A constraint optimization framework for discovery of cellular signaling and regulatory networks

by

Shao-shan Carol Huang

B.Sc., University of British Columbia (2005)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUN 14 2011

LIBRARIES

Submitted to the Computational and Systems Biology Program in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

ARCHIVES

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

© Massachusetts Institute of Technology 2011. All rights reserved.

Author ....

Computational and Systems Biology Program
May 19, 2011

Certified by.......................................

nest Fraenkel
Associate Professor of Biological Engineering
Thesis Supervisor

Accepted by..

Christopher B. Burge
Professor of Biology and Biological Engineering
Director, Ph.D. Graduate Program

A constraint optimization framework for discovery of cellular signaling and regulatory networks

by

Shao-shan Carol Huang

Submitted to the Computational and Systems Biology Program
on May 19, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Cellular signaling and regulatory networks underlie fundamental biological processes such as growth, differentiation, and response to the environment. Although there are now various high-throughput methods for studying these processes, knowledge of them remains fragmentary. Typically, the majority of hits identified by transcriptional, proteomic, and genetic assays lie outside of the expected pathways. In addition, not all components in the regulatory networks can be exposed in one experiment because of systematic biases in the assays. These unexpected and hidden components of the cellular response are often the most interesting, because they can provide new insights into biological processes and potentially reveal new therapeutic approaches. However, they are also the most difficult to interpret. We present a technique, based on the Steiner tree problem, that uses a probabilistic protein-protein interaction network and high confidence measurement and prediction of protein-DNA interactions, to determine how these hits are organized into functionally coherent pathways, revealing many components of the cellular response that are not readily apparent in the original data. We report the results of applying this method to (1) phosphoproteomic and transcriptional data from the pheromone response in yeast, and (2) phosphoproteomic, DNaseI hypersensitivity sequencing and mRNA profiling data from the U87MG glioblastoma cell lines over-expressing the variant III mutant of the epidermal growth factor receptor (EGFRvIII). In both cases the method identifies changes in diverse cellular processes that extend far beyond the expected pathways. Analysis of the EGFRVIII network connectivity property and transcriptional regulators that link observed changes in protein phosphorylation and differential expression suggest a few intriguing hypotheses that may lead to improved therapeutic strategy for glioblastoma.

Thesis Supervisor: Ernest Fraenkel
Title: Associate Professor of Biological Engineering

Acknowledgments

First of all, I would like to acknowledge my generous funding sources. This project was supported by the National Cancer Institute grant to the Integrated Cancer Biology Program (ICBP) at MIT. I was funded by a National Institutes of Health training grant to the Computational and Systems Biology program at MIT, Postgraduate Scholarships from the National Science and Engineering Council of Canada (NSERC PGS), and an ICBP graduate fellowship. Part of the computational work was performed on a computing cluster funded by the National Science Foundation.

One of my favorite parts of being at MIT is the opportunity to meet and work with many amazing individuals. Members of the Fraenkel lab, both former and current, have helped me every day in all aspects of my work and made the lab a fun place to be. Kenzie MacIsaac, Esti Yeger-Lotem and Laura Riva helped me get started in the lab and my project. I began doing experiments in the last two years of my study, and any experimental results would not be possible without the generosity, patience, and advice from Shmulik Motola, Alice Lo, Tatjana Degenhardt, and Ferah Yildirim. Chris Ng and Adam Labadorf offered valuable insights in computational analysis, especially with regard to next-generation sequencing data. Martina Koeva, Sara Gosline and Nuncan Tuncbag, part of the growing network subgroup, have brought their experiences and fresh ideas that I really appreciate. Scott McCallum, Tali Mazor, Candance Chouinard, William Gordon, Jim Zhang, Aparna Kumar, and Deepika Dinesh are lab technicians who I worked with over the years and contributed to this project. I also worked with several very talented undergraduate students, Young Eun Choi, Melissa Gymrek, Jennifer Lai and Oana Ursu, who reminded me the excitement of science and made me realize the joy of mentoring. My collaboration with the lab of Prof. Forest White gave me access to the technical expertise of Kristen Naegle (also from the Lauffenburger lab), Emily Miraldi, Jason Neil, Bryan Owens, and Scott Carlson at some point of the project. Paul Huang, a former member of the White lab, was instrumental to get this project started and provided me with valuable technical and professional advice. I also used equipment in the labs of Doug

Lauffenburger, Tania Baker and also services from the BioMicro Center for collecting the data presented in this thesis.

I would like to thank my thesis committee, Profs. Doug Lauffenburger, Forest White and Marc Vidal, for their guidance and suggestions. They were open to ideas that were very untested at the time and helped me improve the methodology in the course of the project. From their questions I have learned not only the knowledge in the field and also how to approach a scientific problem. These lessons will benefit me greatly in the future.

I have the good fortune to have Prof. Ernest Fraenkel as my advisor in the last six years. Crucial insights for the project, both computational and biological, almost always came from our discussions. He gave me a lot of freedom in developing my ideas but made sure that I always keep in my mind the biological relevance of anything I do, a principle I will always remember in my career in computational biology. His advice, patience and encouragement, both for scientific and career development, helped me realize the possibility of pursuing a career in academia. I will try my best not to disappoint.

Lastly, despite my reluctance to mix my professional and personal life, I need to thank my parents for everything they have done for me. I am forever indebted to my mother for her indulging love to my father and myself, and I cherish the unwavering dreams and hope of my father and his drive to pursue knowledge. I hope I have inherited their optimism and perseverance to face future challenges in life.

Contents

List of Figures

List of Tables

List of Abbreviations

17-AAG	17-allylamino-17-demethoxygeldanamycin
bp	base pair
C/EBPβ	CCAAT/enhancer binding protein beta
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
ChIP-seq	chromatin immunoprecipitation sequencing
DNase-seq	DNaseI hypersensitive site sequencing
EGF	epidermal growth factor
EGFR	epidermal growth factor receptor
EMT	epithelial to mesenchymal transition
ESR1	estrogen receptor 1
GO	Gene Ontology
IMAC	immobilized metal affinity chromatography
iTRAQ	isobaric tags for relative and absolute quantitation
LC	liquid chromatography
MAPK	mitogen activated protein kinase

MS	mass spectrometry
MS/MS	tandem mass spectrometry
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PCST	prize collecting Steiner tree
PI3K	phosphoinositide 3-kinase
pY	phosphotyrosine
RNAi	RNA interference
STAT	signal transducer and activator of transcription
TF	transcription factor
TMRE	tetramethylrhodamine, ethyl ester
TMZ	temozolomide

Chapter 1

Introduction

1.1 Overview

Cellular signaling and transcription are tightly integrated processes that underlie many cellular responses to the environment. A network of signaling events, often mediated by post-translational modification on proteins, can lead to long-term changes in cellular behavior by altering the activity of specific transcriptional regulators and consequently the expression level of their downstream targets. Dysregulation of these molecular events have been implicated in many disease conditions such as neurodegeneration (Gil and Rego, 2008; Imarisio et al., 2008), metabolic disorder (Schinner et al., 2005), and every stage of tumor development and growth (Hanahan and Weinberg, 2011, 2000). The discovery of these events by molecular biology techniques has greatly enhanced the understanding of the causes of these diseases and subsequently the therapeutic strategies. This objective of my thesis is to link together global measurements of signaling and transcription to elucidate how specific signaling events lead to changes in transcription that determine the long-term behavior of the cell.

The first part of this chapter outlines the technologies and resources that enable systematic profiling of signaling and transcription events at the global level, emphasizing the discovery nature of these techniques. These approaches are complementary to hypothesis-driven experiments, many of which have been adopted to be run in high-throughput format but for characterizing pre-defined sets of targets. Since methods focused on discovery are not required to select what to measure *a priori*, there is potential to find novel events and assign new relevance to previously observed events.

The application of computational methods to biological signaling pathways is able to reveal behaviors of systems that cannot be presented by its individual components (Bhalla and Iyengar, 1999), known as the "emergent property". The ability to measure the network components and their connections at the global level, often in a high-throughput format, has created a wealth of data and sparked the development of many computational algorithms in order to gain biological insight. These methods can be generic or specific for the particular type of experimental technique. They represent a spectrum of abstraction of biological entities for which different computational approaches are appropriate with different objectives of modeling outcome (Ideker and Lauffenburger, 2003). The second part of this chapter gives examples of several methodologies that are either popular with the kind of datasets used in this project, or are conceptually similar to the core computational ideas presented in this thesis but used for different kinds of data sets.

The review of current methods is not intended to be exhaustive. Instead, I focused on the unbiased property of the experimental assays, and selected examples that represent major algorithmic approaches for analyzing signaling and gene expression data but are inherently different from my proposed framework. It is with this context that I summarize the motivation and innovation behind this work, where I highlight the distinct features of my method and the unique perspective that it might bring to complement the existing methodologies.

1.2 Datasets for interrogating signaling and transcription at the global level

1.2.1 Phosphoproteomics mass spectrometry

Post-translational modification on proteins is a major mechanism by which the functions and activities of proteins are regulated in response to environmental cues. In particular, phosphorylation on amino acid residues serine, threonine and tyrosine regulates a variety of functions of the affected proteins such as protein-protein interaction (Yaffe, 2002), enzymatic activity (Cole et al., 2003), protein stability (Sears et al., 2000), and also higher level processes such as proliferation (Iyer et al., 2006), apoptosis (Yousefi et al., 1994) and metabolism (Boura-Halfon and Zick, 2009). Although phosphorylation on tyrosine is relatively rare compared to that on serine and threonine (Olsen et al., 2006), it has been linked to many critical cellular functions and appears to display more dynamics than serine and threonine phosphorylation in response to growth factor signaling (Olsen et al., 2006, and Figure 1-1). Therefore, profiling tyrosine phosphorylation on multiple proteins may provide information about the activities of many molecular processes and the components in those processes.

The ability of phospho-specific antibodies to recognize phosphorylated residues but not the non-phosphorylated counterparts (Mandell, 2003; Blaydes et al., 2000) makes it possible to study protein phosphorylation by several experimental techniques. For instance, using an antibody that recognizes a specific phosphorylated residue on a cell surface receptor, one can use western blot to detect the presence of this phosphorylation site in protein lysate transferred to a nitrocellulose membrane (Kurien and Scofield, 2009), or use flow cytometry to isolate single cells that express this phospho-form of the receptor (Oberprieler and Taskn, 2011; Krutzik et al., 2004). There are also antibody microarrays (Chaga, 2008) where a collection of antibodies are conjugated to glass slides and then incubated with cell lysate. In these assays the identities of the phosphorylated proteins are pre-determined and limited by the availability and quality of phospho-specific antibodies. In contrast, phosphoproteomics by mass spectrometry (MS) offers clear advantage in its ability to find phosphorylation sites in the whole proteome in an unbiased manner.

Large-scale quantification of *in vivo* phosphorylation sites at the global level is made possible by recent technological development in key steps of the experimental protocol: enrichment of phosphorylated peptides from complex protein mixtures, separation of the peptide mixture by liquid chromatography (LC) , tandem mass spec-

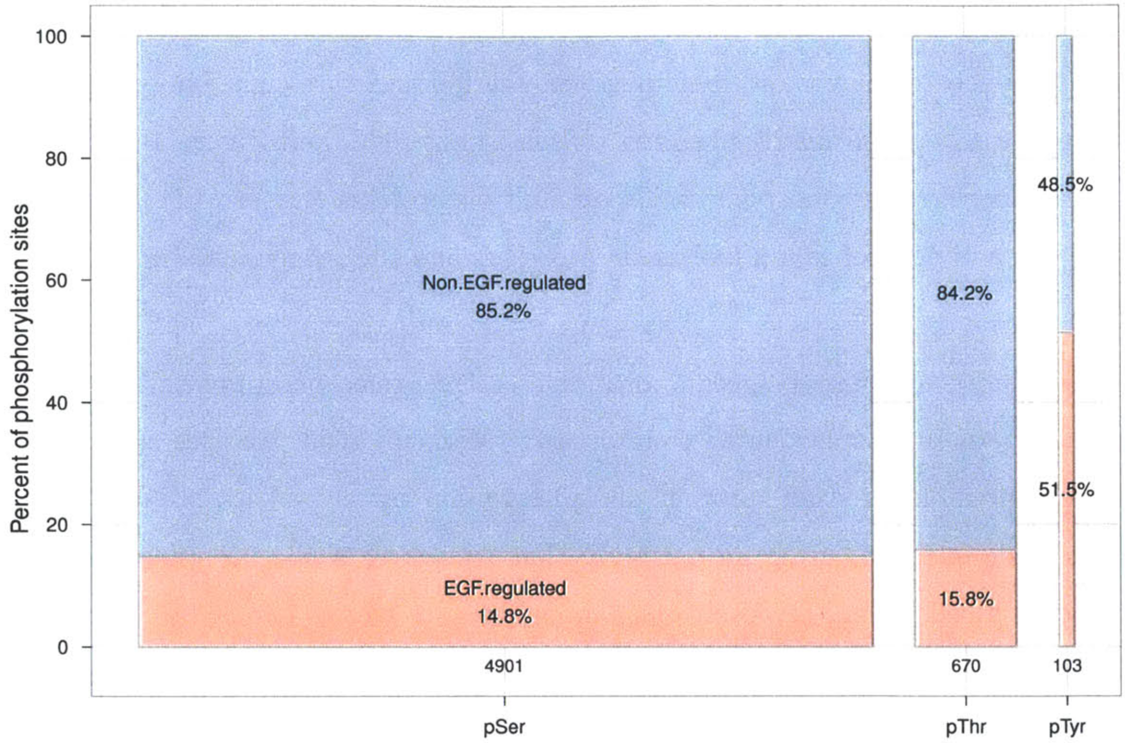


Figure 1-1: Distribution of phosphoserine, phosphothreonine and phosphotyrosine sites identified by mass spectrometry in HeLa cells stimulated by the epidermal growth factor (EGF) (Olsen et al., 2006). Phosphorylation sites that change by more than two fold in a time course of 20 minutes were considered to be regulated by EGF. Although phosphotyrosine sites make up less than 2% of all the phosphorylation sites, a much larger fraction of these sites are regulated by EGF compared to phosphoserine and phosphothreonine sites. pSer: phosphoserine. pThr: phosphothreonine. pTyr: phosphotyrosine.

trometry (MS/MS) for peptide sequence identification, and isotope labeling strategies for peptide quantification (Grimsrud et al., 2010). In particular, the input dataset for the computational modeling in this thesis was collected by our collaborator in Forest White's group that employs immunoprecipitation and immobilized metal affinity chromatography (IMAC) to enrich for tyrosine phosphorylated peptides and iTRAQ (isobaric tags for relative and absolute quantitation) labeling to quantify these peptides in four- or eight-plex format (Zhang et al., 2007b, and Figure 1-2). The resulting dataset is in the form of peptide sequences containing phosphorylated tyrosine and the relative levels of each peptide in the four or eight input samples.

The unbiased approach of mass spectrometry has led to novel insights into the global state of signaling networks. For instance, in the EGFRvIII glioblastoma dataset that I will describe in Section 1.4, the expression EGFRvIII, an oncogenic mutant of EGFR, was found to induce the phosphorylation of an activating tyrosine residue on the c-MET receptor, and combined inhibition of c-MET and EGFR results in enhanced cytotoxicity (Huang et al., 2007). With mounting evidence for the prevalence and functional significance of interconnections between pathways (Bauer-Mehren et al., 2009), the value of this technology will become increasingly appreciated.

As with many systems level datasets, there are many challenges in making interpretation of the phosphoproteomic data, especially in deriving biological meanings beyond validating the top hits. The EGFRvIII dataset mentioned above contains phosphorylated peptides that can be mapped to 85 genes, but only twelve (14%) appear in the human ErbB signaling pathway in the April 11, 2011 version of the KEGG PATHWAY database (Kanehisa et al., 2010), five (5.8%) are in the MAPK signaling pathway, eight (9.4%) are in the phosphatidylinositol signaling system, and 31 (36%) are not found in any of the KEGG pathways. In terms of connecting to transcription, three genes (3.5%) are annotated to have transcription factor activity. These statistics simultaneously show the opportunity for discovery but also the urgent need for new analysis approaches.

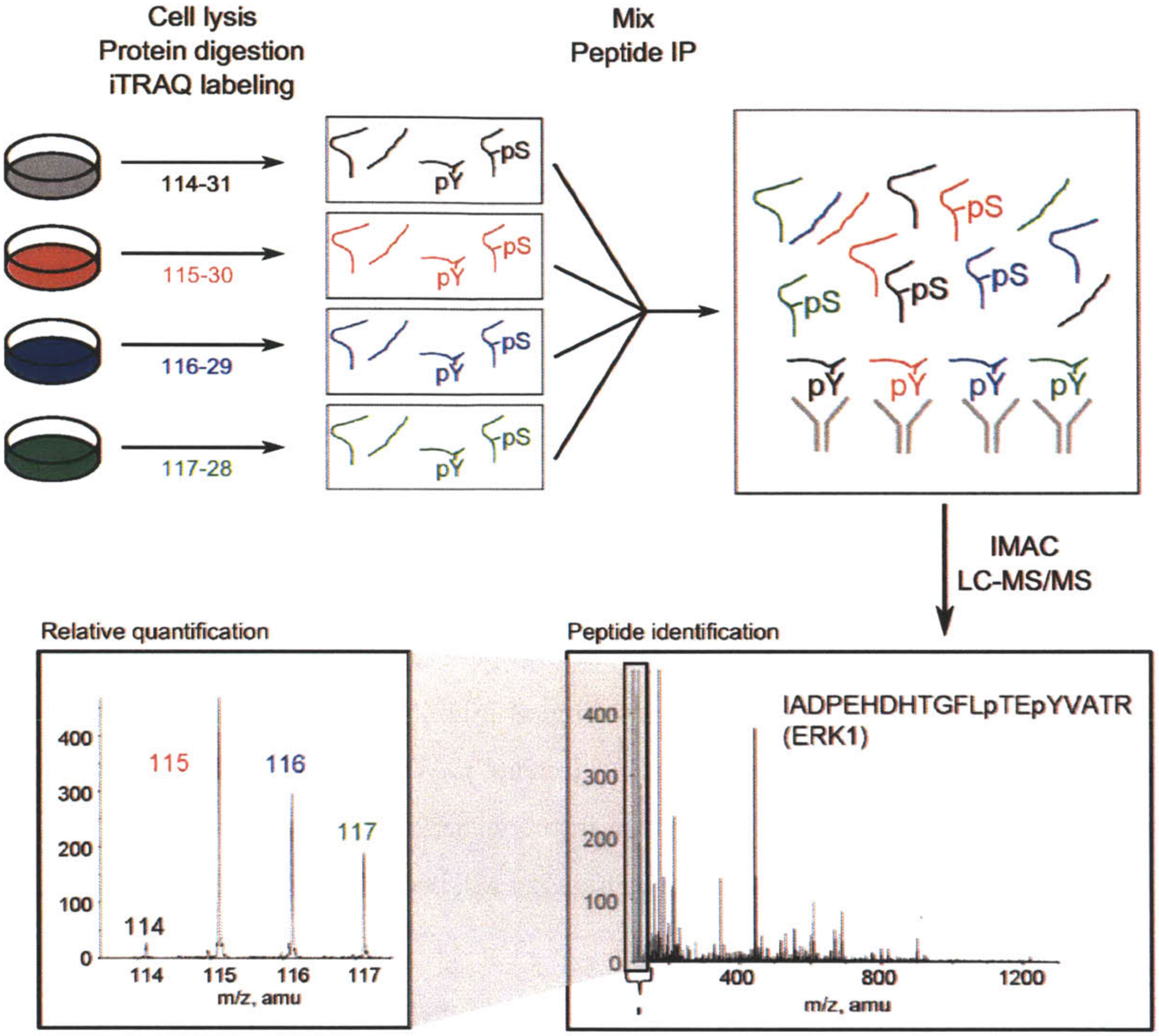


Figure 1-2: Work flow of MS-based phosphoproteomic experiment for quantifying relative tyrosine phosphorylated peptides across multiple conditions. Following cell lysis in urea, the protein lysates from multiple conditions are digested with trypsin. Each sample is labeled with iTRAQ reagent of a different mass tag, which is covalently attached to the N-terminus or side chain amines of tryptic digested peptides in the sample. Currently labeling can be done for four or eight samples. The samples are then mixed, immunoprecipitated by an anti-phosphotyrosine antibody and further enriched for phosphorylated peptides in an IMAC column. The peptides are analyzed by LC-MS/MS, where the peptide sequences and the phosphorylated tyrosine residues are identified by the mass spectra, and the relative levels of the peptides in the original samples are identified by the iTRAQ mass tags. Adapted from Schmelzle et al. (2006). Copyright 2006 American Diabetes Association.

1.2.2 Transcriptome profiling

Identification of proteins *in vivo* presents many challenges due to the need to design antibodies that are capable of targeting the specific three-dimensional configuration of the protein of interest. As such, identifying mRNA levels by sequence as a proxy for protein abundance has become increasingly popular (Lockhart and Winzeler, 2000). During the course of this project several transcriptome profiling technologies have become widely available with different options in terms of cost, sensitivity, and the ability to study transcript isoforms.

An expression microarray (Figure 1-3) consists of thousands to millions of DNA sequences spotted onto glass slides. The sequence at each spot ("probes") may correspond a gene transcript or a short section on the transcript ("targets"). Gene transcripts in a sample are first reversed transcribed to cDNA and labeled. After incubating the array with the cDNA and washing, the signal from the labeled transcripts hybridized to a spot gives a quantitative measure of the abundance of target transcripts complementary to the probe sequence. The format of detection can be in two-channel or one-channel. In two-channel detection, two samples of cDNA, each labeled with a different fluorophore, are hybridized to the same array and the relative signal intensity from the two fluorophores at the same spot represents the relative expression level of the target transcript (Schena et al., 1995). In one-channel detection such as the commercially available Affymetrix GeneChip platform (Lipshutz et al., 1999), one transcriptome sample is labeled with biotin, hybridized to the array and hybridization is detected by fluorophore-conjugated streptavidin that binds to biotin. Relative quantifications of gene transcripts are obtained by downstream processing that performs normalization between arrays and assesses differential expression.

RNA-seq utilizes the ability of next-generation sequencing to sequence millions of short nucleotide pieces in parallel to quantify transcript abundance and detect alternative splicing. Nucleotide sequences in the reversed transcribed cDNA library are sequenced at the end for a fixed number of base pairs. These millions of sequence "tags" can be aligned to a reference transcriptome or used in novel transcriptome

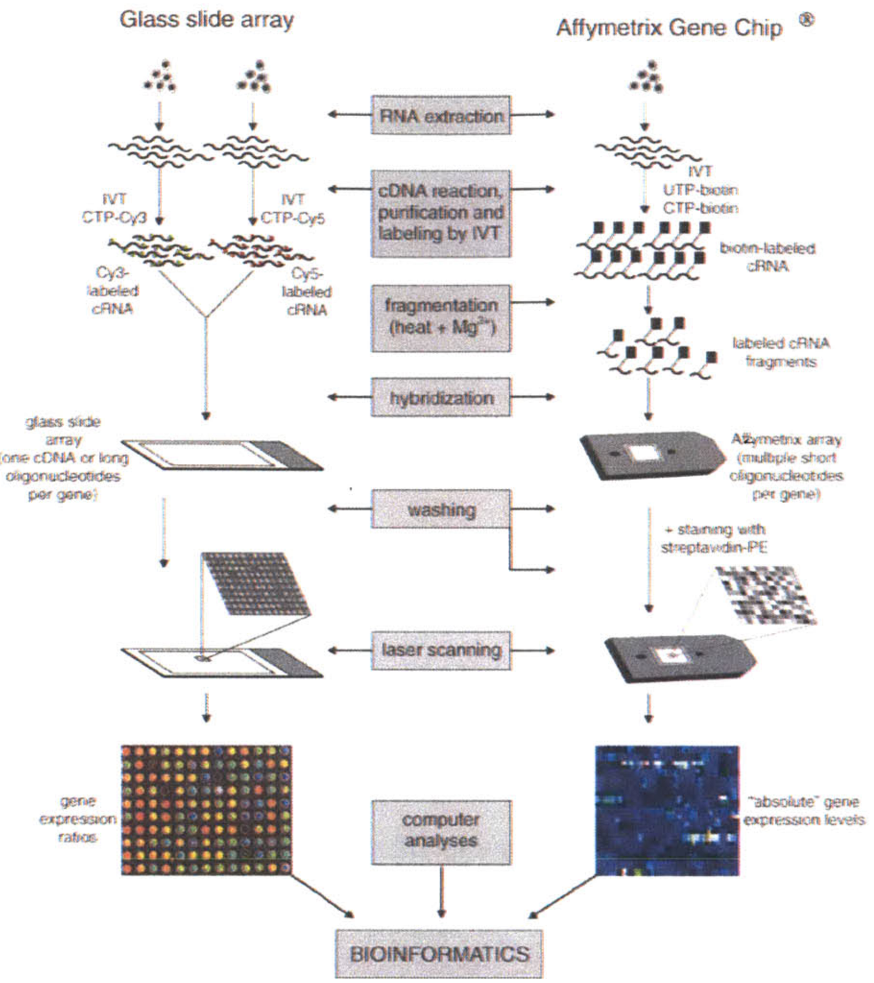

Glass slide array
Affymetrix Gene Chip ®
RNA extraction
cDNA reaction, purification and labeling by IVT
IVT
CTP-Cy3
IVT
CTP-Cy5
Cy3-labeled cRNA
Cy5-labeled cRNA
IVT
UTP-biotin
CTP-biotin
biotin-labeled cRNA
fragmentation (heat + Mg^{2+})
labeled cRNA fragments
hybridization
glass slide array (one cDNA or long oligonucleotides per gene)
Affymetrix array (multiple short oligonucleotides per gene)
washing
+ staining with streptavidin-PE
laser scanning
gene expression ratios
computer analyses
"absolute" gene expression levels
BIOINFORMATICS

Figure 1-3: Procedure of mRNA expression profiling on DNA microarrays. There are two major technical platforms. Both are based on the principle of complementary base pairing between the nucleic acid sequences attached to the array (probes) and the transcript sequences in the sample (targets). In the glass slide array platform, the sequences on the array are usually complementary DNA (cDNA) or long oligonucleotide sequences. Purified RNA from two samples are reversed transcribed to make cDNA libraries, which undergo *in vitro* transcription (IVT) reactions with fluorescence labeled nucleotides (Cy3 dye that emits green for one sample and Cy5 dye that emits red for a second sample). The two samples are mixed and hybridized to the glass slide array. The array is washed and scanned in a laser scanner. The relative intensity between red fluorescence and green fluorescence channels at each spot gives the relative abundance of the gene at that spot between the two samples. In the Affymetrix GeneChip platform, the spots on the array are short oligonucleotide sequences that match part of a gene, and usually one gene is represented by multiple sequences. The purified RNA sample is reversed transcribed into cDNA library, which is used in an IVT reaction with biotinylated ribonucleotides to generate biotin labeled cRNA. The cRNA is fragmented and hybridized to the array. After washing away unbound sequences, the array is stained with phycorerythrin (PE) conjugated streptavidin and scanned in a laser scanner. The intensity at each spot represent an absolute expression value for that gene, but downstream computational processing is required to normalize between multiple arrays or conditions and identify differential expression. Reprinted by permission from Macmillan Publishers Ltd: Leukemia (Staal et al., 2003), copyright 2003.

assembly and differential expression analysis (Trapnell et al., 2010). Furthermore, the sequencing depth achieved is sufficient to identify many novel splicing events (Trapnell et al., 2009). Unlike microarrays, transcript quantification in RNA-seq has a digital readout. And since the actual sequences of transcripts are obtained, it is not limited to detecting targets that complement probe sequences on the array, which are usually designed from known and predicted gene models. As a result RNA-seq has great potential for discovery of novel transcripts and transcript isoforms.

Compared to mass spectrometry techniques, microarray protocols are relatively easy to perform and be standardized. Due to the early demonstration that global transcript level can represent cell states in response to perturbation of signaling pathways (Roberts et al., 2000), and that such a representation can discover and predict cancer subtypes of clinical relevance (Golub et al., 1999), microarrays have become a primary choice for large scale expression profiling projects such as The Cancer Genome Atlas (TCGA) (Verhaak et al., 2010; Cancer Genome Atlas Research Network, 2008). The establishment of central data repository Gene Expression Omnibus (GEO) (Barrett et al., 2011) makes large amount of data easily accessible and has encouraged development of many data analysis methods. In Section 1.3 I will discuss a few techniques that share our goal of finding regulatory relationships from data but use mRNA expression profiling. An obvious drawback of this technique is the inability to investigate regulatory processes outside of transcription, and an obvious solution is to apply these algorithms to proteomic data. I will explain why this remedy is overly simplistic given the reality of the datasets, and therefore a new computational approach is necessary.

1.2.3 Next-generation sequencing technology for transcriptional regulation

In addition to RNA-seq, next-generation sequencing technology has been applied to investigate the many different stages at which gene expression can be regulated (Nat Rev Genet Article Series, 2011). Here we focus on methods that capture interactions between trans-acting protein factors and the DNA genome: chromatin immunoprecipitation sequencing (ChIP-seq) (Park, 2009; Johnson et al., 2007) and DNaseI hypersensitive site sequencing (DNase-seq) (Song and Crawford, 2010; Hesselberth et al., 2009).

ChIP is an experimental technique for investigating interactions between proteins and DNA in the cell (Collas, 2010; Carey et al., 2009). It has been used to identify the localization of transcription factors, co-regulators, and post-translationally modified histones in the genome or to a specific locus. In a ChIP experiment (Figure 1-4), protein and DNA interactions are temporarily fixed, the chromatin is sheared and protein-DNA complexes are selectively immunoprecipitated by an antibody to obtain the DNA fragments associated with the protein. Downstream sequencing or polymerase chain reaction (PCR) amplification of the DNA fragments reveals the *in vivo* binding locations of the protein factor to the genome.

While ChIP-seq finds binding locations of specific proteins, DNase-seq is a general assay for open chromatin regions that may be bound by different protein factors (Figure 1-5). In traditional DNase footprinting assays, protection of DNA regions from digestion by a non-sequence specific endonuclease DNaseI is used as evidence for protein binding (Brenowitz et al., 1986; Galas and Schmitz, 1978). In a genome wide format, the DNase-digested fragments are purified and sequenced. Mapping the sequences back to the reference genome reveals genomic regions that are hypersensitive to digestion. Interestingly, it was observed in the sequencing results that the DNaseI footprints display a distinct cleavage pattern where regions immediately surrounding the protection sites have more aligned reads, indicative of increased sensitivity (Boyle et al., 2011). This is probably due to the disruption of regular histone organization as a result of binding of protein factors. Therefore, searching for regions in the genome that have significantly more reads from the DNaseI treated chromatin compared to naked genomic DNA control gives us a way to identify open chromatin regions.

Protein signaling networks rely on protein-DNA interactions to transmit information to the transcriptional machinery, where these signals are integrated with instructions encoded by the genome and epigenome to create a global transcriptional

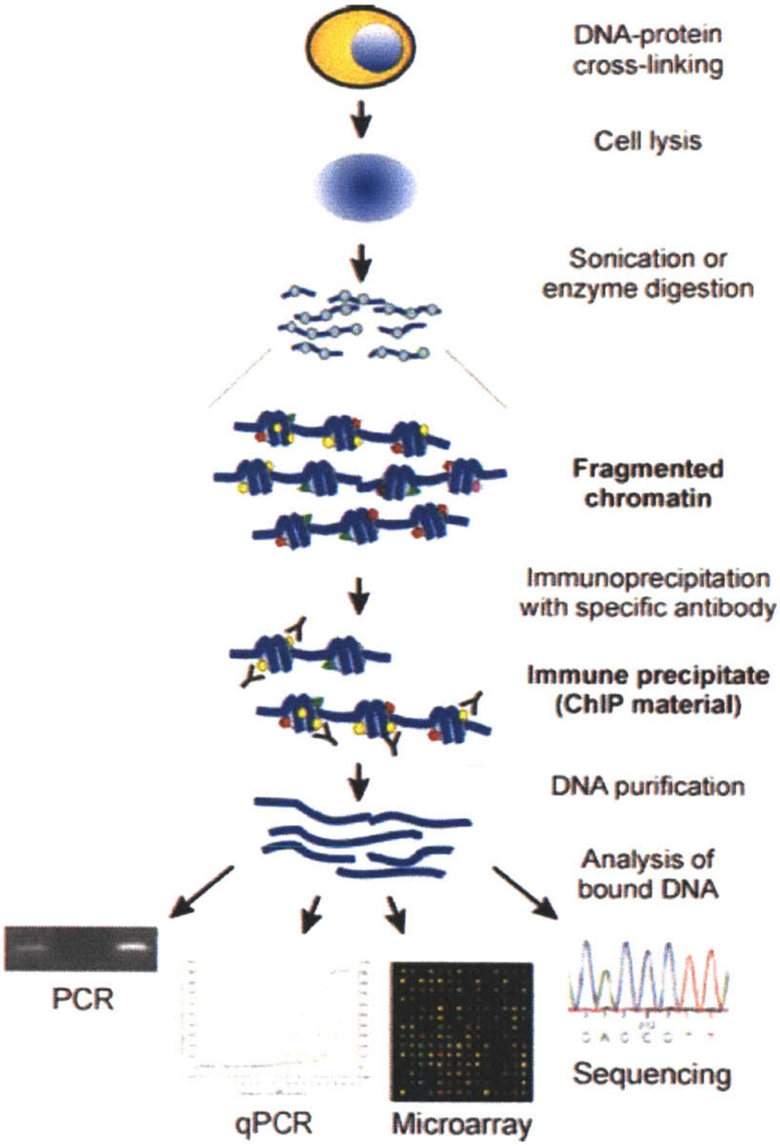

DNA-protein cross-linking
Cell lysis
Sonication or enzyme digestion
Fragmented chromatin
Immunoprecipitation with specific antibody
Immune precipitate (ChIP material)
DNA purification
Analysis of bound DNA
PCR
qPCR
Microarray
Sequencing

Figure 1-4: Overview of ChIP procedure for locating binding sites of specific proteins in the genome. In the first step, *in vivo* protein-DNA interactions are fixed by formaldehyde treatment that cross-links proteins and DNA in close contact. The chromatin is then sonicated into short fragments and immunoprecipitated with an antibody that recognizes a protein of interest, for instance, transcription factors, co-regulators, or modifications on histones. The antibody has been pre-bound to protein-A or protein-G conjugated magnetic beads, so applying a magnetic field to the mixture extracts the antibody along with the protein-DNA complexes that contain the protein of interest and the bound fragments of DNA. The crosslinks are reversed by incubation at high temperature and the DNA fragments are purified. These DNA fragments, now enriched in sequences bound by the protein factor, can be analyzed by PCR amplifying a known locus (ChIP-PCR), hybridization to tiling DNA microarrays (ChIP-chip) or direct sequencing (ChIP-seq). In particular, ChIP-seq generates an unbiased and genome-wide readout of the bound sequences. Mapping the sequence reads back to the genome reveals the binding locations of the protein factor. Reprinted with kind permission from Springer Science+Business Media: Molecular Biotechnology (2010) 45:87-100, Figure 2 (Collas, 2010).

