Identifying Regulatory Subnetworks for a Set of Genes*

Michelle S. Scott‡§¶, Theodore Perkins‡, Scott Bunnell‡§‖, François Pepin‡§‖, David Y. Thomas§, and Michael Hallett‡§‖**

High throughput genomic/proteomic strategies, such as microarray studies, drug screens, and genetic screens, often produce a list of genes that are believed to be important for one or more reasons. Unfortunately it is often difficult to discern meaningful biological relationships from such lists. This study presents a new bioinformatic approach that can be used to identify regulatory subnetworks for lists of significant genes or proteins. We demonstrate the utility of this approach using an interaction network for yeast constructed from BIND, TRANSFAC, SCPD, and chromatin immunoprecipitation (ChIP)-Chip data bases and lists of genes from well known metabolic pathways or differential expression experiments. The approach accurately rediscovers known regulatory elements of the heat shock response as well as the gluconeogenesis, galactose, glycolysis, and glucose fermentation pathways in yeast. We also find evidence supporting a previous conjecture that approximately half of the enzymes in a metabolic pathway are transcriptionally co-regulated. Finally we demonstrate a previously unknown connection between GAL80 and the diauxic shift in yeast. ***Molecular & Cellular Proteomics 4:683–692, 2005.***

High throughput genomic and proteomic strategies generate lists of genes or proteins that are believed to be significant. These distinguished sets, such as sets of genes that are differentially expressed between two conditions (*e.g.* tumor *versus* normal or mutant *versus* wild type), alone provide little insight into *if* and *why* these genes are important in a particular context. One approach to gain deeper insight into such questions is to focus on determining the regulatory relationships between members of the set (1). For example, the relationships between a set of co-expressed genes might be adequately explained if it is determined that they share a small group of transcription factors or if their proteins are known to form a complex.

A common technique to find such regulatory relationships is to examine the distinguished set of genes relative to a so-called *interaction network* (2). High throughput methods, such as ChIP[1]-Chip, combined with computational approaches have generated transcriptional regulatory networks (3, 4). Proteomic approaches, such as tandem affinity purification tagging (5) and yeast two-hybrid assays, have yielded large scale protein-protein interaction maps (6, 7). Interaction networks are formed from the union of these two data types. Whereas the challenge of proteomics is to generate complete interaction networks for many different organisms and to compile this information in data bases, the challenge of bioinformatics is to automate strategies for extracting meaningful information from these massive graphs. Due to the size and complexity of these networks (≈30,000 nodes for human), a manual search for regulatory subgraphs is nearly impossible for even a handful of distinguished genes. Several related approaches exist in the literature for investigating interaction networks.

Oyama *et al.* (8) present a feature selection strategy. Given a protein-protein interaction network and various types of information such as enzyme functionality (EC number), functional annotations (Swiss-Prot keywords), PROSITE motifs, and conserved sequences, they attempt to automatically learn association rules from the data to more accurately characterize protein-protein interactions. A second approach from Lappe *et al.* (9) uses hierarchically organized functional annotations of proteins to contract or "distill" a protein-protein interaction graph into a more manageable but equally informative representation. A third approach from Kelley *et al.* (10) entitled PathBLAST allows investigators to compare interaction networks across species to identify pathways and complexes that have been conserved. The framework takes as input a protein "interaction path" and searches for high scoring alignments between pairs of protein interaction paths for which proteins of the first path are paired with putative orthologs occurring in the same order in the interaction network.

A fourth strategy from Ideker *et al.* (2) essentially maps mRNA expression data on to an interaction network. The weight of a node is a measure of significance of differential expression over one or more microarray hybridizations. The goal is to search for subnetworks that display a statistically significant amount of differential expression. Such subnetworks might correspond to small sets of proteins belonging to

From the ‡McGill Centre for Bioinformatics, 3775 University Street, Montreal H3A 2B4 and the §Department of Biochemistry, 3655 Promenade Sir William Osler, McGill University, Montreal H3G 1Y6, Canada

Received, August 18, 2004, and in revised form, February 5, 2005
Published, MCP Papers in Press, February 18, 2005, DOI 10.1074/mcp.M400110-MCP200

[1] The abbreviations used are: ChIP, chromatin immunoprecipitation; HAP, heme activated protein.

a common metabolic pathway for instance. The underlying computational problem is to find a maximum-weight connected component of the network (a set of proteins in the network in which there is a path between any two members). Although a solution may contain nodes that do not exhibit differential expression, such nodes must lie on a path between nodes that do exhibit differential expression. The non-differentially expressed proteins might then represent important "lost connections" between sets of differentially expressed proteins. In Ideker *et al.* (2) the authors show that these subnetworks succeed in finding proteins that are involved in common biological processes (*e.g.* the same metabolic pathway).

This study presents a new approach using so-called Steiner trees (see "Experimental Procedures") for investigating interaction networks. The goal is to find subnetworks within an interaction network that are plausibly responsible for regulating a given distinguished set of genes/proteins. Whereas the Ideker *et al.* (2) approach is a global search of the interaction network for differentially expressed subnetworks, our approach allows the user to specify a "distinguished" set of nodes. This distinguished set might, for example, represent a list of differentially expressed genes from a microarray experiment, chemical-genetic/synthetic-lethal interactions (11), or any other type of assay that produces lists of interesting genes/proteins and where similar questions regarding regulatory relationships are important. The Steiner tree solution provides a "backbone" consisting of genes that are the most likely candidates to be involved in regulatory relationships between members of the distinguished set. Additional nodes and edges are added to this backbone to include as many potential regulatory genes as possible while excluding most irrelevant interactions. Although it is likely impossible to infer a complete causal explanation of how/why a gene is important from an interaction network, the approach does produce subnetworks that are typically of a size that allows investigators to perform a more detailed, literature-based analysis. This data reduction technique is analogous to a BLAST search for a given uncharacterized sequence where the goal is to find a relatively small set of likely homologues within a large data base of sequences.

