetic characterization of a new type IV-A pilus gene cluster found in both classical and El Tor biotypes of *Vibrio cholerae*. *Infection and Immunity*, 67(3):1393–1404, 1999.

[40] Gallager. *Low Density Parity Check Codes*. M.I.T. Press, 1963.

[41] E Gari, L Piedrafita, M Aldea, and E Herrero. A set of vectors with a tetracycline-regulatable promoter system for modulated gene expression in *Saccharomyces cerevisiae*. *Yeast*, 13(9):837–848, 1997.

[42] Irit Gat-Viks and Ron Shamir. Refinement and expansion of signaling pathways: The osmotic response network in yeast. *Genome Research*, 17(3):358–367, 2007.

[43] Irit Gat-Viks, Amos Tanay, Daniela Raijman, and Ron Shamir. The factor graph network model for biological systems. *RECOMB*, pages 31–47, 2005.

[44] Irit Gat-Viks, Amos Tanay, Daniela Raijman, and Ron Shamir. A probabilistic methodology for integrating knowledge and experiments on biological networks. *Journal of Computational Biology*, 13(2):165–181, March 2006.

[45] Anne-Claude Gavin, Markus Bosche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jorg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Hofert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck,

Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[46] Pradipta Ghosh, Mikel Garcia-Marcos, Scott J. Bornheimer, and Marilyn G. Farquhar. Activation of Gαi3 triggers cell migration via regulation of GIV. *The Journal of Cell Biology*, 182(2):381–393, 7 2008/7/28.

[47] Guri Giaever, Angela M. Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Veronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno Andre, Adam P. Arkin, Anna Astromoff, Mohamed El Bakkoury, Rhonda Bangham, Rocio Benito, Sophie Brachat, Stefano Campanaro, Matt Curtiss, Karen Davis, Adam Deutschbauer, Karl-Dieter Entian, Patrick Flaherty, Francoise Foury, David J. Garfinkel, Mark Gerstein, Deanna Gotte, Ulrich Guldener, Johannes H. Hegemann, Svenja Hempel, Zelek Herman, Daniel F. Jaramillo, Diane E. Kelly, Steven L. Kelly, Peter Kotter, Darlene LaBonte, David C. Lamb, Ning Lan, Hong Liang, Hong Liao, Lucy Liu, Chuanyun Luo, Marc Lussier, Rong Mao, Patrice Menard, Siew Loon Ooi, Jose L. Revuelta, Christopher J. Roberts, Matthias Rose, Petra Ross-Macdonald, Bart Scherens, Greg Schimmack, Brenda Shafer, Daniel D. Shoemaker, Sharon Sookhai-Mahadeo, Reginald K. Storms, Jeffrey N. Strathern, Giorgio Valle, Marleen Voet, Guido Volckaert, Ching-yun Wang, Teresa R. Ward, Julie Wilhelmy, Elizabeth A. Winzeler, Yonghong Yang, Grace Yen, Elaine Youngman, Kexin Yu, Howard Bussey, Jef D. Boeke, Michael Snyder, Peter Philippsen, Ronald W. Davis, and Mark Johnston. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896):387–391, 2002.

[48] Daniel Gietz, Andrew St. Jean, Robin A. Woods, and Robert H. Schiestl. Improved method for high efficiency transformation of intact yeast cells. *Nucleic Acids Research*, 20(6):1425, 1992.

[49] Hiroko Hatano, Yasusei Kudo, Ikuko Ogawa, Takaaki Tsunematsu, Akira Kikuchi, Yoshimitsu Abiko, and Takashi Takata. IFN-induced transmembrane protein 1 promotes invasion at early stage of head and neck cancer progression. *Clinical Cancer Research*, 14(19):6097–6105, Oct 2008.

[50] Lin He and Gregory J Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5(7):522–531, 2004.