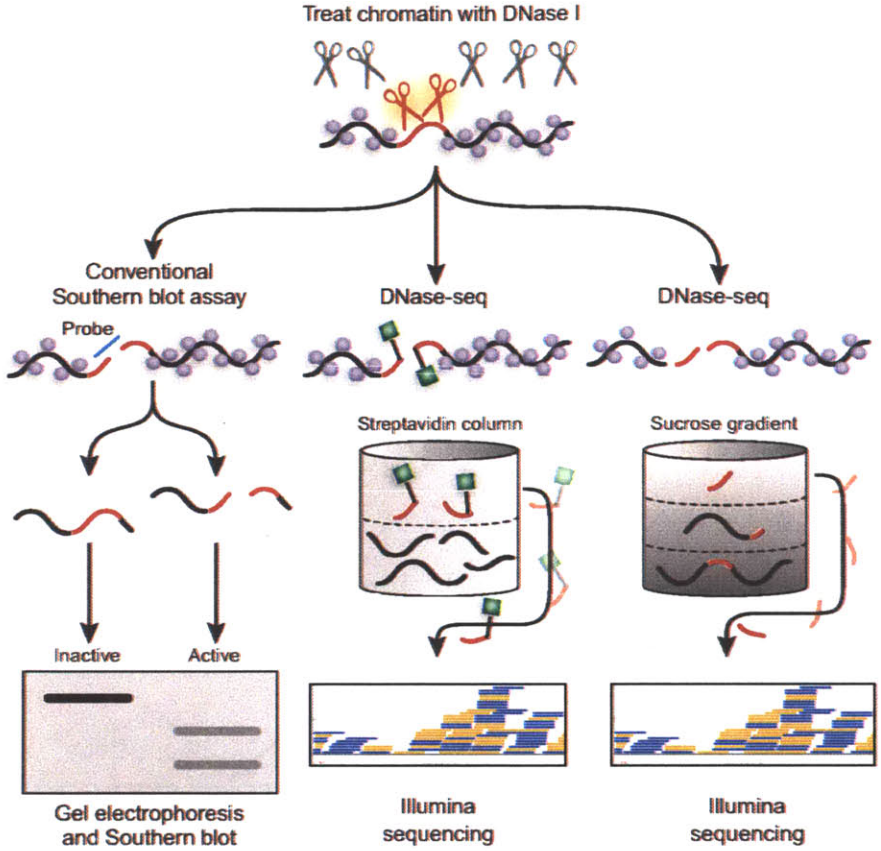


Figure 1-5: Methods for detecting DNaseI hypersensitive regions. Purified nuclei are treated with DNaseI enzyme for a short time so only the most sensitive regions are cleaved by the enzyme ("hypersensitive"), and the cleaved fragments are identified in southern blot or by sequencing. In the southern blot format of the assay, the DNaseI treated chromatin is separated by size in gel electrophoresis and transferred to a membrane. Probe sequence complementary to a genomic region of interest is hybridized to the membrane to detect cleavage of that region by DNaseI. Two alternative protocols exist for downstream sequencing application. In the first method (Song and Crawford, 2010; Boyle et al., 2008), the cleaved ends of DNA are ligated to a biotinylated linker (green squares), the genomic DNA are sheared and the tagged fragments are isolated by binding to streptavidin. The second method relies on a "two hit" assumption that short fragments produced by DNaseI cutting at both ends are more likely from accessible chromatin regions than due to random shearing during sample processing. DNaseI digested chromatin is separated by molecular weight in a sucrose gradient, and DNA fragments from fractions of small molecular weight are purified and sequenced. Adapted by permission from Macmillan Publishers Ltd: Nature Methods (Giresi and Lieb, 2006), copyright 2006.

program. Signaling pathways can target multiple transcription factors (Chang et al., 2003), and transcription factors can respond to multiple activation pathways and carry out a variety of biological functions (Desrivires et al., 2006). Adding to the complexity is the phosphorylation of a transcription factor on the same amino acid residue can both activate and inhibit its activity (Lim and Cao, 1999; Decker and Kovarik, 2000). Therefore, ChIP is still the gold-standard method to determine the *in vivo* targets of transcription factors and consequently their condition-specific functions. However, the assay requires large amount of input material and good quality antibodies to specific proteins, so with few exceptions it is impractical to apply it exhaustively to all the transcriptional regulators of an organism in all cell types and conditions. On the contrary, one DNase-seq experiment, with replicate, can report condition-specific open chromatin genome-wide. Integrating this information with known sequence specificity of transcription factors has enabled accurate predictions of transcription factor binding (Boyle et al., 2011; Pique-Regi et al., 2011). The level of accuracy appears to be dependent on the transcription factor, and not all the factors have known sequence specificity, so DNase-seq and ChIP-seq are two complementary techniques in our quest to characterize transcriptional regulation at the global level.

An often cited limitation of ChIP is that it is not a functional assay and does not directly provide information about the functional significance of observed binding sites (see a list of examples cited in Carey et al., 2009). Correlating binding with transcriptome profiling may establish this connection (Ouyang et al., 2009) and is an area under active research. In my algorithm I adopted this idea with modification for DNase-seq and transcription profiling data, and I reasoned that adding the phosphoproteomic data should be able to further narrow down the search for biological functions.

1.2.4 Protein-protein interactome

While protein-DNA interactions are essential in the regulation of gene expression (Maston et al., 2006), protein-protein interactions are the building blocks of signaling pathways (Pawson and Nash, 2003). Together they define a global regulatory network of the cell. Large collections of protein-protein interactions have been utilized to gain biological insights, starting from the level of individual gene functions and up to the global properties of the entire regulatory network (Bader et al., 2008; Cusick et al., 2005). To be compatible with the discovery nature of the phosphoproteomic and transcriptome datasets, we turn to sources of protein-protein interaction data that are not exclusive to pre-defined protein targets or expected pathways: high-throughput experimental mapping and databases of protein interactions.

Yeast two hybrid (Y2H) and affinity purification mass spectrometry (AP/MS) are two popular experimental methods for large scale mapping of protein interactions (Berggrd et al., 2007). Y2H measures direct physical interaction between pairs of proteins (Uetz et al., 2000; Ito et al., 2001) and AP/MS (Gavin et al., 2002; Ho et al., 2002) identifies protein complexes in which the components may or may not directly interact. When carried out under carefully controlled experimental conditions, these techniques have been shown to generate interaction data of high quality (Yu et al., 2008; Dreze et al., 2010).

Many databases of protein-protein interactions are publicly available. The IntAct molecular interaction database (Kerrien et al., 2007), the Database of Interacting Proteins (DIP) (Salwinski et al., 2004), the Molecular Interaction database (MINT) (Chatr-aryamontri et al., 2007) and the Biological General Repository for Interaction Datasets (BioGRID) (Stark et al., 2011) are examples of independent ongoing efforts to curate interactions from published literature and they recently formed the International Molecular Exchange Consortium (IMEx) to unify curation rules and to coordinate curation to avoid redundancy (Salwinski et al., 2009). There are other databases which focus on signaling and metabolic pathways such as the KEGG PATHWAY (Kanehisa et al., 2010) and Reactome (Matthews et al., 2009) databases, and also "meta" databases, such as the Agile Protein Interaction DataAnalyzer (APID) (Prieto and Rivas, 2006), the Michigan Molecular Interactions database (MiMI) (Tarcea et al., 2009) and the Unified Human Interactome database (UniHI) (Chaurasia et al., 2007), that aim to consolidate interactions from individual databases to provide a comprehensive resource.

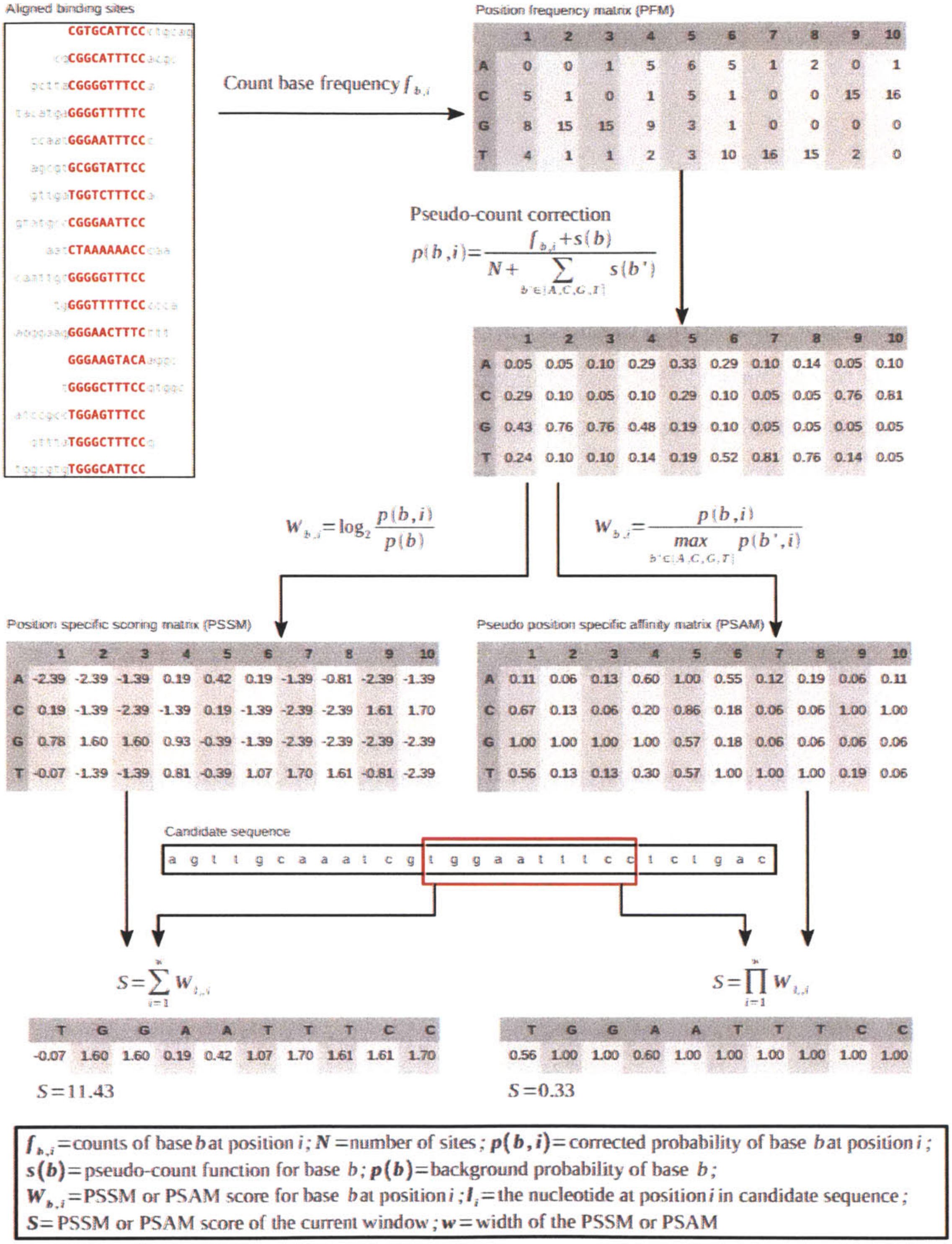
Even with the combination of large experimental efforts and curated databases we are still far from a complete mapping of all possible protein-protein interactions, and thus many computational methods have been developed to predict possible interactions. These methods make use of features such as gene neighborhood (Huynen et al., 2000), gene fusion (Marcotte et al., 1999), sequence co-evolution (Goh et al., 2000), and may incorporate multiple such features in a Bayesian framework (von Mering et al., 2005; Jansen et al., 2003). Predictions of kinase-substrate relationships by NetworKIN (Linding et al., 2007) and the binding interactions by ScanSite (Obenauer et al., 2003) are particularly valuable to complement the curated databases for interpretation of our datasets.

While it is appealing to place the signaling and transcription datasets on the protein interaction network for novel biological insights, care must be taken so the results are interpretable, reliable and biologically relevant. First of all, since not all signaling and regulatory events are mediated by events reported in the phosphoproteomic data, in building a network for these hits we have to consider proteins that they interact with directly and indirectly. Despite being incomplete, the amount of interaction data is still large, so the size of the network explodes exponentially and quickly becomes non-interpretable, as pointed out by previous data integration efforts (Hwang et al., 2005). Secondly, interaction records in databases come from thousands of laboratories and many experimental techniques, so overall the data quality is heterogeneous and should not be treated non-discriminantly. Lastly, pooling these interactions together risks losing the specific context under which they were detected. It is with these issues in mind that I designed my constraint optimization approach, where the interactome edges are weighted probabilistically by confidence, biological contexts are provided by constraining the network to edges that include signaling and transcriptional events, and a simple set of interactions that connect the data is selected by an optimization procedure.

1.2.5 Transcription factor binding motifs

The binding specificity of trans-acting factors to cis-regulatory elements in the genome is determined by the three-dimensional structure of these factors and may be used to predict new binding sites. Commonly used quantitative representation of such binding patterns, also known as sequence motifs, include position weight matrices (PWM)/position specific scoring matrices (PSSM) (D'haeseleer, 2006; Stormo, 2000) with an information theoretic perspective, and position specific affinity matrices (PSAM) with a statistical mechanics perspective (Figure 1-6 and Foat et al., 2006, 2005; Manke et al., 2008; Roider et al., 2007). Experimentally, *in vivo* binding patterns can be determined by applying various motif discovery tools to the DNA sequences obtained from the ChIP-seq datasets discussed above. *In vitro* techniques are also available that take an enrichment then sequencing approach (SELEX - systematic evolution of ligands by exponential enrichment - followed by conventional sequencing or next-generation sequencing; Stoltenburg et al., 2007; Jolma et al., 2010) or by microarray hybridization (PBM - protein binding microarrays; Berger et al., 2006). TRANSFAC (Wingender, 2008; Matys et al., 2006) and JASPAR (Sandelin et al., 2004; Bryne et al., 2008) are two major databases that collect published transcription factor binding motifs from literature that can be used for prediction of regulatory elements.

The sequence motifs are useful for predicting binding of specific factors to genomic regions and associating these factors to nearby genes as their downstream targets. Since the motifs are short and degenerate, scanning for matches in genome sequences, even limited to promoter regions, results in numerous hits, most of which are non-functional *in vivo* (Wasserman and Sandelin, 2004). Restricting the search space to evolutionarily conserved regions can significantly reduce false predictions (Wasserman and Sandelin, 2004) at the expense of missing species-specific binding events that are very prevalent (Odom et al., 2007). Another approach is to search for enriched motifs in the promoter region of differentially expressed genes (Sui et al., 2007; Tavazoie et al., 1999). Neither case takes into account the chromatin acces-



Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10
A	0	0	1	5	6	5	1	2	0	1
C	5	1	0	1	5	1	0	0	15	16
G	8	15	15	9	3	1	0	0	0	0
T	4	1	1	2	3	10	16	15	2	0

Pseudo-count correction

$$p(b,i)=\frac{f_{b,i}+s(b)}{N+\sum_{b'\in\{A,C,G,T\}} s(b')}$$

	1	2	3	4	5	6	7	8	9	10
A	0.05	0.05	0.10	0.29	0.33	0.29	0.10	0.14	0.05	0.10
C	0.29	0.10	0.05	0.10	0.29	0.10	0.05	0.05	0.76	0.81
G	0.43	0.76	0.76	0.48	0.19	0.10	0.05	0.05	0.05	0.05
T	0.24	0.10	0.10	0.14	0.19	0.52	0.81	0.76	0.14	0.05

$$W_{b,i}=\log_2\frac{p(b,i)}{p(b)}$$

$$W_{b,i}=\frac{p(b,i)}{\max_{b'\in\{A,C,G,T\}} p(b',i)}$$

Position specific scoring matrix (PSSM)

	1	2	3	4	5	6	7	8	9	10
A	-2.39	-2.39	-1.39	0.19	0.42	0.19	-1.39	-0.81	-2.39	-1.39
C	0.19	-1.39	-2.39	-1.39	0.19	-1.39	-2.39	-2.39	1.61	1.70
G	0.78	1.60	1.60	0.93	-0.39	-1.39	-2.39	-2.39	-2.39	-2.39
T	-0.07	-1.39	-1.39	0.81	-0.39	1.07	1.70	1.61	-0.81	-2.39

Pseudo position specific affinity matrix (PSAM)

	1	2	3	4	5	6	7	8	9	10
A	0.11	0.06	0.13	0.60	1.00	0.55	0.12	0.19	0.06	0.11
C	0.67	0.13	0.06	0.20	0.86	0.18	0.06	0.06	1.00	1.00
G	1.00	1.00	1.00	1.00	0.57	0.18	0.06	0.06	0.06	0.06
T	0.56	0.13	0.13	0.30	0.57	1.00	1.00	1.00	0.19	0.06

$$S=\sum_{i=1}^{w} W_{l_i,i}$$

T	G	G	A	A	T	T	T	C	C
-0.07	1.60	1.60	0.19	0.42	1.07	1.70	1.61	1.61	1.70

$S=11.43$

$$S=\prod_{i=1}^{w} W_{l_i,i}$$

T	G	G	A	A	T	T	T	C	C
0.56	1.00	1.00	0.60	1.00	1.00	1.00	1.00	1.00	1.00

$S=0.33$

$f_{b,i}$=counts of base b at position i; N=number of sites; $p(b,i)$=corrected probability of base b at position i; $s(b)$=pseudo-count function for base b; $p(b)$=background probability of base b; $W_{b,i}$=PSSM or PSAM score for base b at position i; l_i=the nucleotide at position i in candidate sequence; S= PSSM or PSAM score of the current window; w=width of the PSSM or PSAM

Figure 1-6: Computational representation and discovery of transcription factor binding sites, with an example of the human REL protein binding profile (JASPAR MA0101.1, curated from Kunsch et al., 1992) and NF-κB binding site in the human IL8 promoter (TRANSFAC binding site HS$IL8_21).

sibility of the predicted binding locations or captures distant regulatory elements. The DNase-seq technique (Section 1.2.3) may represent an adequate solution to these problems.

1.3 Computational methods for finding molecular regulatory networks

Choice of computational methods for analysis of biological data is defined by the goal of the modeling and the characteristics of the data. We want to connect signaling events to differential mRNA expression, using measurements of tens of protein phosphorylation sites and thousands of gene transcripts from a handful of experimental conditions. In this section I summarize methods for inferring molecular regulatory relationships that aim for the same goal but start from different data sources. I will explain why these methods may appear to be applicable to our problem but in closer inspection are not suitable for our datasets.

1.3.1 *de novo* learning of regulatory relationships

Many computational algorithms have been created to infer regulatory relationships between genes. A vast number of these construct transcriptional regulatory networks from mRNA profiling data that recently have become widely available. The central premise of these methods is that correlated expression of genes is indicative of co-regulation, and the observed correlation between genes can be explained by the presence of other genes, all measured on the microarrays. Here I adopt the conceptual framework presented in Markowetz and Spang (2007) to organize an overview of current methods and incorporate the review by Bansal et al. (2007) to give examples of publicly available software packages tailored to the properties of input datasets.

In the notation from Markowetz and Spang (2007), let V be a set of p network components, which are genes on the microarrays in this context but can be proteins. The measurements on $v \in V$ are modeled as a random variable X_v and so all the p

components in the network form a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. Measurements of $\mathbf{X}$ in N experiments result in data vectors $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N$. The goal is to build a network $T = (V, E)$ where an edge $e_{ij} \in E$ represents the dependency structure between network components i and j. To draw an edge between i and j, the computational methods ask the question "is X_i independent of X_j given Z?", and the identity of Z defines the specific statistical model. X_i and X_j are conditionally independent given Z, i.e. $X_i \perp X_j \mid Z$, if and only if

$$P(X_i = x_i, X_j = x_j \mid Z = z) = P(X_i = x_i \mid Z = z) P(X_j = x_j \mid Z = z). \qquad (1.3.1)$$

Table 1.1 summarizes a few different definitions of Z, the corresponding methodologies and example applications.

While these methods are capable of discovering new connections between genes without prior knowledge of the network topology, there are a few shortcomings: the cause-effect directions are often unclear, and only transcription regulation is modeled but not other parts of the biologically relevant networks. Key to resolving these issues are introducing controlled perturbations and making protein level measurements, as demonstrated in analysis of data from flow cytometry (Sachs et al., 2005) and microwestern arrays (Ciaccio et al., 2010). Interestingly, in one simulation study Bayesian network models are comparable to simpler correlation networks when applied to observational data but they achieve better performance in interventional data (Werhli et al., 2006). Continuing improvement in the throughput of making precise perturbations and measuring the outcome will realize more potential of Bayesian networks.

Given we have unbiased measurements of the phosphoproteome and transcriptome, it may be possible to apply these statistical inference methods that infer *de novo* relationships, but we encounter three major limitations. First, it is difficult to achieve statistical significance with four samples and hundreds of variables. Secondly, the numerical correlations give little hint for mechanistic relationships. Lastly, as mentioned above, we must account for intermediate signaling nodes not reported in

Z	Meaning of $X_i \not\perp X_j \mid Z$	Method and representative references
$\varnothing$	X_i and X_j are not marginally independent	Co-expression clustering (Eisen et al., 1998; Spellman et al., 1998)
$\mathbf{X}_S$ for $S = V \backslash \{i, j\}$	Correlation between i and j cannot be explained by all the other genes	Markov random field; special case Gaussian graphical models (Schfer and Strimmer, 2005) Dependency networks from sparse regression (Bonneau et al., 2006; Soinov et al., 2003; Rogers and Girolami, 2005) Ordinary differential equation models from regression (Gardner et al., 2003; di Bernardo et al., 2005; Bansal et al., 2006)
X_k for all $k \in V \backslash \{i, j\}$	No third gene can explain the correlation between i and j	First order conditional independence: Gaussian (Wille and Bhlmann, 2006; Wille et al., 2004; Magwene and Kim, 2004) and mutual information (Margolin et al., 2006; Carro et al., 2010)
$\mathbf{X}_S$ for all $S \subseteq V \backslash \{i, j\}$	No subset of all other genes can explain the correlation between i and j	Bayesian networks (Segal et al., 2005; Friedman, 2004; Friedman et al., 2000) Dynamic Bayesian networks (Zou and Conzen, 2005; Perrin et al., 2003; Murphy and Mian, 1999)

Table 1.1: In the conceptual framework by Markowetz and Spang (2007) for inferring genetic regulatory network from gene expression data, an edge is drawn between network components i and j if and only if $X_i \not\perp X_j \mid Z$, and many current methods of network inference can be grouped by the definition of Z. Detailed notations are defined on page 34. This table is a compilation of reviews by Markowetz and Spang (2007); Bansal et al. (2007).

the tyrosine phosphorylation MS, but learning Bayesian network with hidden variables is a theoretically challenging and computationally intensive problem (Chickering and Heckerman, 1996; Friedman, 1997). These factors point to a direction that incorporates prior biological knowledge.

1.3.2 Finding relevant connections from the interactome

One alternative to learning connections between molecules from data *de novo* is to start with a pre-defined network structure from interaction datasets and extract relevant interactions that can explain the data. The interaction datasets are rich in mechanistic information but fall short in context. For instance, the BioGRID database contains entries for protein complexes and phosphorylation reactions, and the TRANSFAC database contains entries for binding of transcription factor to promoter region of a gene. These interactions may define a pathway in the cell types where the experiments were performed, but they may not be applicable in other biological contexts. The predicted protein-protein interactions and transcription factor binding are additionally plagued by false positives (Nguyen and Goodrich, 2006; Wasserman and Sandelin, 2004). Supplementing the interaction data by biological context from high throughput experiments has successfully led to many interesting discoveries. For example, using the yeast protein interactome, Ideker et al. (2002) presented a simulated annealing algorithm to find connected subnetworks of genes that showed unexpectedly high degree of differential expression under a subset of conditions. The resulting subnetworks were consistent with known regulatory circuits and signaling pathways. Yeang et al. (2004) inferred models of transcriptional regulation in yeast by searching for paths of protein-protein and protein-DNA interactions that are consistent with knock-out effects. They were able to provide mechanistic explanations for some knock-out effects and accurately predict the knock-out responses in cross-validation. Scott et al. (2005) found subnetworks that connected a distinguished set of genes (for example, a set of genes differentially regulated under a certain condition) in the most compact way by solving a node-weighted Steiner tree problem, and they argued the resulting subnetworks were likely to contain regulators of the genes in the distinguished set. The algorithm recovered regulatory elements in yeast metabolic pathways. Interestingly, Djebbari and Quackenbush (2008) observed improved performance of Bayesian network learning by starting from a seeded network structure derived from previously known interaction data. However, as the majority of the interactome-based methods till this day have been applied to transcription profiling data in yeast, and to human data in very few cases, there is little evidence to demonstrate the capability of this approach for handling proteomic data from mammalian regulatory networks.

1.4 Biology of EGFRvIII in human glioblastoma

The epidermal growth factor receptor (EGFR; ERBB1; HER1 in human) is a transmembrane protein that is a member of the ErbB family of receptor tyrosine kinases. It is the cell surface receptor for the epidermal growth factor (EGF) protein and other growth factor ligands (Linggi and Carpenter, 2006). Binding of the ligand induces dimerization of the receptor and activation of its tyrosine kinase activity that leads to auto-phosphorylation of several tyrosine residues in the C-terminal domain (Linggi and Carpenter, 2006). These phosphotyrosine residues associate with other signaling proteins to activate downstream pathways such as mitogen activated protein kinase (MAPK), phosphoinositide 3-kinase (PI3K)-Akt, and c-Jun N-terminal kinases (JNK) pathways (Citri and Yarden, 2006; Oda et al., 2005; Yarden and Sliwkowski, 2001), and cellular processes such as DNA synthesis (Roche et al., 1994), cell proliferation (Honegger et al., 1988), apoptosis (Boerner et al., 2004), and cell adhesion (Xie et al., 1998) and migration (Andl et al., 2003). Aberrant signaling by EGFR due to receptor over-expression or mutations has been implicated in many cancers, resulting in poor prognosis and decreased survival (Herbst, 2004; Nicholson et al., 2001). These discoveries have led to active development of anti-cancer therapies targeting EGFR (Modjtahedi and Essapen, 2009; Zhang et al., 2007a; Zandi et al., 2007).

EGFRvIII is a truncated, constitutively active mutant of EGFR (Pedersen et al., 2001). Deletion of exons 2-7 removes most of the extracellular ligand binding domain, so it is unable to bind EGF or other EGFR-binding ligands (Huang et al., 1997). However, this mutant receptor is constitutively phosphorylated (Nishikawa et al., 1994). The receptor is capable of activating downstream signaling pathways, but the low level of phosphorylation appears insufficient to trigger receptor-mediated down-regulation, contributing to the transforming ability of this mutant (Huang et al., 1997). It is the most common deletion mutant of EGFR in human cancer (Pedersen et al., 2001) and is highly correlated with and poor prognosis in glioblastoma multiforme (Pelloski et al., 2007; Heimbèrger et al., 2005; Feldkamp et al., 1999).

The relevance of EGFRvIII in human cancer has motivated much work to elucidate the downstream signaling events activated by this receptor but many questions remain. The mutant receptor displays signaling properties different from the ligand-activated EGFR, with a largely inactive MAPK pathway (Moscatello et al., 1996) and a highly active PI3K pathway (Moscatello et al., 1998). Few report the consequences of these signals on the activity of the transcription factors and the regulated genes, and the results are often contradictory. A microarray experiment on mouse fibroblasts expressing EGFRvIII reports that a group of interferon response genes is up-regulated by EGF stimulus but not by EGFRvIII expression, and the up-regulation is correlated with the activation of STAT3 and STAT5 transcription factors (Pedersen et al., 2005). In the U87 human glioblastoma cell line, STAT3 is persistently active to bind DNA but this binding is negatively regulated by the PI3K-Akt pathway (Ghosh et al., 2005). Finally, activated STAT3 is significantly correlated with EGFRvIII in gliomas (Mizoguchi et al., 2006). While the inconsistencies may be due to cell-type differences, it is possible that the few phosphorylation sites measured on selected signaling molecules cannot fully represent the activation state of the molecules or the pathway. Therefore, systematic measurements and modeling are necessary to provide a clearer picture of the signaling events and transcriptional responses.

The model system in this study is the U87 human glioblastoma cell line engineered to express titrated levels of EGFRvIII and the tumorigenicity is correlated with the expression level of the mutant receptor (Huang et al., 1997; Nishikawa et al., 1994). It is a good starting point for methodology development. The wild-type EGFR signaling network is very well established, but the events and connections downstream of EGFRvIII are poorly characterized. So comparing our results to the wild-type network will give insights to the oncogenic mechanism of this mutant. Mechanisms like the activation of distinct transcription factors by EGFR and EGFRvIII (Fromm et al., 2008) will be informative in design of therapeutics.

1.5 Motivation and innovation

In this thesis I present a computational method for joint analysis of phosphoproteomic and transcriptome data and argue that this conceptual framework provides an intuitive approach that brings together multiple heterogeneous data sources. I will show how we can bridge the gap between signaling and transcription to gain better understanding of an important problem in cancer biology, that is, the long term consequence of aberrant signaling in cancer. From an algorithmic perspective we aim to fill the need of connecting *de novo* discovery of network topology and detailed modeling of regulatory dynamics. Lastly I propose that this framework is especially suitable for generating novel hypotheses given our current experimental capability of measuring many things under few numbers of conditions.

As described above, recent technological advances make it possible for the first time to develop an unbiased and systematic view of the proteomic and transcriptional changes that occur during oncogenesis. Analysis of these two kinds of datasets individually have provided novel insights into the regulatory dynamics at the level of the proteome (Wolf-Yadlin et al., 2006; Zhang et al., 2005) and transcriptome (Amit et al., 2007). Despite these advances, knowledge of the connections between phosphoproteomic signaling changes and transcriptional networks remains fragmentary. Significant uncertainties remain even for some pathways that have been very well characterized. The STAT DNA-binding proteins, for example, are a well studied part of the EGF pathway. At least twelve distinct STAT isoforms are present in humans, which can be activated by kinases from the JAK, STAT or EGFR families (Lim and Cao, 2006; Quesnelle et al., 2007). There are a multitude of potential interactions among all these proteins that may have distinct roles in signaling and transcriptional regulation. Undoubtedly, there are also many other connections between proteomic and transcriptional changes mediated by pathways that remain to be discovered. Jointly analyzing the transcriptional and proteomic data may provide new insights into these questions.

Integrating transcriptional and proteomic data will require novel computational approaches. In particular, because not all regulatory events are mediated by protein phosphorylation, even the most comprehensive phosphoproteomics technologies cannot capture all these events. Computational techniques are needed to discover proteins that participate in the signaling networks but are undetected in the experiments. One approach may be to fill in the paths between phosphorylated proteins by known pathway models. However, using this approach with currently curated pathways ignores the information encoded in a large fraction of the phosphoproteomic data that do not map to the curated pathways. In fact, when these proteomic technologies are applied to well-characterized responses, it is clear that overlap between the data and the components of the known pathways is poor. Many important signaling proteins are absent from the data; at same time, many of the proteins that show proteomic changes have no obvious connection to the process under study. We have observed this in data from yeast (Huang and Fraenkel, 2009) and human (Section 1.2.1).

A recent review by Hyduke and Palsson (2010) for global reconstruction of signaling networks recognizes that methods for filling the missing information in signaling network models are less established compared to metabolic network models and proposes that interactome datasets represent a promising direction for progress in this area. In the context of finding connections between signaling and transcription changes, we start with a collection of protein-protein and protein-DNA interactions, which represent known or experimentally determined signaling and regulatory connections, and consider the observed phosphorylation events and differential gene expression as connectivity constraints that the reconstructed network must satisfy. Additionally, we take into account the different confidence levels among the interaction data sources by preferentially selecting the more reliable interactions. We show that these objectives can be formulated as a constraint network optimization problem, in particular, as a prize collecting Steiner tree (PCST) problem on the interactome network. Since the interactions are not limited to known pathways and the phosphorylation events and differential expressed genes are not limited to known players in these pathways, there is great potential for novel discoveries. On the other hand, all the interactions were experimentally determined and therefore have mechanistic basis that might become relevant in the current context. I believe that these two features of the method strike a good balance between finding novel connections and revealing the relevance of known connections.

The challenge of making sensible interpretation of the vast amount of genomic and proteomic data is such a growing concern that warranted coverage by a special issue of Science in February 2011. In general these high-throughput technologies have become very good at measuring many molecules at very few number of conditions. Therefore, with the exception of the strongest hits, each of these assays is a source of abundant, weak evidence of regulatory events. The core computational method developed in this thesis is an example of how we can take advantage of the heterogeneity of the datasets, including phosphoproteomic MS, ChIP-seq, DNase-seq, transcription profiling and interactions. As each of these techniques provides a different view of the molecular regulatory network, by putting them together we can generate high confidence hypotheses that have biological relevance and can be tested experimentally. This framework may represent a direction in which to organize the data and enhance our understanding of the cell at the systems level.