To investigate the utility of our approach, we performed three sets of experiments using an interaction network for yeast composed of 5,458 proteins and 23,642 interactions from the BIND version 2 (12) (yeast protein-protein interactions), TRANSFAC (13), SCPD (34) (yeast protein-DNA interactions), and yeast ChIP-Chip (protein-DNA) (3) data sets. First, to test the ability of our framework to find plausible regulatory subnetworks, we conducted experiments on the GAL80 deletion expression data set (14). Our results support the claim in Jansen *et al.* (15) that examining co-expression alone may not always be sufficient for finding regulatory subnetworks and allowed for the discovery of a previously unknown connection between GAL80 and the diauxic shift in yeast. Second, we analyzed the advantages of different interaction networks using a heat shock microarray expression data set (16). This allowed for the identification of a heat shock transcriptional response subnetwork that integrates known but dispersed observations and offers new insights into this process. Third, we present evidence in support of a claim of Ihmels *et al.* (17) that approximately half of the enzymes in a metabolic pathway are transcriptionally co-regulated and show that our framework accurately rediscovers known regulatory elements of the gluconeogenesis, galactose, glycolysis, and glucose fermentation pathways in yeast.

EXPERIMENTAL PROCEDURES

To simplify the figures, names such as GAL4 refer to both the gene and the protein it encodes (for example in Fig. 1, the gene GAL4 is shown to be transcriptionally regulated by MIG1, and the GAL4 protein transcriptionally regulates the gene GAL80). We use this convention throughout the text as well.

Interaction Graphs—We represent an *interaction network* as a graph in which nodes represent proteins. The graph is the union of the protein-protein and protein-DNA interaction networks. When two proteins bind directly, we term this a *protein-protein interaction*. A protein-protein interaction is represented by an undirected edge between the two nodes corresponding to the proteins. In the case of protein-DNA interactions, an edge from node u to node v is included in the graph when it is known that the protein u is a transcription factor that regulates the expression of the gene corresponding to v. (Our approach can be modified to use the directional information implicit in protein-DNA interactions but is omitted from the discussion here.) In the following experiments, the graph is built using the BIND data set (12) (restricted to yeast protein-protein interactions), TRANSFAC (13) (yeast protein-DNA interactions), SCPD (yeast protein-DNA interactions), and ChIP-Chip data (yeast protein-DNA interactions) (3). We include protein-DNA interactions from the ChIP-Chip data if their associated p value is 0.001 or less, a conservative threshold (3). In total, this yields 5,458 nodes (proteins) and 23,642 edges (interactions).

Steiner Trees—Suppose we have an interaction network and a distinguished set of nodes S, a set of proteins in the interaction graph whose regulation (and possibly co-regulation) we wish to understand. A natural question is to ask how to connect the nodes of S together. Ideally there would be short paths between any two nodes in S. Moreover we would like to "highlight" nodes that are used by many paths between pairs of elements in S. The notion from mathematics that captures these criteria is the *Steiner tree*. A Steiner tree is a connected subgraph of the interaction network that includes all the distinguished nodes. That is, for any pair of nodes in the distinguished set, there is a path between the nodes in the Steiner tree. Fig. 1*b* depicts a Steiner tree for the distinguished set of vertices (*dark shaded*) given in Fig. 1*a*. The nodes in the tree that are not in S are called *Steiner points*. Note that nodes such as HAP4, PUT3, and CAT8 lie on the path between many pairs of proteins.

Finding a Steiner tree is trivial. However, what one seeks is a Steiner tree that is optimal in some sense. In the "classical" Steiner tree problem, each edge e is assigned a positive real number $w(e)$ called the *weight* or *cost* of the edge. The *cost of a Steiner tree* T is simply the sum of the weight of each edge in the tree. The problem is to find a Steiner tree with minimal cost. To address the problem of finding regulatory subnetworks, we consider a modified version of this problem called the *node-weighted* Steiner tree problem. Rather than weighting edges, each node v is assigned cost $w(v)$. This cost typically represents a measure of differential expression associated with

v. The problem is to find a Steiner tree T for which the sum of node costs is minimal. Intuitively we are asking to connect the distinguished vertices in the "most compact" way possible. Each edge in the resultant network could then be examined for a regulatory role with respect to the elements of the distinguished set.

Both of these problems have been studied extensively in the computer science literature (18–21), and there is likely no efficient algorithm for solving these problems optimally. However, there is an algorithm for the node-weighted Steiner tree problem that guarantees a solution no more than $2\log|S|$ times optimal (20). This means that although we cannot be guaranteed that the algorithm will find the best solution, it will most often produce a sufficiently good solution. A contribution of this study is the observation that it is feasible to solve the node Steiner tree problem if the distinguished set S is small. By modifying the Dreyfus-Wagner algorithm (18) we can solve the node Steiner problem in time on the order of $3^{k-1} \cdot n + n^3$ time where k is the size of the distinguished set and n is the number of nodes in the interaction network. In practice, k tends to be as small as 10, but n is over 5,000. When k is too large for exact solution to be feasible, we default to the approximation algorithm from Klein *et al.* (20). In our experiments, this has always given solutions that are optimal or very close to optimal. Moreover there are many small simplifications one can automate that speed up the computation further. Succinctly our software is fast and gives accurate solutions; it is capable of handling the size of distinguished sets typically created by microarray, chemical-genetic, or synthetic-lethal assays.

Weight Functions—Our framework considers two types of weight functions denoted w_d and w_1. The weight function w_d measures the amount of differential expression over a set of m microarray hybridization experiments. Any package for deciding differential expression can be used as long as it returns a p value, p_u, as the measure of differential expression for gene u. We assign $w_d(u) = -\log(1 - p_u)$. Further background corrections can be applied to improve w_d, but we omit this discussion here. Intuitively this cost function corresponds to choosing those nodes in the network that connect the distinguished set and show the most differential expression. The weight function w_1 simply assigns the weight of every node to be 1. This cost function corresponds to finding the shortest way to connect the nodes in the distinguished set. Intuitively w_d corresponds to finding the most likely regulatory explanation, whereas w_1 corresponds to the most parsimonious explanation.

Steiner Trees as Backbones to Be Augmented—It is not likely that the Steiner tree alone constitutes a regulatory subnetwork for the distinguished set. Instead it is hoped that the Steiner tree prunes down the search for regulatory networks to a small, manageable number of alternative nodes. In this sense, the Steiner tree is a backbone that we can augment with additional nodes that are also likely to be relevant. Our software is implemented in the VisANT interactive visualization software (22). Beyond the computation of the Steiner tree, we offer several routines that search the neighborhood of the Steiner tree in a rational manner.