[51] Henning Hermjakob, Luisa Montecchi-Palazzi, Gary Bader, Jerome Wojcik, Lukasz Salwinski, Arnaud Ceol, Susan Moore, Sandra Orchard, Ugis Sarkans,

Christian von Mering, Bernd Roechert, Sylvain Poux, Eva Jung, Henning Mersch, Paul Kersey, Michael Lappe, Yixue Li, Rong Zeng, Debashis Rana, Macha Nikolski, Holger Husi, Christine Brun, K Shanker, Seth G N Grant, Chris Sander, Peer Bork, Weimin Zhu, Akhilesh Pandey, Alvis Brazma, Bernard Jacq, Marc Vidal, David Sherman, Pierre Legrain, Gianni Cesareni, Ioannis Xenarios, David Eisenberg, Boris Steipe, Chris Hogue, and Rolf Apweiler. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22(2):177–183, 2004.

[52] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, Lingyun Yang, Cheryl Wolting, Ian Donaldson, Soren Schandorff, Juanita Shewnarane, Mai Vo, Joanne Taggart, Marilyn Goudreault, Brenda Muskat, Cris Alfarano, Danielle Dewar, Zhen Lin, Katerina Michalickova, Andrew R Willems, Holly Sassi, Peter A Nielsen, Karina J Rasmussen, Jens R Andersen, Lene E Johansen, Lykke H Hansen, Hans Jespersen, Alexandre Podtelejnikov, Eva Nielsen, Janne Crawford, Vibeke Poulsen, Birgitte D Sorensen, Jesper Matthiesen, Ronald C Hendrickson, Frank Gleeson, Tony Pawson, Michael F Moran, Daniel Durocher, Matthias Mann, Christopher W V Hogue, Daniel Figeys, and Mike Tyers. Systematic identification of protein complexes in *Saccharomyces cerevisiae* mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[53] Christine E Horak and Michael Snyder. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods in Enzymology*, 350:469–483, 2002.

[54] Linda S. Huang and Paul W. Sternberg. Genetic dissection of developmental pathways. In The *C. elegans* Research Community, editor, *WormBook*. WormBook, 2006.

[55] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[56] Falk Huffner, Sebastian Wernicke, and Thomas Zichner. FASPAD: fast signaling pathway detection. *Bioinformatics*, 23(13):1708–1709, 2007.

[57] T R Hughes, M J Marton, A R Jones, C J Roberts, R Stoughton, C D Armour, H A Bennett, E Coffey, H Dai, Y D He, M J Kidd, A M King, M R Meyer, D Slade,

P Y Lum, S B Stepaniants, D D Shoemaker, D Gachotte, K Chakraburtty, J Simon, M Bard, and S H Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–26, 2000.

[58] Albert Y Hung and Morgan Sheng. PDZ domains: structural modules for protein complex assembly. *The Journal of Biological Chemistry*, 277(8):5699–5702, 2002.

[59] Gyorgy Hutvagner and Phillip D. Zamore. RNAi: nature abhors a double-strand. *Current Opinion in Genetics & Development*, 12(2):225–232, 2002.

[60] Phillip J Iaquinta and Jacqueline A Lees. Life and death decisions by the E2F transcription factors. *Current opinion in cell biology*, 19(6):649–657, Dec 2007.

[61] Soren Impey, Sean R McCorkle, Hyunjoo Cha-Molstad, Jami M Dwyer, Gregory S Yochum, Jeremy M Boss, Shannon McWeeney, John J Dunn, Gail Mandel, and Richard H Goodman. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell*, 119(7):1041–1054, 2004.

[62] Rosalyn B Irby, Renae L Malek, Greg Bloom, Jennifer Tsai, Noah Letwin, Bryan C Frank, Kathleen Verratti, Timothy J Yeatman, and Norman H Lee. Iterative microarray and RNA interference-based interrogation of the SRC-induced invasive phenotype. *Cancer Research*, 65(5):1814–1821, Mar 2005.