Bibliography

I. Amit, A. Citri, T. Shay, Y. Lu, M. Katz, F. Zhang, G. Tarcic, D. Siwak, J. Lahad, J. Jacob-Hirsch, N. Amariglio, N. Vaisman, E. Segal, G. Rechavi, U. Alon, G. B. Mills, E. Domany, and Y. Yarden. A module of negative feedback regulators defines growth factor signaling. *Nat Genet*, 39(4):503–512, Apr 2007.

C. D. Andl, T. Mizushima, H. Nakagawa, K. Oyama, H. Harada, K. Chruma, M. Herlyn, and A. K. Rustgi. Epidermal growth factor receptor mediates increased cell proliferation, migration, and aggregation in esophageal keratinocytes in vitro and in vivo. *J Biol Chem*, 278(3):1824–1830, Jan 2003.

S. Bader, S. Khner, and A.-C. Gavin. Interaction networks for systems biology. *FEBS Lett*, 582(8):1220–1224, Apr 2008.

M. Bansal, G. D. Gatta, and D. di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, Apr 2006.

M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol Syst Biol*, 3:78, 2007.

T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res*, 39(Database issue):D1005–D1010, Jan 2011.

A. Bauer-Mehren, L. I. Furlong, and F. Sanz. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol Syst Biol*, 5:290, 2009.

M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, and M. L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–1435, Nov 2006.

T. Berggrd, S. Linse, and P. James. Methods for the detection and analysis of protein-protein interactions. *Proteomics*, 7(16):2833–2842, Aug 2007.

U. S. Bhalla and R. Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, Jan 1999.

J. P. Blaydes, B. Vojtesek, G. B. Bloomberg, and T. R. Hupp. The development and use of phospho-specific antibodies to study protein phosphorylation. *Methods Mol Biol*, 99:177–189, 2000.

J. L. Boerner, M. L. Demory, C. Silva, and S. J. Parsons. Phosphorylation of y845 on the epidermal growth factor receptor mediates binding to the mitochondrial protein cytochrome c oxidase subunit ii. *Mol Cell Biol*, 24(16):7059–7071, Aug 2004.

R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol*, 7(5):R36, 2006.

S. Boura-Halfon and Y. Zick. Phosphorylation of IRS proteins, insulin action, and insulin resistance. *Am J Physiol Endocrinol Metab*, 296(4):E581–E591, Apr 2009.

A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, Jan 2008.

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

M. Brenowitz, D. F. Senear, M. A. Shea, and G. K. Ackers. Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Methods Enzymol*, 130:132–181, 1986.

J. C. Bryne, E. Valen, M.-H. E. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, 36(Database issue):D102–D106, Jan 2008.

Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, Oct 2008.

M. F. Carey, C. L. Peterson, and S. T. Smale. Chromatin immunoprecipitation (ChIP). *Cold Spring Harb Protoc*, 2009(9):pdb.prot5279, Sep 2009.

M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape, A. Califano, and A. Iavarone. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, Jan 2010.

G. S. Chaga. Antibody arrays for determination of relative protein abundances. *Methods Mol Biol*, 441:129–151, 2008.

F. Chang, L. S. Steelman, J. T. Lee, J. G. Shelton, P. M. Navolanic, W. L. Blalock, R. A. Franklin, and J. A. McCubrey. Signal transduction mediated by the Ras/Raf/MEK/ERK pathway from cytokine receptors to transcription factors: potential targeting for therapeutic intervention. *Leukemia*, 17(7):1263–1293, Jul 2003.

A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni. MINT: the Molecular INTeraction database. *Nucleic Acids Res*, 35(Database issue):D572–D574, Jan 2007.

G. Chaurasia, Y. Iqbal, C. Hnig, H. Herzel, E. E. Wanker, and M. E. Futschik. UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res*, 35(Database issue):D590–D594, Jan 2007.

D. M. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. In *Machine Learning*, pages 181–212, 1996.

M. F. Ciaccio, J. P. Wagner, C.-P. Chuu, D. A. Lauffenburger, and R. B. Jones. Systems analysis of EGF receptor signaling dynamics with microwestern arrays. *Nat Methods*, 7(2):148–155, Feb 2010.

A. Citri and Y. Yarden. Egf-erbb signalling: towards the systems level. *Nat Rev Mol Cell Biol*, 7(7):505–516, Jul 2006.

P. A. Cole, K. Shen, Y. Qiao, and D. Wang. Protein tyrosine kinases Src and Csk: a tail's tale. *Curr Opin Chem Biol*, 7(5):580–585, Oct 2003.

P. Collas. The current state of chromatin immunoprecipitation. *Mol Biotechnol*, 45 (1):87–100, May 2010.

M. E. Cusick, N. Klitgord, M. Vidal, and D. E. Hill. Interactome: gateway into systems biology. *Hum Mol Genet*, 14 Spec No. 2:R171–R181, Oct 2005.

T. Decker and P. Kovarik. Serine phosphorylation of STATs. *Oncogene*, 19(21): 2628–2637, May 2000.

S. Desrivires, C. Kunz, I. Barash, V. Vafaizadeh, C. Borghouts, and B. Groner. The biological functions of the versatile transcription factors STAT3 and STAT5 and new strategies for their targeted inhibition. *J Mammary Gland Biol Neoplasia*, 11 (1):75–87, Jan 2006.

P. D'haeseleer. What are DNA sequence motifs? *Nat Biotechnol*, 24(4):423–425, Apr 2006.

D. di Bernardo, M. J. Thompson, T. S. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. E. Schaus, and J. J. Collins. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol*, 23(3):377–383, Mar 2005.

A. Djebbari and J. Quackenbush. Seeded Bayesian Networks: constructing genetic networks from microarray data. *BMC Syst Biol*, 2:57, 2008.

M. Dreze, D. Monachello, C. Lurin, M. E. Cusick, D. E. Hill, M. Vidal, and P. Braun. High-quality binary interactome mapping. *Methods Enzymol*, 470:281–315, 2010.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25): 14863–14868, Dec 1998.

M. M. Feldkamp, P. Lala, N. Lau, L. Roncari, and A. Guha. Expression of activated epidermal growth factor receptors, Ras-guanosine triphosphate, and mitogen-activated protein kinase in human glioblastoma multiforme specimens. *Neurosurgery*, 45(6):1442–1453, Dec 1999.

B. C. Foat, S. S. Houshmandi, W. M. Olivas, and H. J. Bussemaker. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci U S A*, 102(49):17675–17680, Dec 2005.

B. C. Foat, A. V. Morozov, and H. J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22(14):e141–e149, Jul 2006.

N. Friedman. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 125–133. Morgan Kaufmann, 1997.

N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, Feb 2004.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *J Comput Biol*, 7(3-4):601–620, 2000.

J. A. Fromm, S. A. S. Johnson, and D. L. Johnson. Epidermal growth factor receptor 1 (egfrl) and its variant egfrviii regulate tata-binding protein expression through distinct pathways. *Mol Cell Biol*, 28(20):6483–6495, Oct 2008.

D. J. Galas and A. Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res*, 5(9):3157–3170, Sep 1978.

T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301 (5629):102–105, Jul 2003.

A.-C. Gavin, M. Bsche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Hfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, Jan 2002.

M. K. Ghosh, P. Sharma, P. C. Harbor, S. O. Rahaman, and S. J. Haque. PI3K-AKT pathway negatively controls EGFR-dependent DNA-binding activity of Stat3 in glioblastoma multiforme cells. *Oncogene*, 24(49):7290–7300, Nov 2005.

J. M. Gil and A. C. Rego. Mechanisms of neurodegeneration in Huntington's disease. *Eur J Neurosci*, 27(11):2803–2820, Jun 2008.

P. G. Giresi and J. D. Lieb. How to find an opening (or lots of them). *Nat Methods*, 3(7):501–502, Jul 2006.

C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen. Co-evolution of proteins with their interaction partners. *J Mol Biol*, 299(2):283–293, Jun 2000.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.

P. A. Grimsrud, D. L. Swaney, C. D. Wenger, N. A. Beauchene, and J. J. Coon. Phosphoproteomics for the masses. *ACS Chem Biol*, 5(1):105–119, Jan 2010.

D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000.

D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, Mar 2011.

A. B. Heimberger, R. Hlatky, D. Suki, D. Yang, J. Weinberg, M. Gilbert, R. Sawaya, and K. Aldape. Prognostic effect of epidermal growth factor receptor and EGFRvIII in glioblastoma multiforme patients. *Clin Cancer Res*, 11(4):1462–1466, Feb 2005.

R. S. Herbst. Review of epidermal growth factor receptor biology. *Int J Radiat Oncol Biol Phys*, 59(2 Suppl):21–26, 2004.

J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields, and J. A. Stamatoyannopoulos. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, 6(4):283–289, Apr 2009.

Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Srensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868): 180–183, Jan 2002.

A. Honegger, T. J. Dull, F. Bellot, E. V. Obberghen, D. Szapary, A. Schmidt, A. Ullrich, and J. Schlessinger. Biological activities of egf-receptor mutants with individually altered autophosphorylation sites. *EMBO J*, 7(10):3045–3052, Oct 1988.

H. S. Huang, M. Nagane, C. K. Klingbeil, H. Lin, R. Nishikawa, X. D. Ji, C. M. Huang, G. N. Gill, H. S. Wiley, and W. K. Cavenee. The enhanced tumorigenic activity of a mutant epidermal growth factor receptor common in human cancers is mediated by threshold levels of constitutive tyrosine phosphorylation and unattenuated signaling. *J Biol Chem*, 272(5):2927–2935, Jan 1997.

P. H. Huang, A. Mukasa, R. Bonavia, R. A. Flynn, Z. E. Brewer, W. K. Cavenee, F. B. Furnari, and F. M. White. Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. *Proc Natl Acad Sci U S A*, 104(31):12867–12872, Jul 2007.

S.-S. C. Huang and E. Fraenkel. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal*, 2(81):ra40, 2009.

M. Huynen, B. Snel, W. Lathe, and P. Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*, 10(8): 1204–1210, Aug 2000.

D. Hwang, J. J. Smith, D. M. Leslie, A. D. Weston, A. G. Rust, S. Ramsey, P. de Atauri, A. F. Siegel, H. Bolouri, J. D. Aitchison, and L. Hood. A data integration methodology for systems biology: experimental verification. *Proc Natl Acad Sci U S A*, 102(48):17302–17307, Nov 2005.

D. R. Hyduke and B. . Palsson. Towards genome-scale signalling-network reconstructions. *Nat Rev Genet*, 11(4):297–307, Feb 2010.

T. Ideker and D. Lauffenburger. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol*, 21(6):255–262, Jun 2003.

T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1: S233–S240, 2002.

S. Imarisio, J. Carmichael, V. Korolchuk, C.-W. Chen, S. Saiki, C. Rose, G. Krishna, J. E. Davies, E. Ttofi, B. R. Underwood, and D. C. Rubinsztein. Huntington's disease: from pathology and genetics to potential therapies. *Biochem J*, 412(2): 191–209, Jun 2008.

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

D. Iyer, D. Chang, J. Marx, L. Wei, E. N. Olson, M. S. Parmacek, A. Balasubramanyam, and R. J. Schwartz. Serum response factor MADS box serine-162 phosphorylation switches proliferation and myogenic gene programs. *Proc Natl Acad Sci U S A*, 103(12):4516–4521, Mar 2006.

R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644): 449–453, Oct 2003.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, Jun 2007.

A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanp, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen, and J. Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, 20(6):861–873, Jun 2010.

M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*, 38(Database issue):D355–D360, Jan 2010.

S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. IntAct–open source resource for molecular interaction data. *Nucleic Acids Res*, 35 (Database issue):D561–D565, Jan 2007.

P. O. Krutzik, J. M. Irish, G. P. Nolan, and O. D. Perez. Analysis of protein phosphorylation and cellular signaling events by flow cytometry: techniques and clinical applications. *Clin Immunol*, 110(3):206–221, Mar 2004.

C. Kunsch, S. M. Ruben, and C. A. Rosen. Selection of optimal kappa B/Rel DNA-binding motifs: interaction of both subunits of NF-kappa B with DNA is required for transcriptional activation. *Mol Cell Biol*, 12(10):4412–4421, Oct 1992.

B. T. Kurien and R. H. Scofield. Introduction to protein blotting. *Methods Mol Biol*, 536:9–22, 2009.

C. P. Lim and X. Cao. Serine phosphorylation and negative regulation of Stat3 by JNK. *J Biol Chem*, 274(43):31055–31061, Oct 1999.

C. P. Lim and X. Cao. Structure, function, and regulation of STAT proteins. *Mol Biosyst*, 2(11):536–550, Nov 2006.

R. Linding, L. J. Jensen, G. J. Ostheimer, M. A. T. M. van Vugt, C. Jrgensen, I. M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J. G. Park, L. D. Samson, J. R. Woodgett, R. B. Russell, P. Bork, M. B. Yaffe, and T. Pawson. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–1426, Jun 2007.

B. Linggi and G. Carpenter. Erbb receptors: new insights on mechanisms and biology. *Trends Cell Biol*, 16(12):649–656, Dec 2006.

R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, 21(1 Suppl):20–24, Jan 1999.

D. J. Lockhart and E. A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405(6788):827–836, Jun 2000.

P. M. Magwene and J. Kim. Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol*, 5(12):R100, 2004.

J. W. Mandell. Phosphorylation state-specific antibodies: applications in investigative and diagnostic pathology. *Am J Pathol*, 163(5):1687–1698, Nov 2003.

T. Manke, H. G. Roider, and M. Vingron. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Comput Biol*, 4(3): e1000039, Mar 2008.

E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, Jul 1999.

A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.

F. Markowetz and R. Spang. Inferring cellular networks–a review. *BMC Bioinformatics*, 8 Suppl 6:S5, 2007.

G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7:29–59, 2006.

L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, 37 (Database issue):D619–D622, Jan 2009.

V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, Jan 2006.

M. Mizoguchi, R. A. Betensky, T. T. Batchelor, D. C. Bernay, D. N. Louis, and C. L. Nutt. Activation of STAT3, MAPK, and AKT in malignant astrocytic gliomas: correlation with EGFR status, tumor grade, and survival. *J Neuropathol Exp Neurol*, 65(12):1181–1188, Dec 2006.

H. Modjtahedi and S. Essapen. Epidermal growth factor receptor inhibitors in cancer treatment: advances, challenges and opportunities. *Anticancer Drugs*, 20(10):851–855, Nov 2009.

D. K. Moscatello, R. B. Montgomery, P. Sundareshan, H. McDanel, M. Y. Wong, and A. J. Wong. Transformational and altered signal transduction by a naturally occurring mutant EGF receptor. *Oncogene*, 13(1):85–96, Jul 1996.

D. K. Moscatello, M. Holgado-Madruga, D. R. Emlet, R. B. Montgomery, and A. J. Wong. Constitutive activation of phosphatidylinositol 3-kinase by a naturally occurring mutant epidermal growth factor receptor. *J Biol Chem*, 273(1):200–206, Jan 1998.

K. Murphy and S. Mian. Modelling Gene Expression Data using Dynamic Bayesian Networks. Technical report, 1999.

Nat Rev Genet Article Series. Applications of next-generation sequencing. http://www.nature.com/nrg/series/nextgeneration/index.html, 2011.

T. N. Nguyen and J. A. Goodrich. Protein-protein interaction assays: eliminating false positive interactions. *Nat Methods*, 3(2):135–139, Feb 2006.

R. I. Nicholson, J. M. Gee, and M. E. Harper. Egfr and cancer prognosis. *Eur J Cancer*, 37 Suppl 4:S9–15, Sep 2001.

R. Nishikawa, X. D. Ji, R. C. Harmon, C. S. Lazar, G. N. Gill, W. K. Cavenee, and H. J. Huang. A mutant epidermal growth factor receptor common in human glioma confers enhanced tumorigenicity. *Proc Natl Acad Sci U S A*, 91(16):7727–7731, Aug 1994.

J. C. Obenauer, L. C. Cantley, and M. B. Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13):3635–3641, Jul 2003.

N. G. Oberprieler and K. Taskn. Analysing phosphorylation-based signalling networks by phospho flow cytometry. *Cell Signal*, 23(1):14–18, Jan 2011.

K. Oda, Y. Matsuoka, A. Funahashi, and H. Kitano. A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol*, 1:2005.0010, 2005.

D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford, and E. Fraenkel. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, 39(6):730–732, Jun 2007.

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

Z. Ouyang, Q. Zhou, and W. H. Wong. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A*, 106(51):21521–21526, Dec 2009.

P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, Oct 2009.

T. Pawson and P. Nash. Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618):445–452, Apr 2003.

M. W. Pedersen, M. Meltorn, L. Damstrup, and H. S. Poulsen. The type III epidermal growth factor receptor mutation. Biological significance and potential target for anti-cancer therapy. *Ann Oncol*, 12(6):745–760, Jun 2001.

M. W. Pedersen, N. Pedersen, L. Damstrup, M. Villingshj, S. U. Snder, K. Rieneck, L. F. Bovin, M. Spang-Thomsen, and H. S. Poulsen. Analysis of the epidermal growth factor receptor specific transcriptome: effect of receptor expression level and an activating mutation. *J Cell Biochem*, 96(2):412–427, Oct 2005.

C. E. Pelloski, K. V. Ballman, A. F. Furth, L. Zhang, E. Lin, E. P. Sulman, K. Bhat, J. M. McDonald, W. K. A. Yung, H. Colman, S. Y. Woo, A. B. Heimberger, D. Suki, M. D. Prados, S. M. Chang, F. G. Barker, J. C. Buckner, C. D. James, and K. Aldape. Epidermal growth factor receptor variant III status defines clinically distinct subtypes of glioblastoma. *J Clin Oncol*, 25(16):2288–2294, Jun 2007.

B.-E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alch Buc. Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, 19 Suppl 2: ii138–ii148, Oct 2003.

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

C. Prieto and J. D. L. Rivas. APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res*, 34(Web Server issue):W298–W302, Jul 2006.

K. M. Quesnelle, A. L. Boehm, and J. R. Grandis. STAT-mediated EGFR signaling in cancer. *J Cell Biochem*, 102(2):311–319, Oct 2007.

C. J. Roberts, B. Nelson, M. J. Marton, R. Stoughton, M. R. Meyer, H. A. Bennett, Y. D. He, H. Dai, W. L. Walker, T. R. Hughes, M. Tyers, C. Boone, and S. H. Friend. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*, 287(5454):873–880, Feb 2000.

S. Roche, M. Koegl, and S. A. Courtneidge. The phosphatidylinositol 3-kinase alpha is required for dna synthesis induced by some, but not all, growth factors. *Proc Natl Acad Sci U S A*, 91(19):9185–9189, Sep 1994.

S. Rogers and M. Girolami. A Bayesian regression approach to the inference of regulatory networks from gene expression data. *Bioinformatics*, 21(14):3131–3137, Jul 2005.

H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, Jan 2007.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308 (5721):523–529, Apr 2005.

L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32 (Database issue):D449–D451, Jan 2004.

L. Salwinski, L. Licata, A. Winter, D. Thorneycroft, J. Khadake, A. Ceol, A. C. Aryamontri, R. Oughtred, M. Livstone, L. Boucher, D. Botstein, K. Dolinski, T. Berardini, E. Huala, M. Tyers, D. Eisenberg, G. Cesareni, and H. Hermjakob. Recurated protein interaction datasets. *Nat Methods*, 6(12):860–861, Dec 2009.

A. Sandelin, W. Alkema, P. Engstrm, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, Jan 2004.

M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270 (5235):467–470, Oct 1995.

S. Schinner, W. A. Scherbaum, S. R. Bornstein, and A. Barthel. Molecular mechanisms of insulin resistance. *Diabet Med*, 22(6):674–682, Jun 2005.

K. Schmelzle, S. Kane, S. Gridley, G. E. Lienhard, and F. M. White. Temporal dynamics of tyrosine phosphorylation in insulin signaling. *Diabetes*, 55(8):2171–2179, Aug 2006.

J. Schfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, Mar 2005.

M. S. Scott, T. Perkins, S. Bunnell, F. Pepin, D. Y. Thomas, and M. Hallett. Identifying regulatory subnetworks for a set of genes. *Mol Cell Proteomics*, 4(5):683–692, May 2005.

R. Sears, F. Nuckolls, E. Haura, Y. Taya, K. Tamai, and J. R. Nevins. Multiple Ras-dependent phosphorylation pathways regulate Myc protein stability. *Genes Dev*, 14(19):2501–2514, Oct 2000.

E. Segal, D. Pe'er, A. Regev, D. Koller, and N. Friedman. Learning Module Networks. *J. Mach. Learn. Res.*, 6:557–588, December 2005. ISSN 1532-4435.

L. A. Soinov, M. A. Krestyaninova, and A. Brazma. Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biol*, 4(1):R6, 2003.

L. Song and G. E. Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, 2010(2):pdb.prot5384, Feb 2010.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, Dec 1998.

F. J. T. Staal, M. van der Burg, L. F. A. Wessels, B. H. Barendregt, M. R. M. Baert, C. M. M. van den Burg, C. van Huffel, A. W. Langerak, V. H. J. van der Velden, M. J. T. Reinders, and J. J. M. van Dongen. DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia*, 17(7):1324–1332, Jul 2003.

C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. V. Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*, 39(Database issue):D698–D704, Jan 2011.

R. Stoltenburg, C. Reinemann, and B. Strehlitz. SELEX–a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng*, 24(4):381–403, Oct 2007.

G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16 (1):16–23, Jan 2000.

S. J. H. Sui, D. L. Fulton, D. J. Arenillas, A. T. Kwon, and W. W. Wasserman. oposum: integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res*, 35(Web Server issue):W245–W252, Jul 2007.

V. G. Tarcea, T. Weymouth, A. Ade, A. Bookvich, J. Gao, V. Mahavisno, Z. Wright, A. Chapman, M. Jayapandian, A. Ozgr, Y. Tian, J. Cavalcoli, B. Mirel, J. Patel, D. Radev, B. Athey, D. States, and H. V. Jagadish. Michigan molecular interactions r2: from interacting proteins to pathways. *Nucleic Acids Res*, 37(Database issue): D642–D646, Jan 2009.

S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nat Genet*, 22(3):281–285, Jul 1999.

C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010.

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, Feb 2000.

R. G. W. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, D. N. Hayes, and C. G. A. R. Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110, Jan 2010.

C. von Mering, L. J. Jensen, B. Snel, S. D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M. A. Huynen, and P. Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 33 (Database issue):D433–D437, Jan 2005.

W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–287, Apr 2004.

A. V. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531, Oct 2006.

A. Wille and P. Bhlmann. Low-order conditional independence graphs for inferring genetic networks. *Stat Appl Genet Mol Biol*, 5:Article1, 2006.

A. Wille, P. Zimmermann, E. Vranov, A. Frholz, O. Laule, S. Bleuler, L. Hennig, A. Prelic, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bhlmann. Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol*, 5(11):R92, 2004.

E. Wingender. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform*, 9(4):326–332, Jul 2008.

A. Wolf-Yadlin, N. Kumar, Y. Zhang, S. Hautaniemi, M. Zaman, H.-D. Kim, V. Grantcharova, D. A. Lauffenburger, and F. M. White. Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Mol Syst Biol*, 2:54, 2006.

H. Xie, M. A. Pallero, K. Gupta, P. Chang, M. F. Ware, W. Witke, D. J. Kwiatkowski, D. A. Lauffenburger, J. E. Murphy-Ullrich, and A. Wells. Egf receptor regulation of cell motility: Egf induces disassembly of focal adhesions independently of the motility-associated plcgamma signaling pathway. *J Cell Sci*, 111 (Pt 5):615–624, Mar 1998.

M. B. Yaffe. Phosphotyrosine-binding domains in signal transduction. *Nat Rev Mol Cell Biol*, 3(3):177–186, Mar 2002.

Y. Yarden and M. X. Sliwkowski. Untangling the erbb signalling network. *Nat Rev Mol Cell Biol*, 2(2):127–137, Feb 2001.

C.-H. Yeang, T. Ideker, and T. Jaakkola. Physical network models. *J Comput Biol*, 11(2-3):243–262, 2004.

S. Yousefi, D. R. Green, K. Blaser, and H. U. Simon. Protein-tyrosine phosphorylation regulates apoptosis in human eosinophils and neutrophils. *Proc Natl Acad Sci U S A*, 91(23):10868–10872, Nov 1994.

H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabsi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, Oct 2008.

R. Zandi, A. B. Larsen, P. Andersen, M.-T. Stockhausen, and H. S. Poulsen. Mechanisms for oncogenic activation of the epidermal growth factor receptor. *Cell Signal*, 19(10):2013–2023, Oct 2007.

H. Zhang, A. Berezov, Q. Wang, G. Zhang, J. Drebin, R. Murali, and M. I. Greene. Erbb receptors: from oncogenes to targeted cancer therapies. *J Clin Invest*, 117 (8):2051–2058, Aug 2007a.

Y. Zhang, A. Wolf-Yadlin, P. L. Ross, D. J. Pappin, J. Rush, D. A. Lauffenburger, and F. M. White. Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Mol Cell Proteomics*, 4(9):1240–1250, Sep 2005.

Y. Zhang, A. Wolf-Yadlin, and F. M. White. Quantitative proteomic analysis of phosphotyrosine-mediated cellular signaling networks. *Methods Mol Biol*, 359:203–212, 2007b.

M. Zou and S. D. Conzen. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79, Jan 2005.

Chapter 2

Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks: case study of the yeast pheromone response network

2.1 Summary

This chapter presents a test case where the PCST approach was applied to the phosphoproteomic and transcriptional data from yeast pheromone response. We chose this system because there is good coverage of the protein-protein and protein-DNA interactions in yeast compared to mammalian systems, and the signaling and transcriptional components of the pheromone response network are well characterized. Nonetheless, the results are still very interesting, especially considering the fact that many of the hits from these datasets do not fall into annotated signaling pathways.

This work was previously published in *Science Signaling* July 28 2009, Vol. 2, Issue 81, page ra40. The American Association for the Advancement of Science (AAAS), the publisher of Science Signaling, grants automatic permission to manuscript authors to reprint the work for inclusion in a dissertation of the author. The details are described in the License to Publish agreement posted on the *Science* website (`http://www.sciencemag.org/site/feature/contribinfo/prep/license.xhtml`).

2.2 Manuscript: Huang and Fraenkel, *Sci Signal* 2: ra40 (2009)

NIH Public Access
Author Manuscript
Sci Signal. Author manuscript; available in PMC 2010 June 22.

Published in final edited form as:
Sci Signal. ; 2(81): ra40. doi:10.1126/scisignal.2000350.

Integration of Proteomic, Transcriptional, and Interactome Data Reveals Hidden Signaling Components

Shao-shan Carol Huang[1] and **Ernest Fraenkel**[2,3]

[1]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA 02139, USA.

Abstract

Cellular signaling and regulatory networks underlie fundamental biological processes such as growth, differentiation, and response to the environment. Although there are now various high-throughput methods for studying these processes, knowledge of them remains fragmentary. Typically, the vast majority of hits identified by transcriptional, proteomic, and genetic assays lie outside of the expected pathways. These unexpected components of the cellular response are often the most interesting, because they can provide new insights into biological processes and potentially reveal new therapeutic approaches. However, they are also the most difficult to interpret. We present a technique, based on the Steiner tree problem, that uses previously reported protein-protein and protein-DNA interactions to determine how these hits are organized into functionally coherent pathways, revealing many components of the cellular response that are not readily apparent in the original data. Applied simultaneously to phosphoproteomic and transcriptional data for the yeast pheromone response, it identifies changes in diverse cellular processes that extend far beyond the expected pathways.

INTRODUCTION

High-throughput experimental techniques provide unprecedented views of the molecular changes that occur in cells as they respond to stimuli. Because many of these techniques are not dependent on prior knowledge of the relevant pathways, they provide a systematic view of signaling and regulatory changes that can uncover previously unrecognized components of these responses (1–4). For example, high-throughput genetic screening identifies sets of genes whose expression changes lead to altered phenotype and, therefore, the products of these genes are likely to be involved in the regulatory pathways (4,5). Mass-spectrometry techniques can provide quantitative measurements of signaling events in the form of peptide or phosphopeptide abundance (6–9). At the level of transcription, changes in the expression of thousands of genes are readily obtained by microarrays. At the interface of protein and transcription, chromatin immunoprecipitation (ChIP) followed by array hybridization or sequencing reports whole genome protein-DNA binding interactions (10,11).

These system-wide datasets often reveal that our current understanding of regulatory networks at the systems level remains incomplete, even in extremely well-characterized systems. For

Correspondence should be addressed to E.F. (fraenkel-admin@mit.edu).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

example, the mitogen-activated protein kinase (MAPK) cascade in the yeast *Saccharomyces cerevisiae* that responds to mating pheromone has been extensively studied and the most important transcription factors regulated by this process are known (12). However, when cells are exposed to pheromone, differentially phosphorylated sites are detected on more than 100 proteins (7), only about 10% of which are known components in the MAPK cascade, and more than 70% are not present in any of the yeast pathways annotated in the KEGG Pathway database (13). Of the hundreds of genes that are differentially transcribed (3), a majority of them are not known to be regulated by the transcription factors included in the MAPK cascade.

The number of unexpected components of the cellular response even in such a well-studied system presents both a challenge and an opportunity for systems biology approaches. Computational methods that can give context to these observations have the potential to reveal more comprehensive views of cellular responses. Any computational approach for this purpose must overcome the fact that not all components in the regulatory networks can be exposed in one experiment due to systematic biases in the assays. For example, compensatory mechanisms can mask the consequences of genetic manipulations. Thus, despite their important roles in mating type signaling, the yeast MAPK-encoding genes *FUS3* and *KSS1* are not detected in genetic screens for mating defects, because they are functionally redundant (14). Similarly, due to many posttranslational regulation mechanisms that do not affect protein concentrations, changes in many important components of signaling pathways escape detection by even the most comprehensive proteomic technologies. For instance, after stimulation by the pheromone alpha-factor (α-factor), the yeast α-factor receptor STE2 activates the trimeric G protein (composed of the subunits GPA1, STE4, and STE18) through conformational changes, so it was not surprising that these proteins were not detected by a mass-spectrometry experiment (7). Although not reported by the assays, these "hidden" components are critical for understanding the cellular response of interest.

We present a method for constructing a network of protein-protein and protein-DNA interactions, including hidden components, that explains the functional context of genes and proteins detected in these assays. This approach takes advantage of the large number of reported protein-protein and protein-DNA interactions present in the interactome. An interactome-based method is attractive, because it not only contains molecular pathways known to be relevant, but also expands beyond these pathways for novel biological insights. Clearly, reconstructing response pathways in the cell from the interactome is more complicated than simply assembling all the interactions that link the proteins or genes reported by the experiments. Because not all molecules in the regulatory networks are detected, the hits identified may be connected by direct or indirect interactions. The "hidden nodes" that were not experimentally detected but that link the proteins or genes detected are often critical for interpreting the functional significance of the data. However, allowing for such indirect connections between proteins and genes in the interactome quickly leads to a combinatorial explosion of potential paths that are not informative.

To discover meaningful regulatory networks linking the identified genes, a few previous studies combine information from phenotypic or expression experiments with a protein-protein interaction network and search for regions that are enriched for the phenotype or differential expression (15–19). Methods interested in transcriptional regulation search for paths less than a predefined length from the stimulus to transcription factor binding activity (20,21). However, most of these techniques do not explicitly consider the dramatically different reliability of the interaction data, which is especially problematic when the interactome is built from multiple databases or experimental sources. In addition to the varying quality of interaction datasets, we recognize that some of the input proteins or genes should not be connected either because they are false positives or because the true pathways that link them to the rest of the dataset are not present in the currently known interactome.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

These two issues were taken into account previously in the context of connecting genetic data and differentially transcribed genes by starting from the interactome and applying a flow-based approach (4,22) or building a physical network model (21). The flow-based approach is designed to find connections linking a set of differentially transcribed genes to a second set of genetic hits that represent the upstream signal. However, applying this approach to phosphoproteomic and transcriptional data is likely to miss many functionally relevant connections within the proteomic data because these connections lack a direct link to transcriptional changes. The physical network model algorithm requires the phenotypic and transcriptional response of the genetic knockouts as input, so it cannot be applied when such data is not available or to other types of signaling data.

Here, we propose to address the problems outlined above by taking a constrained optimization view of the overall objective (Fig. 1). The proteins and genes that are detected in the experiments should guide the selection of relevant pathways from the interactome. To avoid forcing a solution that integrates false positives from the experiments and to preferentially include the most reliable interactions, we treat the goal of connecting the data as a constraint that we attempt to satisfy through an optimization procedure. We show that this problem can be modeled as a prize-collecting variant of the Steiner tree problem.

The Steiner tree problem begins with a weighted graph and a set of "terminal" nodes in the graph. The algorithm constrains the solution to link these termini directly or indirectly through the edges of the graph. The prize-collecting variant of the Steiner tree problem relaxes these constraints so that not all the termini are required to be included in the solution. Rather, the algorithm balances two costs: (i) It pays a penalty for leaving a terminal out of the network; (ii) it pays a price for using edges to include a terminal in the network. In addition, we control the size of the solution network by introducing a single parameter β that weights the penalties of excluding terminal nodes relative to the cost of including edges. We define the cost of the edges so that more reliable edges have lower cost than less reliable ones and we define penalties for excluding each terminal node to reflect the relative importance of that terminal in the experimental data. The solution to the prize-collecting Steiner tree (PCST) problem is a minimum-weighted subtree that connects a subset of the termini to each other through the edges of the interactome graph and additional nodes not in the terminal set.

We demonstrate the utility of our approach by relating mRNA expression changes to two classes of upstream regulatory data from *S. cerevisiae*: One is derived from curated genetic interactors (23) and the other is from phosphoproteomics mass-spectrometry (7). We show that our method reports compact networks that connect the experimental data through high-confidence interactions. We present evidence that the proteins in the networks predicted by the algorithm are functionally relevant, provide a clear context to interpret the experimental observations, and uncover diverse pathways not obvious from the input.