The general flavor of how Steiner backbones are extended is captured by the notion of an (i,l)-augmented graph. The intuition is as follows. If a node is within distance i of the Steiner tree (the shortest path to the Steiner tree has length at most i) and if there are at least l different such short paths, then this node is included in the subnetwork. Fig. 3 depicts the (1,2)-augmented tree from Fig. 1. Additional rules for augmenting backbones use functional annotations (including Gene Ontology annotations) to help decide the relevance of the gene/protein.

Other Strategies for Discovering Regulatory Subnetworks—In addition to our work based on Steiner trees, we have experimented with simpler, more efficient strategies for finding small regulatory subnetworks. For example, one approach we have used is to compute the shortest paths between all pairs of nodes in the distinguished set. The graph formed as the union of these paths can then serve as a backbone. Alternatively one can choose any one of the distinguished nodes and take the union of paths from that node to all the other distinguished nodes. This is a sensible approach, for example, for connecting a gene to the proteins that are differentially expressed in a knock-out expression experiment. These routines will also be made available within our implementation in VisANT. We plan to also implement them in CytoScape (23) in the near future.

RESULTS

Most Likely Versus Smallest Steiner Trees—Two important and conflicting observations have emerged recently from studies of interaction networks and high throughput protein-protein interaction data from Ideker *et al.* (2) and Jansen *et al.* (15), respectively. The strategy of Ideker *et al.* (2) requires that many of the proteins involved in a common pathway either as enzymes or regulatory transcription factors exhibit co-expression. If this assumption is valid, these proteins should belong to subnetworks whose members are significantly differentially expressed. Ideker *et al.* (2) and Ihmels *et al.* (17) provide evidence that such statistically significant co-expression does exist. However, studies such as Jansen *et al.* (15) find that mRNA expression levels do not correlate strongly to known protein-protein interactions unless the interactions are part of a so-called *stable* complex (complexes that are maintained throughout most cellular conditions) and not simply *transient* complexes. In such cases, looking for subgraphs that exhibit common differential expression will miss key proteins participating in a common biological process or proteins responsible for regulation via either protein-DNA or protein-protein interactions. To investigate these conflicting observations, we used the results of GAL80 deletion microarray expression experiments (14). We noticed that in the experiments where GAL80 is deleted and in the absence of galactose a surprisingly large set of genes is highly overexpressed (compared with wild-type yeast grown in the presence of galactose), and many of these genes are involved in sugar metabolism.

We investigated the connectivity between GAL80 and a set of genes S that represents all genes whose average log ratio of expression is greater than $+2.0$ in these conditions. To test the ability of our framework to find plausible regulatory subnetworks, two new experiments were conducted. In the first experiment, we weighted the nodes of the interaction graph using the function w_d (nodes are weighted as a function of the p value of their differential expression, see "Experimental Procedures"). If the proteins that play a regulatory role "between" GAL80 and the other members of S show sufficient co-expression, then intuitively the algorithm should find Steiner trees containing plausible regulatory proteins. In the second experiment, we used function w_1 (all nodes are equally weighted regardless of their expression ratio, see "Experimental Procedures"). If it is not the case that there is sufficient differential expression among the regulatory proteins, then intuitively it may be more effective to simply look

TABLE I

Cost of shortest and most likely paths connecting GAL80 to the remaining elements of S in an optimal Steiner tree under two weighting schemes ω_1 and ω_d

M = metabolic is yes (Y) if the protein is an enzyme in a metabolic pathway. $|V|$ is the number of nodes in the path connecting the protein to GAL80. B denotes edges unique to BIND. S denotes edges unique to SCPD. T denotes edges unique to TRANSFAC. C denotes edges unique to the ChIP-Chip data at $p = 0.001$. TS denotes edges in both T and S. Row "Average" denotes the average path length of elements of *S* to GAL80 in the interaction network. Row "Random" denotes the average path length of all yeast proteins to GAL80 in the interaction network.

Name		M	Node weight w_1							Node weight w_d						
			$\|V\|$	B	S	T	C	TS	TSC	$\|V\|$	B	S	T	C	TS	TSC
ATO3	YDR384C	Y	5	0	0	0	3	0	1	8	4	0	1	2	0	0
ACS1	YAL054C	Y	5	2	0	0	1	0	1	11	5	0	3	2	0	0
ICL1	YER065C	N	5	3	0	1	0	0	0	14	10	0	2	1	0	0
PHO89	YBR296C	N	6	1	0	0	3	0	1	13	9	0	1	1	0	1
JEN1	YKL217W	N	6	0	0	2	0	2	1	9	5	0	1	2	0	0
PUT1	YLR142W	Y	5	0	0	1	2	0	1	9	5	0	0	3	0	0
Average			5.3	1.0	0.0	0.7	1.5	0.3	0.8	10.7	6.3	0.0	1.3	1.8	0.0	0.2
Random			5.3	2.1	0.0	0.5	1.0	0.1	0.5	8.0	5.0	0.0	1.0	0.9	0.0	0.0

for the smallest Steiner trees connecting the elements of *S*.

Table I depicts the lengths of paths between GAL80 and each remaining element of *S* in an optimal Steiner tree using either w_1 or w_d. The columns labeled $|V|$ contain the number of nodes between GAL80 and elements of *S*. The first observation is that the average length of such paths is 10.7 when the w_d function is used but only 5.3 when w_1 is chosen. We take a 2-fold increase to be significant. Additionally we note that the algorithm includes many more edges from the BIND data base when using w_d (~65% of the edges) as compared with w_1 (~25% of the edges). It tends to be very difficult to construct plausible arguments establishing that the paths found under w_d are biologically interesting. BIND contains many interactions from high throughput projects, and there have been several studies suggesting that more than 40% of the interactions in these data sets are not physiologically relevant (6). Although it would be very interesting to find paths that use one or two protein-protein interactions (and thus data from BIND should be included), we are less confident that larger linear chains of protein-protein interactions reflect biological reality; larger chains increase the probability that a false positive interaction pairing has been included. Conversely it is interesting to note that the inclusion of ChIP-Chip data with a conservative *p* value threshold of 0.001 does not seem to cause the same types of problems. Many of the edges in the networks we found that originate from this data set seem plausible or are at least good targets for further study.

Because the set *S* includes all proteins that exhibit the most up-regulation when GAL80 is deleted, it had seemed plausible that the average length of a path from an element of *S* to GAL80 (under either w_1 or w_d weighting functions) would be shorter than the average length of a path from a randomly chosen node. We calculated the average length of a path for each yeast protein to GAL80 and found that the length of these paths is the same as for elements of *S*. This was somewhat surprising, although in retrospect it may simply be a result of the scale-free nature of the interaction graph.