[63] T Ito, T Chiba, R Ozawa, M Yoshida, M Hattori, and Y Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[64] F Jacob and J Monod. Genentic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–56, 1961.

[65] P James, J Halladay, and E A Craig. Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics*, 144(4):1425–1436, 1996.

[66] Hongkai Ji, Steven A Vokes, and Wing H Wong. A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors. *Nucleic Acids Research*, 34(21):e146, 2006.

[67] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl. 1):D428–432, 1 2005.

[68] J T Kadonaga. Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell*, 92(3):307–313, 1998.

[69] Kimberly A Kafadar, Heng Zhu, Michael Snyder, and Martha S Cyert. Negative regulation of calcineurin signaling by Hrr25p, a yeast homolog of casein kinase i. *Genes & Development*, 17(21):2698–2708, Nov 2003.

[70] P N Kanabar, C J Vaske, C H Yeang, F H Yildiz, and J M Stuart. Inferring disease-related pathways using a probabilistic epistasis model. *Pacific Symposium Biocomputing*, pages 480–491, 2009.

[71] Ingrid M. Keseler, Julio Collado-Vides, Socorro Gama-Castro, John Ingraham, Suzanne Paley, Ian T. Paulsen, Martin Peralta-Gil, and Peter D. Karp. EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Research*, 33(suppl. 1):D334–337, 2005.

[72] M C King and A C Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, 1975.

[73] Kenneth S. Kosik. The neuronal microRNA system. *Nature Reviews Neuroscience*, 7(12):911–920, 2006.

[74] Kschischang, Frey, and Loeliger. Factor graphs and the sum-product algorithm. *IEEETIT: IEEE Transactions on Information Theory*, 47, 2001.

[75] Tong Ihn Lee, Nicola J Rinaldi, Francois Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, Julia Zeitlinger, Ezra G Jennings, Heather L Murray, D Benjamin Gordon, Bing Ren, John J Wyrick, Jean-Bosco Tagne, Thomas L Volkert, Ernest Fraenkel, David K Gifford, and Richard A Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[76] Ivica Letunic, Richard R Copley, Birgit Pils, Stefan Pinkert, Jorg Schultz, and Peer Bork. Smart 5: domains in the context of genomes and networks. *Nucleic Acids Research*, 34(Database issue):D257–60, 2006.

[77] Noah E. Letwin, Neri Kafkafi, Yoav Benjamini, Cheryl Mayo, Bryan C. Frank, Troung Luu, Norman H. Lee, and Greg I. Elmer. Combined application of behavior genetics and microarray analysis to identify regional expression themes and gene-behavior associations. *Journal of Neuroscience*, 26(20):5277–5287, 5 2006.

[78] Benjamin P. Lewis, Richard E. Green, and Steven E. Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences*, 100(1):189–192, 2003.

[79] Chang Gong Li, Mike Gruidl, Steven Eschrich, Susan McCarthy, Hong-Gang Wang, Mark G. Alexandrow, and Timothy J. Yeatman. Insig2 is associated with

colon tumorigenesis and inhibits Bax-mediated apoptosis. *International Journal of Cancer*, 123(2):273–282, 2008.

[80] Bentley Lim, Sinem Beyhan, James Meir, and Fitnat H. Yildiz. Cyclic-diGMP signal transduction systems in *Vibrio cholerae*: modulation of rugosity and biofilm formation. *Molecular Microbiology*, 60(2):331–348, 2006.

[81] Joanne S. Luciano. Pax of mind for pathway researchers. *Drug Discovery Today*, 10(13):937–942, 2005.

[82] David J. C. MacKay and Radford M Neal. Good codes based on very sparse matrices. In C. Boyd, editor, *Cryptography and Coding 5th IMA Conf.*, number 1025 in Lecture Notes in Computer Scienc, pages 100–111. Berlin, Germany: Springer, 1995.

[83] Florian Markowetz, Jacques Bloch, and Rainer Spang. Non-transcriptional pathway features reconstructed from secondary effects of rna interference. *Bioinformatics*, 21(21):4026–4032, 2005.