RESULTS

Linking genetic and transcriptional data recovers relevant biological processes

The results of genetic screens generally share very little overlap with genes differentially expressed in response to the same perturbations (4). One strategy to address this gap is a flow-based algorithm that links the genetic hits and differentially transcribed genes (4). We evaluated our approach by applying the Steiner tree algorithm to the same problem. We tested solving the PCST on five sets of genetic hits and the associated mRNA profiles: Four were from genetic interactors of a few components in well-characterized signaling pathways, such as MAPK signaling (23,24) and the DNA damage response (25–27), and one was from overexpression screen of alpha-synuclein (α-syn) (4), a protein implicated in Parkinson's disease. In order to derive connections between the genetic hits and differentially expressed genes, we followed

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

the approach of the flow-based algorithm and supplemented the protein-protein interactome (28,29) with protein-DNA interaction data (30,31). In this interactome, each protein and the transcript that encodes it are represented as separate nodes. The nodes representing transcripts are only linked to DNA-binding proteins that have been shown to bind the corresponding promoter (Fig. 1).

In all four input datasets for the known signaling pathways, the nodes discovered by the network are highly enriched in the relevant biological processes (Fig. 2). As expected, putting heavier weights on the node penalties (larger β) forces more terminal nodes to be included and produces larger solution networks. In some cases, this leads to marked decrease in the fraction of nodes in the solution that have the expected annotation [the *STE12* deletion (STE12Δ) and *STE2* deletion (STE2Δ) datasets] and even results in the loss of significant enrichment (*STE2*Δ), demonstrating the benefit gained by the exclusion of terminal nodes by the PCST. We then compared our method to the flow-based approach and to two simpler methods of building networks: (i) assembling the shortest paths between all pairs of nodes in the set of the genetic hits and differentially transcribed genes, and (ii) expanding from the genetic hits to the nodes that directly interact with them (first neighbors). Because the PCST algorithm excludes some genetic hits and differentially expressed genes, we used solutions from the flow-based approach that contain approximately equal number of nonterminal nodes and constructed the networks for the other two methods with those terminal nodes included in the PCST solution. Although all these methods predict hidden nodes that are significantly enriched for the relevant biological process, the PCST solutions contain higher fraction of nodes with the expected annotation and the networks are much smaller than the shortest path and first neighbor networks (Fig. 2). And by these two measures the PCST solutions are comparable to the networks reported by the flow-based algorithm. This suggests that the PCST approach reconstructs compact networks that nevertheless retain the functionally relevant connections.

For the α-syn overexpression dataset, we compared our results to the reported cellular pathways implicated in Parkinson's disease and additional processes uncovered by the flow-based approach. We observed that the solution network partitions into clusters that are biologically coherent. To formally evaluate this observation, we used a previously reported algorithm for partitioning a network into local clusters (32,33) and tested the Gene Ontology (GO) (34) enrichment of each cluster (fig. S1). The most enriched biological process GO terms from the clusters include vesicle trafficking [FDR (False Discovery Rate)-corrected P-value<1E-09] and ubiquitin-dependent protein degradation (FDR-corrected P-value<3E-06). Both of these processes have been associated with Parkinson's disease (4). In addition, the network contains smaller clusters of genes in the heat shock response and the target of rapamycin pathways (fig. S1), two biological processes first identified in the flow-based approach as responsive to α-syn expression and subsequently validated by biological experiments (4). This shows that the PCST approach can uncover new mechanisms when applied to connect genetic data and transcriptional data.

The pheromone response network linking proteomic and transcriptional data reveals diverse biological processes

Having demonstrated that the PCST approach can identify relevant interactions from regulatory proteins that may not directly interact, we tested whether it could be applied to find regulatory networks from phosphoproteomic and transcriptional data. We used published mass-spectrometry (7) and mRNA profiling (3) datasets for yeast responding to the mating pheromone α-factor. As noted above, only a small fraction of the proteins with differentially phosphorylated sites are mapped to the MAPK pathway or any annotated signaling pathways in yeast. Among them only four proteins are annotated to have transcription factor activity. Of the 201 genes differentially expressed by more than threefold at the mRNA level, only six encode proteins in the MAPK pathway and ten encode proteins in the cell cycle pathway in the KEGG database.

We asked whether the PCST approach could provide a functional context for the many proteins that lie outside of the expected pathways. We used the same interactome presented above that contains protein-protein interactions with added transcription factor to target gene relationships, and we defined the terminal nodes to include the proteins with differentially phosphorylated sites in the protein-protein interaction layer and the genes with differentially expressed mRNA transcripts in the transcription factor to target gene layer (Fig. 1). The penalties reflected the magnitudes of the changes in phosphorylation or mRNA expression (see Materials and Methods).

The resulting network (Fig. 3) reveals that the algorithm recovers the expected pathways, as well as many other components of the cellular response that are not immediately apparent from the input. The algorithm connects 56 of the 112 proteins with α-factor-responsive phosphorylation sites and 100 of the 201 differentially expressed genes through 94 intermediate proteins. The solution contains a subnetwork that resembles the known pheromone-induced MAPK pathway (labeled "pheromone core" in Fig. 3). It is noteworthy that those components in this pathway that were not detected by mass-spectrometry, GPA1, STE11, and BEM1, are correctly recovered. We confirmed that the solution is relatively stable for a wide range of β values (fig. S2) and is robust to noise in the interactome (fig. S3).

One of the principal benefits of our approach is that by placing the data in a functional context it reveals broad changes in cellular processes beyond those that might have been expected. For example, the solution features two other yeast MAPK pathways: the protein kinase C (PKC) pathway and the filamentous growth pathway. The PKC pathway is activated during pheromone induction to promote polarized cell growth for mating projection formation (35, 36). In the PCST solution, it is represented by the MAPK SLT2, the transcription factor RLM1 and the SWI4/SWI6 transcription factor complex. SLT2 is activated by PKC (37), and RLM1 and the SWI4/SWI6 complex are activated by SLT2 (38,39). Components of the filamentous growth pathway appear alongside the pheromone core in Fig. 3, which is not surprising as filamentous growth and pheromone MAPK pathways are known to share multiple signaling components (40). SHO1 is an osmosensor in the high-osmolarity glycerol (HOG) pathway (41) that also activates the filamentous growth pathway through the STE11 MAPKKK (42, 43), leading to the phosphorylation and inhibition of transcription factors DIG1 and DIG2 and subsequently the activation of the transcription factors STE12 and TEC1 (44,45). In the PCST solution, the proteins STE11, DIG1, DIG2, and STE12 are common between the pheromone-induced MAPK pathway and the filamentous growth pathway, as expected (40). The decrease in phosphorylation on SHO1 suggests a mechanism by which the specificity of mating signal is achieved in response to pheromone. This may be similar to the situation where the HOG pathway shares a few components with the mating response pathway and inhibiting SHO1 prevents the crosstalk between them (43).

Outside of the core pheromone signaling pathway, a diverse array of mating-related biological functions are apparent, including regulation of cell cycle, transcription control, cellular polarization, and cellular transport. The nonterminal nodes in the PCST solution provide mechanistic insights into these processes. For instance, the algorithm included two proteins that did not contain differentially phosphorylated sites, CDC5 (a protein kinase) and DBF4 (the regulatory subunit for the protein kinase CDC7), in a subnetwork related to DNA replication that contains several phosphorylated proteins. CDC5 is a CDC28 substrate (46) and is recruited to origin of replication by DBF4. DBF4 acts to initiate DNA replication in late G1 phase (47), the point at which pheromone-stimulated cells arrest their cell cycle (48). In the shmoo (mating projection) formation subnetwork, the algorithm highlights the involvement of

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

AFR1, a protein required for forming pheromone-induced projections, in regulating the septin proteins through interaction with the septin protein CDC12, which is not phosphorylated [reviewed in (49)]. The role of the molecular chaperone HSP82 is also made clear by its interactors in the protein folding subnetwork. Differentially phosphorylated heat shock proteins SIS1 and SSB2 are connected to HSP82, which is required for pheromone signaling (50), and to the HSP82 co-chaperone SSE1. Neither HSP82 nor SSE1 has pheromone-responsive phosphorylation sites that were detected in the mass-spectrometry experiment. These observations demonstrate the rich range of biological knowledge represented by the hidden nodes that are not present in the experimental datasets.

Similar to the results obtained from the α-syn datasets, the network can be partitioned into functionally coherent clusters by an automated procedure (Table 1, and fig. S4) (32), and these clusters represent many of the biological functions altered in response to mating factor.

The reconstructed network is enriched in genes implicated in mating defects

To further assess the relevance of the genes in the reconstructed network to pheromone response, we asked whether the network was enriched in genes reported to display mating defects in two whole-genome deletion screens (51,52) (Fig. 4). The sets of genes encoding the protein nodes in the PCST solution network, excluding terminal nodes, are significantly enriched in genes that are involved in the mating-specific transcriptional response (51) or in changes in cellular morphology induced by pheromone (52). The PCST solution is smaller and has a higher fraction of the genes implicated in these mating defects compared to the network constructed by three other approaches: the flow-based approach, the network composed of pairwise shortest paths between the terminal nodes, as well as the network of the set of immediate neighbors of the phosphorylated proteins. Because these two screens are independent of the data sources incorporated in the algorithm, the fact that the PCST solution includes a high percentage of genes necessary to produce a normal mating phenotype is strong evidence that our method identifies signaling nodes that are perturbed in pheromone response.

Targets of transcription factors in the solution show significant expression coherence

The transcription factors in the solution network are included because of constraints from both the upstream phosphorylation events and the downstream target genes that are differentially expressed. Many of the transcription factors in the solution are indeed known to be induced by pheromone, such as DIG1, DIG2, MCM1, and STE12, or have functions in mating related-processes, such as the cell cycle regulators SWI4, SWI6 and MBP1, but the algorithm also includes many others that are not previously known to be involved in pheromone response. To quantitatively assess the relevance of these transcription factors, we computed the expression coherence scores under different conditions (53) for targets of each transcription factor and used these scores as a condition-specific measure of the similarity of the mRNA expression profiles of the targets. After stimulus by α-factor, the previously reported targets (30,31) of the transcription factors included in the PCST solution are more likely to show significant expression coherence than the transcription factors that were excluded. In addition, we show that such coherence, as expected, is specific to pheromone signaling but not to unrelated conditions, such as when yeast cells undergo metabolic shift from fermentation to respiration (diauxic shift) (Fig. 5). Additionally, some transcription factors that function cooperatively are placed in close proximity to the expected upstream signaling pathways. Examples include the DIG1/DIG2/STE12 complex in the core pheromone signaling pathway and the SWI4/SWI6 and SWI6/MBP1 complexes in the PKC pathway (Fig. 3).

Most proteins in the network are not coordinately expressed

It has been proposed that genes in regulatory pathways tend to be coordinately expressed, and this has been evaluated by several techniques, including the expression coherence score (53)

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

and the expression activity score (18). This rationale inspires many of the network inference algorithms to search for local neighborhoods in the interactome that have this property. Because our network was constructed from phosphoproteomic data and represents proteins and transcripts separately, it provides an opportunity to examine these assumptions in an unbiased way. Overall, the proteins identified by our approach do not have significantly correlated expression as measured by the significance of the expression coherence score or the significance of the expression activity score (Table 1). We then examined the individual clusters in our network produced by the clustering algorithm (33). Despite the high degree of functional coherence, these clusters show a large variability in the significance of expression coherence score and the significance of expression activity score (Table 1). For example, although the cell cycle-related cluster (cluster 9) has a significant expression activity score, the score for the cellular transport cluster (cluster 1) is not significant, and therefore this cluster would not have been recovered by expression-based methods. These observations are consistent with the fact that many biological processes are regulated posttranscriptionally and highlight the critical role of proteomic data in revealing the full extent of the proteins involved in biological responses.

DISCUSSION

We describe a computational method based on constrained optimization for discovery of regulatory networks from high-throughput data and apply it to reconstruct pathways linking transcriptional data with proteomic or genetic data. The objective of finding relevant mechanistic connections is formulated as solving a PCST problem on the weighted interactome graph. We reasoned that this approach would be well-suited to overcome noise in the input data and in the interactome. Because the algorithm does not require all terminal nodes to be included in the solution, it should handle false positives in the input data well. False positives in the interactome correspond to reported interactions that do not occur in the cell. These may be eliminated by choosing a cost function that penalizes edges based on the probability that they represent real interactions (4).

Application of the algorithm to yeast genetic, phosphoproteomics, and transcription profiling datasets reveals highly coherent, global views of the many cellular processes involved in creating the response of interest, and identifies transcription factors that connect differentially expressed genes to upstream regulatory events. In the reconstructed networks, the hidden nodes, which are not present in the genetic, mass-spectrometry or transcriptional datasets, give biological context for understanding the functions of the terminal nodes, while providing a systematic view of the biological processes at the global level. Many of the functionally coherent clusters that we identified are not coordinately expressed, and so could not have been recovered by mapping mRNA expression data onto the interactome.

We note that our method is distinct from many existing computational techniques that are typically applied to discover regulatory relationships from high-throughput signaling and expression measurements. Approaches such as probabilistic graphical models (54) and partial least-squares regression (55) can reveal the presence of correlated events across diverse datasets, but it is often difficult to discern why these events are correlated. The biological interaction network provides valuable context for interpretation of these events.

Previous studies using the Steiner tree formulation to analyze biological networks mapped the mRNA abundance onto the protein-protein interaction network and searched for regions that show high degree of differential mRNA expression (16,19). Despite the Steiner tree formulation, this problem is inherently different from our objective of connecting signaling and expression through intermediate nodes in the interactome. In addition to the distinct objectives, the input data are also treated differently in these prior studies. Differential mRNA

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

expression was used as a proxy for subsequent changes at the protein level. Here, we provided evidence that mRNA expression measurements alone are insufficient to capture many of the relevant cellular processes. In contrast, by modeling proteins and transcripts as separate entities, our approach uses the mRNA data as evidence of upstream changes in signaling and reveals biological processes not captured by measurable changes in mRNA abundance. Furthermore, the optimization functions in these two studies (16,19) include only the weights on the nodes but not the reliability of the edges in the interactome graph.

Another computational method that takes a constrained optimization approach constructs functional protein networks from genetic hits by finding optimal paths on an interactome weighted by interaction reliability (56). Our results are consistent with their observation that the Steiner tree approach recovers pathways that are functionally coherent, but our approach differs in two critical ways. First, by using a prize-collecting variant of the Steiner tree problem, we can handle noise in the experimental data and in the interactome and avoid producing unnecessarily large networks that include irrelevant nodes. Second, we demonstrate that our approach effectively integrates expression data with proteomic and genetic data. As a result, we can discover a coherent view of the links between the biological processes from diverse experimental data sources.

This method represents a general framework for building models of regulatory networks from high-throughput measurements of signaling and transcription. It can be applied when there are suitably defined constraints and in different species where the interactome are available. The constraints can be defined in multiple ways to focus on different aspects of the regulatory networks. For example, we can easily extend our approach to use time-courses of proteomic and expression measurements to examine the time-dependent changes in the signaling network. We expect that our framework will be increasingly useful and accurate as the interactome becomes more complete.

MATERIALS AND METHODS

Overview

We consider the goal of finding a network that explains the regulatory data as a constrained optimization problem on an interactome graph, in particular, as solving a PCST problem. Input to the algorithm consists of two components: terminal nodes and a weighted interactome. The terminal nodes are derived from a list of molecules reported in some experiments as potential components in the regulatory network, for instance, hits from genetic screens, proteins with differentially phosphorylated sites, or genes with altered mRNA expression. Each interaction is associated with a weight to indicate the confidence of the interaction. Solving the PCST problem on the weighted interactome is equivalent to trying to find a set of most confident interactions that connect the terminal nodes while possibly leaving some unconnected.

The PCST formulation

We use the Goemans and Williamson Minimization (GW) definition of the PCST problem (57).

Given an undirected graph of nodes V and edges E, a function $p(v)\geq 0$ that assigns a penalty to each node $v\in V$, and a function $c(e)\geq 0$ that assigns a cost to each edge $e\in E$, the PCST problem is to find a subtree T of nodes $V_T\subseteq V$ and edges $E_T\subseteq E$ that minimizes the objective

$$GW(T)=\sum_{v\notin V_T} p(v)+\sum_{e\in E_T} c(e).$$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Note that we incur penalties for excluding nodes while paying costs for including edges. Although this problem is NP-hard (58,59), exact solutions for the datasets presented here can be found by a published algorithm (60).

The probabilistic interactome

The interactome graph of *S. cerevisiae* and probabilistic weights on the edges were constructed as previously described (4). Briefly, experimentally determined protein-protein interactions and the experimental evidence for each interaction were collected from publicly available databases such as BioGRID (29) and MIPS (28). With a naïve Bayes probabilistic model where the probability of each evidence is conditioned on whether two proteins interact, we computed the conditional probability tables from published gold standard set of positive (61) and negative (62) interactions. By applying Bayes rule to the experimental evidence of individual edges in the interactome graph, we obtained the reliability of the interaction represented by the edge. To this protein-protein interaction graph we added protein-mRNA edges that represent transcription factor to target gene relationships. The mRNA node of a gene was represented separately from the protein node of the same gene (Fig. 1). These transcription factor target data were collected from literature and published ChIP-chip assays (30,31), and the edge weights were computed to reflect the reliability of binding events.

Because the optimization objective is to minimize the sum of the edge costs, we took the negative log of the probability weights on the edges as the edge costs. Furthermore, this general interactome graph was slightly modified when the node penalties were defined for specific mRNA expression datasets (see the section on node penalties).

Node penalties

Although the weighted interaction graph was generic, the node penalties were specific for each dataset. We used a formulation such that the optimization would preferentially include nodes that show the largest experimental signals. For example, the experimental signal can be the severity of defect of the genetic hits in genetic screens or the fold change in phosphorylation in the phosphoproteomics data. We will refer to the experimental signal generally as "strength." Let *prot* be the set of proteins with the experimental signal. For all $v \in prot$ we have *strength* $(v)>0$ as a measure of the importance of v in the network. We computed the node penalty as the normalized absolute log of the strength:

$$p(v)=\frac{\left|\log(strength(v))\right|}{\sum_{v'\in prot}\left|\log(strength(v'))\right|}.$$

To connect mRNA profiling datasets to upstream regulatory events, we need to make some modification to the interactome. Let *mrna* be the set of differentially expressed transcripts, and *fc*(v) be the fold change in mRNA abundance of each gene $v \in mrna$. For each $v \in mrna$, we searched the interactome for the set of upstream transcription factors *F*, removed v from the interactome, and added one node v_f for each transcription factor $f \in F$ and one edge between *f* and v_f. The fold change of v was transferred to all the v_f and normalized so the penalties were

$$p(v_f)=\frac{\left|\log(fc(v))\right|}{\sum_{v'\in mrna}\left|\log(fc(v'))\right|},\ \forall f \in F.$$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

All nodes in the interactome not in the *prot* or *mrna* set are given a penalty of zero.

Solving and analyzing the PCST

We introduced a scaling factor β for the node penalties described in the previous section. The PCST minimization objective becomes

$$\sum_{v \notin V_T} \beta\, p(v) + \sum_{e \in E_T} c(e).$$

Intuitively, the objective function represents a trade-off between excluding nodes and including edges. In a given problem with defined penalty and cost values, the larger the value of β, the greater the penalty to exclude a node, making the optimization procedure exclude fewer nodes at the expense of including more edges (with higher total edge cost) and generating a larger network. The parameter β thus controls the size of the solution. We used a published algorithm (60) to find the exact solution to the PCST and experimented with a wide range of β values. For the pheromone response data, we show results for β=4 because it produces a midsize solution network that includes most of the terminal nodes present in solutions of larger β values (fig. S2). The solution networks were visualized in Cytoscape (63). GO enrichment statistics were computing using BiNGO (64).

Yeast genetic and matching mRNA profiling data

Genetic interactors for *STE2*, *STE5*, and *STE12* deletions were downloaded from the *Saccharomyces cerevisiae* genome database (SGD) (23). Differentially expressed genes are defined as genes that show at least a twofold change with P-value ≤ 0.05 (24). For the DNA damage response, 91 genetic hits common to two independent screens (25,26) and the DNA damage signature genes from mRNA profiling (27) were used. α-syn genetic and transcriptional data were from (4).

Yeast pheromone response data

For termini that were proteins, we used the set of phosphorylation sites that change by least two fold after 2 μM α-factor treatment for 120 minutes (7). Termini that were mRNA represented genes that were differentially expressed by greater than three fold in wild-type cells after 50 nM α-factor treatment for 120 minutes (3). Although the treatment concentrations were different between these two datasets, there was evidence that the transcriptional response to α-factor saturates at concentrations above 15.8 nM [fig. S5 and (3)]. The gene sets used in calculating enrichment of mating-related genes were the Group III, pheromone-unresponsive set from (51) and the ASD set from (52). Only the genes tested in each screen were used as background in the calculation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES AND NOTES

1. de Chassey B, Navratil V, Tafforeau L, Hiet MS, Aublin-Gex A, Agaugue S, Meiffren G, Pradezynski F, Faria BF, Chantier T, Le Breton M, Pellet J, Davoust N, Mangeot PE, Chaboud A, Penin F, Jacob Y, Vidalain PO, Vidal M, Andre P, Rabourdin-Combe C, Lotteau V. Hepatitis C virus infection protein network. Mol Syst Biol 2008;4:230. [PubMed: 18985028]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

2. Huang PH, Mukasa A, Bonavia R, Flynn RA, Brewer ZE, Cavenee WK, Furnari FB, White FM. Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. Proc Natl Acad Sci U S A 2007;104:12867–12872. [PubMed: 17646646]

3. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR, Tyers M, Boone C, Friend SH. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science 2000;287:873–880. [PubMed: 10657304]

4. Yeger-Lotem E, Riva L, Su LJ, Gitler AD, Cashikar AG, King OD, Auluck PK, Geddie ML, Valastyan JS, Karger DR, Lindquist S, Fraenkel E. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat Genet 2009;41:316–323. [PubMed: 19234470]

5. Cooper AA, Gitler AD, Cashikar A, Haynes CM, Hill KJ, Bhullar B, Liu K, Xu K, Strathearn KE, Liu F, Cao S, Caldwell KA, Caldwell GA, Marsischky G, Kolodner RD, Labaer J, Rochet JC, Bonini NM, Lindquist S. Alpha-synuclein blocks ER-Golgi traffic and Rab1 rescues neuron loss in Parkinson's models. Science 2006;313:324–328. [PubMed: 16794039]

6. Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proc Natl Acad Sci U S A 2007;104:5860–5865. [PubMed: 17389395]

7. Gruhler A, Olsen JV, Mohammed S, Mortensen P, Faergeman NJ, Mann M, Jensen ON. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. Mol Cell Proteomics 2005;4:310–327. [PubMed: 15665377]

8. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 2006;127:635–648. [PubMed: 17081983]

9. Ballif BA, Villen J, Beausoleil SA, Schwartz D, Gygi SP. Phosphoproteomic analysis of the developing mouse brain. Mol Cell Proteomics 2004;3:1093–1101. [PubMed: 15345747]

10. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 2002;298:799–804. [PubMed: 12399584]

11. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-wide location and function of DNA binding proteins. Science 2000;290:2306–2309. [PubMed: 11125145]

12. Chen RE, Thorner J. Function and regulation in MAPK signaling pathways: lessons learned from the yeast Saccharomyces cerevisiae. Biochim Biophys Acta 2007;1773:1311–1340. [PubMed: 17604854]

13. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. Nucleic Acids Res 2008;36:D480–D484. [PubMed: 18077471]

14. Elion EA, Brill JA, Fink GR. FUS3 represses CLN1 and CLN2 and in concert with KSS1 promotes signal transduction. Proc Natl Acad Sci U S A 1991;88:9392–9396. [PubMed: 1946350]

15. Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD. Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 2004;101:18006–18011. [PubMed: 15608068]

16. Scott MS, Perkins T, Bunnell S, Pepin F, Thomas DY, Hallett M. Identifying regulatory subnetworks for a set of genes. Mol Cell Proteomics 2005;4:683–692. [PubMed: 15722371]

17. Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno V, Kirchhoff T, Gold B, Assmann V, Elshamy WM, Rual JF, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Sole X, Hernandez P, Lazaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, Livingston DM, Gruber SB, Parvin JD, Vidal M. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet 2007;39:1338–1349. [PubMed: 17922014]

18. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 2002;18 Suppl 1:S233–S240. [PubMed: 12169552]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

19. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 2008;24:i223–i231. [PubMed: 18586718]

20. Bromberg KD, Ma'ayan A, Neves SR, Iyengar R. Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. Science 2008;320:903–909. [PubMed: 18487186]

21. Yeang CH, Ideker T, Jaakkola T. Physical network models. J Comput Biol 2004;11:243–262. [PubMed: 15285891]

22. Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T. eQED: an efficient method for interpreting eQTL associations using protein networks. Mol Syst Biol 2008;4:162. [PubMed: 18319721]

23. SGD Project. "Saccharomyces Genome Database"

24. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. Cell 2000;102:109–126. [PubMed: 10929718]

25. Begley TJ, Rosenbach AS, Ideker T, Samson LD. Hot spots for modulating toxicity identified by genomic phenotyping and localization mapping. Mol Cell 2004;16:117–125. [PubMed: 15469827]

26. Chang M, Bellaoui M, Boone C, Brown GW. A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage. Proc Natl Acad Sci U S A 2002;99:16934–16939. [PubMed: 12482937]

27. Gasch AP, Huang M, Metzner S, Botstein D, Elledge SJ, Brown PO. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. Mol Biol Cell 2001;12:2987–3003. [PubMed: 11598186]

28. Mewes HW, Frishman D, Mayer KF, Munsterkotter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stumpflen V. MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res 2006;34:D169–D172. [PubMed: 16381839]

29. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–D539. [PubMed: 16381927]

30. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. Nature 2004;431:99–104. [PubMed: 15343339]

31. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 2006;7:113. [PubMed: 16522208]

32. Dunn R, Dudbridge F, Sanderson CM. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics 2005;6:39. [PubMed: 15740614]

33. Girvan M, Newman ME. Community structure in social and biological networks. Proc Natl Acad Sci U S A 2002;99:7821–7826. [PubMed: 12060727]

34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–29. [PubMed: 10802651]

35. Buehrer BM, Errede B. Coordination of the mating and cell integrity mitogen-activated protein kinase pathways in Saccharomyces cerevisiae. Mol Cell Biol 1997;17:6517–6525. [PubMed: 9343415]

36. Zarzov P, Mazzoni C, Mann C. The SLT2(MPK1) MAP kinase is activated during periods of polarized cell growth in yeast. Embo J 1996;15:83–91. [PubMed: 8598209]

37. Lee KS, Irie K, Gotoh Y, Watanabe Y, Araki H, Nishida E, Matsumoto K, Levin DE. A yeast mitogen-activated protein kinase homolog (Mpk1p) mediates signalling by protein kinase C. Mol Cell Biol 1993;13:3067–3075. [PubMed: 8386319]

38. Madden K, Sheu YJ, Baetz K, Andrews B, Snyder M. SBF cell cycle regulator as a target of the yeast PKC-MAP kinase pathway. Science 1997;275:1781–1784. [PubMed: 9065400]

39. Watanabe Y, Irie K, Matsumoto K. Yeast RLM1 encodes a serum response factor-like protein that may function downstream of the Mpk1 (Slt2) mitogen-activated protein kinase pathway. Mol Cell Biol 1995;15:5740–5749. [PubMed: 7565726]

40. Schwartz MA, Madhani HD. Principles of MAP kinase signaling specificity in Saccharomyces cerevisiae. Annu Rev Genet 2004;38:725–748. [PubMed: 15568991]

41. Posas F, Saito H. Osmotic activation of the HOG MAPK pathway via Ste11p MAPKKK: scaffold role of Pbs2p MAPKK. Science 1997;276:1702–1705. [PubMed: 9180081]

42. Mosch HU, Roberts RL, Fink GR. Ras2 signals via the Cdc42/Ste20/mitogen-activated protein kinase module to induce filamentous growth in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 1996;93:5352–5356. [PubMed: 8643578]

43. O'Rourke SM, Herskowitz I. The Hog1 MAPK prevents cross talk between the HOG and pheromone response MAPK pathways in Saccharomyces cerevisiae. Genes Dev 1998;12:2874–2886. [PubMed: 9744864]

44. Cook JG, Bardwell L, Kron SJ, Thorner J. Two novel targets of the MAP kinase Kss1 are negative regulators of invasive growth in the yeast Saccharomyces cerevisiae. Genes Dev 1996;10:2831–2848. [PubMed: 8918885]

45. Madhani HD, Fink GR. Combinatorial control required for the specificity of yeast MAPK signaling. Science 1997;275:1314–1317. [PubMed: 9036858]

46. Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, Shokat KM, Morgan DO. Targets of the cyclin-dependent kinase Cdk1. Nature 2003;425:859–864. [PubMed: 14574415]

47. Hardy CF, Pautz A. A novel role for Cdc5p in DNA replication. Mol Cell Biol 1996;16:6775–6782. [PubMed: 8943332]

48. Bardwell L. A walk-through of the yeast mating pheromone response pathway. Peptides 2005;26:339–350. [PubMed: 15690603]

49. Douglas LM, Alvarez FJ, McCreary C, Konopka JB. Septin function in yeast model systems and pathogenic fungi. Eukaryot Cell 2005;4:1503–1512. [PubMed: 16151244]

50. Louvion JF, Abbas-Terki T, Picard D. Hsp90 is required for pheromone signaling in yeast. Mol Biol Cell 1998;9:3071–3083. [PubMed: 9802897]

51. Chasse SA, Flanary P, Parnell SC, Hao N, Cha JY, Siderovski DP, Dohlman HG. Genome-scale analysis reveals Sst2 as the principal regulator of mating pheromone signaling in the yeast Saccharomyces cerevisiae. Eukaryot Cell 2006;5:330–346. [PubMed: 16467474]

52. Narayanaswamy R, Niu W, Scouras AD, Hart GT, Davies J, Ellington AD, Iyer VR, Marcotte EM. Systematic profiling of cellular phenotypes with spotted cell microarrays reveals mating-pheromone response genes. Genome Biol 2006;7:R6. [PubMed: 16507139]

53. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 2001;29:153–159. [PubMed: 11547334]

54. Friedman N. Inferring cellular networks using probabilistic graphical models. Science 2004;303:799–805. [PubMed: 14764868]

55. Janes KA, Albeck JG, Gaudet S, Sorger PK, Lauffenburger DA, Yaffe MB. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. Science 2005;310:1646–1653. [PubMed: 16339439]

56. Yosef N, Ungar L, Zalckvar E, Kimchi A, Kupiec M, Ruppin E, Sharan R. Toward accurate reconstruction of functional protein networks. Mol Syst Biol 2009;5:248. [PubMed: 19293828]

57. Goemans, MX.; Williamson, DP. Approximation algorithms for NP-hard problems. Hochbaum, DS., editor. Boston, MA: PWS Publishing Co; 1997. p. 144-191.

58. Garey, MR.; Johnson, DS. Computers and Intractability: A Guide to the Theory of NP-completeness. San Francisco: Freeman; 1979.

59. Karp, RM. Reductibility among combinatorial problems. Univ. of California; 1972.

60. Ljubic I, Weiskircher R, Pferschy U, Klau GW, Mutzel P, Fischetti M. An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem. Mathematical Programming 2006;105:427–449.

61. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

S, Svrzikapa N, Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, Barabasi AL, Tavernier J, Hill DE, Vidal M. High-quality binary protein interaction map of the yeast interactome network. Science 2008;322:104–110. [PubMed: 18719252]

62. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003;302:449–453. [PubMed: 14564010]

63. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–2504. [PubMed: 14597658]

64. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005;21:3448–3449. [PubMed: 15972284]

65. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 1995:289–300.

66. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–686. [PubMed: 9381177]

67. We thank L. Riva, E. Yeger-Lotem, P. Huang and F. White for data and helpful discussions. This work is supported by National Cancer Institute (NCI) Grant U54-CA112967. S.S.H. is supported by an NCI integrative cancer biology program graduate fellowship and a National Science and Engineering Research Council of Canada Postgraduate Scholarship. E.F. is the recipient of the Eugene Bell Career Development Chair. S.S.H. and E.F. have applied for a patent for using the PCST algorithm to construct cellular regulatory networks

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

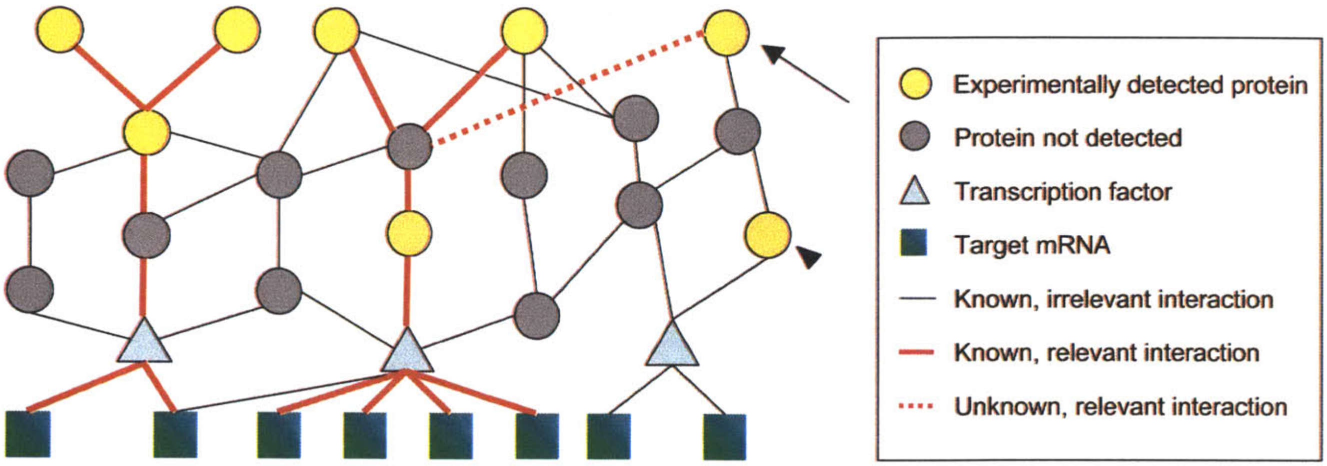

Experimentally detected protein
Protein not detected
Transcription factor
Target mRNA
Known, irrelevant interaction
Known, relevant interaction
Unknown, relevant interaction

Fig. 1.

