Published in final edited form as:
Methods Cell Biol. 2012 ; 110: . doi:10.1016/B978-0-12-388403-9.00003-5.

Swimming upstream: identifying proteomic signals that drive transcriptional changes using the interactome and multiple "-omics" datasets

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

[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Abstract

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. As many high-throughput, "-omics" methods are now available that can simultaneously measure changes in hundreds of proteins and thousands of transcripts, it should be possible to systematically reconstruct cellular responses to perturbations in order to discover previously unrecognized signaling pathways.

This chapter describes a computational method for discovering such pathways that aims to compensate for the varying levels of noise present in these diverse data sources. Based on the concept of constraint optimization on networks, the method seeks to achieve two conflicting aims: (1) to link together many of the signaling proteins and differential expressed transcripts identified in the experiments ("constraints") using previously reported protein-protein and protein-DNA interactions, while (2) keeping the resulting network small and ensuring it is composed of the highest confidence interactions ("optimization"). A further distinctive feature of this approach is the use of transcriptional data as evidence of upstream signaling events that drive changes in gene expression, rather than as proxies for downstream changes in the levels of the encoded proteins.

We recently demonstrated that by applying this method to phosphoproteomic and transcriptional data from the pheromone response in yeast, we were able to recover functionally coherent pathways and to reveal many components of the cellular response that are not readily apparent in the original data. Here we provide a more detailed description of the method, explore the robustness of the solution to the noise level of input data and discuss the effect of parameter values.

I. Introduction

One of the central challenges for systems biology is the reconstruction of cellular processes from high-throughput experimental data. Much of the early work in this area was driven by the development of microarray technologies that allowed relatively comprehensive measurement of changes in mRNA expression. Using these data as proxies for changes at the protein level has generated many insights into the regulatory networks of the cell

*Corresponding author. fraenkel-admin@mit.edu.

(Spellman et al. 1998; Segal et al. 2005; Ozsolak et al. 2011). However, the actual correlation between the transcriptome and the proteome is unclear (Schwanhässer et al. 2011; Maier et al. 2009; de Sousa Abreu et al. 2009), and more direct proteomic data are likely to provide a more reliable and thorough view of cellular processes.

Recently, technological advances have made it possible to directly measure proteomic changes at the global level. Mass-spectrometry (MS) techniques can quantify the relative levels of hundreds of peptides across multiple biological conditions (Choudhary et al. 2010; White 2008) and focused data collection on phosphoproteins was able to reveal the regulatory dynamics of cellular signaling networks at the level of the proteome (Grimsrud et al. 2010; Macek et al. 2009; Yi Zhang et al. 2007).

With new data come new challenges. Even in the best characterized responses there is poor overlap between hits identified by phosphoproteomics technologies and known pathway components. For example, in a study of phosphorylation changes that occur in response to mating pheromone in yeast (Gruhler et al. 2005), 112 proteins contain differentially phosphorylated sites; of these, only 11 are known components of the expected mitogen-activated protein kinase (MAPK) cascade that responds to pheromone, and 76 were not present in any of the yeast pathways annotated in the KEGG PATHWAY database (Kanehisa et al. 2010). Finding new ways to interpret these data could reveal previously unrecognized cellular pathways.

A second important challenge is to integrate transcriptional and proteomic data in order to observe the interplay between different layers of cellular signaling. For example, it may be possible to detect proteomic changes in signal transduction cascades that drive expression and also to reveal the resulting feedback of transcription on the proteome. But integrating these data will require novel computational approaches. Because regulation is mediated by diverse mechanisms, even the most comprehensive proteomics technologies cannot capture all these events. For example, MS based methods focusing on protein phosphorylation will fail to detect changes in other post-translational modifications such as acetylation, ubiquitination and sumoylation. Computational techniques are needed to discover proteins that participate in the signaling networks but are undetected in the experiments and also to provide insight into their functional roles. One successful approach has been to map these proteins onto known metabolic and regulatory pathways such as those curated in the KEGG PATHWAY (Kanehisa et al. 2010) and Reactome (Matthews et al. 2009) databases. This approach can reveal functional coherence and relevant biological processes from the data. However, as mentioned above, a large fraction of the phosphoproteomic data do not map to known pathway models, so we must turn to other approaches.

The interactome provides an alternative to using well-studied pathways. Advances in high-throughput experimental mapping of protein-protein interactions as well as efforts to extract known interactions from the literature have produced a number of large databases of protein interactions (selected examples are listed in Table I). Despite being incomplete, especially for higher organisms, the amount of interaction data in these databases is still very large. Thus, it may be possible to discover unknown pathways among these interactions. While utilizing these large interactome datasets improves our ability to find connections among a set of proteins of interest, it also presents several challenges. First, the size of the potential 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 hundreds of laboratories and many experimental techniques of varying degrees of reliability (von Mering et al. 2002), so overall the data quality is heterogeneous and should not be treated indiscriminately. Lastly, pooling these interactions together risks losing the specific context under which they were detected. It is with these issues in mind

that we propose a constraint optimization approach for finding regulatory networks that are interpretable, reliable and biologically relevant.

Our method starts with a collection of protein-protein and protein-DNA interactions, which represent known or experimentally determined signaling and regulatory connections. It considers 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 graph. 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. These two features of the method strike a balance between finding novel connections and revealing the relevance of known connections. We hypothesize that since each of our input data sources 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 serves to organize these heterogeneous datasets and enhance our understanding of the cell at the systems level.

II. Computational methods

Network optimization is an area of computer science that has recently become very useful for analyzing biological problems, and a variety of algorithms are available to solve specific optimizations. The problem we have posed consists of finding a set of edges of minimum weight in order to connect a defined set of nodes (known as termini) in a weighted network. This problem is called the Steiner tree problem. An important generalization that allows some terminal nodes to be excluded is known as the prize-collecting Steiner tree (PCST) problem. For our purpose, we will use a network in which edge weights reflect our confidence in the interactions and where terminal nodes represent hits from the experiments, i.e., phosphorylated proteins and differentially expressed transcripts. In this setting, the solution to the PCST optimization is a set of most confident interactions that link together the hits while possibly leaving some unconnected [Figure 1(A)].

Although the concept of Steiner tree has been previously applied to biological networks (Dittrich et al. 2008; Scott et al. 2005), we note that our approach is distinctive in multiple aspects. First, instead of mRNA transcript abundance we use protein level measurements on nodes in the interactome, which provides a much more accurate representation of the underlying biological processes. Second, we explicitly model the confidence of individual edges in the interactome to account for the uncertainties in the interaction data. Third, we do not require all nodes in the solution to be detected in the experiments, allowing our approach to compensate for multiple sources of noise. This last feature is absent in an application of a Steiner tree like algorithm to build a high confidence network with genetic screening hits as terminal nodes (Yosef et al. 2009). A minimum-cost flow optimization approach connects genetic hits to differentially expressed genes (Lan et al. 2011; Yeger-Lotem et al. 2009) but the result is less compact than the PCST (Huang et al. 2009). We now describe the process of constructing the optimization problem, solving it and analyzing the results. We also offer some advice on practical matters such as tuning the parameter values and visualizing the network.

A. Setting up the prize-collecting Steiner tree

We treat the interactome as an undirected graph $G = (V, E)$ where nodes are proteins or genes and edges represent the known interactions. Each node $v \in V$ is associated with a penalty $\pi_v \geq 0$. Protein nodes to which experimental data are mapped receive positive penalty values and therefore are termini for the PCST. All other nodes receive zero penalties. As the magnitude of the penalty value increases the more confident we are that the protein/gene was experimentally detected as relevant in the signaling response. The algorithm is forced to pay a penalty each time it leaves a terminal out of its final network. This constraint causes the network to include as many high-confidence nodes as possible. However, this constraint alone would lead to very large networks that might contain many unreliable edges. So we also assign to each edge $e \in E$ a cost $c_e \geq 0$ that is inversely related to our confidence in each interaction.