Fig. 1 depicts an optimal Steiner tree for *S* using the w_1 function. (There can be more than one Steiner tree with the same cost. We depict one arbitrarily chosen tree from the small set of optimal trees.) Note that all paths between elements of *S* and GAL80 "pass through" GAL4. This makes some intuitive sense: because GAL80 is known to inhibit GAL4 activity in the absence of galactose, the deletion of GAL80 allows unfettered GAL4 function. It can be hypothesized that GAL4 in turn is somehow responsible for the up-regulation of the remaining elements in *S*.

The Steiner tree of Fig. 1 contains several known interaction subnetworks. For example, the subnetwork formed by CAT8 and its interactors corresponds to the well studied diauxic shift response of yeast. This response is known to require CAT8, which up-regulates genes involved in gluconeogenesis (such as PCK1 and FBP1) and the glyoxylate cycle (ICL1) as well as in the production of acetyl-CoA from acetate (ACS1), the import of succinate into mitochondria (SFC1), and the import of lactate in the cell (JEN1) (24). Three of these diauxic shift CAT8-activated genes are part of the distinguished set *S* (ICL1, JEN1, and ACS1), and three additional genes (FBP1, PCK1, and SFC1) are found to be highly overexpressed in the conditions under study. These latter three genes have in fact log expression ratios between 1.7 and 1.9, marginally below our 2.0 threshold for differential expression (depicted as *shaded* nodes in Fig. 1).

The HAP activator complex is also known to increase the expression of its target genes, which are mostly involved in the tricarboxylic acid cycle and the respiratory chain during the diauxic shift. HAP4, the subunit that provides the principal activation function of the HAP complex, is a Steiner point in the solution for the distinguished set. It allows for the connection of two highly overexpressed genes PHO89 and PUT1. These are likely both required during a diauxic shift (to ensure the appropriate concentration of inorganic phosphate and nitrogen, respectively, in the cell during this cellular response (25)). Taken together, the solution found by the Steiner pro-

(a)

Systematic Name	Common Name	Differential Expression
YML051W	GAL80	−1.6
YER065C	ICL1	2.2
YKL217W	JEN1	2.2
YAL054C	ACS1	2.2
YDR384C	ATO3	2.2
YBR296C	PHO89	2.1
YLR142W	PUT1	2.0
YLR377C	FBP1	1.7
YKR097W	PCK1	1.7
YJR095W	SFC1	1.8
...	...	...

(b)

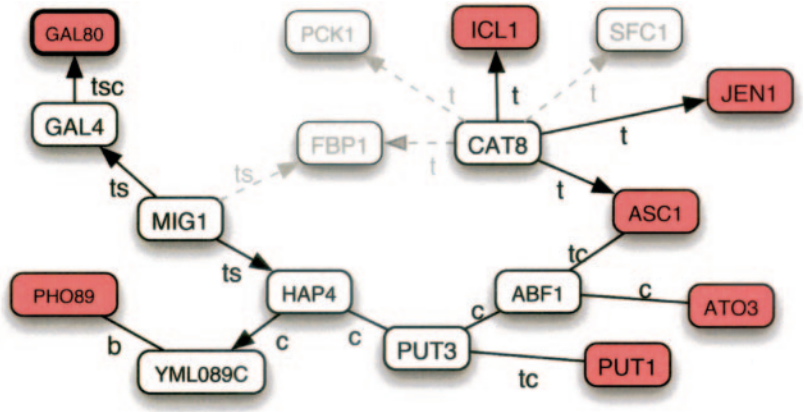

FIG. 1. *a*, a list of distinguished genes with their corresponding measure of differential expression. *b*, a Steiner tree backbone connecting target GAL80 (*red*, *highlighted*) with the set of proteins exhibiting ≥2-fold differential expression (*red nodes*). *t* denotes an arc from TRANSFAC. *c* denotes an arc from the ChIP-Chip data. *b* denotes an edge from BIND. *s* denotes an arc from SCPD. *Shaded nodes* represent proteins where the corresponding genes show slightly less than a factor 2 differential expression. *GAL80*, transcriptional regulator involved in the repression of GAL genes in the absence of galactose, inhibits transcriptional activation by Gal4p; *GAL4*, DNA-binding transcription factor required for the activation of the GAL genes in response to galactose; *MIG1*, transcription factor involved in glucose repression; *HAP4*, subunit of the glucose-repressed Hap2p/3p/4p/5p CCAAT-binding complex, a transcriptional activator of respiratory gene expression; *YML089C*, hypothetical; *PHO89*, Na^+/P_i cotransporter; *PUT3*, positive regulator of PUT (proline utilization) genes; *PUT1*, proline oxidase; *ABF1*, DNA binding protein involved in transcriptional activation and gene silencing; *ATO3*, plasma membrane protein, possible role in export of ammonia from the cell; *ACS1*, acetyl-CoA synthetase isoform, expressed during growth on nonfermentable carbon sources; *CAT8*, zinc cluster protein involved in activating gluconeogenic genes; *JEN1*, lactate transporter whose expression is derepressed by transcriptional activator Cat8p under nonfermentative growth conditions; *SFC1*, mitochondrial succinate-fumarate transporter, required for ethanol and acetate utilization; *ICL1*, isocitrate lyase, responsible for a key reaction of the glyoxylate cycle; *PCK1*, phosphoenolpyruvate carboxykinase, key enzyme in gluconeogenesis; *FBP1*, fructose-1,6-bisphosphatase, required for glucose metabolism.

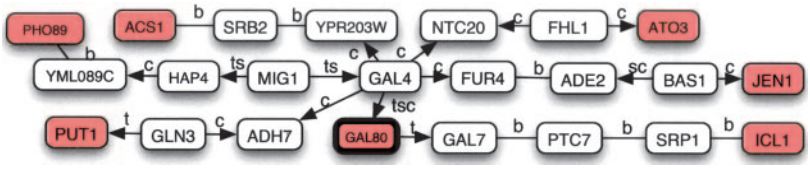

FIG. 2. **The backbone created for the GAL80 distinguished set using the union of the shortest paths to GAL80.** *Dark nodes* represent members of the distinguished set. *Edges* are labeled according to the conventions followed in Fig. 1.

gram suggests that, under these particular growth conditions, the yeast underwent a diauxic shift. This is remarkable because the yeast were exponentially growing in a 2% raffinose solution (14). Under such conditions, glycolysis is the major energy-providing pathway. It has been shown that activation of diauxic shift genes in wild-type yeast occurs well after the initial exponential growth phase (26). Because our solution is based upon GAL80 deletion expression data, this suggests that GAL80 may play a role in controlling the shift from glucose to ethanol metabolism.