Finding relevant interactions as a constraint optimization problem. We seek a set of high-confidence edges present in the interactome that directly or indirectly link the proteins and genes identified in the experimental assays. Because some of the input data may be false positives (arrowhead) or may not be explained by currently known interactome (arrow), our approach does not require that all the input data be connected, but rather uses these data as constraints. Note that the protein product and mRNA transcript of the same gene are represented as separate nodes.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

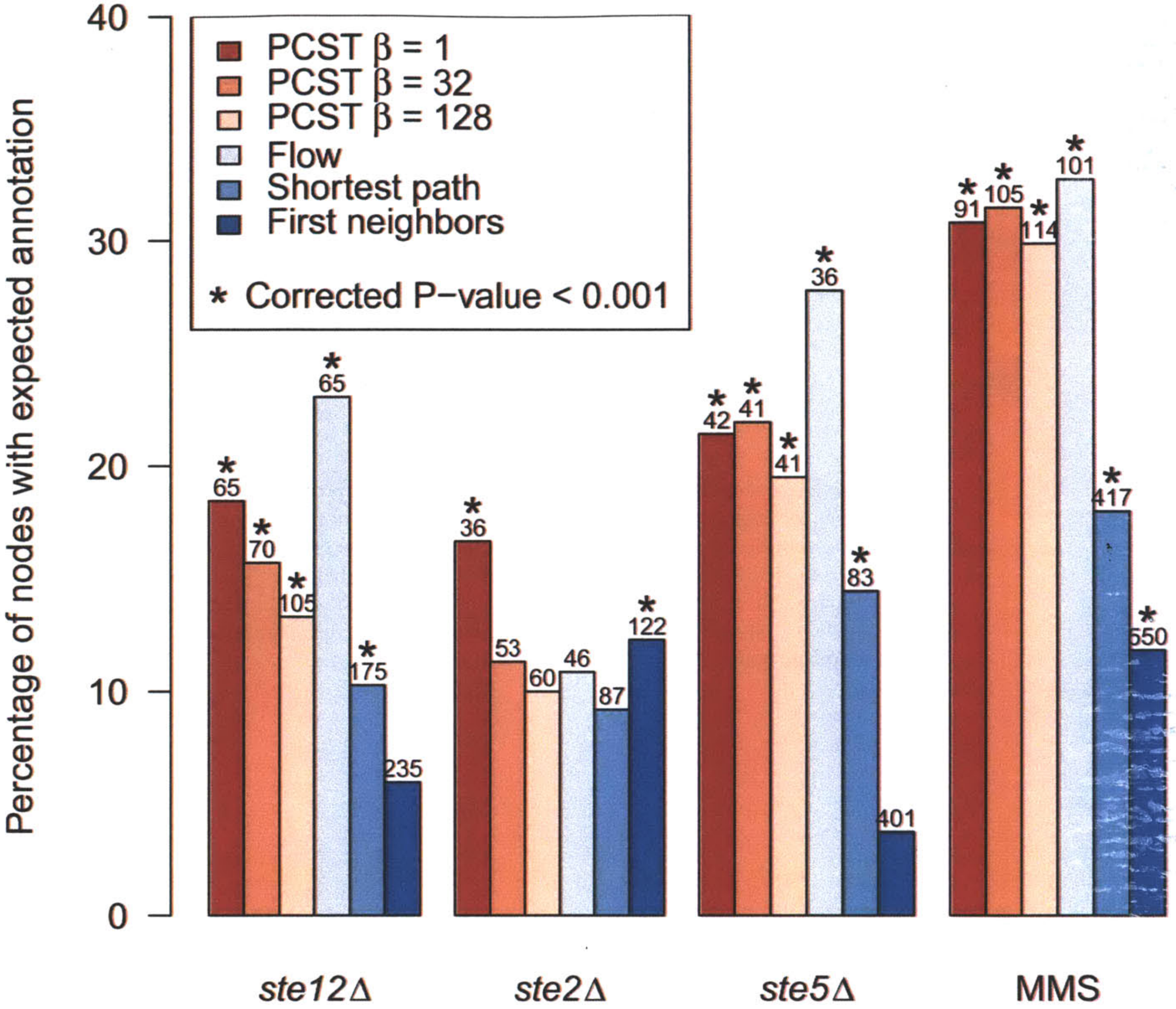


Fig. 2.

The PCST solution recovers compact networks. The fraction of nodes associated with the expected biological process is comparable to the networks from flow-based approach that include approximately equal number of nonterminal nodes, but this fraction is higher than the first neighbor and shortest path networks connecting the same set of terminal nodes. Perturbations for the genetic hits are *STE12*Δ (*STE12* deletion), *STE2*Δ (*STE2* deletion), *STE5*Δ (*STE5* deletion), and, MMS (methyl methanesulfonate treatment). The number above each bar denotes the number of nonterminal nodes in the respective network. The GO annotations tested are response to pheromone (GO:0019236) for *STE12*Δ, *STE2*Δ, and *STE5*Δ, and response to DNA damage stimulus (GO:0006974) for MMS. The evidence code IGI (Inferred from Genetic Interaction) was excluded from the calculation. Statistical significance of the GO term enrichment was computed by hypergeometric test followed by FDR correction (65).

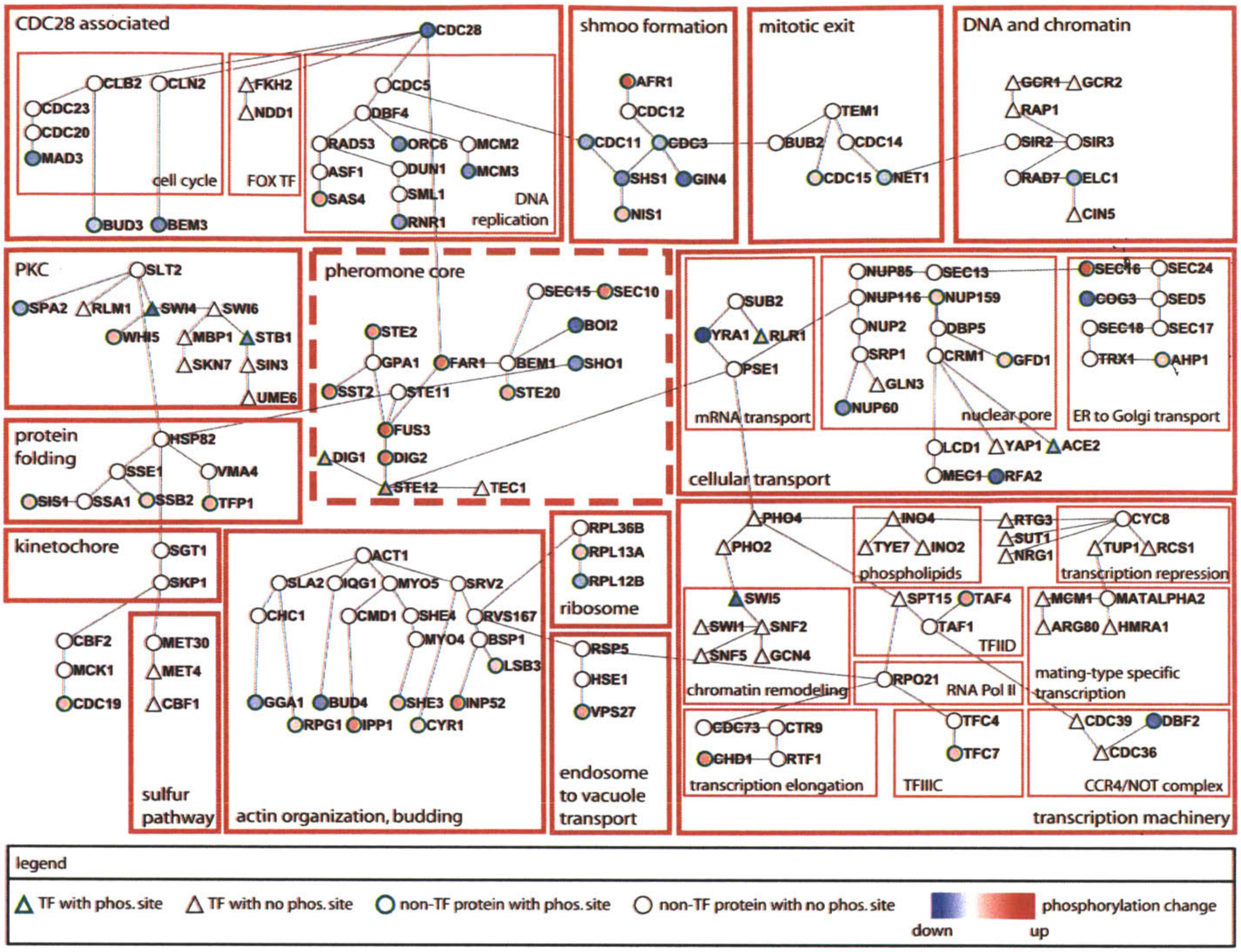


Fig. 3.

The protein components of the pheromone response network constructed by the PCST approach. Note that the canonical pheromone response pathway (enclosed by dashed lines) is but a small component of the broad cellular changes revealed by applying the algorithm to the mass spectrometry and expression data. For clarity the differentially transcribed genes included in the network are not presented. Functional groups based on GO annotation are outlined with red boxes. PKC, protein kinase C; TF with phos. site, transcription factor with at least one differentially phosphorylated sites; TF with no phos. site, transcription factor with no differentially phosphorylated sites; non-TF protein with phos. site, a protein that is not a transcription factor and with at least one differentially phosphorylated sites; non-TF with no phos. site, a protein that is not a transcription factor and with no differentially phosphorylated sites.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

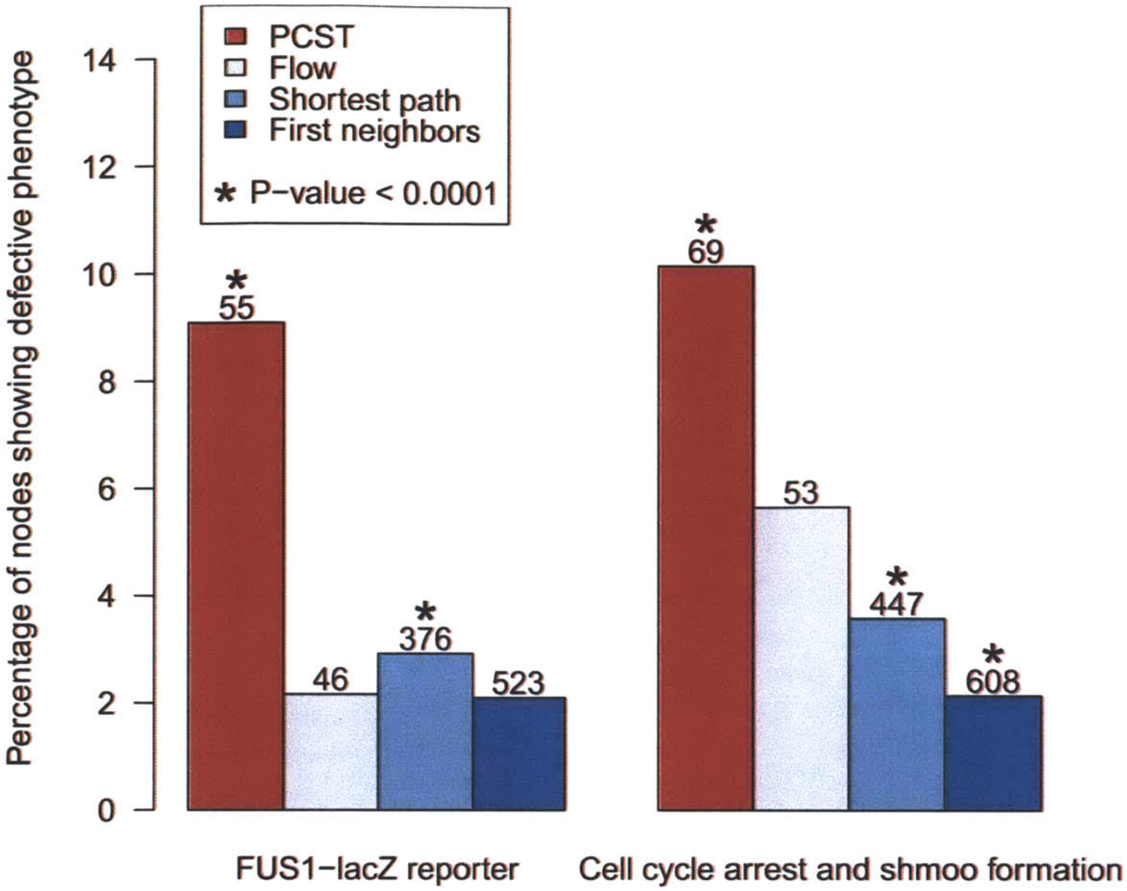


Fig. 4.

The PCST pheromone response network is compact, and, when compared to networks predicted by other methods, it contains higher fraction of genes that are implicated in mating responses, measured by defects in activating a FUS1-lacZ reporter gene (51) and defects in cell cycle arrest and shmoo formation (52). Enrichment P-values were computed by hypergeometric tests using all the genes tested in the respective genetic screen as background. The number above each bar denotes the number of nodes in the network.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

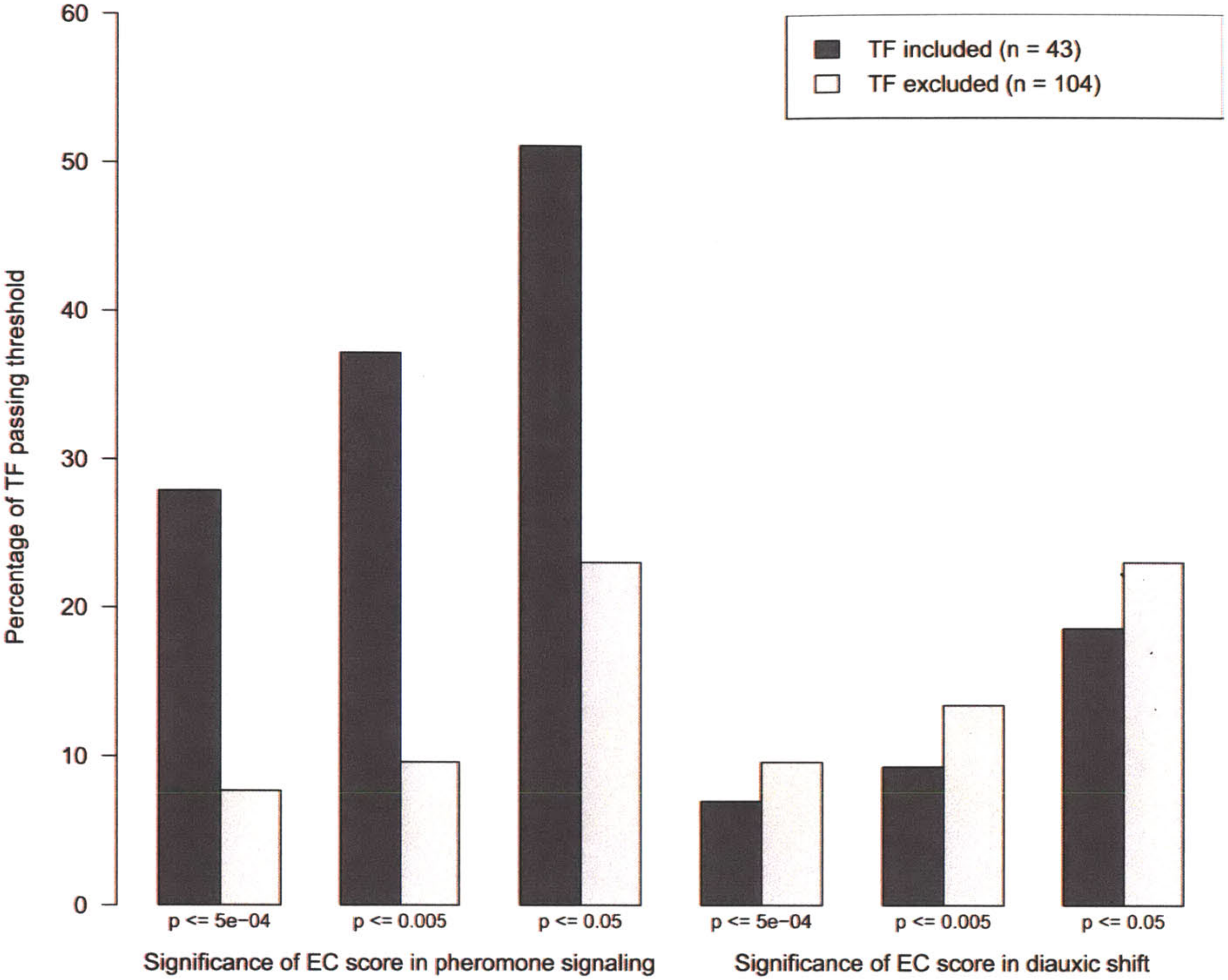


Fig. 5.

Percentage of transcription factors (TF) with targets that show significant expression coherence (EC) scores computed from 50 nM α-factor time course (3) and diauxic shift conditions (66), for transcription factors included in and excluded from the PCST solution network. The P-values indicate thresholds on the significance of the expression coherence score of the target genes.

Table 1

Biological functions and measures of coordinated mRNA expression of the clusters in the pheromone network (fig. S4). EC, expression coherence (53). EA, expression activity (18).

Cluster	Top three enriched GO biological process terms		Corrected P-value	P-value of EC score	P-value of EA score
1	GO:0046907	intracellular transport	1.23E-09	0.711	1
	GO:0051649	establishment of cellular localization	1.23E-09		
	GO:0051641	cellular localization	1.71E-09		
2	GO:0006457	protein folding	1.41E-04	0.251	0.735
	GO:0042026	protein refolding	1.41E-04		
	GO:0000069	kinetochore assembly	8.35E-04		
3	GO:0016193	endocytosis	1.73E-06	0.128	1
	GO:0007114	cell budding	1.26E-05		
	GO:0051301	cell division	1.26E-05		
4	GO:0000074	regulation of progression through cell cycle	2.68E-06	0.421	0.453
	GO:0051726	regulation of cell cycle	2.68E-06		
	GO:0006270	DNA replication initiation	3.44E-06		
5	GO:0006350	transcription	8.00E-14	0.863	1
	GO:0045449	regulation of transcription	1.94E-12		
	GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	7.15E-12		
6	GO:0007096	regulation of exit from mitosis	3.52E-07	0.063	1
	GO:0007088	regulation of mitosis	4.45E-07		
	GO:0000074	regulation of progression through cell cycle	1.05E-05		
7	GO:0048856	anatomical structure development	3.19E-14	0.35	0
	GO:0007148	cell morphogenesis	3.19E-14		
	GO:0019236	response to pheromone	1.26E-11		
8	GO:0006350	transcription	1.89E-09	0.504	0.35
	GO:0006351	transcription, DNA-dependent	7.90E-09		
	GO:0032774	RNA biosynthesis	7.90E-09		
9	GO:0000082	G1/S transition of mitotic cell cycle	2.15E-04	0.272	0.008
	GO:0051325	interphase	1.07E-03		
	GO:0051329	interphase of mitotic cell cycle	1.07E-03		
Full network	GO:0006350	transcription	2.67E-23	0.729	1

NIH PA Author Manuscript NIH PA Author Manuscript NIH PA Author Manuscript

Cluster	Top three enriched GO biological process terms	Corrected P-value	P-value of EC score	P-value of EA score
	GO:0019222 regulation of metabolism	2.73E-21		
	GO:0050791 regulation of physiological process	1.16E-20		

2.3 Supplemental material for manuscript Huang and Fraenkel, *Sci Signal* 2: ra40 (2009)

Supplementary Material

Hidden components of signaling and regulatory networks are revealed by integrating proteomic, transcriptional, and interactome data

Shao-shan Carol Huang[1], Ernest Frankel[2,3]

[1]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA 02139, USA.
Correspondence should be addressed to E.F. (fraenkel-admin@mit.edu).

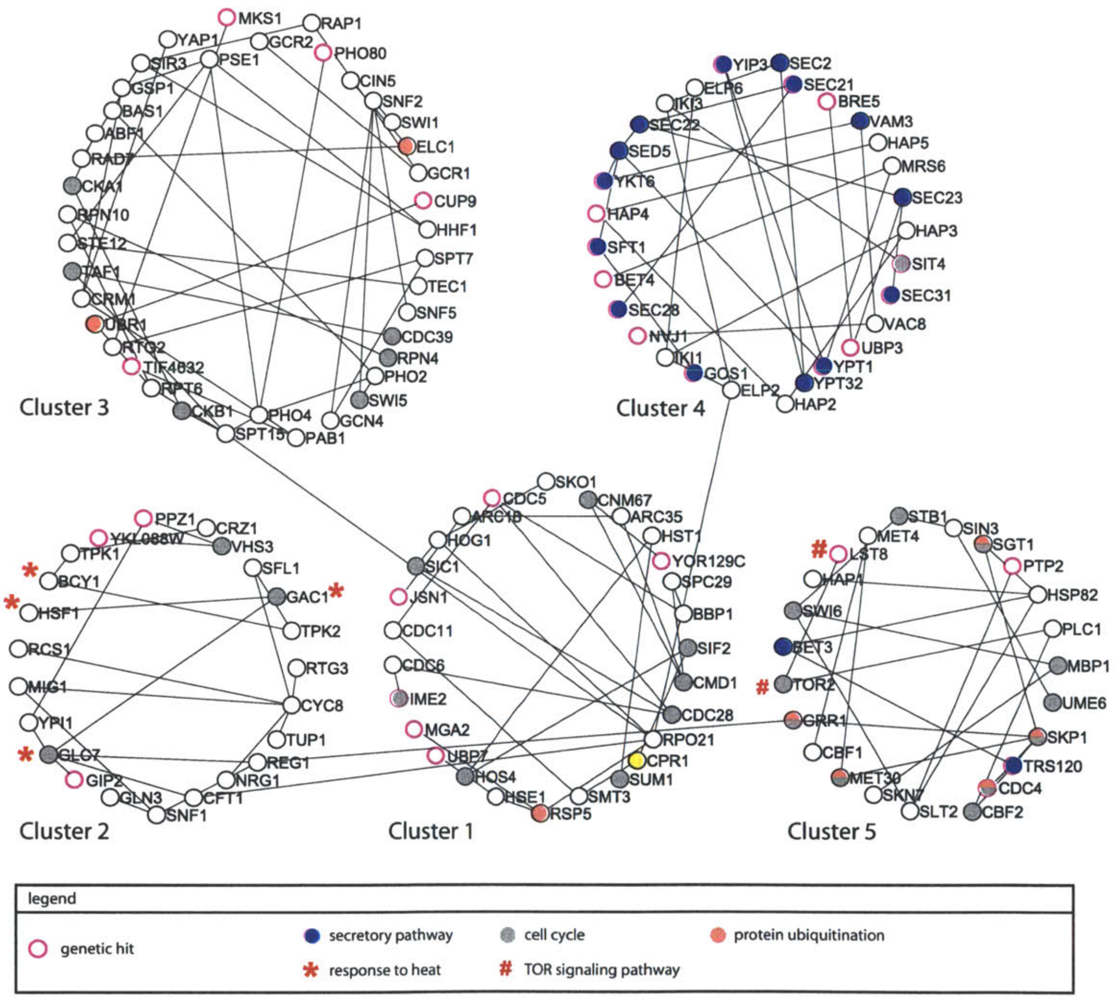


Fig. S1. The clusters in the protein-protein interaction part of the α-syn PCST solution. Nodes are colored or marked by GO biological process. TOR: target of rapamycin.

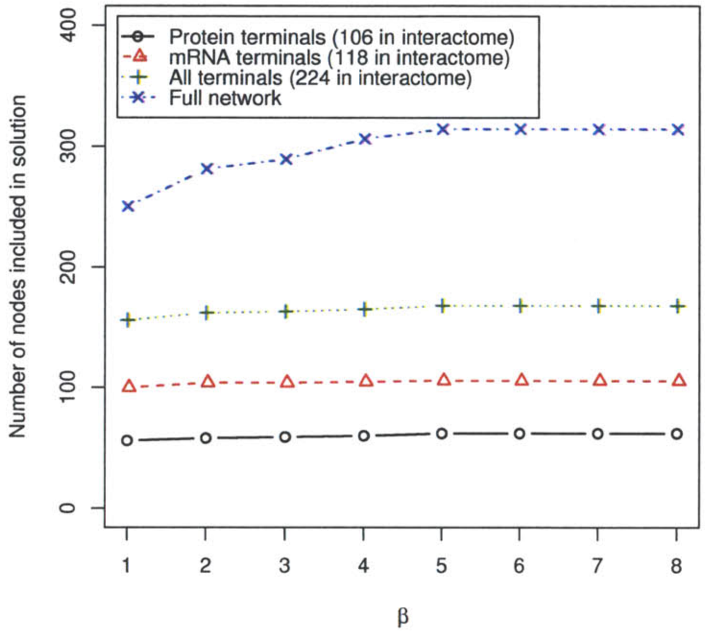


Fig. S2. Statistics of the yeast pheromone PCST solution network for different values of β. The PCST solution constructed from the pheromone response datasets is relatively stable with respect to the parameter β, as measured by the number of terminal nodes included in the solution that represent proteins with differentially phosphorylated sites (protein terminals) and genes that are differentially transcribed (mRNA terminals). The number of terminals indicated in the figure legend counts only the ones present in the interactome.

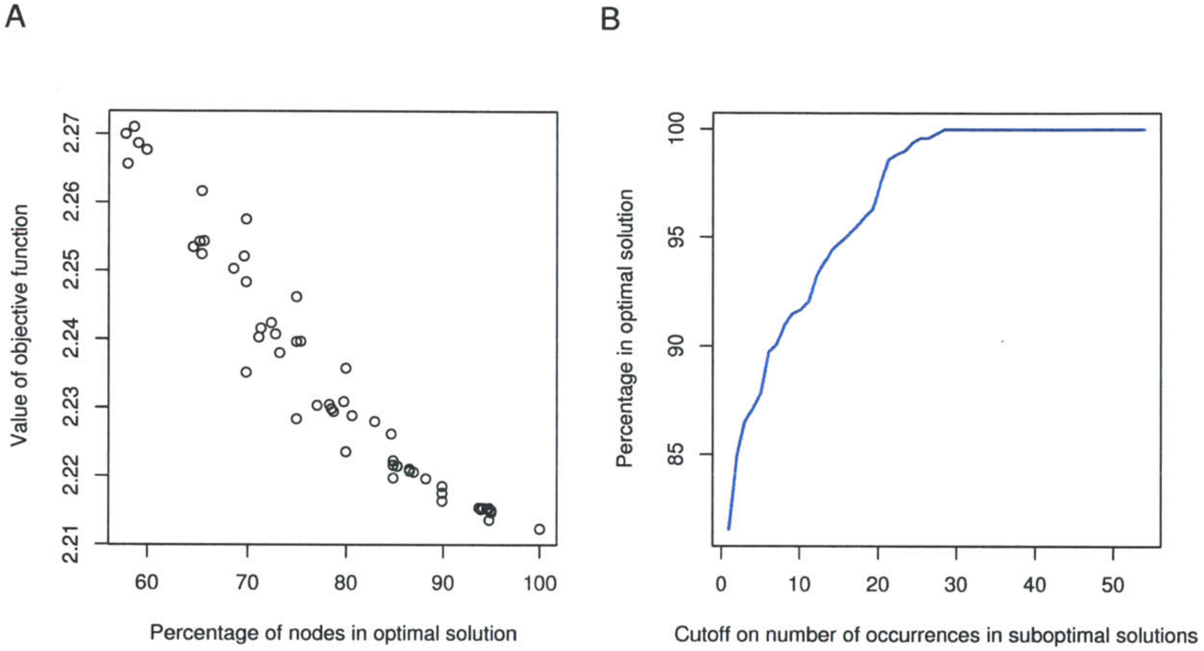


Fig. S3. Alternative or suboptimal solutions to the yeast pheromone response dataset. Because we use an optimization approach to analyze inherently noisy data, we asked whether the network we obtained was stable - are there very different networks that explain the data almost as well? For this, we compared the optimal solution network to a set of alternative solution networks obtained by finding networks that are different from the optimal one by at least a specific percentage of nodes. **(A)** No alternative solutions in the neighborhood of the optimal solution achieves the same objective function value. **(B)** Of the nodes that appear at least once in the 54 suboptimal solutions, at least 80% also appear in the optimal solution.

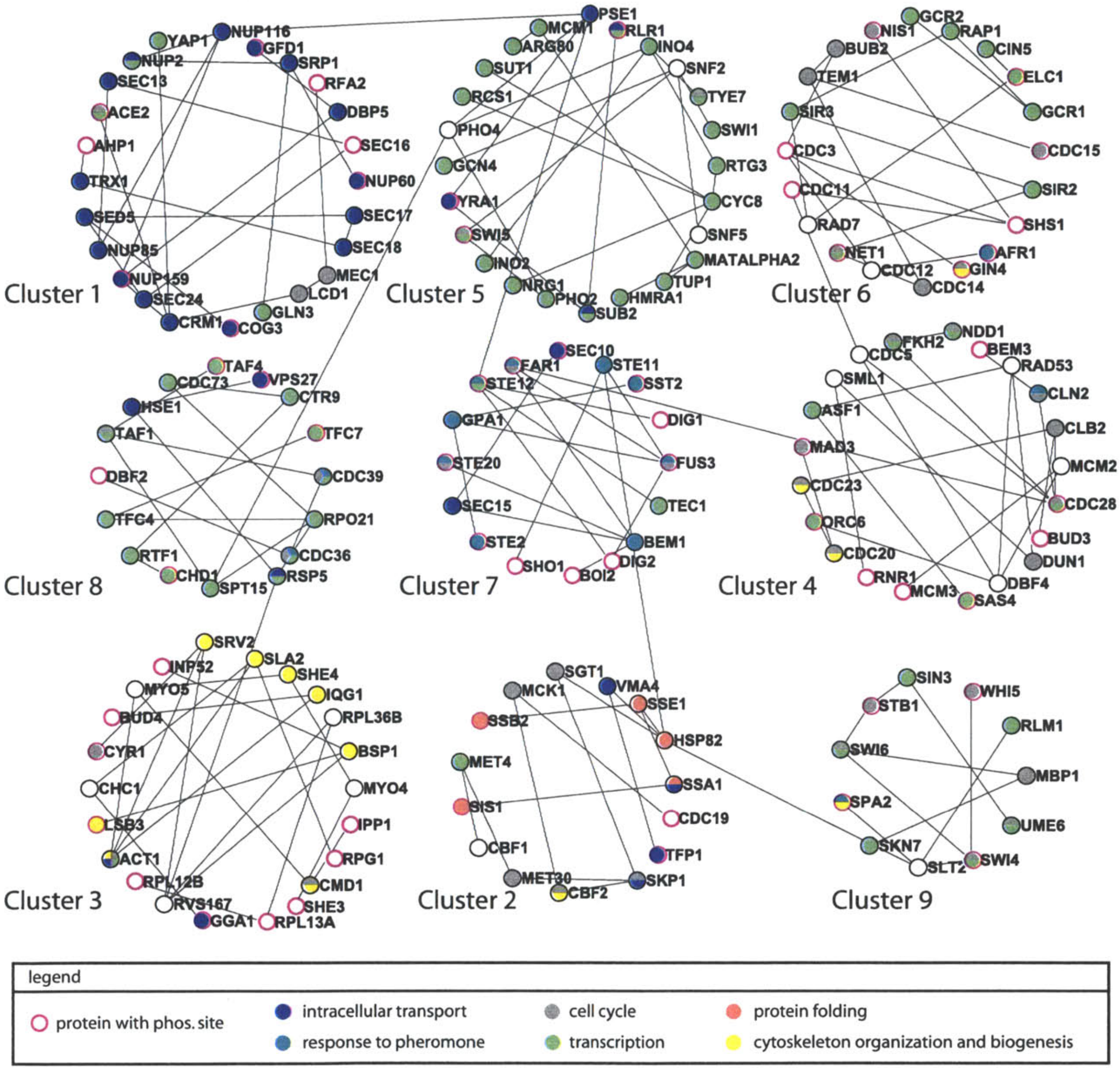


Fig. S4. The clusters in the protein-protein interaction part of the yeast pheromone response PCST solution. Nodes are colored by GO biological process. Cluster labels correspond to those in Table 1.

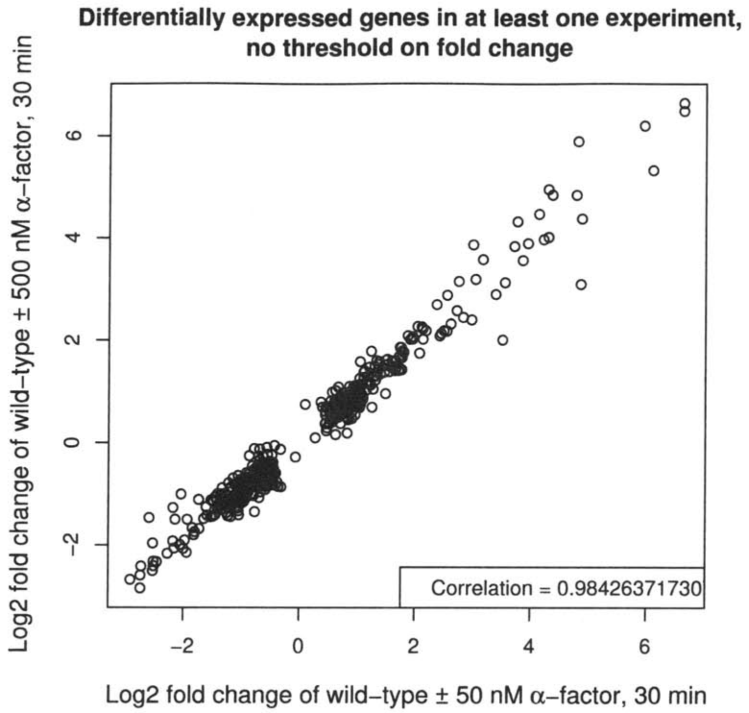


Fig. S5. Scatter plot of gene expression changes following 50 nM and 500nM α-factor treatment. Wild-type yeast cells were treated with 50 nM and 500 nM α-factor for 30 minutes (*1*). Fold changes were calculated with respect to wild-type, untreated cells.