We aim to find a subtree $F = (V_F, E_F)$ of G that minimizes the objective function

$$\sum_{e \in E_F} c_e + \sum_{v \notin V_F} \pi_v.$$

Because we incur penalties for excluding nodes while paying costs for including edges, the algorithm will be forced to favor connecting high-confidence data with high-confidence interactions. We further introduce a scaling parameter β to balance the penalties and the edge costs:

$$\sum_{e \in E_F} c_e + \sum_{v \notin V_F} \beta \pi_v.$$

We may solve this optimization problem exactly by using the branch-and-cut approach (Ljubić et al. 2005) implemented in the `dhea-code` software program that calls the ILOG CPLEX mathematical programming solver. As an alternative to solving it as an integer linear program, an approach from statistical physics (Bayati et al. 2008) has resulted in new heuristic algorithms based on message-passing techniques (Bailly-Bechet et al. 2011). We now describe how the experimental data are transformed into input for the algorithm. An overview of the work flow is in Figure 1(B).

B. A probabilistic interactome

This is Step 1 in Figure 1(B). The set of edges E of the input graph G consists of direct (physical) protein-protein interactions found in databases of molecular interactions such as those listed in Table I. To assign confidence values for these interactions, a few methods have been previously published (Razick et al. 2008; Orchard et al. 2007; Jansen et al. 2003). Here we use a naïe Bayes probabilistic model (Jansen et al. 2003). Interaction between two proteins is modeled as random variable $i \in \{0,1\}$ with $i = 1$ when two proteins interact and $i = 0$ otherwise, and each kind of experimental evidence is modeled as a random variable $f_j \in \{0,1\}$ where $f_j = 1$ indicates f_j is observed and $f_j = 0$ otherwise. From published gold standard sets of positive (Yu et al. 2008) and negative interactions (Jansen et al. 2003) we can compute the conditional probability table for each kind of evidence, $P(f_j|i)$. Then, for each interaction e supported by a set of experimental evidence $F_e = \{f_{e,j} | j = 1,\ldots, n\}$, assuming independence between the evidence we have

$$P(F_e|i)=\prod_j P(f_{e,j}|i),$$

and a straight forward application of Bayes rule gives the probability that this interaction is real:

$$P(i=1|F_e)=\frac{P(F_e|i=1)P(i=1)}{\sum_{i'\in\{0,1\}}P(F_e|i')}.$$

The cost c_e on edge e that is input into the PCST objective function is

$$c_e=-\log P(i=1|F_e), \forall e \in E.$$

C. Determining transcription factor targets

Transcription factor to mRNA target relationships are added to the protein-protein interactome to form the total interactome [Step 2 in Figure 1(B)]. A variety of experimental, computational techniques and combinations of both are possible. For yeast, there are published genome-wide binding sites for almost all the transcriptional regulators under multiple conditions measured by chromatin immunoprecipitation (ChIP) experiments (Harbison et al. 2004; MacIsaac, T. Wang, et al. 2006). The human and mouse ENCODE projects (Birney et al. 2007) represent systematic efforts to generate ChIP profiles for multiple transcription factors in a variety of human cell lines and mouse tissues. Computationally, transcription factors often have sequence specificities that allow binding sites to be predicted to some extent (Figure 6). 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 (Foat et al. 2006, 2005; Manke et al. 2008; Roider et al. 2007). Motifs from the TRANSFAC (Wingender 2008; Matys et al. 2006) and JASPAR (Sandelin et al. 2004; Bryne et al. 2008) databases, which collect published transcription factor binding motifs from the literature, can be used for predicting regulatory elements. Once a genomic region is determined to be bound by a transcription factor based on experimental and/or computational evidence, nearby genes can be associated with this factor as its potential downstream targets, and we add to the interactome edges going from the transcription factor (a protein node) to these target mRNA nodes.

D. Node penalties

This is Step 3 in Figure 1(B). We define two kinds of penalties for proteins in the interactome: one at the protein level derived from the phosphoproteomics MS data, and the other at the mRNA level derived from mRNA expression data.

Although published phosphoproteomic MS datasets often provide the identities of the proteins that contain the peptide sequences inferred from the MS spectra, it is still advisable to map the peptides to a database of protein sequences from which the interactome dataset is derived in order to avoid issues such as inconsistencies in mapping gene identifiers and in treating protein isoforms. This can be achieved by finding protein sequences in a database that contain matches to the peptide sequences, for example, by the sequence alignment search tool BLAST (Altschul et al. 1990) with parameter settings optimized for matching

short peptide sequences. In an analysis comparing two conditions, proteins that contained perfect alignment to a peptide sequence receive a positive penalty value that is proportional to the absolute value of log fold change in phosphorylation between the conditions of interest. If one peptide sequence is aligned to multiple proteins in the interaction graph, all these proteins receive the same penalty value. If multiple phosphorylated peptide sequences are perfectly aligned to one protein, the maximum fold change in phosphorylation of these peptides is used to calculate the penalty value for this protein. Other methods of assigning penalties are also possible and are discussed below.

For penalty values on mRNA nodes, some modifications to the interactome are required to make the resulting network more biologically realistic. If we simply put penalty values on the mRNA nodes, the tree structure of the solution network means that any one mRNA node is connected to at most one upstream transcription factor. Such a network cannot capture one gene being targeted by multiple transcription factors, which is a common feature of transcriptional regulation. Instead, we represent multiple transcription factors bound to the same gene with separate nodes. Let M be the set of differentially expressed transcripts, and $fc(m)$ be the fold change in mRNA abundance of each gene $m \in M$. For each m, we searched the interactome for the set of upstream transcription factors F that target m, remove m from the interactome, and add one node m_f for each transcription factor $f \in F$ and one edge between f and m_f. The fold change of m is transferred to all the m_f to compute the penalty values on m_f. Each new terminal node m_f may be interpreted as a binding site of f on m.

E. Sensitivity analysis

Applying an optimization approach to inherently noisy biological data makes it necessary to explore the alternative or sub-optimal solution space surrounding the reported optimal solution. This is to ensure that the nodes and edges selected by the algorithm, from which significant efforts will be invested to extract biological meaning, are relatively stable to possible sources of noise. Figure 3 presents two ways to quantify this stability at the global level. First, starting from the optimal solution reported by the algorithm, we can re-formulate the optimization problem to find a number of sub-optimal solutions - networks that are optimal under the additional constraint that they must differ from the original optimal solution by a pre-defined percentage of nodes. We can then compare these sub-optimal solutions to the optimal one in terms of the objective function value [Figure 3(A)] and the frequency at which the nodes in the original optimal solution are preserved in the sub-optimal solutions [Figure 3(B)] in order to decide whether the solution is robust to noise.

F. Practical advice

Parameters: Tuning the value of parameter β essentially controls the size of the PCST solution output. With larger β values it becomes more expensive to exclude each terminal node (i.e., making the objective function larger), so the optimization algorithm will include more edges in the PCST solution. Although a larger network may include more hits from the experimental data, it is more difficult to interpret and also more likely to include false positive hits that may connect to the real underlying network via tenuous interactions. To find a suitable value of this parameter, it is advisable to run the algorithm with a range of values and choose a solution that (1) includes any expected pathways based on prior biological knowledge, (2) is stable for the neighborhood of β values, and (3) contains as many of the hits as possible. One can also start with a small value of β to build a core network and gradually increase β to explore how more hits are connected to the core network.

It may be possible to use cross-validation to objectively choose β. In such an approach, one would randomly partition the terminal nodes into two complementary subsets, build a PCST using one subset (training set) and compute the recovery of the second subset (validation set) in that PCST. To reduce the effect of random variations, for each value of β, multiple rounds of such cross-validation can be performed and one average performance value is reported. Based on this performance measure, a β value can be selected.

While this approach has a certain appeal, we urge caution since the assumptions and requirements of cross-validation may not be satisfied by the biological datasets. First, in order for the recovery of the validation set by a PCST to be a good indicator of its performance, the training set and the validation set must be drawn from the same distribution. This criterion requires the terminal sets to be sufficiently large that each random sample contains termini from all the underlying biological processes. Since the current datasets are subject to many limitations such as the sensitivity of the MS instrument depending on protein abundance and the coverage of the interactome, we do not know *a priori* whether this assumption is appropriate. Second, it is unclear which of the conventionally used measures of predictor performance is suitable in this setting. We aim to recover intermediate nodes that are undetected in experiments, so we cannot count such nodes included in the PCST as false positives. In the absence of a false positive definition, counting the recovery of the terminal nodes makes little sense since the optimal value of β will be the one that produces a PCST that include the most terminal nodes (weighted by penalty values).