[84] Florian Markowetz, Dennis Kostka, Olga G. Troyanskaya, and Rainer Spang. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23(13):i305–312, 2007.

[85] David C Martinelli and Chen-Ming Fan. The role of Gas1 in embryonic development and its implications for human disease. *Cell Cycle*, 6(21):2650–2655, Nov 2007.

[86] Suresh Mathivanan, Balamurugan Periaswamy, T K B Gandhi, Kumaran Kandasamy, Shubha Suresh, Riaz Mohmood, Y L Ramachandra, and Akhilesh Pandey. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7 Suppl 5:S19, 2006.

[87] John McCullough, Robert D Fisher, Frank G Whitby, Wesley I Sundquist, and Christopher P Hill. ALIX-CHMP4 interactions in the human ESCRT pathway. *Proceedings of the National Academy of Sciences*, 105(22):7687–7691, Jun 2008.

[88] Britt Mellstrom, Valentin Cena, Monica Lamas, Carlos Perales, Carmen Gonzalez, and Jose R Naranjo. Gas1 is induced during and participates in excitotoxic neuronal death. *Molecular and Cellular Neurosciences*, 19(3):417–429, Mar 2002.

[89] M. Mezard, G. Parisi, and R. Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

[90] R Moll. Cytokeratins in the histological diagnosis of malignant tumors. *The International journal of biological markers*, 9(2):63–69, Apr-Jun 1994.

[91] H Naora. Involvement of ribosomal proteins in regulating cell growth and apoptosis: translational modulation or recruitment for extraribosomal activity? *Immunology and Cell Biology*, 77(3):197–205, Jun 1999.

[92] H Ogata, S Goto, K Sato, W Fujibuchi, H Bono, and M Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, 1 1999.

[93] Oved Ourfali, Tomer Shlomi, Trey Ideker, Eytan Ruppin, and Roded Sharan. SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, 23(13):i359–366, 2007.

[94] P Pavlidis and W S Noble. Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology*, 2(10), 2001.

[95] J. Pearl. *Causality: Models, Reasoning, and Inference*. Causality, by Judea Pearl, pp. 400. ISBN 0521773628. Cambridge, UK: Cambridge University Press, March 2000., March 2000.

[96] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[97] Suraj Peri, J Daniel Navarro, Ramars Amanchy, Troels Z Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan, Babylakshmi Muthusamy, T K B Gandhi, Mads Gronborg, Nieves Ibarrola, Nandan Deshpande, K Shanker, H N Shivashankar, B P Rashmi, M A Ramya, Zhixing Zhao, K N Chandrika, N Padma, H C Harsha, A J Yatish, M P Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Shubha Suresh, Neelanjana Ghosh, R Saravana, Sreenath Chandran, Subhalakshmi Krishna, Mary Joy, Sanjeev K Anand, V Madavan, Ansamma Joseph, Guang W Wong, William P Schiemann, Stefan N Constantinescu, Lily Huang, Roya Khosravi-Far, Hanno Steen, Muneesh Tewari, Saghi Ghaffari, Gerard C Blobe, Chi V Dang, Joe G N Garcia, Jonathan Pevsner, Ole N Jensen, Peter Roepstorff, Krishna S Deshpande, Arul M Chinnaiyan, Ada Hamosh, Aravinda Chakravarti, and Akhilesh Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371, 2003.

[98] Jeffrey A. Pleiss, Gregg B. Whitworth, Megan Bergkessel, and Christine Guthrie. Transcript specificity in Yeast Pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biology*, 5(4), 2007.