References

1. C. J. Roberts, *et al.*, *Science* **287**, 873 (2000).

Chapter 3

Integrating proteomic, transcriptional, and interactome data reveals key components of signaling network and transcriptional response in a cell line model of human glioblastoma

3.1 Summary

Applying our PCST method to the yeast pheromone response datasets produced some promising results, but taking it to the mammalian system presented a significantly more challenging problem. A fundamental component of the algorithm is the interactome network. Despite large scale efforts to measure and curate protein-protein interactions in mammalian cells, coverage of the network connections at the global level are still very sparse. The connections between transcription factors and target genes are intimately linked to the context which alters the activity of these factors and which consequently determines the downstream effect of transcriptional activity. With current technology, it is difficult, if not impossible, to experimentally identify the targets of all mammalian transcriptional regulators genome-wide for all tissue types and conditions. These considerations motivated our choice of the experimental model and the new experimental and computational techniques that we additionally incorporated for modeling the phosphoproteomic and transcriptional data from mammalian systems.

The U87MG EGFRvIII over-expressing cells (Huang et al., 1997; Nishikawa et al., 1994) makes a nice model for studying the effects of oncogenic mutations in glioblastoma. There are opportunities for novel development in both cancer biology and computational modeling. The wild-type EGFR network is very well studied but the events downstream of the EGFRvIII mutant, the relationships among them and the functional significance are largely unexplored. We focus on the functional significance of those events that lead to mRNA transcript changes. Since transcriptional regulation is context dependent, evidence of changes at the protein level or phosphorylation level of the transcriptional regulators is often insufficient to predict the nature of downstream differential expression. For example, activation of one transcription factor by wild-type EGFR compared to EGFRvIII may change the expression of different sets of targets. With the vast amount of knowledge about the wild-type EGFR signaling, we can be confident that the connections among the important signaling events are well established and therefore may alleviate some issues with the sparsity of the total interactome data. Furthermore, it is possible to make qualitative and quantitative comparisons with the EGFR network, which will give us better understanding of oncogenic signaling and transcription in cancer. Therefore, this is a good starting point to develop the method before tackling other disease models.

In order to meet the challenges in applying the PCST method to mammalian datasets, we made some major modifications to both the experimental data collection and computational analysis for generating the components of the PCST method. Unlike the yeast dataset, the phosphoproteomics MS data was for tyrosine phosphorylation only, and there were no global measurements of transcription factor targets for all the transcriptional regulators. As a consequence, the transcription factor to target layer was poorly constrained. To address this issue, we adopted DNase-seq, a recently developed method that could profile condition-specific open chromatin regions genome-wide, and combined these data with sequence motif libraries and mRNA expression data in a regression-based technique to identify potential transcriptional regulators. This approach, which constrains the transcriptional regulators from the upstream signaling layer and downstream differential expression layer, is a promising direction for discovery of relevant transcriptional regulators.

In the last part of this chapter I show how to generate testable hypotheses from the PCST results. At the subnetwork level, since the method puts the phosphorylation events in the context of protein-protein interactions, the connections participated by these events or groups of events are suggestive of their cellular functions. The transcription factors included in the network and the connections among them point to the functional consequence of the upstream signals. Also interesting are the global network connectivity properties that reveal key signaling nodes in the network that are not immediately apparent from the experimental data. I will present experimental results of testing some of the hypotheses generated from the PCST solution of EGFRvIII datasets.

This work was done in collaboration with Paul Huang in the laboratory of Forest White, who generously provided the U87MG EGFRvIII cell lines and collected the EGFRvIII phosphoproteomics data, and William Gordon in the laboratory of Ernest Fraenkel, who collected the mRNA expression microarray data. The DNase-seq data was collected with the help of Tatjana Degenhardt in the Fraenkel lab.

3.2 Materials and methods: experimental

3.2.1 Cell culture

The human glioblastoma cell line U87MG expressing high level of EGFRvIII (U87H; 2×10^6 receptors per cell) and a kinase dead version of EGFRvIII (U87DK; 2×10^6 receptors per cell) were cultured in complete media (DMEM (Mediatech) supplemented with 10% fetal bovine serum, 100 units/mL penicillin, 100 mg/mL streptomycin (Invitrogen), 4 mM L-glutamine) and in a 95% air/5% CO_2 humidified atmosphere at 37 °C. Expression of EGFRvIII and DK receptors were selected by 400 mg/mL G418

(Calbiochem). To enhance cell attachment tissue culture vessels with the Corning CellBIND surface (Corning) were used.

3.2.2 Mass spectrometry phosphoproteomics

Quantitative phosphotyrosine proteomic data on the U87MG cells expressing titrated levels of the EGFRvIII receptor (U87M, U87H, and U87SH) and U87DK cells were collected by Paul Huang in the laboratory of Forest White and published previously (Huang et al., 2007). Briefly, the four cell lines were serum starved for 24 hours, lysed in 8 M urea and digested in trypsin. Each sample was then labeled with iTRAQ reagent with a different mass tag. The samples were mixed and tyrosine-phosphorylated peptides were immunoprecipitated with an antibody specific to phosphorylated tyrosine. The phosphorylated peptides were further enriched in an IMAC column and analyzed by LC-MS/MS. The peptide sequences were identified from the MS/MS spectra and the relative levels of the phosphorylated peptides were computed from the area under the peak of the iTRAQ marker ions and normalized with respect to the U87DK cell line.

3.2.3 Transcription profiling

Total RNA was prepared from the four U87MG derived cell lines by the RNeasy Plus Mini Kit (Qiagen) and quantified on the Affymetrix Human Genome U133 Plus 2.0 arrays. Labeling, hybridization, washing and staining were performed following the standard Affymetrix GeneChip protocol. The arrays were hybridized in an Affymetrix GeneChip Hybridization Oven 640 at 45 °C 45C at 60 rpm for 16 hours, washed and stained in Affymetrix Fluidics Station 450, and scanned with Affymetrix GeneChip Scanner 3000 7G. Two biological replicates were done for each cell line. The intensity values were normalized using the GC Robust Multi-array Average (`gcrma`) package (Wu et al., 2004) in the R BioConductor library (Smyth, 2005) and differential gene expression was calculated by the Linear Models for Microarray Data method (Smyth, 2004) implemented as the `limma` package in BioConductor.

3.2.4 DNaseI hypersensitivity sequencing (DNase-seq)

The U87DK and U87H cells were seeded in parental media (complete media without G418). After 24 hours, the cells were washed gently with PBS and cultured in serum free media for 24 hours. To collect nuclei, the cells were washed with cold PBS, scraped from the flasks and pelleted in 50 mL tubes at 500 g 4 °C for 8 min. The cell pellet was washed in 25 mL cold PBS and pelleted again at 500 g 4 °C for 8 min. The cell pellet was re-suspended in 20 mL cold buffer A (15 mM Tris HCl (pH 8), 15 mM NaCl, 60 mM KCl, 1 mM EDTA (pH 8), 0.5 mM EGTA (pH 8), into which 0.5 mM spermidine was added immediately before use), and cytoplasmic membrane was lysed by adding 20 mL 2× NP-40 buffer (buffer A supplemented with 0.2% IGEPAL) and incubated on ice for 8 min. The nuclei were pelleted at 500 g for 8 min at 4 °C, washed in 20 mL cold buffer A and pelleted again at 500 g for 8 min at 4 °C. After removal of the supernatant, the nuclei pellet was gently re-suspended in nuclei storage buffer (20 mM Tris HCl, 75 mM NaCl, 0.5 mM EDTA, 50% glycerol by volume, 0.85 mM DTT, into which 0.125 mM PMSF was added immediately before use). 5×10^7 nuclei were collected for each of the two biological replicates of U87DK and U87H cells. The pellet was flash frozen in liquid nitrogen and stored in −80 °C or proceeded to DNaseI digestion.

Prior to digestion, 10× digestion buffer (60 mM $CaCl_2$, 750 mM NaCl) was diluted 1 : 10 in buffer A to make 1× digestion buffer, and stop buffer was made by adding RNase A (final concentration 10 µg/mL), spermidine (final concentration 1 mM) and spermine (final concentration 0.3 mM) to stock stop buffer (50 mM Tris HCl (pH 8), 100 mM NaCl and 0.1% SDS). The digestion buffer and stop buffer were warmed up for 30 min in a 37 °C water bath. Frozen nuclei were thawed on ice and washed twice with cold buffer A. Each digestion reaction occurred in a 2 mL centrifuge tube for 10^7 nuclei so usually one biological replicate of 5×10^7 nuclei required five to six digestion tubes. 100 µL of DNaseI (100 units; Promega) was added to one tube and put on a 37 °C hot plate for 2 min. The nuclei were gently re-suspended in digestion buffer to a concentration of 10^7 per 850 µL, and aliquots of 850 µL were added to digestion tubes containing the DNaseI enzyme. After 2 min of digestion at 37 °C, the reaction was stopped by adding 950 µL pre-warmed stop buffer and inverting the tubes multiple times. The digested samples were then incubated at 55 °C for 15 min, after which 4 µL Proteinase K was added to each digestion tube and incubated overnight at 55 °C. Aliquots of the same biological replicate were combined and DNA was isolated by phenol-chloroform extraction and NaCl was added to the aqueous phase to make a final concentration of 0.8 M. The purified DNA was size separated in a step sucrose gradient (sucrose solutions of 40%, 35%, 30%, 25%, 20%, 17.5%, 15%, 12.5% and 10% in layers of 3 mL in a SW28 ultracentrifuge tube with about 6 mL sample as the top layer) centrifuged in a SW28 rotor for 24 hours at 25000 rpm at 25 °C. Fractions of 1.8 mL were taken from the top of the solution. 30 µL of each fraction was run on a 1.2% agarose gel with SYBR green I nucleic acid gel stain (Invitrogen) at 100 V for two hours and scanned in a Typhoon Imager. Fractions with faint fragments primarily in the 500 to 1000 bp range were purified using Qiagen MinElute columns and sequenced. To prepare control digestion samples, genomic DNA was extracted from one 10 cm tissue culture plate (usually containing 5×10^6 cells) with Promega Genomic DNA extraction kit (Promega), digested with 0.3 units of DNaseI enzyme in one 2 mL digestion tube, size separated and purified in the same way as the nuclei samples.

Sequencing library was prepared by the Illumina sample preparation kit, specifically selecting the 100 to 300 bp fragments by gel electrophoresis. Each biological replicate was sequenced in one lane on a Genome Analyzer II sequencer (Illumina). The sequencing reads of 36 bp were aligned to the hg18 genome by Illumina's Eland extended software with maximum two mismatches in the first 25 bp. The sequencing and alignment statistics are listed in Table 3.1.

3.2.5 Chromatin immunoprecipitation sequencing (ChIP-seq)

The U87DK and U87H cells were seeded in media without G418. After 24 hours, the cells were washed gently with PBS and cultured in serum free media for 24 hours. Crosslinking was initiated by adding formaldehyde directly to the culture media to

Sample name	Total reads	Aligned unique		Aligned repeat		Aligned none	
	million	million	percent	million	percent	million	percent
U87DK control	29.9	16.6	55.4	5.8	19.2	7.6	25.4
U87DK replicate 1	12.7	10.5	83.0	2.0	15.4	0.2	1.6
U87DK replicate 2	14.8	12.4	83.7	2.2	14.9	0.2	1.4
U87H control	29.8	14.6	48.9	5.2	17.5	10.0	33.6
U87H replicate 1	15.1	12.4	81.8	2.5	16.8	0.2	1.4
U87H replicate 2	30.0	11.1	37.0	9.0	29.9	9.9	33.0

Table 3.1: Illumina sequencing statistics of the DNase-seq samples. Aligned unique: reads that are uniquely aligned to the genome. Aligned repeat: reads that are aligned to more than one location in the genome. Aligned none: reads that are not aligned to any location in the genome.

a final concentration of 1%, rocked at room temperature for 10 min and stopped by adding glycine to a final concentration of 0.125 M and incubating for 5 min. The crosslinked cells were washed twice with cold PBS, scraped into a 50 mL centrifuge tube, pelleted at 1500 rpm for 5 min at 4 °C. Cell pellets were transferred to 15 mL centrifuge tubes, flash frozen in liquid nitrogen and stored in −80 °C or proceeded to sonication. 5×10^7 cells were used for each of the two biological replicates of U87DK and U87H cells.

For sonication, each tube of 5×10^7 crosslinked cells was thawed on ice. Then the cell pellet was re-suspended in 10 mL of lysis buffer 1 (50 mM Hepes-KOH (pH 7.5), 140 mM NaCl, 1 mM EDTA, 10% glyerol, 0.5% NP-40 and 0.25% Triton X-100) and rocked at 4 °C for 10 min. The cells were pelleted in a centrifuge at 4 °C at 2500 rpm for 3 min. The pellet was re-suspended in 10 mL of lysis buffer 2 (200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA and 10 mM Tris (pH 7.5)) and rocked at 4 °C for 5 min. The nuclei were pelleted at 4 °C at 1500 rcf for 3 min. The pellet was re-suspended in 1.5 mL lysis buffer 3 (1 mM EDTA, 0.5 mM EGTA, 10 mM Tris-HCl (pH 7.5), 100 mM NaCl, 0.1% Na-Deoxycholate and 0.5% N-lauroyl sarcosine), transferred to two 15 mL polystyrene centrifuge tubes with 0.75 mL each, and sonicated in a Bioruptor NextGen sonication system (Diagenode) with 10 cycles of 30 s on, 30 s off at high power setting.

Chromatin IP was done on the SX-8G IPStar Automated System (Diagenode) with buffers from the Auto Transcription ChIP kit (Diagenode) following instruction manual version V1_07-10-10. The pre-set IP protocol "ChIP 22hr IPure16 200vol" was used with 5 hours of antibody coating and 16 hours of ChIP reaction at 4 °C. ChIP of ESR1 used 3 µg of the Diagenode antibody Mab-009-050 Lot NR-010 and 100 µL of the sonicated chromatin diluted with 100 µL of ChIP Buffer T. ChIP of p300 used 3 µg of the Santa Cruz antibody sc-585x Lot#E2610 and 25 µL of the sonicated chromatin diluted with 175 µL of ChIP Buffer T. The ChIP products were reverse crosslinked at 65 °C for 6 hours with occasional vortexing. ChIP DNA was purified by reagents in the Auto IPure kit (Diagenode) but done manually following the IPure kit (Diagenode) instruction manual version V2_12-05-10.

Sequencing library was prepared from the purified DNA by the SPRI-te Nucleic Acid extractor (Beckman Coulter) with SPRIworks Fragment Library System I cartridges according to manufacturer's protocol. Enrichment was done with 2x Phusion Master Mix, PE PCR primer 1.0 (Illumina) and a barcoded paired-end PCR primer 2.0. Each biological replicate was sequenced in one paired-end (PE) lane on Illumina Genome Analyzer II. The sequencing reads of 36 bp were aligned to the hg18 genome by the short reads aligner bowtie version 0.12.5 suppressing all alignments for reads that align to more than one location (`-m 1`). The sequencing and alignment statistics are listed in Table 3.2.

Sample name	Total reads	Aligned unique		Aligned repeat		Aligned none	
	million	million	percent	million	percent	million	percent
U87H p300 PE1	39.7	19.8	50.0	18.8	18.1	12.7	32.0
U87H p300 PE2	39.7	19.6	49.5	18.6	18.0	12.6	31.8
U87H ESR1 PE1	39.6	20.2	51.1	19.2	19.2	11.7	29.7
U87H ESR1 PE2	39.6	20.1	50.9	19.1	19.1	11.7	29.6

Table 3.2: Illumina sequencing statistics of the ChIP-seq samples. PE1: paired-end read 1. PE2: paired-end read2. Aligned unique: reads that are uniquely aligned to the genome. Aligned repeat: reads that are aligned to more than one location in the genome. Aligned none: reads that are not aligned to any location in the genome.

3.2.6 Cell viability assays

WST-1 assay. 4,000 cells in 100 µL of parental media were seeded per well in a 96 well clear plate. 24 hours later, the media was removed and each well was washed with 150 µL of PBS and replaced with 100 µL of fresh serum free media (DMEM with no phenol red) containing indicated concentrations of AG1478, 17β-Estradiol, or 17-AAG. Six to eight technical replicates were done for at least three biological replicates for each treatment of each cell line. AG1478 and 17-AAG were purchased from A.G. Scientific (San Diego, CA) and dissolved in DMSO to make 10 mM stock solution, stored at −20 °C in the dark and diluted to the desired concentration immediately prior to adding to the culture media. 17β-Estradiol was purchased from Sigma-Aldrich (product number E1024) and dissolved in pure ethanol 200 proof to make 10 mM stock solution, stored at −20 °C in the dark and diluted to the desired concentration immediately prior to adding to the culture media. After 72 hours of drug treatment cell viability was measured by the WST-1 reagent (Roche Applied Science). 10 µL of WST-1 was added to each well. The plates were incubated at 37 °C for three hours and absorbance at 450 nm was measured by Varioskan Flash Multimode Reader (Thermo Scientific). Background intensities were obtained from wells that were treated identically but without cells and were subtracted from the readings of wells on the same plate. Relative cell numbers were computed by taking ratios between the background subtracted readings of the drug treated cells and vehicle control cells, and statistical significance was computed by paired Student's t-test between the treatment and control conditions for each cell line. The Bliss independence effect was calculated as $F_{AB,expect} = F_A + F_B(1 - F_A)$, where F_A is the observed growth inhibition of 5 µM AG1478, F_B is the observed growth inhibition of various concentrations of 17β-Estradiol or 17-AAG, and $F_{AB,expect}$ is the expected growth inhibition of 5 µM AG1478 with a second drug at that concentration if the two drugs were independent. The expected survival fraction $1 - F_{AB,expect}$ was plotted.

To enable comparison with Sauvageot et al. (2009), the data from treatment of 5 µM AG1478 with and without 17-AAG was replotted in the form of interaction ratio, defined as the ratio of the observed growth inhibition of the drug combination and the expected growth inhibition computed from the observed inhibition by either drug alone: $\frac{F_{AB,observed}}{F_{AB,expect}}$.

TMRE imaging. Cell seeding and treatment were done in the same way as the WST-1 assay with the exception that the cells were seeded at a density of 20,000 cells per well in 500 µL of media in a 24 well clear plate. At the end of treatment, tetramethylrhodamine, ethyl ester (TMRE) (Invitrogen) was added to the cell culture media to a final concentration of 100 nM. After incubating at 37 °C for 20 min, the media was removed and replaced with fresh, warm serum free media (DMEM with no phenol red). Fluorescence photomicrographs were taken on a Leica DM IL LED tissue culture microscope connected to a Leica EC3 digital camera.

3.3 Materials and methods: computational

3.3.1 Overview of the prize collecting Steiner tree

We used the Goemans-Williamson formulation of the prize-collecting Steiner tree (PCST) problem. Given an undirected graph $G = (V, E)$ where nodes $i \in V$ are associated with penalties $\pi_i \geq 0$ and edges $e \in E$ are associated with costs $c_e \geq 0$, we aim to find a subtree $F = (V_F, E_F)$ of G such that

$$\sum_{i \notin V_F} \pi_i + \sum_{e \in E_F} c_e \tag{3.3.1}$$

is minimized.

Nodes that have positive penalty values are called terminals. For our application, the nodes and edges were obtained from protein-protein interaction network datasets. Protein nodes to which experimental data could be mapped received positive penalty values (and therefore they were terminals; see Section 3.3.3) and other nodes received zero penalties. The cost on edges was inversely related to the confidence on each interaction based on available evidence (see Section 3.3.2) so that high confidence edges had lower costs and therefore were preferentially selected to be in the solution. We further introduced a scaling parameter β to balance the penalties paid to exclude nodes with experimental observations and the costs of including edges to connect these nodes:

$$\sum_{i \notin V_F} \beta \pi_i + \sum_{e \in E_F} c_e \,. \tag{3.3.2}$$

We solved this optimization problem using the branch-and-cut approach (Ljubi et al., 2006) implemented in the `dhea-code` software program that called the ILOG CPLEX linear programming solver version 12.1 (IBM). We now describe how the experimental data were transformed into input for the algorithm. An overview of the work flow is in Figure 3-1.

3.3.2 Interactome graph

This is Step 1 in Figure 3-1. The set of edges E of the input graph G consisted of direct (physical) and indirect (functional) protein-protein interactions for human in the STRING database version 8.2 (Szklarczyk et al., 2011), which also assigned confidence score for each interaction based on multiple evidence types: conserved neighborhood, gene fusions, phylogenetic co-occurrence, co-expression, database, large-scale experiments, and literature co-occurrence. Since we were only interested in experimentally determined interactions, we computed a new score S_e for each interaction e combining the scores from evidence type database $S_{e,database}$ and evidence type experiment $S_{e,experiment}$:

$$S_e = 1 - S_{e,database} S_{e,experiment}, \forall e \in E. \tag{3.3.3}$$

The cost $c(e)$ on each edge e was

$$c(e) = -\log(S_e), \forall e \in E. \tag{3.3.4}$$

3.3.3 Node penalties

We defined two kinds of penalties for proteins in the STRING interaction graph: one at the signaling level from the phosphoproteomics MS data, the other at the transcription regulation level from the DNase-seq and mRNA expression data. This

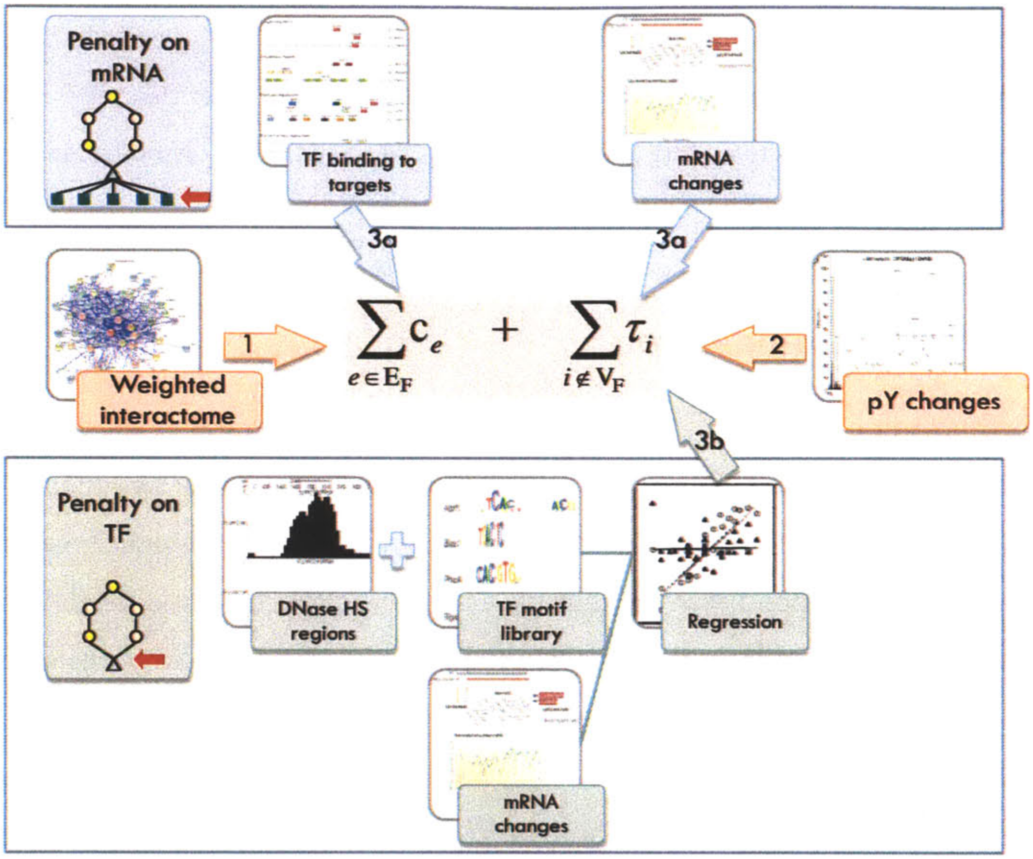


Figure 3-1: Work flow diagram for defining the optimization objective function from input datasets. There are two streams of analysis differed by the way in which the transcription data is included. In both streams, interaction weights go into the edge cost summation term (Step 1) and the changes in tyrosine phosphorylation from MS data go into the node penalty summation term (Step 2). From there, either Step 3a or Step 3b is followed. In Step 3a, the transcription factor to mRNA target relationships are added to the edges to form the total interactome, and the mRNA nodes are assigned penalty values. In Step 3b, the DNaseI hypersensitive regions are scored for matches to transcription factor motifs, and these scores are assigned to nearby genes. The changes in the mRNA transcript level of genes are regressed against the motif scores and the significance of the regression is used as penalty on the transcription factors corresponding to that motif in the protein interactome.

is Step 2 and Step 3 in Figure 3-1. The node penalties from the phosphorylation data corresponded to Step 2. Step 3a and 3b are two alternatives for including the transcriptional data. The yeast pheromone response network in Chapter 2 followed 3a, while the U87 EGFRvIII analysis here followed 3b.

Phosphorylated peptide sequences from MS/MS data were matched to the protein sequences provided by the STRING database download using the stand-alone peptide BLAST program (`blastall`) version 2.2.17 with the following parameters recommended for matching short amino acid sequences: type of search blastp, exception value 20000, do not filter low complexity regions, gap opening cost 9, gap extension cost 1, protein scoring matrix PAM30, word size 2, multiple hits window size 40 (`-p blastp -e 200000 -F F -G 9 -E 1 -M PAM30 -W 2 -A 40`). 100 alignments were requested for each peptide in BLAST XML format report (`-b 100 -m 7`), which were parsed by the `Bio.Blast` module in BioPython. Proteins that contained perfect alignment to a peptide sequence received a positive penalty value that was proportional to the absolute value of log fold change in phosphorylation between the U87H and U87DK cells. If one peptide sequence was aligned to multiple proteins in the STRING interaction graph, all these proteins received the same penalty value. If multiple phosphorylated peptide sequences were perfectly aligned to one protein, the maximum fold change in phosphorylation of these peptides was used to calculate the penalty value for this protein.

We derived the penalty values for transcription factors in the protein interaction network from the inferred activity of these transcription factors in inducing changes of mRNA expression. Specifically, we used a regression method to find the correlation between the differential mRNA expression to the sequence specific transcription factor binding motifs in nearby open chromatin regions. Filtering the `limma` analysis results by a maximum p-value of 0.001 adjusted by the Benjamini and Hochberg method (Benjamini and Hochberg, 1995) gave 1292 probe sets differentially expressed between U87DK and U87H cells. These probe sets were mapped to 1624 genes using annotation from the Ensembl Project release 54 (`http://may2009.archive.ensembl.org` and Flicek et al., 2011). From the DNase-seq data of U87DK and U87H cells, we found genomic regions that were differentially hypersensitive between these two cell lines by using the peak caller MACS (Zhang et al., 2008) version 1.4.0beta. For each cell line, aligned reads from the two biological replicates were concatenated. With the U87H read file as the `treatment` parameter and U87DK read file as the `control` parameter, a p-value cutoff of 1×10^{-6} (`-p 1e-6`) and also calling subpeaks (`--call-subpeaks`), 7760 peaks, further divided into 13141 subpeaks, were identified to be more hypersensitive in U87H cells than in U87DK cells. By reversing the `treatment` and `control` read files, 5047 peaks, divided into 9683 subpeaks, were identified to be more hypersensitive in U87DK cells than in U87H cells. Each of these subpeak summits was mapped to the Ensembl 54 annotated human transcripts that have transcription start sites within 40,000 bp of the peak, using functionalities in the `ChIPpeakAnno` package in BioConductor. Sequences from 100 bp upstream and downstream of the subpeak summits were retrieved and the transcription factor affinity scores (Figure 1-6 and Foat et al., 2006) were computed for the 572 good quality matrices in release 2009.1 of the TRANSFAC database (Matys et al., 2006). Let $S_{g,H}$ be the set of sequences whose summits are mapped to the gene transcript g in U87H cells, $S_{g,DK}$ be the set of sequences whose summits are mapped to g in U87DK cells. Let T be the set of TRANSFAC matrices and and $x_{i,g,\tau,c}$ be the affinity score for matrix $\tau \in T$ for the i^{th} sequence in $S_{g,c}$ that is mapped to gene g in condition $c \in \{\text{DK, H}\}$. The affinity score of transcription factor matrix τ for gene transcript g is

$$x_{g,\tau} = \sum_{i=1}^{|S_{g,H}|} x_{i,g,\tau,H} - \sum_{j=1}^{|S_{g,DK}|} x_{j,g,\tau,DK} \ . \tag{3.3.5}$$

Let G be the set of differential expressed genes described previously and and y_g be the log base 2 fold change in expression of transcript $g \in G$ comparing U87H and U87DK cells. For each matrix τ we fit the differential expression of g and the affinity score by a univariate linear model:

$$y_g = \beta_\tau x_{g,\tau} + \epsilon_g. \tag{3.3.6}$$

We selected the matrices for which the coefficients of the linear regression were significantly different from zero by a p-value threshold of 0.01 after Bonferroni correction, and used the t-statistic values of the regression coefficients as the penalties on proteins that were mapped to these binding matrices by TRANSFAC (Lee and Bussemaker, 2010; Foat et al., 2008, 2005). In cases where one binding matrix corresponded to multiple transcription factors (as commonly found in transcription factor families), all these transcription factors received the same penalty value.

An alternative to the univariate regression of individual motifs is to fit a multivariate regression model of differential gene expression to the motif scores. The elastic net regression algorithm (Zou and Hastie, 2005) was chosen because it balances sparse and complex models and is thought to be good for selecting correlated features. Under a cross-validation setting, the motif features selected were not significantly different from those resulting from the univariate regression method (Figure 3-2).

3.3.4 Post-processing of PCST solutions

Solution of the PCST optimization problem was visualized on the Cytoscape program (Cline et al., 2007) and all network diagrams were exported graphics from the software. To compute the node betweenness centrality of the nodes in the PCST solution, we first augmented the tree structure of the solution by adding back the edges in the input interactome graph that contain the nodes in the solution, and used the `centrality.betweenness_centrality` function in the `networkx` Python package (Hagberg et al., 2008) version 0.35.

3.3.5 ChIP-seq data analysis

Peak calling of the ChIP-seq aligned reads was done by MACS version 1.4 (Zhang et al., 2008) with the following parameter values: `--mfold=10,30 --tsize=35 --bw=150`. The reported peaks were filtered by a p-value threshold of 1×10^{-7} and associated with genes whose transcription start sites lied within 10,000 bp of any peaks. This list of genes was then ranked by fold enrichment (also reported by MACS) and used as input to the web tool GOrilla (Eden et al., 2009) to identify enriched Gene Ontology

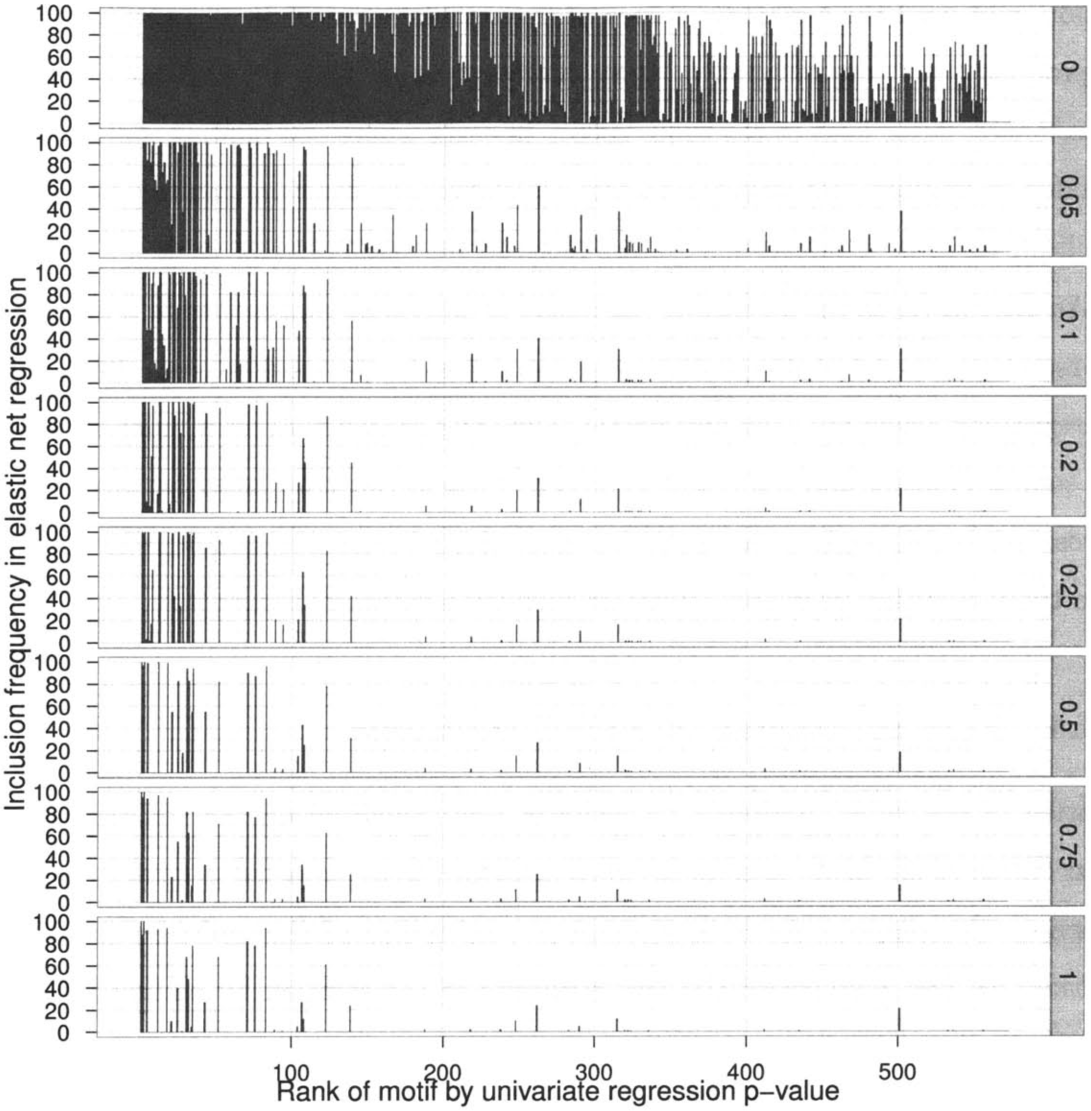


Figure 3-2: In regression of differential gene expression with respect to motif scores, the motif features that are more frequently included in the elastic net regression are also ranked high by univariate regression. A univariate regression fit to the differential gene expression was performed for each motif feature, and the motifs were then ranked by the p-value from this regression. From the set of differential expression and the associated motif scores, 100 random subsets were generated, each consisting of 80% of the original dataset. For each subset, a four-fold cross-validation procedure was used to fit an elastic net regression model that selected a set of motif features to be included in the regression. The x axis is the rank of motifs sorted by p-values from the univariate regression procedure, and the height of each bar is the frequency at which that motif feature was included in the 100 multivariate regressions by elastic net. Each panel shows the feature selection results for the indicated α value parameter of elastic net.