Implementation: There are various approximation algorithms to solve the PCST problem. These have recently been reviewed (Archer et al. 2011). The `dhea-code` program (Ljubić et al. 2005), which can be downloaded from Dr. Ljubić's website (Ljubić 2008), uses a branch-and-cut approach to obtain exact, optimal solutions. This program requires the ILOG CPLEX (IBM) optimization library that is available at no-charge for teaching and non-commercial research as part of the IBM Academic Initiative (IBM 2010). In the supplement of this article we provide a simple Python script for creating the input file for `dhea-code` from tab delimited text files of the weighted interactome and terminal nodes. The output files of `dhea-code` include the PCST solution in a DOT file [a plain text format for specifying graphs; (Graphviz 2011b)]. From there the solution can be rendered and viewed by the tools in Graphviz (Graphviz 2011a), or further manipulated and analyzed by the Python library NetworkX (Hagberg et al. 2008). One standard operation is to convert the DOT file to one of the files formats supported by Cytoscape (Smoot et al. 2011; Cline et al. 2007) in order to utilize Cytoscape's many visualization capabilities for biological networks.

A recently published message passing algorithm, although taking a heuristic approach, is able to find solutions with objective values comparable to `dhea-code` under much less computing time and memory (Bailly-Bechet et al. 2011). It requires a depth parameter to be specified *a priori* to control the length of paths in the solution network. This appears to have the consequence of eliminating long braches in the solution. The effect of this difference on the identities and functional relevance of the recovered nodes remains to be investigated.

III. Biological insights

The PCST solution connects together the phosphorylation events and transcriptional changes using a compact set of interactions. 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. These are certainly of great interest for elucidating the role of individual

hits. Also interesting are the properties that emerge from the network at the systems level, and we will describe a few computational techniques for such analyses using the yeast pheromone response PCST solution as an example (Figure 2).

A. Properties of the full network

The PCST solution in Figure 2 was constructed from published phosphoproteomic (Gruhler et al. 2005) and transcription profiling (Roberts et al. 2000) datasets of the yeast *Saccharomyces cerevisiae* in response to the mating pheromone α-factor. This network was first reported in (Huang et al. 2009). The network 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 network shows a few notable features at the global level. First, the MAPK cascade known to be induced by pheromone (labeled “pheromone core” in Figure 2) is recovered by the algorithm. In particular, it correctly identifies the proteins GPA1, STE11 and BEM1, where no phosphorylation sites were detected, as well as their connections to other proteins in the pheromone signaling pathway. In addition, only proteins that are present in the pheromone response pathway are included. Secondly, beyond the MAPK cascade, the solution network partitions into highly coherent subnetworks with biological functions relevant to mating. At the transcription level, phosphorylated proteins seem highly informative in selecting interacting transcription factors. Examples include DIG1/DIG2/STE12 complex in the pheromone signaling pathway, SWI4/SWI6 and SWI6/MBP1 in the PKC pathway, and FKH2/NDD1 complex regulated by CDC28. These observations suggest the constraints imposed by the phosphorylated proteins and differentially expressed genes are sufficient to guide the selection of important players that contribute to the response.

To assess the functional significance of the intermediate nodes from the PCST solution in mating response, we examined two independent whole-genome deletion screen datasets that screen for genes whose deletion result in mating defects. One screen measures a molecular phenotype in the form of activation of FUS1-lacZ reporter (Chasse et al. 2006) and the other screen measures a morphological phenotype in the form of cell cycle arrest and shmoo formation (Narayanaswamy et al. 2006). For each screen we counted the number of hits that overlap with the intermediate nodes in the PCST solution, and using all the screening genes as background we computed a hypergeometric P-value for which such overlap would appear by chance. As seen in Figure 4, compared to networks constructed from shortest paths and first neighbors of the terminal nodes, the PCST solution is more compact while achieving higher enrichment of genes implicated in mating defects.

B. Biological functions of subnetwork/modules

To objectively quantify the empirical observation that the PCST solution is partitioned into functional coherent subnetworks, we applied the Girvan-Newman algorithm (Dunn et al. 2005; Girvan et al. 2002) to cluster the solution. This algorithm is used for detecting clusters in an interaction network that contain dense connections between nodes in the same cluster but less dense connections to nodes in other clusters. Gene Ontology enrichment analysis of the resulting clusters reveals that all the clusters have high degree of functional coherence (Table II). It is interesting to note that many of the clusters are not coordinately expressed at the mRNA level, as quantified by the significance of expression coherence score (Pilpel et al. 2001) or by the significance of expression activity score (Ideker et al. 2002). Notably, the clusters that show significant coordinated expression are involved in cell cycle processes.

Being able to recover functionally coherent clusters that are not coherent at the transcript level is a significant result. Transcriptional data, which are more readily available than

proteomics data, are the focus of many computational methods for regulatory network construction. Our results suggest that methods that rely solely on expression data, including a prior Steiner tree approach (Dittrich et al. 2008), will be unable to recover the full extent of a biological response.

C. Quantifying the relevance of the transcription factors

In addition to the transcription factors mentioned above that are known to be induced by pheromone or function in related biological processes, the PCST solution network features many other transcriptional regulators not previously implicated in pheromone response. We use expression coherence score as a metric to quantify the significance of these transcription factors at the global level. For each transcription factor with targets in the interactome, we obtained the expression values of those targets across a set of conditions that stimulate pheromone signaling, and computed the significance P-value of the expression coherence score. Then we set a threshold on the significance P-value, and compared the percentage of transcription factors included and excluded in the PCST that pass this threshold. As shown in Figure 5, the transcription factors included in the network are more likely to have a set of targets that are coherently expressed than the factors excluded from the network. To check if these transcription factors are condition specific, we did a similar calculation for the expression values from a set of conditions that are unrelated to pheromone: when yeast undergoes the metabolic shift from fermentation to respiration (diauxic shift). We found that coherence is specific to the conditions related to pheromone signaling but not to diauxic shift.

IV. Open challenges

A. Improving the input data

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. It is important to note, however, that many of these interactions may only occur under specific conditions that are not relevant to the problem being studied. It is not yet practical to collect condition-specific interaction data on a large scale. Nevertheless, there are a few strategies to ensure the selected interactions are indeed relevant. First, as a pre-processing step, the input interaction network can be filtered to remove nodes that are not believed to be expressed under the condition of interest, based on transcript or protein assays. With the improved sensitivity of RNA-seq to detect low abundance transcripts compared to microarrays, this step may now be done with higher confidence. However, expression data are still noisy, and removing nodes completely risks missing important components of a network. Alternatively, 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 ones without the node capacities (Ahuja et al. 1995).

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 multiple 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. We distinguish two kinds of significance: statistical significance and biological significance. The former requires the development of robust error models (Yi 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)]. But these two need not to be exclusive: once statistical significance is established, penalty values can be defined by analyzing for potential biological significance.

As phosphorylation sites are the starting point from which the PCST network solution is built, it is critical to have a good coverage of interactions involving these proteins in the interactome graph. Phosphorylation sites participate in interactions with other proteins in two ways: as substrates of kinase and phosphatases, and as binding partners of proteins that recognize the phosphorylated residues. Many of these interactions are transient and context specific and thus difficult to capture in some interaction assays. In particular, among the various high-throughput interactome mapping techniques, a modified affinity capture MS method is the most informative in identifying kinase targets, with yeast two-hybrid being second (Sharifpoor et al. 2011). Many *in vivo* methods are available to link kinases to phosphorylation substrates [reviewed in (Sopko et al. 2008)] but only for specific kinases. Taking these efforts to the global level, and using other information such as sequence motifs integrated within a computational framework such as NetworKIN (Linding et al. 2007), will produce interaction datasets that greatly enhance the ability of our algorithm to connect the phosphorylated proteins.

Beyond the focused mapping of interactions involving phosphorylated proteins, the ability to discover novel signaling pathways also depends on the coverage of other parts of the interactome. 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, especially in less well-studied organisms. Therefore, many computational methods have been developed to predict possible interactions. These methods make use of features such as gene neighborhood (M. Huynen et al. 2000), gene fusion (Marcotte et al. 1999), sequence co-evolution (Goh et al. 2000), and may incorporate several such features in a Bayesian framework (Jansen et al. 2003). The probabilistic nature of edge weights in our PCST formulation provides a natural way to include these computational predictions.

