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).

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.

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 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

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) 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 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

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

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 $\nu \in prot$ we have *strength* $(\nu)>0$ as a measure of the importance of ν in the network. We computed the node penalty as the normalized absolute log of the strength:

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

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 $fc(\nu)$ be the fold change in mRNA abundance of each gene $\nu \in mrna$. For each $\nu \in mrna$, we searched the interactome for the set of upstream transcription factors F, removed ν from the interactome, and added one node ν_f for each transcription factor $f \in F$ and one edge between f and ν_f. The fold change of ν was transferred to all the ν_f and normalized so the penalties were

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

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

$$\sum_{v \notin \mathrm{V}_{\mathrm{T}}} \beta\, p(v) + \sum_{e \in \mathrm{E}_{\mathrm{T}}} c(e).$$

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]

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]

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

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

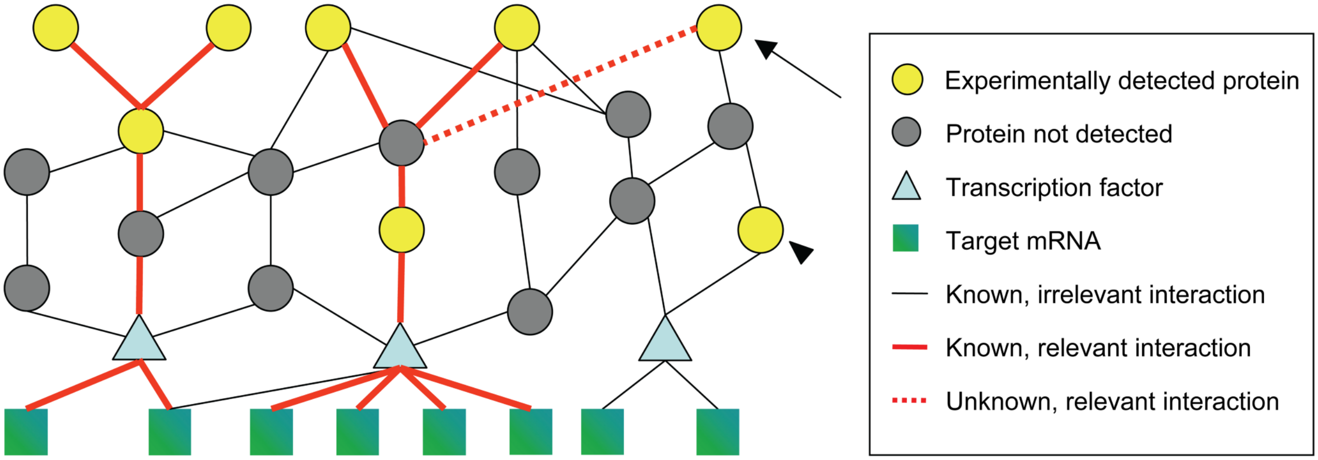
Experimentally detected protein
Protein not detected
Transcription factor
Target mRNA
Known, irrelevant interaction
Known, relevant interaction
Unknown, relevant interaction

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.