(GO) terms.

3.4 Results

3.4.1 The PCST solution provides a global view of the EGFRvIII signaling network

The PCST network we constructed by comparing the U87H and U87DK cells using phosphoproteomics, transcriptional profiling and DNase-seq data gives a high-level view of the overall biological processes involved (Figure 3-3). Some of these processes, such as the PI3K/PTEN pathway and the focal adhesion signaling pathway, have been previously implicated in EGFRvIII induced migration in U87 cells (Cai et al., 2005; Liu et al., 2010). At the level of transcriptional regulation, the transcription factor STAT3 mediates EGFRvIII induced transformation (de la Iglesia et al., 2008), and the CDK inhibitor p27 is lower in EGFRvIII tumors, leading to hyperphosphorylation on Rb and activation of the E2F transcription factors (Narita et al., 2002). Although no phosphorylation sites on the E2F transcription factors were reported by mass spectrometry, and the tyrosine phosphorylation site on STAT3 (Y705) shows less than 10% change between U87DK and U87H, these two proteins are featured prominently in our network solution, demonstrating the value of the transcriptional datasets.

3.4.2 ESR1 and HSP90 are key nodes in the PCST solution and important components for cell viability

Many proteins included in the network solution do not contain phosphorylation sites reported by mass spectrometry. We can quantify the importance of each node in the overall network connectivity by a node-betweenness centrality measure (Table 3.3). The proto-oncogene SRC is ranked among highest in this analysis. The Src family kinases FYN and SRC are known effectors of EGFRvIII signaling in glioblastoma, leading to tumor growth and motility (Liu et al., 2010). The estrogen receptor ESR1 is also ranked among the top, which presents an intriguing hypothesis. Glial tumors

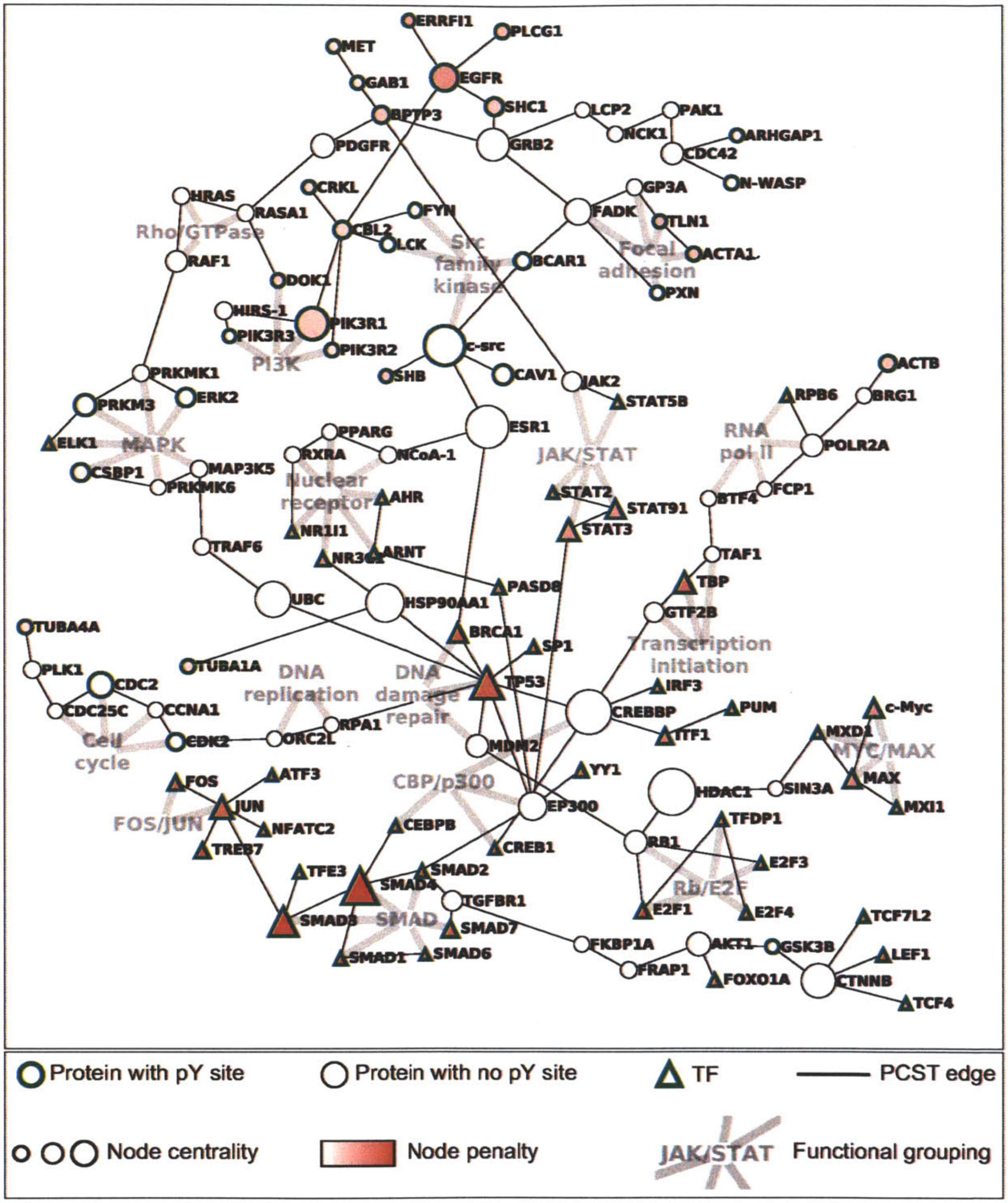


Figure 3-3: The PCST solution network constructed from phosphoproteomic, mRNA expression array and DNase-seq data comparing the U87H and U87DK cells. pY: phosphotyrosine. TF: transcription factor.

are more prevalent in males (Schwartzbaum et al., 2006; McKinley et al., 2000), the risk increases with later menarche and menopause (Cowppli-Bony et al., 2011), and young age and female gender are two favorable factors for long-term survival (Krex et al., 2007; Shinojima et al., 2004). A recent study of glioblastoma in a rat model shows estrogen treatment increases survival in a gender specific manner although it is unclear whether the estrogen receptor is involved (Barone et al., 2009). On the other hand, the selective estrogen receptor modulator tamoxifen can inhibit proliferation and induce apoptosis in several glioma cell lines (Pollack et al., 1990; Zhang et al., 2000; Kim et al., 2005), and a subgroup of malignant glioma patients responded or stabilized with high dose of tamoxifen (Couldwell et al., 1996). There are several ongoing clinical trials for tamoxifen treatment in combination with other agents or radiation therapies (with EGFR inhibitors in breast cancer only). We further analyze our PCST network by comparing it to networks constructed from genes discovered in RNAi screens that are (1) essential for tumor cell growth in multiple tumor types (Luo et al., 2008), (2) essential for growth in cell lines of glioma lineage (Luo et al., 2008), and (3) sensitizing cells to EGFR inhibition (Astsaturov et al., 2010), ESR1 is one of the few hits that appear in all three of the networks. To test this hypothesis, I treated the U87DK and U87H cells with the EGFR inhibitor AG1478 with or without 17β-Estradiol, and assayed for cell viability by the WST-1 reagent that measures the metabolic activity of viable cells. The viability of both cell lines are reduced with increasing concentrations of 17β-Estradiol compared to vehicle control (Figure 3-4). Adding AG1478 further reduces viability to a greater extent than if these two drugs acted independently. These suggest that the estrogen receptor pathway is an important component in the survival of these cells.

HSP90 was selected from the list of nodes with the highest node betweenness centrality as a second target for validation as there are well characterized small molecular inhibitors for this protein. 17-allylamino-17-demethoxygeldanamycin (17-AAG) binds to and inhibits HSP90 and may lead to degradation of HSP90 client proteins. To test this hypothesis, I treated the U87DK and U87H cells with AG1478 with or without 17-AAG. Nanomolar concentrations of 17-AAG greatly reduce the viability of both

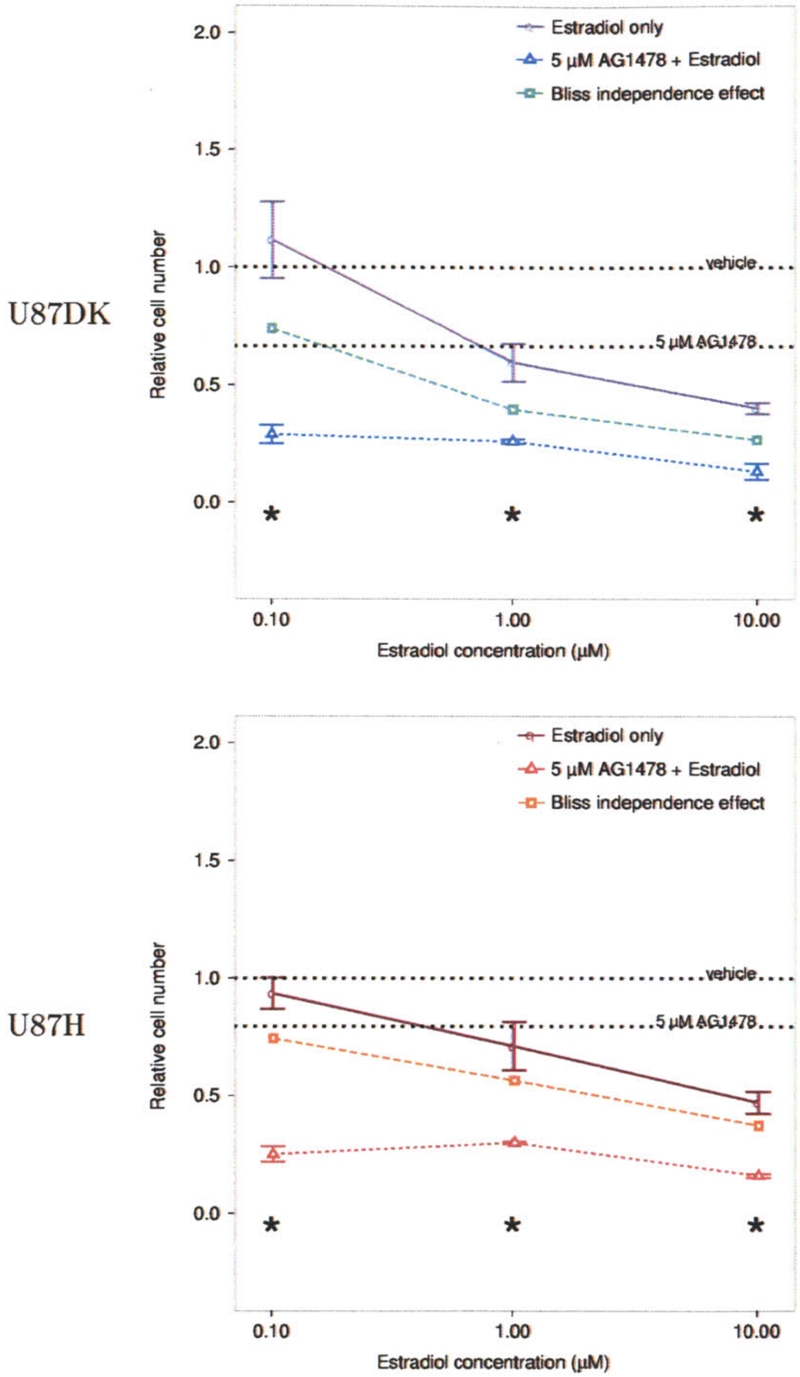


Figure 3-4: Cell viability assay by WST-1 following estradiol and AG1478 treatment of U87DK and U87H cells. The cells were serum starved for 24 hours and treated with the indicated concentration of estradiol with or without AG1478 for 72 hours. The Bliss independence effect (BLISS, 1956) was computed using measurements from single agent treatment of 5 µM AG1478 and estradiol. Error bars are standard error ($n = 3$) and asterisks (*) mark the concentrations of estradiol that in combination with 5 µM AG1478 reduce cell viability significantly compared to the expected effect if the two drugs acted independently ($p < 0.05$).

Protein name	STRING v8.2 ID	Node betweenness centrality
HDAC1	9606.ENSP00000362649	0.098600
CREBBP	9606.ENSP00000371502	0.092185
ESR1	9606.ENSP00000206249	0.086484
SRC	9606.ENSP00000362659	0.084406
SMAD4	9606.ENSP00000341551	0.080064
HSP90AA1	9606.ENSP00000335153	0.073390
UBC	9606.ENSP00000344818	0.064671
TP53	9606.ENSP00000269305	0.063942
PIK3R1	9606.ENSP00000274335	0.060804
CTNNB	9606.ENSP00000344456	0.059800

Table 3.3: Top ten nodes in the EGFRvIII PCST network ranked by node betweenness centrality.

cell lines, as assayed by a metabolic activity readout WST-1 (Figure 3-5) and visualized by the mitochondria membrane potential stain TMRE (Figure 3-6). Adding AG1478 at a concentration that has moderate effect by itself further reduces the viability. Using the Bliss independence measure for drug synergy (BLISS, 1956), we computed the expected effects of the AG1478 and 17-AAG combinations if they were independent. Compared to the experimentally observed effect, this combination is synergistic in U87H cells at a much lower concentration of 17-AAG than in U87DK cells.

3.4.3 Transcriptional regulators p300 and SMAD proteins may contribute to the mesenchymal-like phenotype of U87H

In the PCST solution, the transcriptional co-activator p300 appears as a central point that links together multiple transcription factors. ChIP-seq experiment with p300 was performed in the U87H cells to elucidate its biological function. By GO enrichment analysis (Eden et al., 2009) (Table 3.4), the p300 target genes are enriched for neuronal development and differentiation processes. This is consistent with the recent hypothesis that the cell of origin for gliomas is a neural stem cell that was

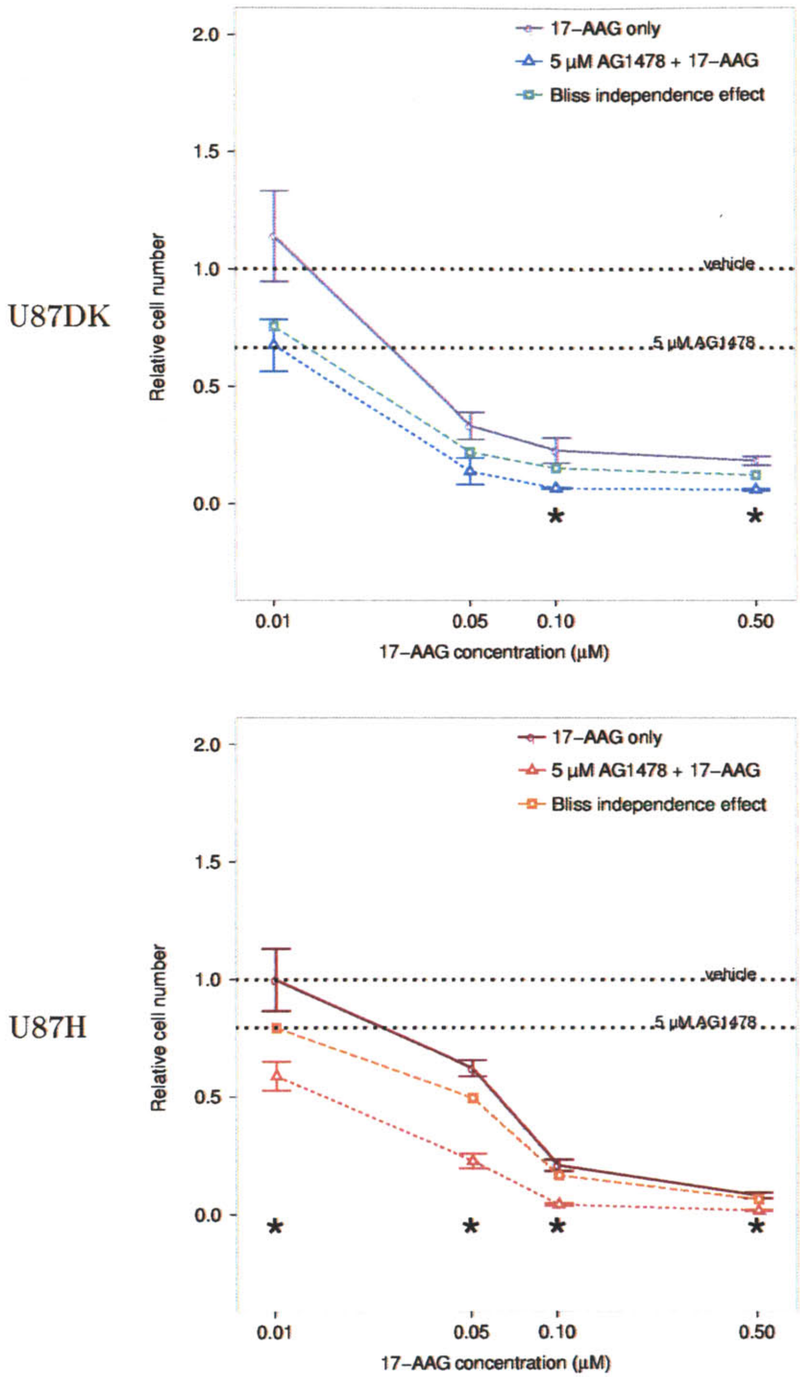


Figure 3-5: Cell viability assay by WST-1 following 17-AAG and AG1478 treatment of U87DK and U87H cells. The cells were serum starved for 24 hours and treated with the indicated concentration of 17AAG with or without AG1478 for 72 hours. The Bliss independence effect (BLISS, 1956) was computed using measurements from single agent treatment of 5 µM AG1478 and 17-AAG. Error bars are standard error ($n = 3$) and asterisks (*) mark the concentrations of 17-AAG that in combination with 5 µM AG1478 reduce cell viability significantly compared to the expected effect if the two drugs act independently ($p < 0.05$).

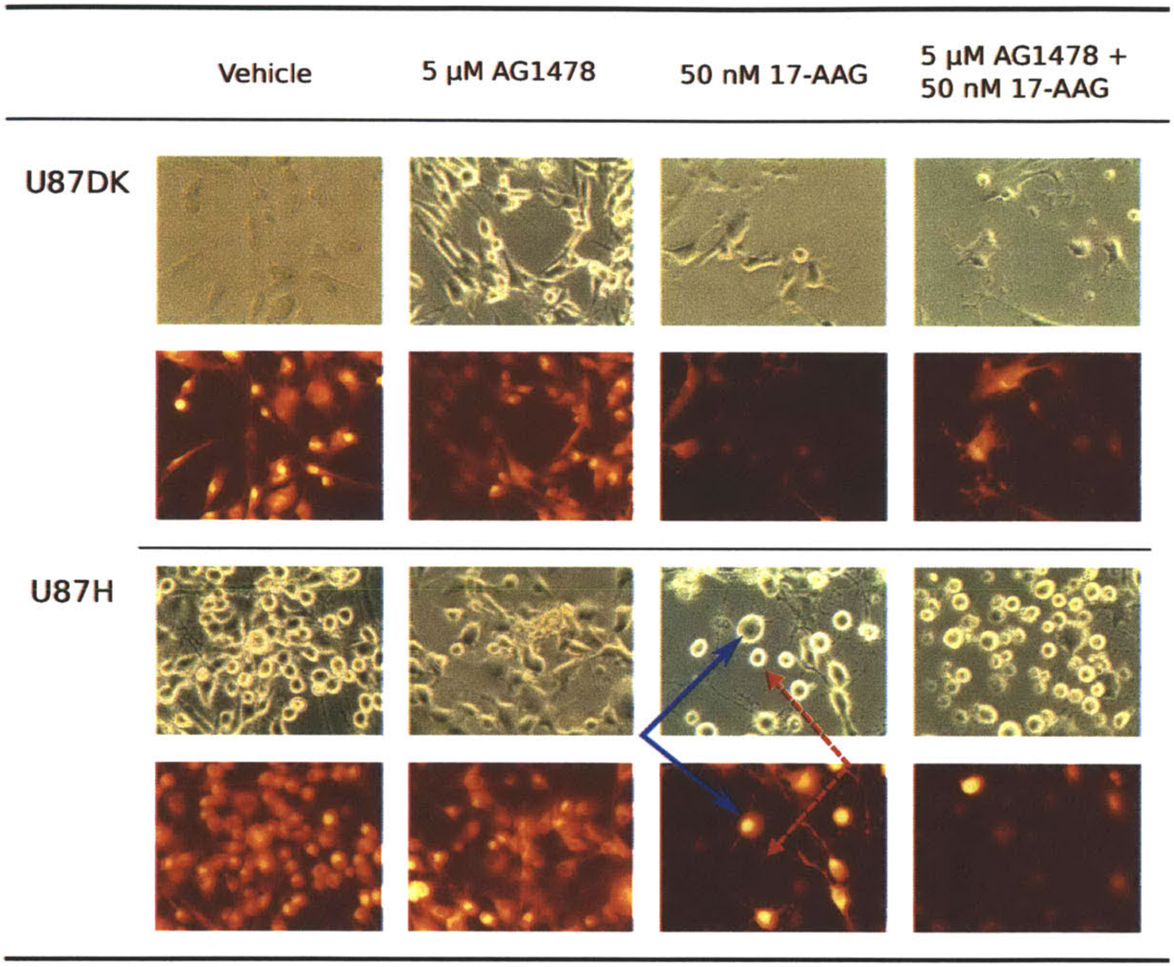


Figure 3-6: Microphotographs of TMRE staining of U87DK and U87H cells following 17-AAG and AG1478 treatment. The top row for each cell type is the bright field image and the bottom row is the fluorescence image of the same field of view. A live cell can be seen in the bright field and appears in red and orange color under fluorescence due to active mitochondria (blue solid arrow). A dead cell can be seen in the bright field but is not visible under fluorescence (red dashed arrow).

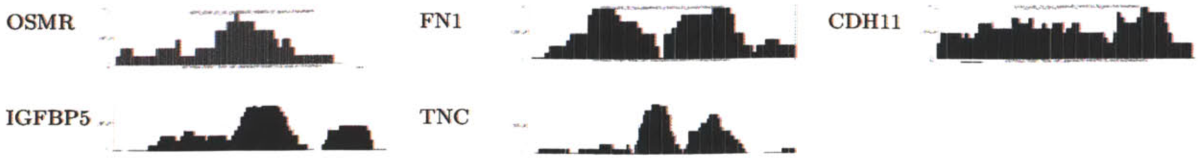


Figure 3-7: ChIP-seq of p300 reveals that many EMT marker genes are bound by p300 in U87H cells. Shown here are the ChIP-seq peaks viewed in UCSC genome browser.

transformed to produce progeny cells with limited differentiation potential (Sanai et al., 2005). Also notable are the functional categories of cellular adhesion and Wnt receptor signaling, which leads to a hypothesis for the mechanism of possible mesenchymal transformation of the U87H cells.

At the global level the p300 target genes are enriched for biological processes implicated in epithelial to mesenchymal transition (EMT), and in the PCST solution p300 links together the SMAD proteins, CEBPB (C/EBPβ), and STAT transcription factors, all are terminals selected by the expression regression method. Several lines of evidence further suggest that these associations contribute to the mesenchymal-like phenotype of U87H cells: (1) the U87H cells have poor attachment in tissue culture compared to the U87DK cells, which resembles mesenchymal cells; (2) CEBPB and STAT3 synergistically induce mesenchymal transformation of glioma cells (Carro et al., 2010), but the level of CEBPB transcript and activated STAT3 protein do not change significantly between the U87H and U87DK cells; (3) the physical interaction between SMAD4 and CEBPB represses the transactivation function of CEBPB (Choy and Derynck, 2003), and the SMAD4 mRNA level is reduced by 5-fold in the U87H cells compared to U87DK. There are several experimental approaches for testing this hypothesis. First, we can use quantitative Western blots to compare the protein level of SMAD4 between U87H and U87DK. ChIP-seq of SMAD4 in U87DK and U87H cells will reveal its binding locations in relation to EMT marker genes. To confirm the functional role of SMAD4, we can over-express SMAD4 or stimulate its activation by the TGF-β pathway, and as phenotypic readout measure the expression of mesenchymal markers and observe whether it restores the attachment of U87H cells.

GO Term	Description	p-value
GO:0051128	regulation of cellular component organization	8.32×10^{-6}
GO:0022610	biological adhesion	1.28×10^{-5}
GO:0007155	cell adhesion	1.28×10^{-5}
GO:0032940	secretion by cell	4.88×10^{-5}
GO:0046903	secretion	1.18×10^{-4}
GO:0031345	negative regulation of cell projection organization	1.88×10^{-4}
GO:0030177	positive regulation of Wnt receptor signaling pathway	1.89×10^{-4}
GO:0015850	organic alcohol transport	2.13×10^{-4}
GO:0010975	regulation of neuron projection development	3.29×10^{-4}
GO:0045664	regulation of neuron differentiation	3.34×10^{-4}
GO:0051960	regulation of nervous system development	4.65×10^{-4}
GO:0046165	alcohol biosynthetic process	5.15×10^{-4}
GO:0050795	regulation of behavior	5.23×10^{-4}
GO:0031344	regulation of cell projection organization	5.82×10^{-4}
GO:0051239	regulation of multicellular organismal process	6.07×10^{-4}
GO:0042133	neurotransmitter metabolic process	8.37×10^{-4}

Table 3.4: Enriched GO categories of p300 bound genes in U87H cells

3.5 Discussion

3.5.1 Linking signaling and transcription data by molecular interactions can generate hypotheses that are not immediately obvious from the experimental data

We showed that the PCST solution network not only provided a high level view of the biological processes in the EGFRvIII network of U87 cells, but it was also useful in generating hypotheses that have functional significance. These proposed hypotheses, some of which were validated experimentally, are not easily apparent from simple inspection of the phosphoproteomic and transcription data. One might argue that the importance of ESR1 and HSP90 have been previously implicated in oncogenesis. However, since neither appears in the phosphoproteomic and transcriptional profiling data, they would be buried in thousands of other proteins in the interactome graph that interact directly or indirectly with the hits from the experiments. Our approach provides a way to prioritize these hypotheses. For transcriptional regulation in this system, standard promoter analysis of the differential expressed genes yielded little information about potential regulators, so bringing in additional data from upstream signaling allowed us to identify promising candidates. This will be especially useful for systems that lack strong signals from "master regulators", either due to the inherent feature of the system or limitations of technology.

3.5.2 Possible mechanism of ESR1 in cell survival

In terms of mechanism, although ESR1 is a well known transcription factor, it was not selected to be a terminal from the expression regression step. Its inclusion in the network may be due to its genomic actions independent of estrogen response elements (ERE) or its non-genomic role in activating protein kinase cascades (Bjrnstrm and Sjberg, 2005) such as the ERK MAPK and PI3K signaling (Levin, 2005). Interestingly, non-genomic signaling by 17β-Estradiol can both stimulate and inhibit apoptosis (Lewis-Wambi and Jordan, 2009). Our experiment suggests that 17β-Estradiol reduces cell viability and is synergistic with EGFR inhibition. Further experiment is necessary to distinguish the two possible mechanisms.

At the population level, the effect of steroid hormone in the etiology of glioma is still unclear (reviewed in Kabat et al., 2010): while epidemiology analysis of glioma incidence shows evidence of relative protection of female to male prior to menopause, suggesting a protective role of estrogens, the relative reduction in glioma risk in women from the use of exogenous hormone is small and inconsistent. It remains to be seen whether there are distinct mechanisms that lead to the development of glioma and the response of glioma to hormonal treatment.

3.5.3 Synergy of inhibiting HSP90 and EGFRvIII

The HSP90 protein is a molecular chaperon that maintains the stability of many oncogenic signaling proteins, and inhibitors to HSP90 inhibitors have shown promising anti-tumor activities (Neckers, 2002; Goetz et al., 2003). Although HSP90 is known to specifically increase the *in vivo* efficacy of chemotherapeutic agents (Neckers, 2007), previous studies offer contradictory predictions on the effect of HSP90 inhibition in glioblastoma in the context of EGFRvIII. Lavictoire et al. (2003) reported that the interaction between HSP90 and nascent EGFRvIII was necessary for EGFRvIII expression but did not give evidence of the consequence of HSP90 inhibition to cell survival. On the other hand, Cao et al. (2011) showed that EGFR inhibition induced translocation of EGFRvIII to the mitochondria and this translocation was responsible for tumor drug resistance. Furthermore, the current standard-of-care agent for glioblastoma therapy temozolomide (TMZ) was not synergistic with 17-AAG in U87 glioma cells (Sauvageot et al., 2009, and Figure 3-8). We present evidence that AG1478 and 17-AAG achieve synergy in the U87H cells at low concentrations of 17-AAG and they may be good candidates for combination therapy for EGFRvIII expressing glioblastoma.

3.5.4 Mechanism of p300 and SMAD4 in glioma EMT

Identifying the possible role of p300 and SMAD4 in EMT of glioma demonstrates the power of joint analysis of proteomic and transcriptional data. CEBPB and STAT3 were revealed to be master regulators of mesenchymal transformation of glioma cells from a large project of transcription profiling of tumors (Carro et al., 2010), but in our data we did not observe changes at the mRNA expression level of these two factors. In the PCST solution these two proteins are connected to p300 and SMAD4 by protein-protein interactions, where SMAD4 was selected by the expression regression analysis. p300 has been implicated in EMT in human colon carcinoma (Pea et al., 2006; Krubasik et al., 2006) and SMAD4 in breast cancer in response to TGF-β signaling (Deckers et al., 2006). The initial ChIP-seq experiment with p300 shows many EMT marker genes are bound in the U87H cells and the p300 bound genes are enriched for EMT related biological processes. Additional ChIP-seq experiments of SMAD4 and from U87DK cells, together with functional assays of EMT phenotype, will help establish the connection between these two factors and EGFRvIII signaling in the context of EMT.

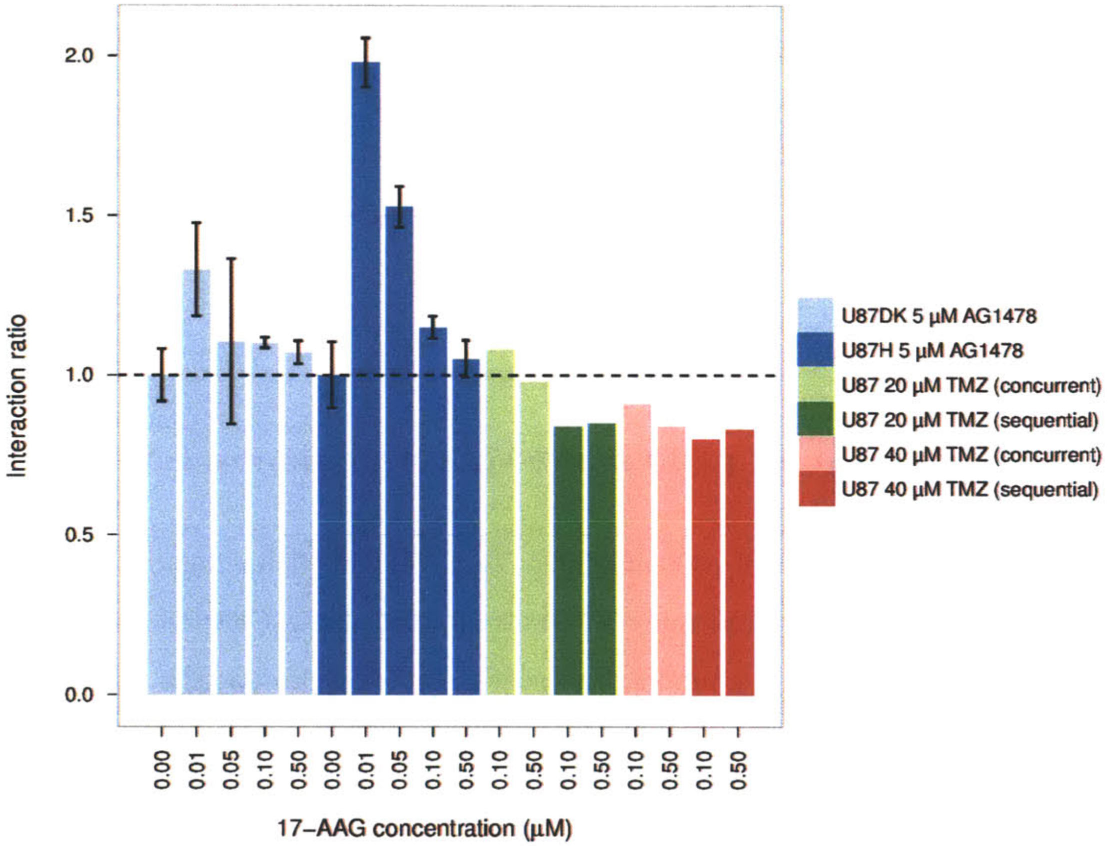


Figure 3-8: The Bliss interaction ratio of cell viability measurements for treatment of 17-AAG with AG1478 and treatment of 17-AAG with TMZ. The ratio was calculated as the observed inhibitory effect achieved by the drugs in combination divided by the expected inhibitory effect computed from each drug alone. Error bars are standard errors ($n = 3$). The 17-AAG with TMZ data is taken from Sauvageot et al. (2009).