B. Other applications and potentials

The PCST approach can be used to analyze jointly a wide variety of types of data. Cellular functions are operated by networks of molecular interactions, which include a lot more than phosphorylation mediated signaling and transcription factor binding to target genes. But regardless of the data type, there are many situations in which we see to find a parsimonious, high-confidence interaction network satisfying a defined set of constraints. Therefore, this approach can be applied to many other levels of regulation, depending on the source of the constraints and the molecular interactions. For example, we may model the global effect of a microRNA by using the microRNA targets as constraints and including microRNA to target relationships in the interactome. Metabolomics data is another area of great interest and may become an entry point to link together protein signaling networks with metabolic networks. The detected metabolites can be used as constraints in a network of metabolic reactions catalyzed by enzymes that are also part of the protein interaction network. For all these datasets, taking a network approach such as the PCST will yield more insight than simply following up on the top hits.

One disadvantage of the PCST method is the tree structure of the resulting network: all the included terminal nodes must be connected to each other. However, it is possible that the terminal nodes belong to multiple, separate signaling pathways that are not connected to each other, either because there is no cross-talk biologically or the cross-talk interactions are not in the known interactome. Adopting a forest formulation, where multiple trees may be used to connect the terminal nodes, may remedy this drawback.

Finally, it is useful to consider this approach in the context of other types of network modeling. The strengths of our method lie in the ability to identify previously unrecognized components of a cellular response and to discover functionally coherent subsets of proteins. However, this approach is not designed to capture the dynamics of a system, including feedback regulation. A natural way to describe such feedback mathematically is by differential equations, which can be simulated numerically or analyzed. Differential equation-based models have been applied genome-wide in a comprehensive transcriptional and translational network for *Escherichia coli* (Thiele et al. 2009) and have been applied extensively to relatively small networks of mammalian proteins (Eungdamrong et al. 2004; Aldridge et al. 2006; Tyson et al. 2003). However, such approaches are not suitable for very large networks where there are not enough data to sufficiently constrain the necessary parameters of the models.

We believe that these two approaches may ultimately be used together to develop dynamic models of previously uncharacterized biological systems. In a first phase, proteomic, transcriptional or other "-omics" datasets would be analyzed using constraint optimization to identify a set of proteins that seem most relevant to the biological process. With the size of the problem now reduced to a more manageable level, more focused experiments together with differential equation-based modeling could reveal the dynamics of the system.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Nurcan Tuncbag for comments on the manuscript, and Cindy Woolley for help with editing. This work was supported by National Cancer Institute (NCI) Grant U54-CA112967. S.S.H. was 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.

References

Ahuja, RK.; Magnanti, TL.; Orlin, JB.; Weihe, K. Network flows: theory, algorithms and applications. Wurzburg: Physica-Verlag; 1995. 1972–1995.

Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK. Physicochemical modelling of cell signalling pathways. Nat Cell Biol. 2006; 8:1195–1203. [PubMed: 17060902]

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]

Aranda B, et al. PSICQUIC and PSISCORE: accessing and scoring molecularinteractions. Nat Methods. 2011; 8:528–529. [PubMed: 21716279]

Archer A, Bateni M, Hajiaghayi M, Karloff H. Improved Approximation Algorithms for Prize-Collecting Steiner Tree and TSP. SIAM journal on computing. 2011; 40:309.

Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. BIND--The Biomolecular Interaction Network Database. Nucleic Acids Res. 2001; 29:242–245. [PubMed: 11125103]

Bailly-Bechet M, Borgs C, Braunstein A, Chayes J, Dagkessamanskaia A, François J-M, Zecchina R. Finding undetected protein associations in cell signaling by belief propagation. Proc Natl Acad Sci U S A. 2011; 108:882–887. [PubMed: 21187432]

Bayati M, Borgs C, Braunstein A, Chayes J, Ramezanpour A, Zecchina R. Statistical mechanics of steiner trees. Phys Rev Lett. 2008; 101:37208.

Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]

Bryne JC, Valen E, Tang M-HE, Marstrand T, Winther O, Piedade I, da Krogh A, Lenhard B, Sandelin A. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 2008; 36:D102–D106. [PubMed: 18006571]

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]

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. MINT: the Molecular INTeraction database. Nucleic Acids Res. 2007; 35:D572–D574. [PubMed: 17135203]

Chaurasia G, Iqbal Y, Hänig C, Herzel H, Wanker EE, Futschik ME. UniHI: an entry gate to the human protein interactome. Nucleic Acids Res. 2007; 35:D590–D594. [PubMed: 17158159]

Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol. 2010; 11:427–439. [PubMed: 20461098]

Cline MS, et al. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc. 2007; 2:2366–2382. [PubMed: 17947979]

DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science (80-). 1997; 278:680–686.

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

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]

D'haeseleer P. What are DNA sequence motifs? Nat Biotechnol. 2006; 24:423–425. [PubMed: 16601727]

Eungdamrong NJ, Iyengar R. Modeling cell signaling networks. Biol Cell. 2004; 96:355–362. [PubMed: 15207904]

Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. Proc Natl Acad Sci U S A. 2005; 102:17675–17680. [PubMed: 16317069]

Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics. 2006; 22:e141–e149. [PubMed: 16873464]

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

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Co-evolution of proteins with their interaction partners. J Mol Biol. 2000; 299:283–293. [PubMed: 10860738]

Graphviz. Graphviz - Graph Visualization Software. 2011a Available at: http://graphviz.org/.

Graphviz. The DOT Language | Graphviz - Graph Visualization Software. 2011b Available at: http://www.graphviz.org/content/dot-language.

Grimsrud PA, Swaney DL, Wenger CD, Beauchene NA, Coon JJ. Phosphoproteomics for the masses. ACS Chem Biol. 2010; 5:105–119. [PubMed: 20047291]

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]

Hagberg, AA.; Schult, DA.; Swart, PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux, G.; Vaught, T.; Millman, J., editors. Proceedings of the 7th Python in Science Conference. Pasadena: 2008. p. 11-15.

Harbison CT, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004; 431:99–104. [PubMed: 15343339]

Huang S-SC, Fraenkel E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci Signal. 2009; 2:ra40. [PubMed: 19638617]

Huynen M, Snel B, Lathe W, Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. 2000; 10:1204–1210. [PubMed: 10958638]

Hwang D, et al. A data integration methodology for systems biology: experimental verification. Proc Natl Acad Sci U S A. 2005; 102:17302–17307. [PubMed: 16301536]

IBM. IBM IBM ILOG Optimization Academic Initiative - United States. 2010

Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2003; 18(Suppl 1):S233–S240. [PubMed: 12169552]

Jansen R, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science (80-). 2003; 302:449–453.

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010; 38:D355–D360. [PubMed: 19880382]

Kerrien S, et al. IntAct--open source resource for molecular interaction data. Nucleic Acids Res. 2007; 35:D561–D565. [PubMed: 17145710]

Keshava Prasad TS, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009; 37:D767–D772. [PubMed: 18988627]

Klingström T, Plewczynski D. Protein-protein interaction and pathway databases, a graphical review. Brief Bioinform. 2010

Kunsch C, Ruben SM, Rosen CA. 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. 1992; 12:4412–4421. [PubMed: 1406630]

Lan A, Smoly IY, Rapaport G, Lindquist S, Fraenkel E, Yeger-Lotem E. ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. Nucleic Acids Research. 2011; 39:W424–W429. [PubMed: 21576238]

Las Rivas, J De; Fontanillo, C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol. 2010; 6:e1000807. [PubMed: 20589078]

Linding R, et al. Systematic discovery of in vivo phosphorylation networks. Cell. 2007; 129:1415–1426. [PubMed: 17570479]

Ljubić I. Prize-Collecting Steiner Tree. 2008 Available at: http://homepage.univie.ac.at/ivana.ljubic/research/pcstp/.

Ljubić 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. 2005; 105:427–449.

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]

MacIsaac KD, Fraenkel E. Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol. 2006; 2:e36. [PubMed: 16683017]

Macek B, Mann M, Olsen JV. Global and site-specific quantitative phosphoproteomics: principles and applications. Annu Rev Pharmacol Toxicol. 2009; 49:199–221. [PubMed: 18834307]

Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Lett. 2009; 583:3966–3973. [PubMed: 19850042]

Manke T, Roider HG, Vingron M. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. PLoS Comput Biol. 2008; 4:e1000039. [PubMed: 18369429]

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science (80-). 1999; 285:751–753.