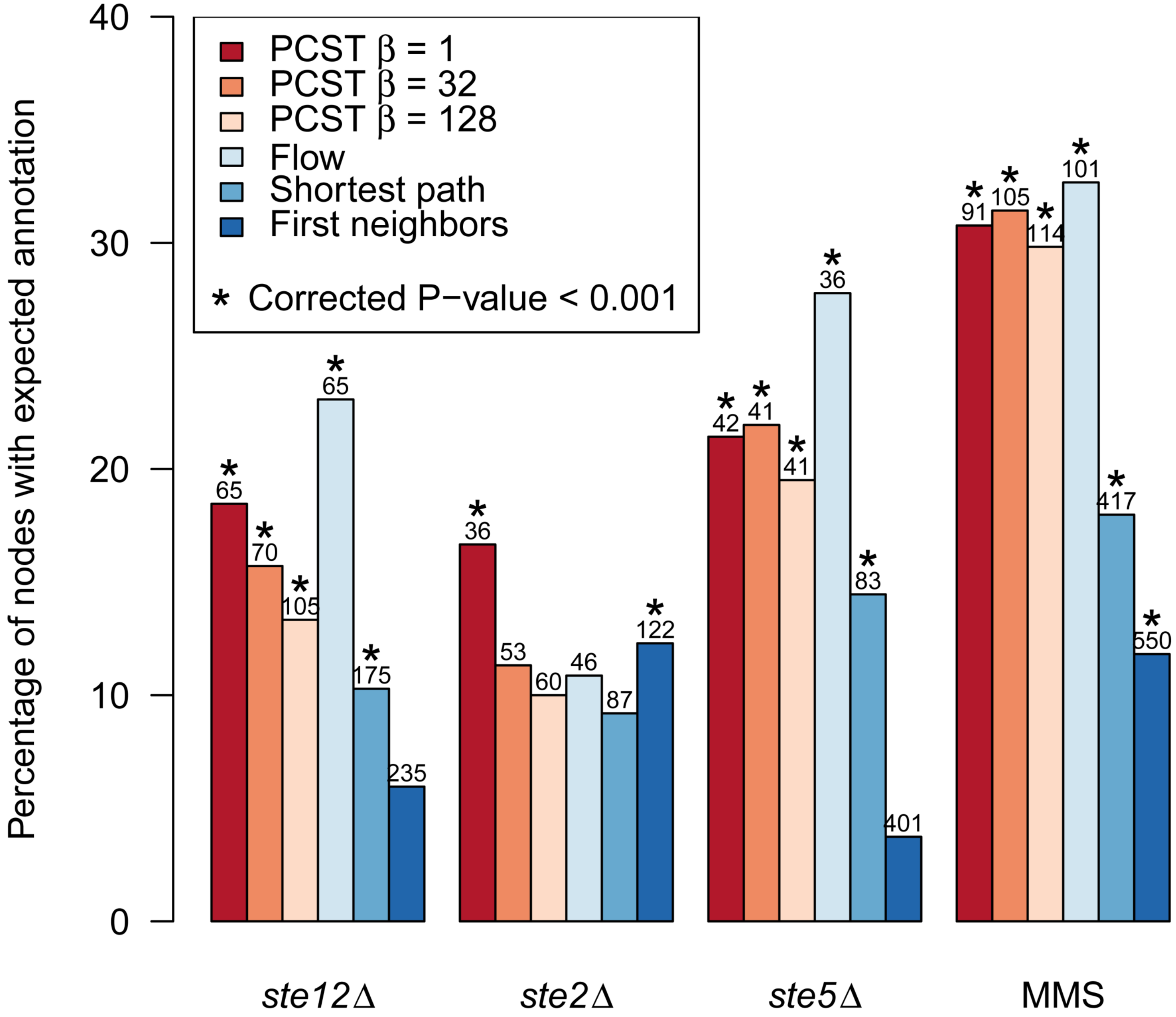

Fig. 2.
The PCST solution recovers compact networks. The fraction of nodes associated with the expected biological process is comparable to the networks from flow-based approach that include approximately equal number of nonterminal nodes, but this fraction is higher than the first neighbor and shortest 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).