[99] Jason Ptacek, Geeta Devgan, Gregory Michaud, Heng Zhu, Xiaowei Zhu, Joseph Fasolo, Hong Guo, Ghil Jona, Ashton Breitkreutz, Richelle Sopko, Rhonda R McCartney, Martin C Schmidt, Najma Rachidi, Soo-Jung Lee, Angie S Mah, Lihao Meng, Michael J R Stark, David F Stern, Claudio De Virgilio, Mike Tyers, Brenda Andrews, Mark Gerstein, Barry Schweitzer, Paul F Predki, and Michael Snyder.

Global analysis of protein phosphorylation in yeast. *Nature*, 438(7068):679–684, 2005 Dec 1.

[100] W Reik and N D Allen. Genomic imprinting: Imprinting with and without methylation. *Current Biology*, 4(2):145–147, 1994.

[101] Sara Ricardo and Ruth Lehmann. An ABC transporter controls export of a drosophila germ cell attractant. *Science*, 323(5916):943–946, 2 2009.

[102] Jean-Francois Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amelie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005.

[103] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

[104] Mirko H H Schmidt, Baihua Chen, Lisa M Randazzo, and Oliver Bogler. SETA/CIN85/Ruk and its binding partner AIP1 associate with diverse cytoskeletal elements, including FAKs, and modulate cell adhesion. *Journal of Cell Science*, 116(Pt 14):2845–2855, Jul 2003.

[105] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[106] Jacob Scott, Trey Ideker, Richard M. Karp, and Roded Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *Journal of Computational Biology*, 13(2):133–144, 2006.

[107] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19:i273–282, 2003.

[108] Nicola Soranzo, Ginestra Bianconi, and Claudio Altafini. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics*, 23(13):1640–1647, 2007.

[109] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher.

Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[110] A Stathopoulos-Gerontides, J J Guo, and M S Cyert. Yeast calcineurin regulates nuclear localization of the Crz1p transcription factor through dephosphorylation. *Genes & Development*, 13(7):798–803, Apr 1999.

[111] Martin Steffen, Allegra Petti, John Aach, Patrik D'haeseleer, and George Church. Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3(1), 2002.

[112] Lena Stromback and Patrick Lambrix. Representations of molecular pathways: an evaluation of SBML, PSI-MI and BioPAX. *Bioinformatics*, 21(24):4401–4407, 2005.

[113] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jung-Sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, 2009 Jan 1.

[114] TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008 Oct 23.

[115] Achim Tresch and Florian Markowetz. Structure learning in nested effects models. *Statistical Applications in Genetics and Molecular Biology*, 7:Article9, 2008.

[116] 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, Apr 2001.

[117] P Uetz, L Giot, G Cagney, T A Mansfield, R S Judson, J R Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, A Qureshi-Emili, Y Li, B Godwin, D Conover, T Kalbfleisch, G Vijayadamodar, M Yang, M Johnston, S Fields, and J M Rothberg. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.

[118] Nancy Van Driessche, Janez Demsar, Ezgi O Booth, Paul Hill, Peter Juvan, Blaz Zupan, Adam Kuspa, and Gad Shaulsky. Epistasis analysis with global transcriptional phenotypes. *Nature Genetics*, 37(5):471–477, 2005.

[119] Charles J Vaske, Carrie House, Truong Luu, Bryan Frank, Chen-Hsiang Yeang, Norman H Lee, and Joshua M Stuart. A factor graph nested effects model to identify networks from genetic perturbations. *PLoS Computational Biology*, 5(1):e1000274, Jan 2009.

[120] M Vidal and P Legrain. Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Research*, 27(4):919–929, 1999.

[121] Andreas Wagner. Reconstructing pathways in large genetic networks from genetic perturbations. *Journal of Computational Biology*, 11(1):53–60, 2004.

[122] Gemma Way, Nicholas Morrice, Carl Smythe, and Antony J. O'Sullivan. Purification and identification of secernin, a novel cytosolic protein that regulates exocytosis in mast cells. *Molecular Biology of the Cell*, 13(9):3344–3354, 9 2002/9/1.