Matthews L, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009; 37:D619–D622. [PubMed: 18981052]

Matys V, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006; 34:D108–D110. [PubMed: 16381825]

Mering C, von Jensen Lars J, Snel Berend, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork Peer. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005; 33:D433–D437. [PubMed: 15608232]

Mering C, von Krause R, Snel Berend, Cornell M, Oliver SG, Fields Stanley, Bork Peer. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 2002; 417:399–403. [PubMed: 12000970]

Naegle KM, Gymrek M, Joughin BA, Wagner JP, Welsch RE, Yaffe MB, Lauffenburger DA, White FM. PTMScout, a Web resource for analysis of high throughput post-translational proteomics studies. Mol Cell Proteomics. 2010; 9:2558–2570. [PubMed: 20631208]

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

Orchard S, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol. 2007; 25:894–898. [PubMed: 17687370]

Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011; 12:87–98. [PubMed: 21191423]

Pagel P, et al. The MIPS mammalian protein-protein interaction database. Bioinformatics. 2005; 21:832–834. [PubMed: 15531608]

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

Prieto C, Las Rivas J De. APID: Agile Protein Interaction Data Analyzer. Nucleic Acids Res. 2006; 34:W298–W302. [PubMed: 16845013]

Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics. 2008; 9:405. [PubMed: 18823568]

Roberts CJ, et al. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science (80-). 2000; 287:873–880.

Roider HG, Kanhere A, Manke T, Vingron M. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics. 2007; 23:134–141. [PubMed: 17098775]

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg David. The Database of Interacting Proteins:2004 update. Nucleic Acids Res. 2004; 32:D449–D451. [PubMed: 14681454]

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004; 32:D91–D94. [PubMed: 14681366]

Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–342. [PubMed: 21593866]

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]

Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: understanding cancer using microarrays. Nat Genet. 2005; 37(Suppl):S38–S45. [PubMed: 15920529]

Sharifpoor S, et al. A quantitative literature-curated gold standard for kinase-substrate pairs. Genome Biol. 2011; 12:R39. [PubMed: 21492431]

Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011; 27:431–432. [PubMed: 21149340]

Sopko R, Andrews BJ. Linking the kinome and phosphorylome--a comprehensive review of approaches to find kinase targets. Mol Biosyst. 2008; 4:920–933. [PubMed: 18704230]

Sousa Abreu, R de; Penalva, Lo; Marcotte, EM.; Vogel, C. Global signatures of protein and mRNA expression levels. Mol Biosyst. 2009; 5:1512–26. [PubMed: 20023718]

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Cell Biol. 1998; 9:3273–3297.

Stark C, et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2011; 39:D698–D704. [PubMed: 21071413]

Stoltenburg R, Reinemann C, Strehlitz B. SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. Biomol Eng. 2007; 24:381–403. [PubMed: 17627883]

Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000; 16:16–23. [PubMed: 10812473]

Tarcea VG, et al. Michigan molecular interactions r2: from interacting proteins to pathways. Nucleic Acids Res. 2009; 37:D642–D646. [PubMed: 18978014]

Thiele I, Jamshidi N, Fleming RMT, Palsson BØ. Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. PLoS Comput Biol. 2009; 5:e1000312. [PubMed: 19282977]

Turinsky AL, Razick S, Turner B, Donaldson IM, Wodak SJ. Interaction databases on the same page. Nat Biotechnol. 2011; 29:391–393. [PubMed: 21552234]

Tyson JJ, Chen KC, Novak B. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. Curr Opin Cell Biol. 2003; 15:221–231. [PubMed: 12648679]

White FM. Quantitative phosphoproteomic analysis of signaling network dynamics. Curr Opin Biotechnol. 2008; 19:404–409. [PubMed: 18619541]

Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief Bioinform. 2008; 9:326–332. [PubMed: 18436575]

Yeger-Lotem E, et al. Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat Genet. 2009; 41:316–323. [PubMed: 19234470]

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]

Yu H, et al. High-quality binary protein interaction map of the yeast interactome network. Science. 2008; 322:104–110. (80-). [PubMed: 18719252]

Zhang, Yi; Askenazi, M.; Jiang, J.; Luckey, CJ.; Griffin, JD.; Marto, JA. A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia. Mol Cell Proteomics. 2010; 9:780–790. [PubMed: 20019052]

Zhang, Yi; Wolf-Yadlin, A.; White, FM. Quantitative proteomic analysis of phosphotyrosine-mediated cellular signaling networks. Methods Mol Biol. 2007; 359:203–212. [PubMed: 17484120]

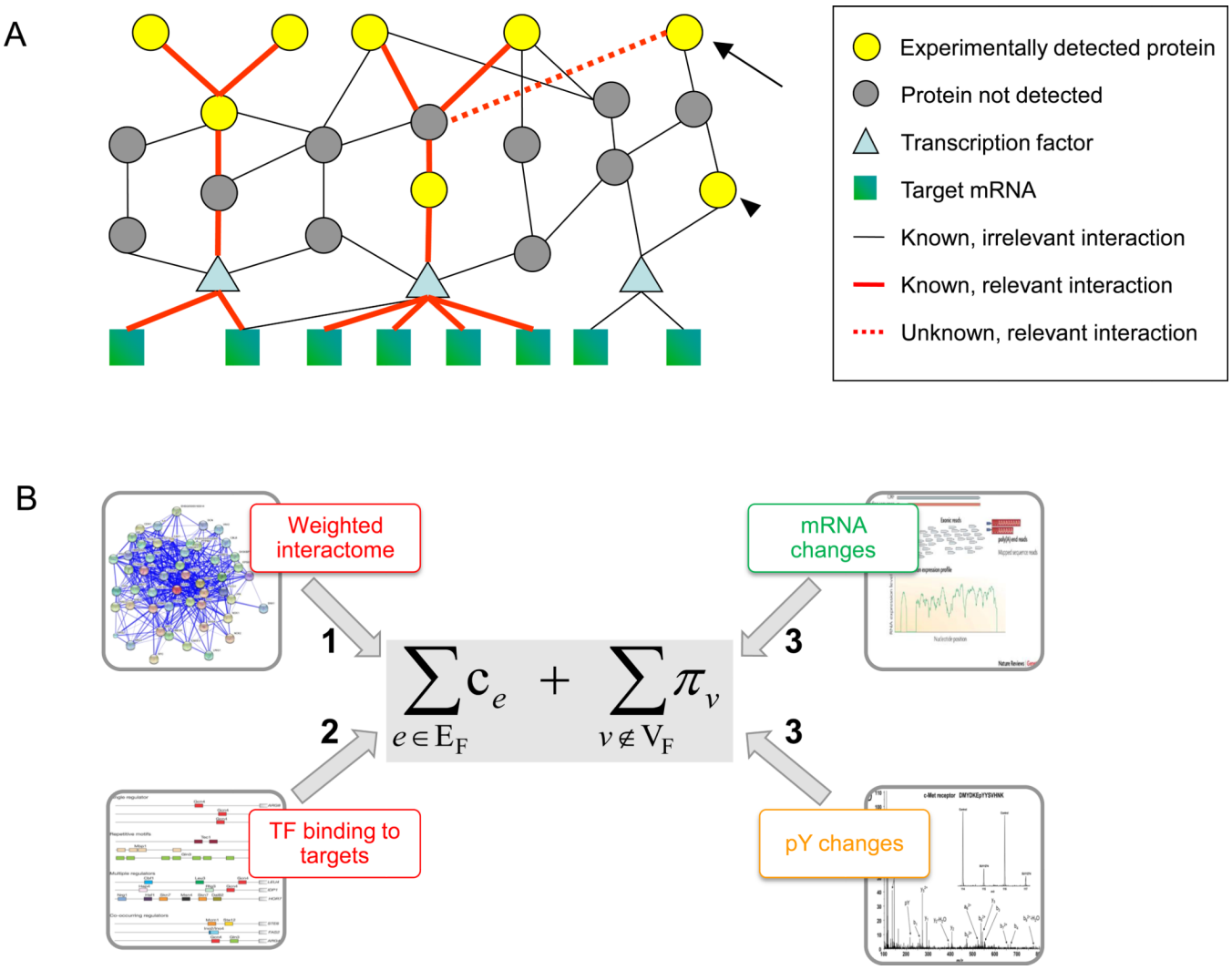


Figure 1.