Alternative and Augmented Backbones for S—We also experimented with backbones created by the unions of the shortest path from GAL80 to each member of the distinguished set (see "Other Strategies for Discovering Regulatory Subnetworks" under "Experimental Procedures"). Fig. 2 depicts the results. In this analysis, it is interesting to note that apart from GAL4 no other node is common to more than one path between GAL80 and elements of the distinguished set *S*. In several cases, this alternative strategy provided very little insight into the regulatory relationships between the elements of *S*. (This approach, which essentially studies each element of *S* individually with respect to GAL80, is more similar to traditional non-automated analyses one would perform using only the literature.) As such, it renders difficult the task of explaining why the elements of *S* are simultaneously highly up-regulated under the conditions studied. For example, the backbone described in Fig. 2 does not allow us to conclude that yeast is undergoing a diauxic shift under these conditions. This illustrates an important advantage of the full Steiner analysis when querying relationships among a group of proteins.

To further illustrate the power of our approach for querying interaction networks, we explored the "neighborhood" of the Steiner tree in Fig. 1. If we were to include all nodes that are adjacent to the tree ((1,1)-augmented tree), the subgraph would have over 437 nodes. The analysis of so large a number of interactions certainly would be arduous. The (2,1)-augmented tree contains 1,600 nodes. However, when we exam-

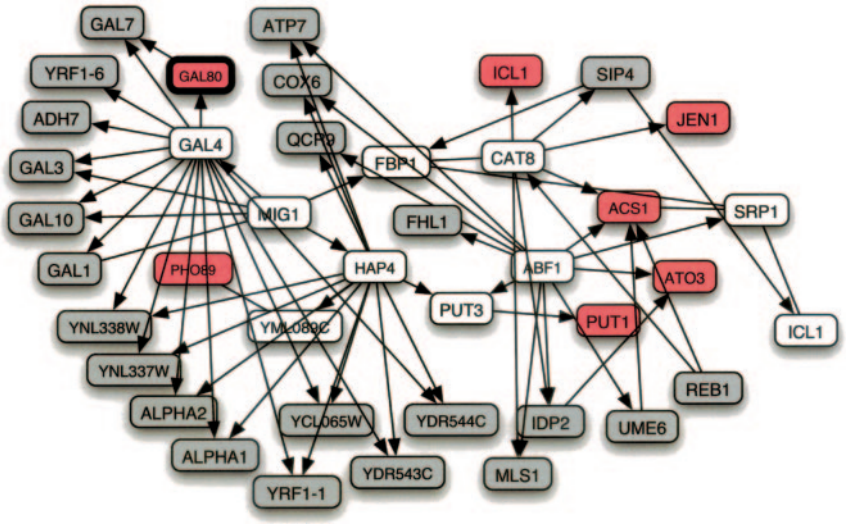

FIG. 3. **The (1,2)-augmented tree for the GAL80 backbone.** *Dark gray nodes* represent proteins not in the backbone but having a distance of 1 from the backbone and at least two distinct (undirected) paths to the backbone.

TABLE II
Distinguished set for a 29 to 37 °C heat shock experiment

Systematic name	Common name	log expression ratio
YFL014W	HSP12	7.608
YER103W	SSA4	6.844
YMR105C	PGM2	6.445
YHR087W	YHR087W	6.271
YGR248W	SOL4	5.718
YPR160W	GPH1	5.665
YML128C	MSC1	5.43
YFR053C	HXK1	5.265
YLR178C	TFS1	5.265
YML100W	TSL1	5.251
YMR250W	GAD1	5.246
YLL026W	HSP104	5.224

ine the (1,2)-augmented tree, we are left with only 25 additional nodes. The (1,3)-augmented tree contains only one additional node. This augmented backbone is depicted in Fig. 3. Note that some of these nodes have extremely high degrees of connectivity in the full interaction network and that many GAL genes appear in this set.

Transcriptional Regulatory Networks Versus Interaction Networks—We used a heat shock microarray data set available from Gene Expression Omnibus (platform GPL51, series GSE18 (16)) to investigate subnetworks found by the Steiner approach when only protein-DNA or both protein-DNA and protein-protein interactions are used. This data set is comprised of samples collected 20 min after yeast cells are exposed to a temperature shift from either 17, 21, 25, 29, or 33 to 37 °C. The goal is to identify plausible regulatory subnetworks that connect the most highly overexpressed genes in these experiments. Results shown here use the data from the 29 to 37 °C temperature shift experiment (other temperature shift experiments generate similar subnetworks, data not shown). The distinguished set is composed of the 12 most highly overexpressed genes between these conditions (Table II). Two types of experiments were conducted to study the usefulness of using different underlying networks.

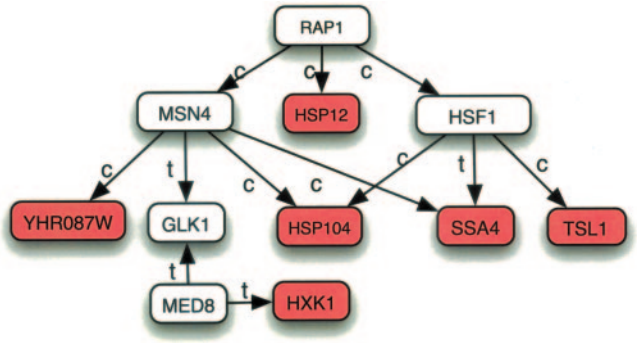

FIG. 4. **Transcriptional subnetwork for the 12 most highly overexpressed yeast genes in the temperature shift from 29 to 37 °C (16).** Here the interaction network is restricted to only protein-DNA interactions. *Edges* are labeled according to the conventions followed in Fig. 1. *RAP1*, repressor activator protein; *MSN4*, transcription factor activated in stress conditions; *HSF1*, heat shock transcription factor; *HSP12*, plasma membrane-localized protein that protects membranes from desiccation, induced by heat shock; *YHR087W*, hypothetical; *HSP104*, responsive to heat shock; *SSA4*, member of 70-kDa heat shock protein family; *TSL1*, subunit of trehalose-6-phosphate synthase/phosphatase complex; *GLK1*, glucokinase; *MED8*, member of RNA polymerase II transcriptional regulation mediator; *HXK1*, hexokinase isoenzyme 1. The remaining six differentially expressed genes from Table II were not present in the interaction network restricted to protein-DNA interactions.