[123] Chaoliang Wei, Xianhua Wang, Min Chen, Kunfu Ouyang, Long-Sheng Song, and Heping Cheng. Calcium flickers steer cell migration. *Nature*, 457(7231):901–905, Feb. 2009.

[124] E A Winzeler, D D Shoemaker, A Astromoff, H Liang, K Anderson, B Andre, R Bangham, R Benito, J D Boeke, H Bussey, A M Chu, C Connelly, K Davis, F Dietrich, S W Dow, M El Bakkoury, F Foury, S H Friend, E Gentalen, G Giaever, J H Hegemann, T Jones, M Laub, H Liao, N Liebundguth, D J Lockhart, A Lucau-Danila, M Lussier, N M'Rabet, P Menard, M Mittmann, C Pai, C Rebischung, J L Revuelta, L Riles, C J Roberts, P Ross-MacDonald, B Scherens, M Snyder, S Sookhai-Mahadeo, R K Storms, S Veronneau, M Voet, G Volckaert, T R Ward, R Wysocki, G S Yen, K Yu, K Zimmermann, P Philippsen, M Johnston, and R W Davis. Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906, 1999.

[125] Ira G Wool. The structure and function of eukaryotic ribosoms. *Annual Review of Biochemistry*, 48:719–754, 1979.

[126] BioPAX working group. BioPAX–biological pathways exchange language. Documentation, 2004.

[127] Hao-Peng Xu, Liming Yuan, Jidong Shan, and Huail Feng. Localization and expression of TSP50 protein in human and rodent testes. *Urology*, 64(4):826–832, Oct 2004.

[128] Young Yang, Jeong-Hyung Lee, Kun Yong Kim, Hyun Keun Song, Jae Kwang Kim, Suk Ran Yoon, Daeho Cho, Kyu Sang Song, Young Ho Lee, and Inpyo Choi. The interferon-inducible 9-27 gene modulates the susceptibility to natural killer cells and the invasiveness of gastric cancer cells. *Cancer Letters*, 221(2):191–200, Apr 2005.

[129] Chen-Hsiang Yeang, Trey Ideker, and Tommi Jaakkola. Physical network models. *Journal of Computational Biology*, 11(2-3):243–262, 2004.

[130] Chen-Hsiang Yeang and Tommi Jaakkola. Physical network models and multi-source data integration. In *RECOMB*, pages 312–321, 2003.

[131] Chen-Hsiang Yeang, H Craig Mak, Scott McCuine, Christopher Workman, Tommi Jaakkola, and Trey Ideker. Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biology*, 6(7):R62, 2005.

[132] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 689–695. M.I.T. Press, 2001.

[133] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282–2312, 2005.

[134] L Yuan, J Shan, D De Risi, J Broome, J Lovecchio, D Gal, V Vinciguerra, and H P Xu. Isolation of a novel gene, TSP50, by a hypomethylated DNA fragment in human breast cancer. *Cancer Research*, 59(13):3215–3221, Jul 1999.

[135] Darwin Pinheiro Machado Zacharias, Manuela Maria Ramos Lima, Alcione Lescano Jr Souza, Ivan Dunshee de Abranches Oliveira Santos, Milvia Enokiara, Nilceo Michalany, and Rui Curi. Human cutaneous melanoma expresses a significant phosphate-dependent glutaminase activity: a comparison with the surrounding skin of the same patient. *Cell biochemistry and function*, 21(1):81–84, Mar 2003.

[136] S Y Zhang, S C Liu, D G Johnson, and A J Klein-Szanto. E2F-1 gene transfer enhances invasiveness of human head and neck carcinoma cell lines. *Cancer research*, 60(21):5972–5976, Nov 2000.

[137] Zhengdong D Zhang, Alberto Paccanaro, Yutao Fu, Sherman Weissman, Zhiping Weng, Joseph Chang, Michael Snyder, and Mark B Gerstein. Statistical analysis of the genomic distribution and correlation of regulatory elements in the encode regions. *Genome Research*, 17(6):787–797, 2007.