(A) 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. Image reproduced with permission from (Huang et al. 2009). (B) Work flow diagram for defining the optimization objective function from input datasets. 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 3). The transcription factor to mRNA target relationships are added to the edges to form the total interactome (Step 2), and the mRNA nodes are assigned penalty values (Step 3).

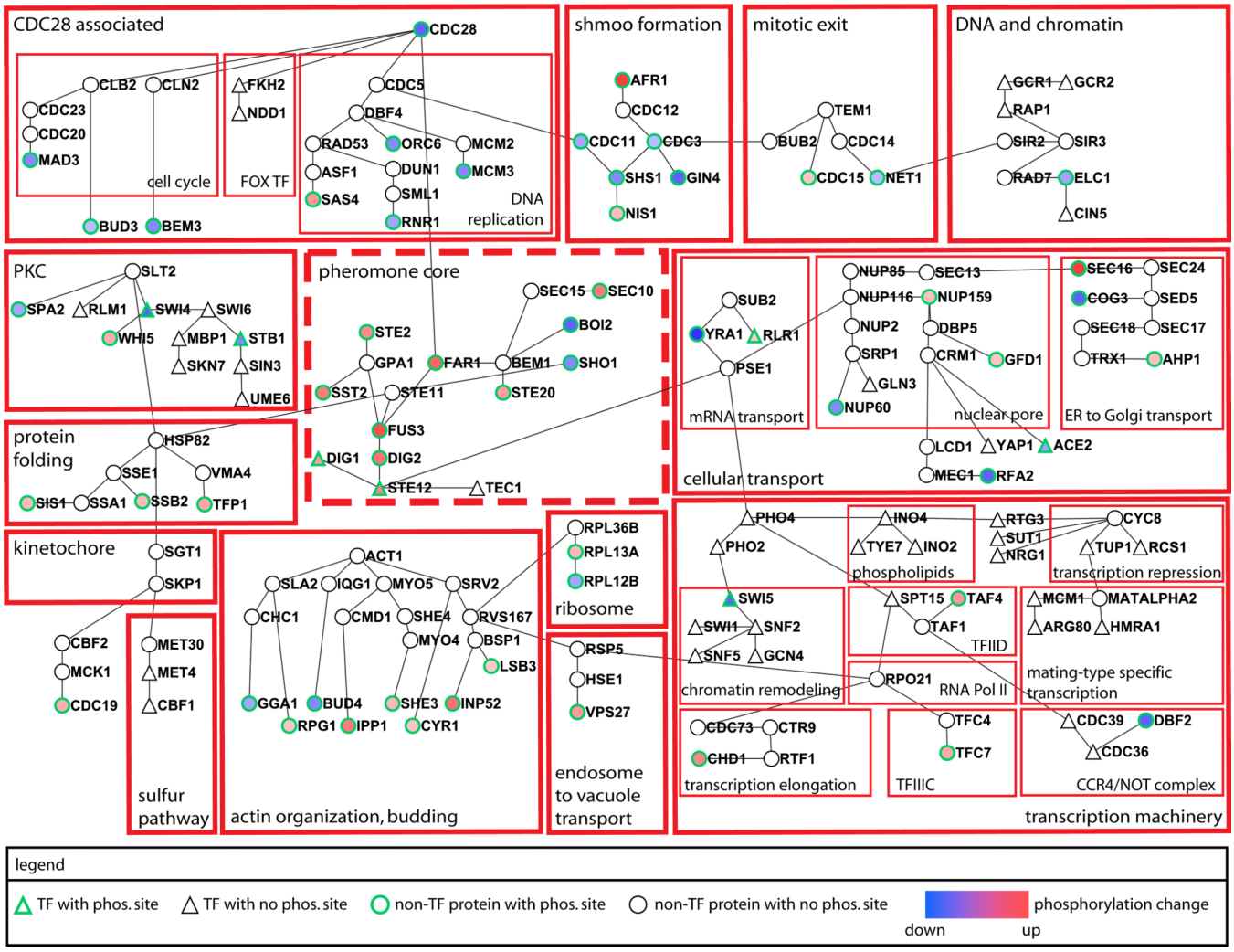


Figure 2.

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. Image reproduced with permission from (Huang et al. 2009).

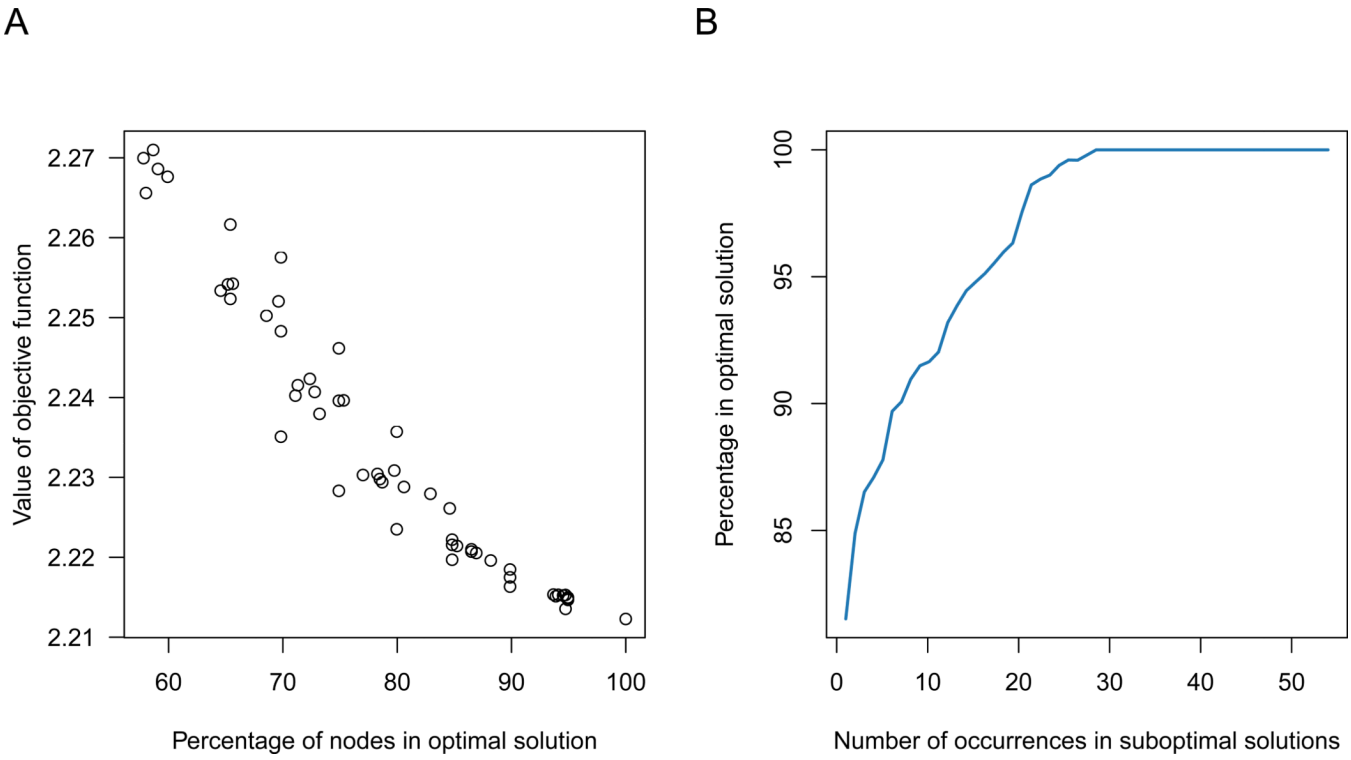


Figure 3.

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 achieve 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. Image reproduced with permission from Huang et al. (2009).