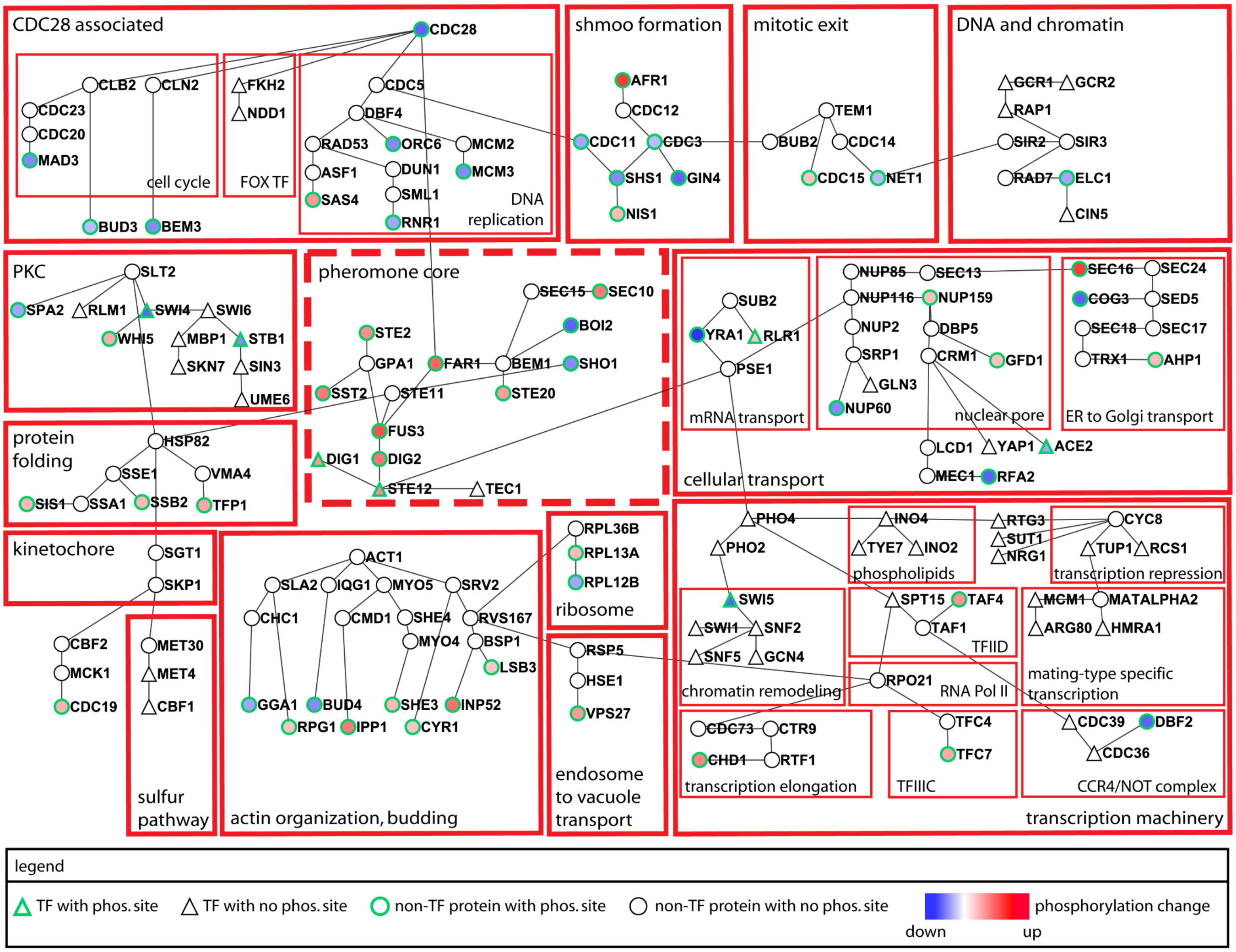

Fig. 3.

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.

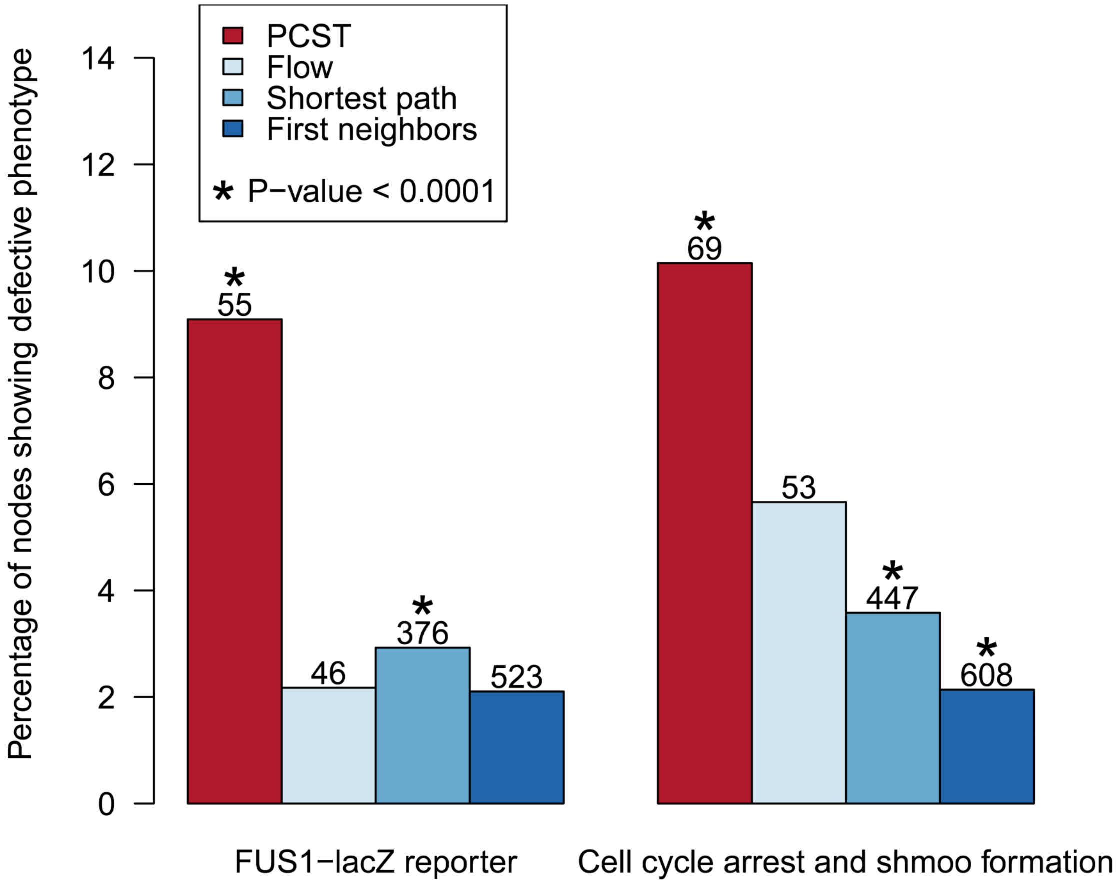

Fig. 4.