If we restrict our attention to only protein-DNA interactions, the Steiner tree approach searches for plausible transcriptional subnetworks regulating the distinguished set of genes. One such transcriptional subnetwork is depicted in Fig. 4.

In this subnetwork, the general transcription factor RAP1 is shown to influence the transcription of the HSP12 gene (which encodes a plasma membrane-located heat shock protein) and of the genes encoding the MSN4 and HSF1 proteins. Both MSN4 and HSF1 are transcription factors known to be up-regulated under some stress conditions (27). Although

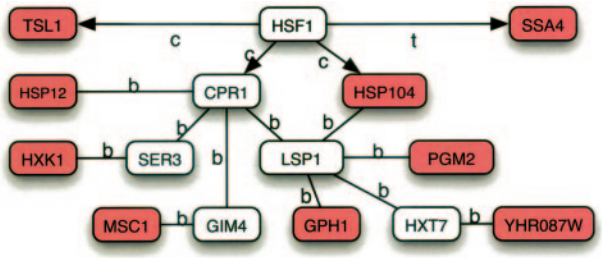

FIG. 5. **Interaction subnetwork for the 12 most highly overexpressed yeast genes in the temperature shift from 29 to 37 °C (16).** The interaction network includes both protein-protein and protein-DNA interactions. The remaining three differentially expressed genes from Table II were not present in the interaction network. *Edges* are labeled according to the conventions followed in Fig. 1.

both transcription factors have unique transcriptional binding targets, they share two targets in this subnetwork: the genes encoding proteins HSP104 and SSA4. The coordination between these two transcription factors has been noted in the case of the transcriptional regulation of HSP104 (28, 29) but not in the case of SSA4, suggesting that this mechanism of stress response might be more widely used than previously thought. This subnetwork shows a highly structured transcriptional response to heat shock in yeast. A previous study reported an MSN4- and HSF1-dependent increase in protein abundance of several members of our subnetwork following heat shock treatment (27). More specifically, increases in the protein levels of HXK1, GLK1, HSP12, and enzymes involved in trehalose metabolism (although not TSL1) were shown to be dependent upon MSN4. Increases in the protein levels of HSP104, SSA4, and HSP12 were shown to be dependent upon HSF1 following heat shock treatment. Our subnetwork suggests a structured transcriptional response that can explain these dependences. We note that although it would have been possible (but probably long and arduous) to identify some parts of this Steiner solution by literature searches, other key links would have been very difficult to identify because they originate from ChIP-Chip experiments. The Steiner approach thus provides a rapid way of constructing plausible regulatory networks of a manageable size that can be subsequently expanded and used to direct effective literature searches and further experiments (much more so than simply using the distinguished set directly). Furthermore the Steiner solution provides additional information when hypothetical proteins are part of the distinguished set. For example, the hypothetical protein YHR087W, which is up-regulated under conditions of heat shock, seems to be the transcriptional target of MSN4 as are the HSP104, SSA4, and GLK1 proteins according to information provided by the Steiner solution.

When both protein-DNA and protein-protein interactions are used in the underlying interaction network, the Steiner tree algorithm provides solutions containing elements of transcriptional regulation as well as direct protein-protein regulation. Fig. 5 shows such a subnetwork for the heat shock-distinguished set of genes. In this subnetwork, the only transcription factor present, HSF1, influences the transcription of four genes that encode the proteins SSA4, TSL1, HSP104, and CPR1. HSP104 and CPR1 are involved in protein-protein interactions with other elements of the subnetwork, including LSP1, which is a central Steiner point in this subnetwork.

The LSP1 protein is believed to play a role in heat stress resistance and to negatively regulate the kinase Pkh1p and downstream signaling pathways PKC1-mitogen-activated protein and YPK1 (30). Several elements of this subnetwork are involved in carbohydrate metabolism including HXK1, GPH1, HXT7, TSL1, and PGM2. This has been noted previously and is possibly caused by the increase in ATP utilization following heat shock (16). Overall this subnetwork suggests an important role for chaperone protein complexes and proteins involved in carbohydrate metabolism pathways following heat shock treatment.

It should be noted that the subnetworks in Figs. 4 and 5 provide useful and different information. Although the transcriptional network in Fig. 4 is easier to interpret and readily suggests a sequence of events (or possibly "causation") following heat shock treatment, Fig. 5 does connect more elements of the distinguished set (because the interaction network is much larger when protein-protein interactions are included) and provides some insight as to which protein complexes are involved in the response.

Finding Regulators of Pathways—Ihmels *et al.* (17) estimate that approximately half of the metabolic proteins in any pathway exhibit co-expression. Our methodology allows us to further evaluate this claim and to identify regulators of these metabolic proteins in the interaction network. We use the 20 microarray results from Ideker *et al.* (14) containing wild-type and single deletions of GAL genes, 10 metabolic pathways from Saccharomyces Genome Database (25), and an interaction network consisting of data from only TRANSFAC, ChIP-Chip, and SCPD (BIND is excluded from this analysis because we want to only look at transcriptional regulation).

The pathways chosen can be divided into two groups based on their proximity to the galactose pathway in the yeast metabolic network. Table III lists these pathways. The "Close to galactose" category contains those pathways that are nearly adjacent to the galactose pathway, whereas the "Far from galactose" category contains several pathways that are not directly related to galactose. For each pathway, we begin by identifying a set of co-expressed genes based on their pairwise correlation coefficients of the log ratio expression levels over all 20 conditions. A subset of the metabolic proteins were chosen if their correlation coefficient was sufficiently high (a conservative >0.35). We computed a Steiner tree using the weight function w_1 for the set of co-expressed genes in each pathway. If the co-expression of these genes is due to the fact that the genes share a common set of regulators, then we expect to find relatively small Steiner trees. If so, this may in turn imply that a small number of related

TABLE III
Steiner transcriptional analysis of metabolic pathways

N is the total number of metabolic proteins in the pathway. *C* is the total number of co-expressed genes. $|V|$ is the size of the Steiner tree (number of nodes). If the Steiner tree is disconnected, we examined only the tree in this forest with the maximum number of leaves (*L*). *R* is the average size of a Steiner tree with *L* leaves.