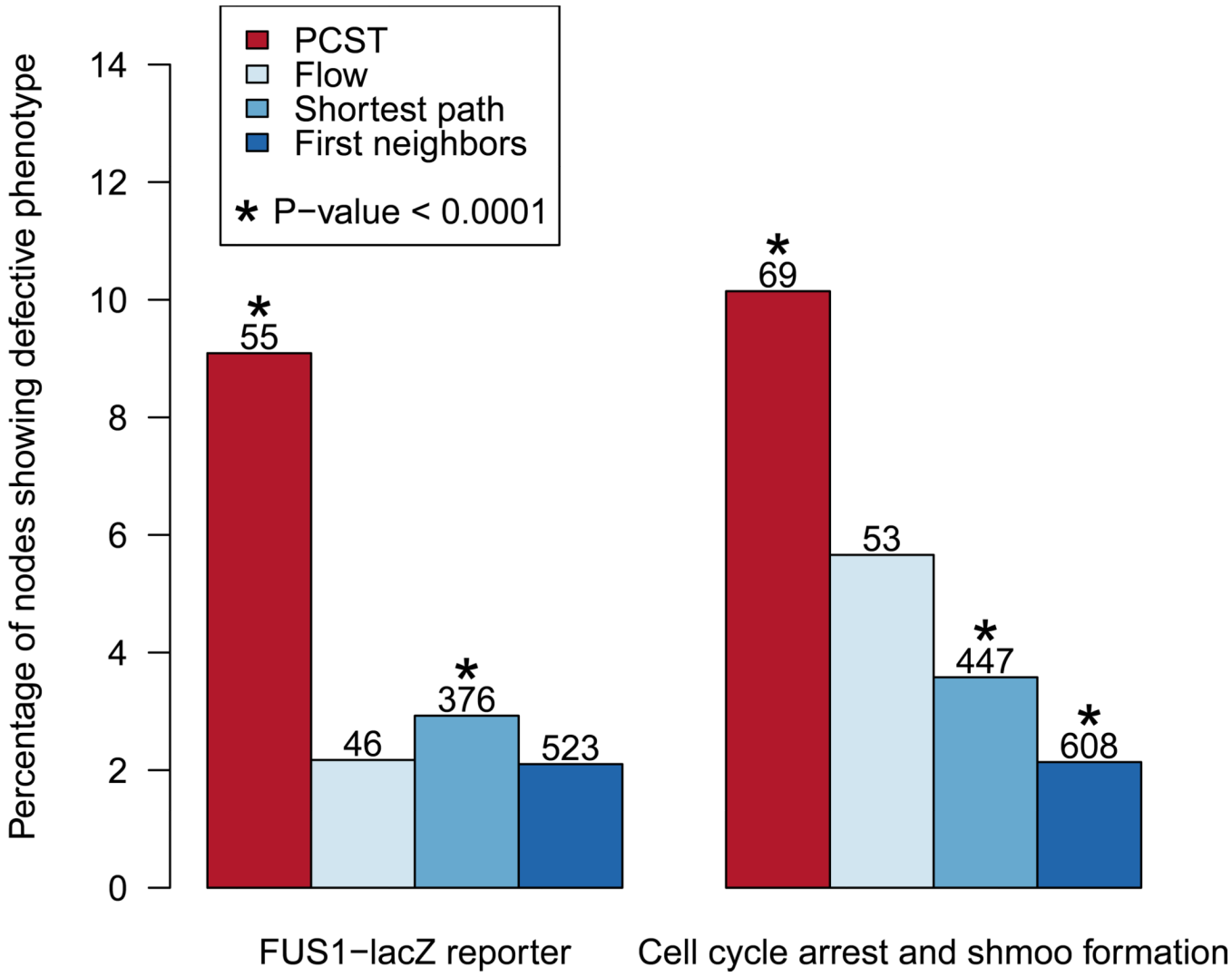


Figure 4.

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 response, measured by defects in activating a FUS1-lacZ reporter gene (Chasse et al. 2006) and defects in cell cycle arrest and shmoo formation (Narayanaswamy et al. 2006). The *Flow* network was constructed from the phosphorylated proteins and differential expressed genes by a previously published algorithm based on network flows (Yeger-Lotem et al. 2009). The *Shortest path* network consists of pairwise shortest paths between the terminal nodes and the *First neighbor* network consists of nodes in the interactome that directly interact with the phosphorylated proteins. 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. Image reproduced with permission from Huang et al. (2009).

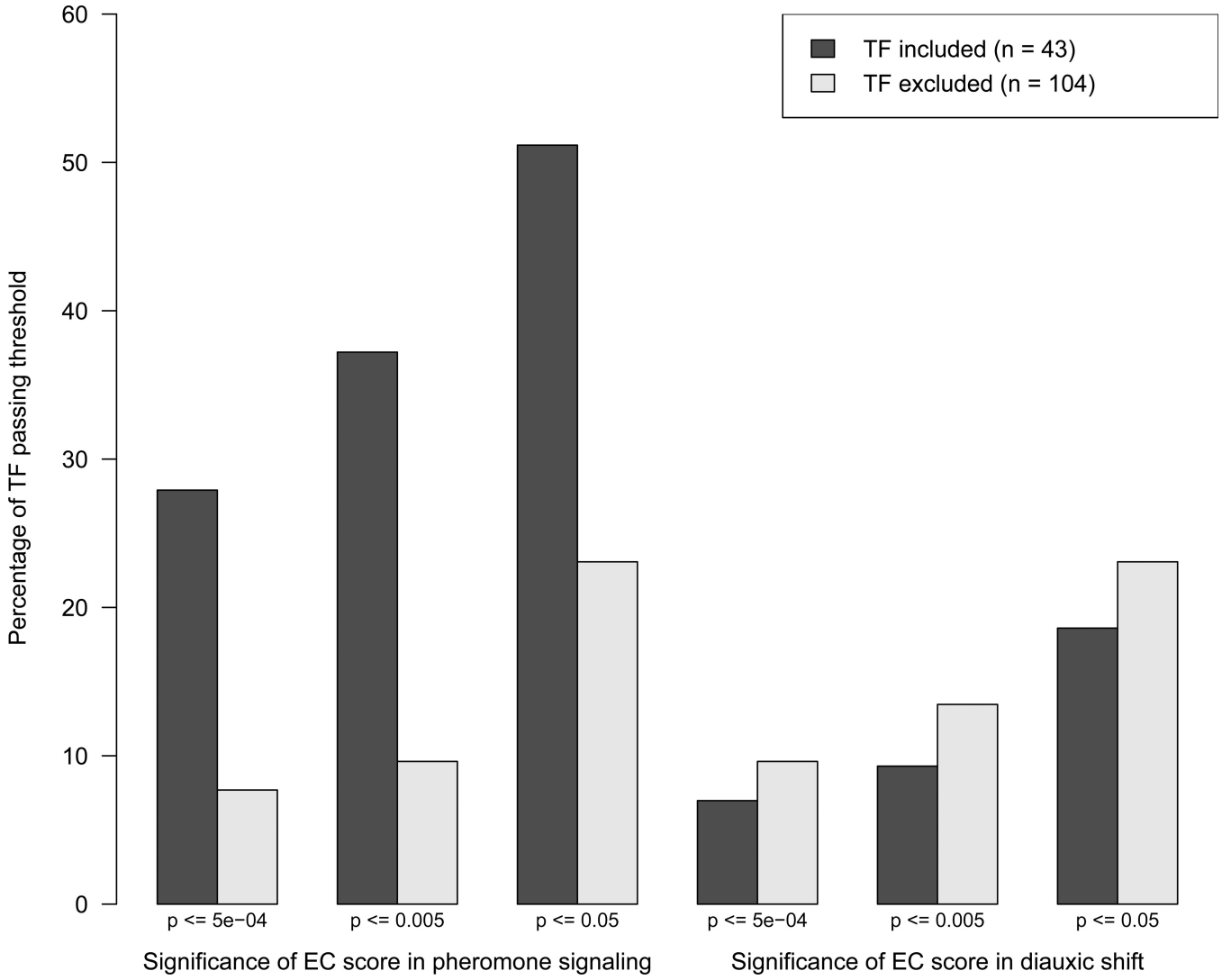
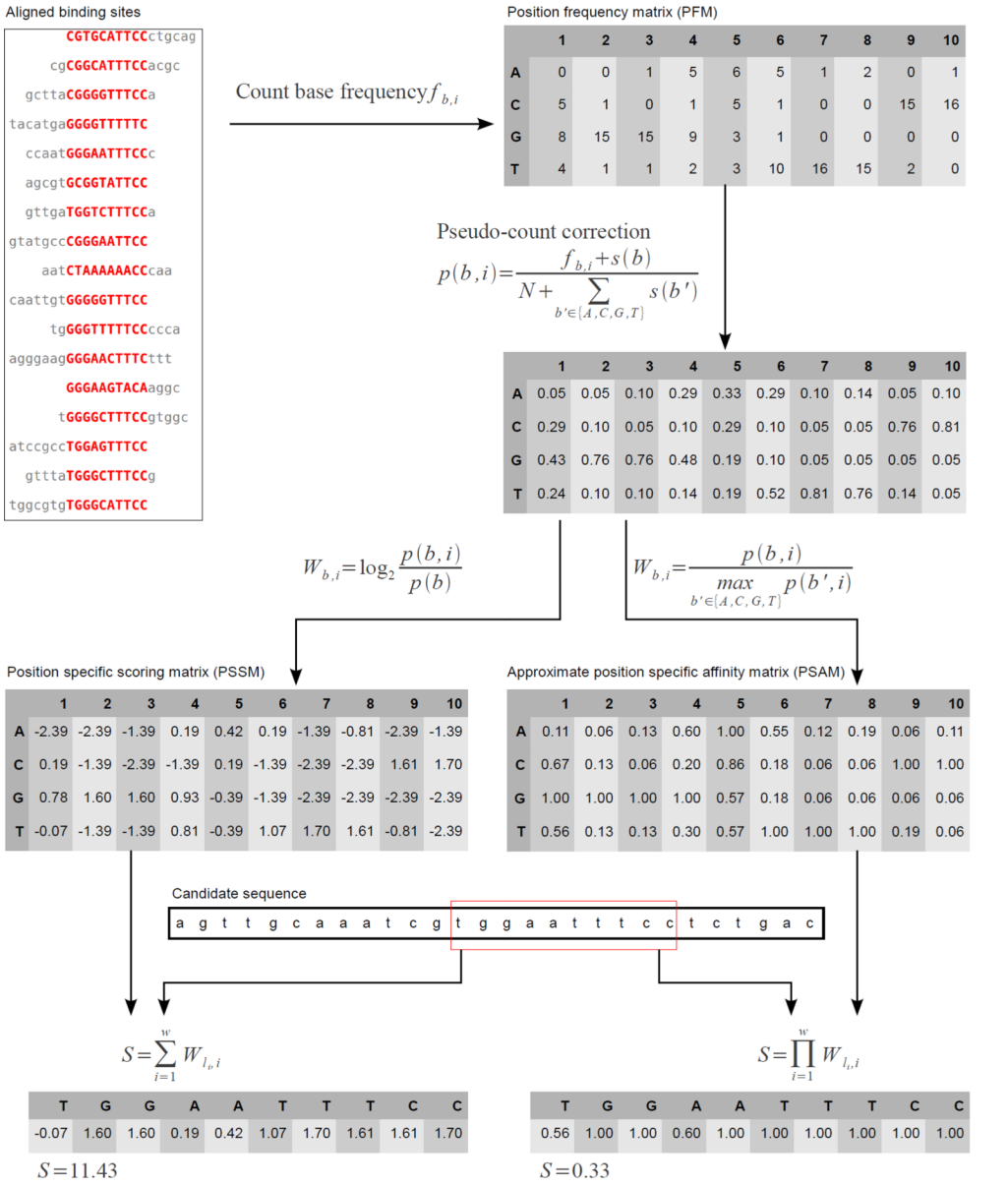