Bibliography

I. Astsaturov, V. Ratushny, A. Sukhanova, M. B. Einarson, T. Bagnyukova, Y. Zhou, K. Devarajan, J. S. Silverman, N. Tikhmyanova, N. Skobeleva, A. Pecherskaya, R. E. Nasto, C. Sharma, S. A. Jablonski, I. G. Serebriiskii, L. M. Weiner, and E. A. Golemis. Synthetic lethal screen of an EGFR-centered network to improve targeted therapies. *Sci Signal*, 3(140):ra67, 2010.

T. A. Barone, J. W. Gorski, S. J. Greenberg, and R. J. Plunkett. Estrogen increases survival in an orthotopic model of glioblastoma. *J Neurooncol*, 95(1):37–48, Oct 2009.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 0035-9246.

L. Bjrnstrm and M. Sjberg. Mechanisms of estrogen receptor signaling: convergence of genomic and nongenomic actions on target genes. *Mol Endocrinol*, 19(4):833–842, Apr 2005.

C. I. BLISS. The calculation of microbial assays. *Bacteriol Rev*, 20(4):243–258, Dec 1956.

X.-M. Cai, B.-B. Tao, L.-Y. Wang, Y.-L. Liang, J.-W. Jin, Y. Yang, Y.-L. Hu, and X.-L. Zha. Protein phosphatase activity of PTEN inhibited the invasion of glioma cells with epidermal growth factor receptor mutation type III expression. *Int J Cancer*, 117(6):905–912, Dec 2005.

X. Cao, H. Zhu, F. Ali-Osman, and H.-W. Lo. EGFR and EGFRvIII undergo stress- and EGFR kinase inhibitor-induced mitochondrial translocalization: A potential mechanism of EGFR-driven antagonism of apoptosis. *Mol Cancer*, 10:26, 2011.

M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, A. Lasorella, K. Aldape, A. Califano, and A. Iavarone. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463(7279):318–325, Jan 2010.

L. Choy and R. Derynck. Transforming growth factor-beta inhibits adipocyte differentiation by Smad3 interacting with CCAAT/enhancer-binding protein (C/EBP) and repressing C/EBP transactivation function. *J Biol Chem*, 278(11):9609–9619, Mar 2003.

M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P.-L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker, and G. D. Bader. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*, 2(10):2366–2382, 2007.

W. T. Couldwell, D. R. Hinton, A. A. Surnock, C. M. DeGiorgio, L. P. Weiner, M. L. Apuzzo, L. Masri, R. E. Law, and M. H. Weiss. Treatment of recurrent malignant gliomas with chronic oral high-dose tamoxifen. *Clin Cancer Res*, 2(4):619–622, Apr 1996.

A. Cowppli-Bony, G. Bouvier, M. Ru, H. Loiseau, A. Vital, P. Lebailly, P. Fabbro-Peray, and I. Baldi. Brain tumors and hormonal factors: review of the epidemiological literature. *Cancer Causes Control*, 22(5):697–714, May 2011.

N. de la Iglesia, G. Konopka, S. V. Puram, J. A. Chan, R. M. Bachoo, M. J. You, D. E. Levy, R. A. Depinho, and A. Bonni. Identification of a PTEN-regulated STAT3 brain tumor suppressor pathway. *Genes Dev*, 22(4):449–462, Feb 2008.

M. Deckers, M. van Dinther, J. Buijs, I. Que, C. Lwik, G. van der Pluijm, and P. ten Dijke. The tumor suppressor Smad4 is required for transforming growth factor beta-induced epithelial to mesenchymal transition and bone metastasis of breast cancer cells. *Cancer Res*, 66(4):2202–2209, Feb 2006.

E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics*, 10:48, 2009.

P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Khri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H. S. Riat, D. Rios, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y. A. Tang, S. Trevanion, J. Vandrovcova, A. J. Vilella, S. White, S. P. Wilder, A. Zadissa, J. Zamora, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernndez-Suarez, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, J. Vogel, and S. M. J. Searle. Ensembl 2011. *Nucleic Acids Res*, 39(Database issue):D800–D806, Jan 2011.

B. C. Foat, S. S. Houshmandi, W. M. Olivas, and H. J. Bussemaker. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci U S A*, 102(49):17675–17680, Dec 2005.

B. C. Foat, A. V. Morozov, and H. J. Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22(14):e141–e149, Jul 2006.

B. C. Foat, R. G. Tepper, and H. J. Bussemaker. TransfactomeDB: a resource for exploring the nucleotide sequence specificity and condition-specific regulatory activity of trans-acting factors. *Nucleic Acids Res*, 36(Database issue):D125–D131, Jan 2008.

M. P. Goetz, D. O. Toft, M. M. Ames, and C. Erlichman. The Hsp90 chaperone complex as a novel target for cancer therapy. *Ann Oncol*, 14(8):1169–1176, Aug 2003.

A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, Aug. 2008.

H. S. Huang, M. Nagane, C. K. Klingbeil, H. Lin, R. Nishikawa, X. D. Ji, C. M. Huang, G. N. Gill, H. S. Wiley, and W. K. Cavenee. The enhanced tumorigenic activity of a mutant epidermal growth factor receptor common in human cancers is mediated by threshold levels of constitutive tyrosine phosphorylation and unattenuated signaling. *J Biol Chem*, 272(5):2927–2935, Jan 1997.

P. H. Huang, A. Mukasa, R. Bonavia, R. A. Flynn, Z. E. Brewer, W. K. Cavenee, F. B. Furnari, and F. M. White. Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. *Proc Natl Acad Sci U S A*, 104(31):12867–12872, Jul 2007.

G. C. Kabat, A. M. Etgen, and T. E. Rohan. Do steroid hormones play a role in the etiology of glioma? *Cancer Epidemiol Biomarkers Prev*, 19(10):2421–2427, Oct 2010.

Y.-J. Kim, C.-J. Lee, U. Lee, and Y.-M. Yoo. Tamoxifen-induced cell death and expression of neurotrophic factors in cultured C6 glioma cells. *J Neurooncol*, 71(2): 121–125, Jan 2005.

D. Krex, B. Klink, C. Hartmann, A. von Deimling, T. Pietsch, M. Simon, M. Sabel, J. P. Steinbach, O. Heese, G. Reifenberger, M. Weller, G. Schackert, and G. G. Network. Long-term survival with glioblastoma multiforme. *Brain*, 130(Pt 10): 2596–2606, Oct 2007.

D. Krubasik, N. G. Iyer, W. R. English, A. A. Ahmed, M. Vias, C. Roskelley, J. D. Brenton, C. Caldas, and G. Murphy. Absence of p300 induces cellular phenotypic changes characteristic of epithelial to mesenchyme transition. *Br J Cancer*, 94(9): 1326–1332, May 2006.

S. J. Lavictoire, D. A. E. Parolin, A. C. Klimowicz, J. F. Kelly, and I. A. J. Lorimer. Interaction of Hsp90 with the nascent form of the mutant epidermal growth factor receptor EGFRvIII. *J Biol Chem*, 278(7):5292–5299, Feb 2003.

E. Lee and H. J. Bussemaker. Identifying the genetic determinants of transcription factor activity. *Mol Syst Biol*, 6:412, Sep 2010.

E. R. Levin. Integration of the extranuclear and nuclear actions of estrogen. *Mol Endocrinol*, 19(8):1951–1959, Aug 2005.

J. S. Lewis-Wambi and V. C. Jordan. Estrogen regulation of apoptosis: how can one hormone stimulate and inhibit? *Breast Cancer Res*, 11(3):206, 2009.

M. Liu, Y. Yang, C. Wang, L. Sun, C. Mei, W. Yao, Y. Liu, Y. Shi, S. Qiu, J. Fan, X. Cai, and X. Zha. The effect of epidermal growth factor receptor variant III on glioma cell migration by stimulating ERK phosphorylation through the focal adhesion kinase signaling pathway. *Arch Biochem Biophys*, 502(2):89–95, Oct 2010.

I. Ljubi, R. Weiskircher, U. Pferschy, G. W. Klau, P. Mutzel, and M. Fischetti. An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem. *Mathematical Programming*, 105:427–449, 2006. ISSN 0025-5610. 10.1007/s10107-005-0660-x.

B. Luo, H. W. Cheung, A. Subramanian, T. Sharifnia, M. Okamoto, X. Yang, G. Hinkle, J. S. Boehm, R. Beroukhim, B. A. Weir, C. Mermel, D. A. Barbie, T. Awad, X. Zhou, T. Nguyen, B. Piqani, C. Li, T. R. Golub, M. Meyerson, N. Hacohen, W. C. Hahn, E. S. Lander, D. M. Sabatini, and D. E. Root. Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci U S A*, 105(51): 20380–20385, Dec 2008.

V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, Jan 2006.

B. P. McKinley, A. M. Michalek, R. A. Fenstermaker, and R. J. Plunkett. The impact of age and sex on the incidence of glial tumors in New York state from 1976 to 1995. *J Neurosurg*, 93(6):932–939, Dec 2000.

Y. Narita, M. Nagane, K. Mishima, H.-J. S. Huang, F. B. Furnari, and W. K. Cavenee. Mutant epidermal growth factor receptor signaling down-regulates p27 through activation of the phosphatidylinositol 3-kinase/Akt pathway in glioblastomas. *Cancer Res*, 62(22):6764–6769, Nov 2002.

L. Neckers. Hsp90 inhibitors as novel cancer chemotherapeutic agents. *Trends Mol Med*, 8(4 Suppl):S55–S61, 2002.

L. Neckers. Heat shock protein 90: the cancer chaperone. *J Biosci*, 32(3):517–530, Apr 2007.

R. Nishikawa, X. D. Ji, R. C. Harmon, C. S. Lazar, G. N. Gill, W. K. Cavenee, and H. J. Huang. A mutant epidermal growth factor receptor common in human glioma confers enhanced tumorigenicity. *Proc Natl Acad Sci U S A*, 91(16):7727–7731, Aug 1994.

C. Pea, J. M. Garca, V. Garca, J. Silva, G. Domnguez, R. Rodrguez, C. Maximiano, A. G. de Herreros, A. Muoz, and F. Bonilla. The expression levels of the transcriptional regulators p300 and CtBP modulate the correlations between SNAIL, ZEB1, E-cadherin and vitamin D receptor in human colon carcinomas. *Int J Cancer*, 119 (9):2098–2104, Nov 2006.

I. F. Pollack, M. S. Randall, M. P. Kristofik, R. H. Kelly, R. G. Selker, and F. T. Vertosick. Effect of tamoxifen on DNA synthesis and proliferation of human malignant glioma lines in vitro. *Cancer Res*, 50(22):7134–7138, Nov 1990.

N. Sanai, A. Alvarez-Buylla, and M. S. Berger. Neural stem cells and the origin of gliomas. *N Engl J Med*, 353(8):811–822, Aug 2005.

C. M.-E. Sauvageot, J. L. Weatherbee, S. Kesari, S. E. Winters, J. Barnes, J. Dellagatta, N. R. Ramakrishna, C. D. Stiles, A. L.-J. Kung, M. W. Kieran, and P. Y. C. Wen. Efficacy of the HSP90 inhibitor 17-AAG in human glioma cell lines and tumorigenic glioma stem cells. *Neuro Oncol*, 11(2):109–121, Apr 2009.

J. A. Schwartzbaum, J. L. Fisher, K. D. Aldape, and M. Wrensch. Epidemiology and molecular pathology of glioma. *Nat Clin Pract Neurol*, 2(9):494–503; quiz 1 p following 516, Sep 2006.

N. Shinojima, M. Kochi, J. ichiro Hamada, H. Nakamura, S. Yano, K. Makino, H. Tsuiki, K. Tada, J. ichi Kuratsu, Y. Ishimaru, and Y. Ushio. The influence of sex and the presence of giant cells on postoperative long-term survival in adult patients with supratentorial glioblastoma multiforme. *J Neurosurg*, 101(2):219–226, Aug 2004.

G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

G. K. Smyth. *Limma: linear models for microarray data.*, pages 397–420. Springer, New York, 2005.

D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res*, 39(Database issue):D561–D568, Jan 2011.

Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99(468):909–917, 2004.

W. Zhang, W. T. Couldwell, H. Song, T. Takano, J. H. Lin, and M. Nedergaard. Tamoxifen-induced enhancement of calcium signaling in glioma and MCF-7 breast cancer cells. *Cancer Res*, 60(19):5395–5400, Oct 2000.

Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9):R137, 2008.

H. Zou and T. Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

Chapter 4

Conclusions

The specific goal of this project is to bring together global datasets of signaling and transcription to better understand how these processes are altered in response to oncogenic mutations and devise therapeutic strategies accordingly. Studies of signaling and transcription in cancer, often done independently, have created a rich body of knowledge that are changing clinical practices, as evidenced by the numerous clinical trials of kinase inhibitors and many transcriptional signatures for cancer diagnosis and treatment. However, many big questions remain. First, since these regulatory events are context specific, how can one generalize observations in one systems to others, and conversely, how can observations from other systems inform discovery of critical regulatory components in the current system and prioritize experimental validation? Secondly, genomic mutations are of paramount importance in oncogenesis and may manifest functionally in aberrant signaling and transcription, but the mechanistic connections between these molecular events are often unknown. Many recent high throughput technologies can address these questions to a certain degree, but it is challenging to interpret the results beyond following up on the top hits. I believe that the work in this thesis demonstrated a direction for approaching these problems.

The concept of constraint optimization is fundamental to our computational approach. Starting from a context-free but weighted protein-protein and protein-DNA interaction network, we map the experimental measurements from phosphoproteomic and mRNA expression data to the nodes in the network, and use these as constraints to find a set of interactions that connect these hits together. To account for the biases in the experimental techniques we allow nodes that are not detected in the experiments to be included, if doing so helps to include other hits. To account for the noise and incompleteness of the experimental data, we allow observed hits to be excluded. Finding a network that satisfies multiple sets of constraints, imposed by different types of experimental data for different aspects of the cellular processes, increases our confidence in the network solution, and ensures the information in these data are more fully utilized.

We first applied our method to yeast, a model organism with good coverage of protein-protein and protein-DNA interaction data, using datasets from pheromone response, a well-characterized signaling pathway. We showed that the method could recover components in the core pheromone response pathway and additionally many biological processes at the global level. In contrast to algorithms designed for inferring signaling pathways from mRNA transcript data alone, we found subnetworks that were coherent in function but not at the expression level. This is consistent with recent observations from comparing the phosphoproteome and transcriptome of human cell cycle (Olsen et al., 2010), where the steady-state transcript abundance do not correlate well with protein abundance except in some biological processes.

Taking our methods to analyze mammalian datasets required some experimental and computational modifications. Most notably, in order to capture the complexity of mammalian transcription regulation, we introduced the DNase-seq technique that identifies open-chromatin regions genome-wide, and used DNase-seq data with DNA sequence motif and differential expression in an expression regression procedure to select potential transcriptional regulators. When applied to datasets from the U87 EGFRvIII over-expressing cell lines, we obtained a network that gave a high level view of the signaling and transcriptional events downstream of this mutant receptor, and more importantly, it served as a basis for generating testable, biologically relevant hypotheses that might contribute to our understanding of cancer.

Bibliography

J. V. Olsen, M. Vermeulen, A. Santamaria, C. Kumar, M. L. Miller, L. J. Jensen, F. Gnad, J. Cox, T. S. Jensen, E. A. Nigg, S. Brunak, and M. Mann. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal*, 3(104):ra3, 2010.

Chapter 5

Future perspectives

In this thesis I set a foundation for linking global datasets of signaling and transcription and demonstrated it could uncover critical regulatory components in cellular signaling networks. In the course of this study several issues have become apparent that can be attributed to limitations of the computational and experimental techniques. Here I will detail these challenges and attempt to propose some directions in which we can address these problems. In the last section I envision how this framework can go beyond its current capability and be useful for comprehensive modeling of global dynamics.

5.1 Further development of the network optimization approach

In order to focus on making the many datasets fit together, we started with the conventional formulation of PCST to take advantage of the large body of literature for solving this problem, for instance, using local search heuristics (Canuto et al., 2001) and linear programming (Archer et al., 2011; Goemans et al., 1992). Algorithms for solving this formulation were motivated by applications in building networks of telecommunication and heat distribution. In contrast, biological networks have some properties that are not captured by this formulation. First, the edges in the input interaction graph are treated as undirected, while many molecular interactions, such as enzyme substrate reactions, have clear directionality. This may not be a significant issue at present since there are very few interactions with known directions in the current interactome. However, it must be properly handled as we improve our ability to obtain enzyme substrate relationships experimentally and computationally. Secondly, the optimization objective function with binary decision of including edges dictates that the optimal solution is a tree structure, i.e. if the solution contains any cycle, removing any edge in the cycle will reduce the objective function value but still satisfy the connectivity constraints. As a consequence of the tree structure, between any two nodes in the network there is only one path. This property is unrealistic in the context of biological networks where parallel pathways are common. I will now describe two approaches that are currently being taken in the Fraenkel lab to solve this issue.

5.1.1 Multi-commodity flow formulation

In collaboration with Bernhard Haeupler in David Karger's group at the Computer Science and Artificial Intelligence Laboratory of MIT, we devised a linear program formulation of the original PCST that models directed interactions and fractional edge selection.

We continue to use the notation from Section 3.3.1, where the interaction graph is a tuple $G = (V, E)$ with node penalties $\pi_i \geq 0$ and edge costs $c_e \geq 0$, except that the edges in the graph are now directed. We further accommodate a set of nodes S as *sources* from which the observed regulatory events originate (for instance, the cell surface receptor being stimulated), define a variable s_t for each node t with positive penalty value (*terminal set* T) to indicate whether it is on the path of directed interactions from a source, and a variable x_e for each directed edge $e \in E$ to indicate whether it is included in the solution. Then for each pair of edge $e \in E$ and terminal $t \in T$ we use a variable $s_{t,e}$ to represent whether the edge e is used to connect t to a source. We now have this optimization problem

$$\min \sum_{t\in T}(1-s_t)\pi_t + \sum_{e\in E} x_e c_e \tag{5.1.1}$$

$$\text{subject to} \sum_{e=(\cdot,w)} s_{t,e} - \sum_{e=(w,\cdot)} s_{t,e} = 0, \qquad \forall t \in T, w \in V\backslash(S \cup \{t\}), \tag{5.1.2}$$

$$\sum_{e=(\cdot,t)} s_{t,e} - \sum_{e=(t,\cdot)} s_{t,e} = s_t \leq 1, \qquad \forall t \in T, \tag{5.1.3}$$

$$0 \leq s_{t,e} \leq x_e \leq 1, \qquad \forall t \in T, e \in E. \tag{5.1.4}$$

A naive implementation of this formulation as an AMPL model invoking the CPLEX linear programming solver was able to find a solution to the yeast pheromone response dataset (Chapter 2) in two hours. But with the U87MG EGFRvIII dataset on the human interactome, the problem was unsolved after using more than 60GB of memory for two weeks. Since AMPL is an additional layer of communication to the CPLEX solver, I also implemented the linear program with direct functional calls to CPLEX using the Python API and MATLAB API. The pre-processing steps were more efficient in MATLAB where the interaction graph is represented as adjacency matrix. However, in both cases the demand for memory was very high and the problem remained unsolved after a week on our computing cluster. We note that this problem bears resemblance to the multi-commodity flow (MCF) problem if we view the variable $s_{t,e}$ as the flow of commodity t on edge e. In one common variant of the MCF problem that can be solved very efficiently for large scale networks, instead of constraining $s_{t,e}$ on edge e for each t with respect to another variable x_e, the total flow on edge e is upper-bounded by a constant u_e:

$$0 \leq \sum_{t\in T} s_{t,e} \leq u_e, \qquad \forall t \in T, e \in E. \tag{5.1.5}$$

In some preliminary testing, simply adopting the summation but not the constant upper bound makes it practical to solve this problem on human datasets:

$$0 \leq \sum_{t \in T} s_{t,e} \leq x_e \leq 1, \qquad \forall t \in T, e \in E. \tag{5.1.6}$$

One may argue that this formulation is in fact biologically realistic as it limits the extent to which a particular interaction is used by different pathways in total. It is therefore a sound and practical modification to the original formulation. However, the quality of the solution network that it generates has yet to be scrutinized.

5.1.2 Message passing approach

As an alternative to solving the PCST as a linear programming problem, statistical physic analysis of the properties of Steiner trees (Bayati et al., 2008) has resulted in new optimization algorithms based on message-passing techniques. Recently Riccardo Zecchina's group in Politecnico di Torino of Italy and Microsoft Research New England published a belief-propagation method that outperforms linear programming based algorithms computationally (including the one we used). They applied it to analyze gene expression datasets from yeast pheromone response and identified a previously unknown regulator (Bailly-Bechet et al., 2011). We are now in collaboration with this group to explore the potential of this algorithm. Since this algorithm is able to solve the optimization in a fraction of the time of our current method, we will be able to conduct more rigorous study of the quality of the network solutions, for instance, assessing the statistical significance and stability of the solutions by randomization.

5.1.3 Condition specific interactome

The central premise behind our constraint optimization framework is that the experimental measurements at the signaling and transcription level are sufficient for guiding selection of relevant interactions from the interactome of many contexts. In the absence of methods that can generate condition specific interactome for numerous experimental conditions, there are a few strategies to ensure the selected interactions are indeed possible in that condition. First, as a pre-processing step the input interaction network can be filtered by the expression level of the nodes measured by transcript or protein assays under that condition. With the improved sensitivity of RNA-seq to detect rare transcripts compared to microarrays, this step may now be done with confidence. However, if it is not desirable to set a threshold on what is considered expressed, we can add to the PCST formulation capacities on the nodes that represent the expression level. There are well-established procedures that transform node capacitated network flow problems to one without the node capacities (Ahuja et al., 1995).

5.1.4 Analysis of time series and multiple conditions data

The PCST analysis in Chapter 3 focused on comparing steady state measurements from the U87H cells and the U87DK cells. Although this comparison has produced fruitful results, a large amount of phosphoproteomic and gene expression data being generated nowadays follow the time dynamics after certain experimental intervention. As noted previously, intervention data, compared to observational data, is better for revealing connections of regulatory networks (Hyduke and Palsson, 2010), so adopting the PCST framework to time series data may increase the confidence in selecting relevant interactions.

5.2 Improving the input datasets

We postulate that the heterogeneity of our input datasets can provide evidence from multiple angles that enhance the prediction of important regulatory components. Therefore, strengthening the accuracy and scope of each experimental dataset will allow us to better differentiate between signal and noise, especially for prioritizing among multiple weak hypotheses.

5.2.1 DNaseI hypersensitivity footprinting

The DNase-seq dataset for the U87 cells for this thesis was generated using the sucrose gradient purification protocol (Sabo et al., 2006, 2004) and from it we were able to identify candidate regulatory regions and correlate with differential gene expression. A different protocol (Song and Crawford, 2010; Boyle et al., 2008; Crawford et al., 2006) uses linker ligation and restriction enzyme digestion to isolate the DNaseI hypersensitive DNA fragments. When combined with very deep sequencing, a computational method that searches for protein binding "footprints" was able to predict transcription factor binding genome-wide with high accuracy and also good spatial resolution (Pique-Regi et al., 2011). So far we were unable to observe these footprints in our DNase-seq data, so using this alternate DNase-seq protocol may give better prediction of functional transcriptional regulators for inclusion in the PCST solution.

5.2.2 Kinase substrate relationships

The interactions containing proteins with phosphorylation sites are the starting point from which the PCST network solution is built, so it is critical to have a good representation of these interactions in the interactome graph. Functionally the phosphorylation sites participate in interactions with other proteins in two ways: as target substrates of other kinases and phosphatases, and as binding partners of proteins that recognize the phosphorylated residues. Many of these interactions are transient and are difficult to capture in some interaction assays. There is recent evidence that yeast two hybrid technology could detect transient signaling interactions (Yu et al., 2008) and is probably the most sensitive method for screening transient kinase-substrate interactions (Sopko and Andrews, 2008), but it is unclear how to distinguish between kinase binding partners and target substrate in this assay (Manning and Cantley, 2002). Many *in vivo* methods are available to link kinases to phosphorylation substrates (reviewed in Sopko and Andrews, 2008) but only for specific kinases. Taking these efforts to the global level, combined with other information such as sequence motif and integrated within a computational framework such as NetworKIN (Linding et al., 2007), will provide interaction datasets that greatly enhance the ability of our algorithm to connect the phosphorylated proteins.

5.2.3 Transcription factor motif

We used the transcription factor binding profiles curated in the TRANSFAC database, which has a large collection but is known to contain redundant motifs of different qualities. In comparison, an alternative database, JASPAR, is a repository of high quality, non-redundant motifs but has a smaller selection. In either case we are far from getting a comprehensive catalog of binding specificities of close to 2000 transcriptional factors in human (Messina et al., 2004; Babu et al., 2004). Recently protein binding microarray (PBM) has emerged as a possible technique to rapidly generate such a catalog (Berger and Bulyk, 2009; Badis et al., 2009). Since *in vitro* binding specificities determined in these assays may deviate from the *in vivo* specificities, motif discovery methods seeded with the *in vitro* profiles (Li, 2009; Macisaac et al., 2006) on ChIP-seq datasets will help to recover the true *in vivo* binding motifs.

5.2.4 Significance of phosphorylation events

Our current analysis defines node penalties on the phosphorylated proteins in a practical but *ad hoc* manner: the penalty values are proportional to the absolute value of log fold changes of phosphorylation; if there are more than one phosphorylation sites on one protein, the maximum value is used. This reflects the assumption that larger changes in phosphorylation carry higher importance and thus should be given higher priority to be included. There are other, probably more principled, ways of quantifying the significance of the phosphorylation changes. But we distinguish two kinds of significance: statistical significance and biological significance. The former requires the development of robust error models (Zhang et al., 2010) while the latter would benefit from knowledge about the context of the phosphorylation sites, such as the structural domain or binding sequence motif where the sites are located (see examples in Naegle et al., 2010) . We note that these two should not be treated separately as biologically meaningful hits may still arise from contradictory statistical analysis (Uher and McGuffin, 2010; Risch et al., 2009; Munaf et al., 2009).

5.3 Towards modeling the dynamics of phosphorylation and transcription

Analysis of the phosphoproteomic and transcriptional data by our constraint optimization framework has helped us gain insight into the molecular biology of cancer, but we recognize that the potentials of these datasets are still not fully utilized, in particular, the quantitative nature of the measurements and the information about feedback regulation encoded in the quantification. Motivated by simplicity and the observed lack of correlation between changes in phosphorylation and mRNA transcript, in our network the proteins that contain phosphorylation events and mRNA transcripts are treated as separate entities. This models the transcriptional changes as a consequence of signaling changes but ignores any subsequent feedback where changes in transcription affect the protein level of the same gene. A natural way to describe such feedback mathematically is by differential equations, which can be simulated numerically or analyzed. This approach has been applied genome-wide in a comprehensive transcriptional and translational network for *Escherichia coli* (Thiele et al., 2009). Although developing such a model for mammalian networks seems daunting, techniques such as flux balance analysis, as demonstrated in the *E. coli* example, hold the promise to predict network behavior even without detailed knowledge of kinetic parameters. In another example, Muzzey et al. (2009) applied control theory to model and design experiments to study the yeast osmoregulation system, for which exhaustive models are available (Klipp et al., 2005). With only a small set of experiments they found the network location responsible for perfect adaptation. The approach presented in this thesis, which quickly generates a set of candidate backbone networks from genome-wide datasets, and in conjunction with the principles and experiences from engineering complex systems, can achieve the ambitious but realistic goal of finding key mechanisms that regulate biologically relevant dynamics at the systems level.

Bibliography

R. Ahuja, T. Magnanti, J. Orlin, and K. Weihe. *Network flows: theory, algorithms and applications.* Wurzburg, Physica-Verlag, 1972-1995., 1995.

A. Archer, M. Bateni, M. Hajiaghayi, and H. Karloff. Improved Approximation Algorithms for Prize-Collecting Steiner Tree and TSP. *SIAM Journal on Computing*, 40(2):309–332, 2011.

M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, 14(3):283–291, Jun 2004.

G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935): 1720–1723, Jun 2009.

M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. Franois, and R. Zecchina. Finding undetected protein associations in cell signaling by belief propagation. *Proc Natl Acad Sci U S A*, 108(2):882–887, Jan 2011.

M. Bayati, C. Borgs, A. Braunstein, J. Chayes, A. Ramezanpour, and R. Zecchina. Statistical mechanics of steiner trees. *Phys Rev Lett*, 101(3):037208, Jul 2008.

M. F. Berger and M. L. Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc*, 4(3):393–411, 2009.

A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, Jan 2008.

S. A. Canuto, M. G. C. Resende, and C. C. Ribeiro. Local search with perturbations for the prize-collecting Steiner tree problem in graphs. *Networks*, 38:2001, 2001.

G. E. Crawford, S. Davis, P. C. Scacheri, G. Renaud, M. J. Halawi, M. R. Erdos, R. Green, P. S. Meltzer, T. G. Wolfsberg, and F. S. Collins. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods*, 3(7):503–509, Jul 2006.

M. Goemans, David, and P. Williamson. A General Approximation Technique For Constrained Forest Problems. *SIAM Journal on Computing*, 24:296–317, 1992.

D. R. Hyduke and B. . Palsson. Towards genome-scale signalling-network reconstructions. *Nat Rev Genet*, 11(4):297–307, Feb 2010.

E. Klipp, B. Nordlander, R. Krger, P. Gennemark, and S. Hohmann. Integrative model of the response of yeast to osmotic shock. *Nat Biotechnol*, 23(8):975–982, Aug 2005.

L. Li. GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *J Comput Biol*, 16(2):317–329, Feb 2009.

R. Linding, L. J. Jensen, G. J. Ostheimer, M. A. T. M. van Vugt, C. Jrgensen, I. M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J. G. Park, L. D. Samson, J. R. Woodgett, R. B. Russell, P. Bork, M. B. Yaffe, and T. Pawson. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–1426, Jun 2007.

K. D. Macisaac, D. B. Gordon, L. Nekludova, D. T. Odom, J. Schreiber, D. K. Gifford, R. A. Young, and E. Fraenkel. A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics*, 22(4):423–429, Feb 2006.

B. D. Manning and L. C. Cantley. Hitting the target: emerging technologies in the search for kinase substrates. *Sci STKE*, 2002(162):pe49, Dec 2002.

D. N. Messina, J. Glasscock, W. Gish, and M. Lovett. An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res*, 14(10B):2041–2047, Oct 2004.

M. R. Munaf, C. Durrant, G. Lewis, and J. Flint. Gene X environment interactions at the serotonin transporter locus. *Biol Psychiatry*, 65(3):211–219, Feb 2009.

D. Muzzey, C. A. Gmez-Uribe, J. T. Mettetal, and A. van Oudenaarden. A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell*, 138(1):160–171, Jul 2009.

K. M. Naegle, M. Gymrek, B. A. Joughin, J. P. Wagner, R. E. Welsch, M. B. Yaffe, D. A. Lauffenburger, and F. M. White. PTMScout, a Web resource for analysis of high throughput post-translational proteomics studies. *Mol Cell Proteomics*, 9 (11):2558–2570, Nov 2010.

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

N. Risch, R. Herrell, T. Lehner, K.-Y. Liang, L. Eaves, J. Hoh, A. Griem, M. Kovacs, J. Ott, and K. R. Merikangas. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis. *JAMA*, 301(23):2462–2471, Jun 2009.

P. J. Sabo, M. Hawrylycz, J. C. Wallace, R. Humbert, M. Yu, A. Shafer, J. Kawamoto, R. Hall, J. Mack, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci U S A*, 101(48):16837–16842, Nov 2004.

P. J. Sabo, M. S. Kuehn, R. Thurman, B. E. Johnson, E. M. Johnson, H. Cao, M. Yu, E. Rosenzweig, J. Goldy, A. Haydock, M. Weaver, A. Shafer, K. Lee, F. Neri, R. Humbert, M. A. Singer, T. A. Richmond, M. O. Dorschner, M. McArthur, M. Hawrylycz, R. D. Green, P. A. Navas, W. S. Noble, and J. A. Stamatoyannopoulos. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods*, 3(7):511–518, Jul 2006.

L. Song and G. E. Crawford. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, 2010(2):pdb.prot5384, Feb 2010.

R. Sopko and B. J. Andrews. Linking the kinome and phosphorylome–a comprehensive review of approaches to find kinase targets. *Mol Biosyst*, 4(9):920–933, Sep 2008.

I. Thiele, N. Jamshidi, R. M. T. Fleming, and B. . Palsson. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput Biol*, 5(3):e1000312, Mar 2009.

R. Uher and P. McGuffin. The moderation by the serotonin transporter gene of environmental adversity in the etiology of depression: 2009 update. *Mol Psychiatry*, 15(1):18–22, Jan 2010.

H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabsi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, Oct 2008.

Y. Zhang, M. Askenazi, J. Jiang, C. J. Luckey, J. D. Griffin, and J. A. Marto. A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia. *Mol Cell Proteomics*, 9(5): 780–790, May 2010.