Pathway	*N*	*C*	*L*	$\|V\|$	*R*	*p* value
Close to galactose						
Galactose metabolism	5	3	3	4	7.5	0.02
Krebs cycle	21	9	3	8	7.5	0.81
Gluconeogenesis	19	9	8	11	18.3	0
Glycolysis	18	11	9	11	20.3	0
Glucose fermentation	29	15	12	18	25.7	0
Far from galactose						
Pentose phosphate pathway	9	3	2	5	5	0.72
De novo pyrimidine biosynthesis	10	5	4	7	10.2	0.02
Pantothenate and coenzyme A biosynthesis	13	5	2	5	5	0.72
Ergosterol biosynthesis	20	6	2	4	5	0.25
Lysine biosynthesis	7	5	4	9	10.2	0.31

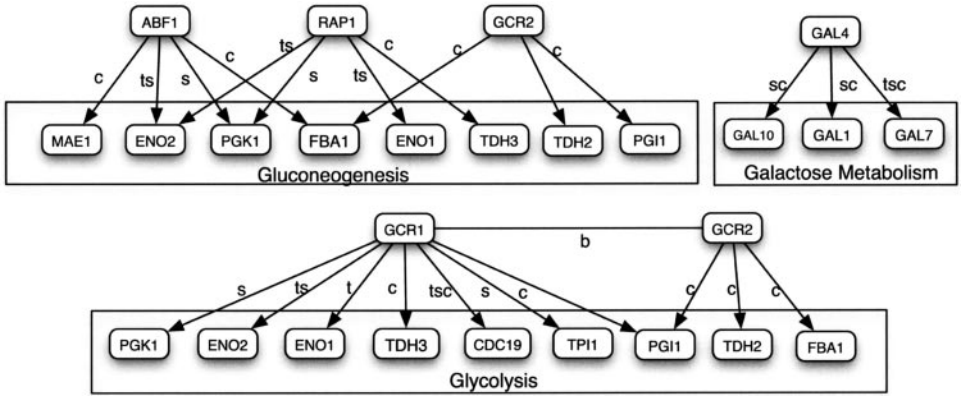

FIG. 6. **A subset of the gluconeogenesis, galactose, and glycolysis pathways exhibiting strong co-expression (*nodes within boxes*) and the resulting Steiner trees.** *Edges* are labeled according to the conventions followed in Fig. 1.

transcription factors are responsible for the expression of these genes.

Our results indicate that the size of the Steiner trees for the set of pathways close to the galactose pathway are significantly below the expected size of a random tree (Table III, columns $|V|$, *R*, and *p* value) with the exception of the Krebs cycle. The *p* values were estimated empirically by generating random sets of size $|V|$. Fig. 6 shows three interesting examples of very small Steiner trees for co-expressed genes in metabolic pathways. In the gluconeogenesis pathway, there exists a Steiner tree for eight of nine of the co-expressed genes with only three Steiner points: ABF1, RAP1, and GCR2. ABF1 is known to be a multifunctional transcription factor with a role in carbon source regulation (25). The RAP1 transcription factor has been shown to target almost 300 yeast genes, including ENO1, ENO2, FBA1, and PGK. GCR2 is an important transcription factor for the activation of glycolysis (FBA1 and TDH2 are involved in both glycolysis and gluconeogenesis). Note that this tree has many fewer Steiner points than expected by a random sampling, and the edges originate from a variety of data sources including the ChIP-Chip source.

A second example is the Steiner tree formed by the co-expressed genes of the galactose pathway. GAL4 is the unique Steiner point. This is as expected because it is known to be the central transcription factor in the regulation of the galactose pathway (2). A third example consists of the co-expressed genes of the glycolysis pathway where nine of 11 are present in the network and can be connected together by only two Steiner points, GCR1 and GCR2. Both Steiner points are known to positively regulate the transcription of glycolytic genes and may function in a complex together (31, 32).

The pathways that are far from the galactose pathway serve as a control and, as expected, have Steiner trees that are close to the expected size of a random tree (Table III, *R* and *p* value columns). Because the expression experiments perturb the galactose pathway (through the deletion of GAL genes), the genes we identify as co-expressed in these pathways are likely unrelated to galactose metabolism and have no obvious physiological relevance (because this set of genes is essentially chosen randomly, we do not expect to detect co-regu-

lation). This explains why the sizes of the optimal Steiner trees are not significantly different from the size of random trees for these pathways.

DISCUSSION

This study develops a method that generates insights into the regulatory relationships between a set of genes/proteins. Such sets of distinguished genes/proteins appear more frequently in the literature as the results of high throughput gene expression, protein expression, chemical-genetic, and genetic interaction platforms are published. Unfortunately, from the sets themselves, it is often difficult to ascertain gene/protein significances and relationships. The interaction networks generated using standard protein interaction visualization tools quickly become too complex for relevant information to be extracted even when the distinguished sets are of moderate size. Our framework allows for a backbone containing likely relevant proteins to be computed quickly. Additional heuristics allow for the augmentation of this backbone in biologically meaningful ways. We believe that our approach provides an extremely useful methodology for the extraction of biologically salient information from interaction networks.

Our framework has identified an interesting subnetwork that suggests a new connection between GAL80 and the diauxic shift. The diauxic shift occurs when glucose and other fermentable carbon sources have been exhausted in the growth medium; this was not the case in the experimental conditions used by Ideker *et al.* (14). Only those cultures with a GAL80 deletion mutant resulted in elevated expression levels of diauxic shift genes. We propose that these cells are prematurely activating or priming the diauxic shift response and that GAL80 contributes to the correct regulation of induction of this response in wild-type cells. It has been observed that GAL80-null cells grow slowly in raffinose (14), and this fact may be a reflection of the competition between two opposing pathways in these cells.

We also show that our framework when applied to graphs containing only protein-DNA interactions is capable of finding plausible transcriptional regulatory subnetworks. In particular, the framework shows that the general transcription factors RAP1 and HSF1 directly regulate a third of the most overexpressed genes in a temperature shift from 29 to 37 °C.