Figure 5.
Percentage of transcription factors (TF) with targets that show significant expression coherence (EC) scores computed from 50 nM α-factor time course (Roberts et al. 2000) and diauxic shift conditions (DeRisi et al. 1997), for transcription factors included in and excluded from the PCST solution network. The P-values indicate thresholds on the significance of the expression coherence score of the target genes. Image reproduced with permission from Huang et al. (2009).



	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

	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

	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

	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

Figure 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 NFκB binding site in the human IL8 promoter (TRANSFAC binding site HS$IL8_21). *In vitro* techniques such as SELEX (systematic evolution of ligands by exponential enrichment) (Stoltenburg et al. 2007) can generate a set of sequences that bind to a specific transcription factor with high affinity. From an alignment of these sequences, a PFM is created to represent the base preference of this factor at each position of the binding site. After pseudo-count correction, the PSSM approach takes the base preference at each position, adjusts for background (usually genome-wide) frequency of that

base, and computes a numerical value for the bases at each position that can be used to score a DNA sequence (D'haeseleer 2006; Stormo 2000). Alternatively, an approximate PSAM for scoring can be created from a pseudo-count corrected PFM by calculating the preference of a base relative to the most frequent base at each position (Foat et al. 2006, 2005; Manke et al. 2008; Roider et al. 2007). See (MacIsaac and Fraenkel 2006) for a more detailed treatment of the topic.

Table I

A selection of publicly available protein-protein interaction databases. For further details see recent summary and reviews in (Turinsky et al. 2011; De Las Rivas et al. 2010; Klingström et al. 2010). Many databases in this table have adopted the Proteomic Standards Initiative Molecular Interaction (PSI-MI) data formats and implemented the PSI Common Query Interface (PSICQUIC) (Aranda et al. 2011) that allows easy, programmatic access and integration of these data.

Type of interactions	Data sources	Name of database and reference[a]
Direct/physical	Curation of primary literature	Biological General Repository for Interaction Datasets (BioGRID) (Stark et al. 2011)
		Human Protein Reference Database (HPRD) (Keshava Prasad et al. 2009)
		Molecular Interaction database (MINT) (Chatr-aryamontri et al. 2007)
		IntAct molecular interaction database (Kerrien et al. 2007)
		Mammalian Protein-Protein Interaction Database (MIPS) (Pagel et al. 2005)
		Database of Interacting Proteins (DIP) (Salwinski et al. 2004)
		Biomolecular Interaction Network Database (BIND) (Bader et al. 2001)
	Collection of multiple primary databases	Interaction Reference Index (iRefIndex) (Razick et al. 2008)
		Agile Protein Interaction DataAnalyzer (APID) (Prieto et al.2006)
		Michigan Molecular Interactions database (MiMI) (Tarcea et al. 2009)
		Unified Human Interactome database (UniHI) (Chaurasia et al. 2007)
Direct/physical + indirect/functional	Collection of multiple primary databases and computational predictions	STRING (von Mering et al. 2005)

[a] References

Aranda, B. et al. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat Methods 8, 528–529.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., and Hogue, C. W. (2001). BIND--The Biomolecular Interaction Network Database. Nucleic Acids Res 29, 242–5.

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

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

Kerrien, S. et al. (2007). IntAct--open source resource for molecular interaction data. Nucleic Acids Res 35, D561–D565.

Keshava Prasad, T. S. et al. (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res 37, D767–72.

Klingström, T., and Plewczynski, D. (2010). Protein-protein interaction and pathway databases, a graphical review. Brief Bioinform.

Las Rivas, J. De, and Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol 6, e1000807.

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

Pagel, P. et al. (2005). The MIPS mammalian protein-protein interaction database. Bioinformatics 21, 832–4.

Prieto, C., and Las Rivas, J. De (2006). APID: Agile Protein Interaction Data Analyzer. Nucleic Acids Res 34, W298–W302.

Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics 9, 405.

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

Stark, C. et al. (2011). The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39, D698–D704.

Tarcea, V. G. et al. (2009). Michigan molecular interactions r2: from interacting proteins to pathways. Nucleic Acids Res 37, D642–D646.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., and Wodak, S. J. (2011). Interaction databases on the same page. Nat Biotechnol 29, 391–3.

Table II

Biological functions and measures of coordinated mRNA expression of the clusters in the pheromone response PCST network generated from edge betweenness clustering. *EC*, expression coherence (Pilpel et al. 2001); *EA*, expression activity (Ideker et al. 2002). Reproduced with permission from (Huang et al. 2009).[b]

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	7.15E-12		
	GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	1.94E-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		

Cluster		Top three enriched GO biological process terms	Corrected P-value	P-value of EC score	P-value of EA score
	GO:0051329	interphase of mitotic cell cycle	1.07E-03		
Full network	GO:0006350	transcription	2.67E-23	0.729	1
	GO:0019222	regulation of metabolism	2.73E-21		
	GO:0050791	regulation of physiological process	1.16E-20		

*b*References

Huang, S.-S. C., and Fraenkel, E. (2009). Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci Signal 2, ra40.
Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 Suppl 1, S233–S240.
Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 29, 153–9.