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.

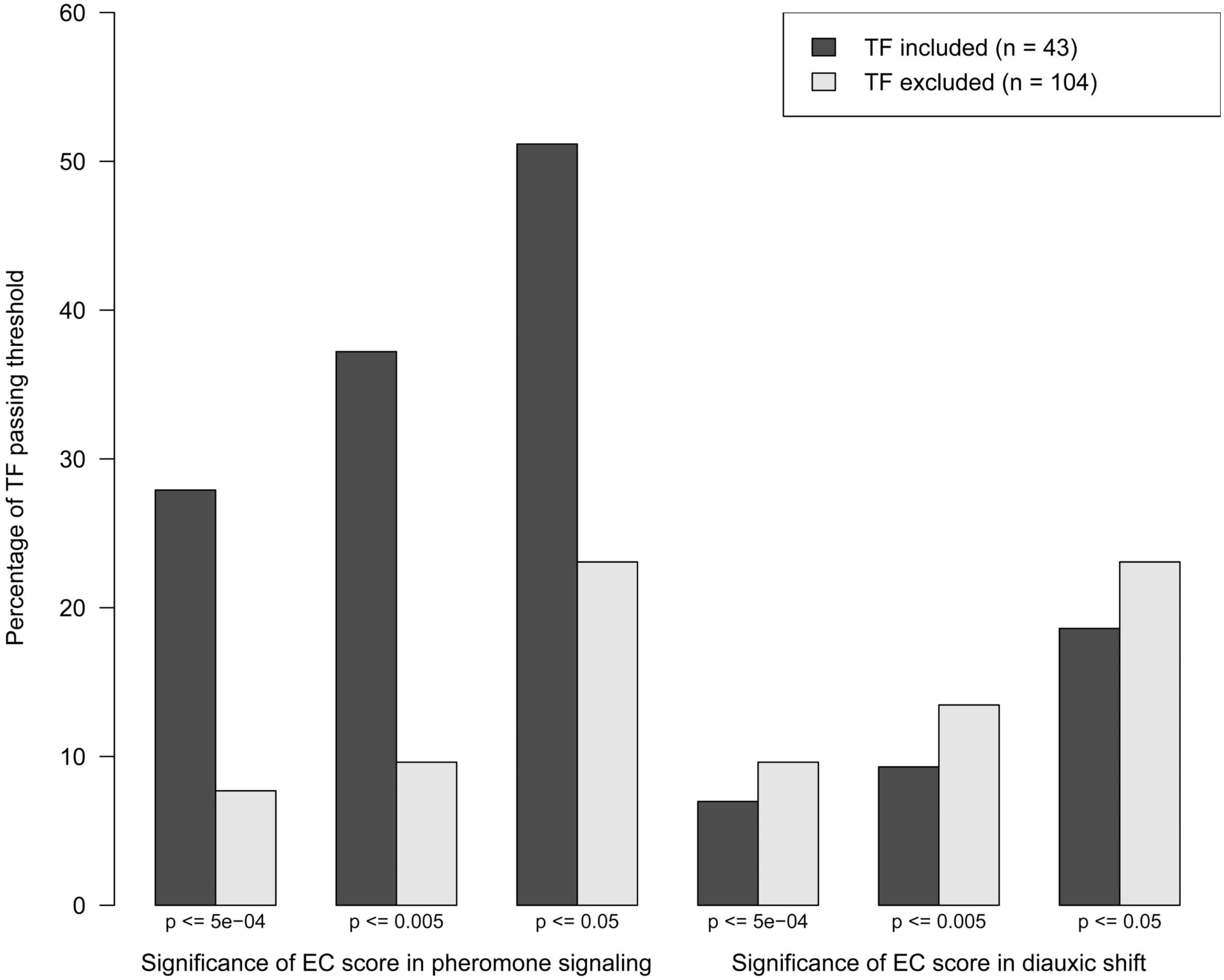

Fig. 5.

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

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		