Gene expression experiments often incorrectly identify genes as being differentially expressed due to the technical and biological errors associated with microarrays. The subnetworks approach in this study can complement current statistical approaches to locating these false positives. If genes from the distinguished set are not closely connected to remaining members (and are therefore not physiologically related to the remaining members), such genes are possibly more likely to be simply experimental artifacts and could be considered less interesting. It is estimated that current transcriptional regulatory networks capture only 10% of existing transcriptional regulatory relationships (33). Such false negatives in the interaction network may result in Steiner trees containing longer paths that are not biologically relevant (or at least very difficult to rationalize). False positive edges in the interaction network may also cause our approach to return irrelevant trees. Our implementation explicitly labels edges with the source of the interaction to help investigators navigate such issues. Nevertheless, as the accuracy of the underlying interaction networks improves, the framework will be able to find more plausible regulatory subnetworks.

The framework presented here is a first step. Better weight functions for the nodes in the interaction network remain to be developed to determine which such functions most closely reflect biological reality. Expression measurements alone are unlikely to return meaningful results in many contexts. It is also possible to weight (with likelihoods) our confidence in each interaction and integrate more types of data into the interaction networks. Weightings that incorporate Gene Ontology classifications and algorithms that use this information to make choices as to how to build and augment the backbone may provide a solid improvement. The challenge will be to find accurate ways to compute these more complex problems.

Acknowledgments—We thank G. Finak, J.-A. Valencia, and N. Betzler for help implementing this work.

* The material presented in this paper is based upon work supported by the National Science Foundation under a grant awarded in 2002. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¶ Supported by Canadian Institutes of Health Research Canada Graduate Scholarship.

‖ Funded by Genome Québec.

** To whom correspondence should be addressed. Tel.: 514-398-5928; E-mail: hallett@mcb.mcgill.ca.

REFERENCES

1. Powell, K. (2004) All systems go. *J. Cell Biol.* **165,** 299–303
2. Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18,** 233–240
3. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002) Transcriptional regulatory networks in *Saccaromyces cerevisiae. Science* **298,** 799–804
4. Shenn-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli. Nat. Genet.* **31,** 64–68
5. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., and Tyers, M. (2002) Systematic identifi-

cation of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415,** 180–183
6. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 4569–4574
7. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* **403,** 623–627
8. Oyama, T., Kitano, K., Satou, K., and Ito, T. (2002) Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics* **18,** 705–714
9. Lappe, M., Park, J., Niggemann, O., and Holm, L. (2001) Generating protein interaction maps from incomplete data: application to fold assignment. *Bioinformatics* **17,** 149–156
10. Kelley, B. P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B., and Ideker, T. (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.* **32,** W83–W88
11. Parsons, A. B., Brost, R. L., Ding, H., Li, Z., Zhang, C., Sheikh, B., Brown, G. W., Kane, P. M., Hughes, T. R., and Boone, C. (2003) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat. Biotechnol.* **22,** 62–69
12. 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,** 242–245
13. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S., and Urbach, S. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* **29,** 281–283
14. Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001) Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science* **292,** 929–934
15. Jansen, R., Greenbaum, D., and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12,** 37–46
16. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11,** 4241–4257
17. Ihmels, J., Levy, R., and Barkai, N. (2004) Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae. Nat. Biotechnol.* **22,** 86–92
18. Dreyfus, S. E., and Wagner, R. A. (1971) The Steiner tree problem in graphs. *Networks* **1,** 195–207
19. Karp, R. (1972) in *Complexity of Computer Computations* (Miller, R. E., and Thatcher, W., eds) pp. 85–103, Plenum, New York
20. Klein, P., and Ravi, R. (1995) A nearly best-possible approximation algorithm for node-weighted Steiner trees. *J. Algorithms* **19,** 104–115
21. Zelikovsky, A. (1993) An 11/6-approximation algorithm for the network Steiner problem. *Algorithmica* **9,** 463–470
22. Hu, Z., Mellor, J., Wu, J., and DeLisi, C. (2004) VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics* **5,** 17
23. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003) CytoScape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504
24. Haurie, V., Perrot, M., Mini, T., Jeno, P., Sagliocco, F., and Boucherie, H. (2001) The transcriptional activator Cat8p provides a major contribution to the reprogramming of carbon metabolism during the diauxic shift in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **276,** 76–85
25. Issel-Tarver, L., Christie, K. R., Dolinski, K., Andrada, R., Balakrishnan, R., Ball, C. A., Binkley, G., Dong, S., Dwight, S. S., Fisk, D. G., Harris, M., Schroeder, M., Sethuraman, A., Tse, K., Weng, S., Botstein, D., and Cherry, J. M. (2002) Saccharomyces Genome Database. *Methods Enzymol.* **350,** 329–346
26. DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278,** 680–686
27. Boy-Marcotte, E., Lagniel, G., Perrot, M., Bussereau, F., Boudsocq, A., Jacquet, M., and Labarre, J. (1999) The heat shock response in yeast: differential regulations and contributions of the Msn2p/Msn4p and Hsf1p regulons. *Mol. Microbiol.* **33,** 274–283
28. Amoros, M., and Estruch, F. (2001) Hsf1p and Msn2/4p cooperate in the expression of *Saccharomyces cerevisiae* genes HSP26 and HSP104 in a gene- and stress type-dependent manner. *Mol. Microbiol.* **39,** 1523–1532
29. Grably, M. R., Stanhill, A., Tell, O., and Engelberg, D. (2002) HSF and Msn2/4p can exclusively or cooperatively activate the yeast HSP104 gene. *Mol. Microbiol.* **44,** 21–35
30. Zhang, X., Lester, R. L., and Dickson, R. C. (2004) Pil1p and Lsp1p negatively regulate the 3-phosphoinositide-dependent protein kinase-like kinase Pkh1p and downstream signaling pathways Pkc1p and Ypk1p. *J. Biol. Chem.* **279,** 22030–22038
31. Uemura, H., Koshio, M., Inoue, Y., Lopez, M. C., and Baker, H. V. (1997) The role of Gcr1p in the transcriptional activation of glycolytic genes in yeast *Saccharomyces cerevisiae. Genetics* **147,** 521–532
32. Uemura, H, and Jigami, Y. (1992) Role of GCR2 in transcriptional activation of yeast glycolytic genes. *Mol. Cell. Biol.* **12,** 3834–3842
33. Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* **431,** 99–104
34. Zhu, J., and Zhang, M. Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae. Bioinformatics* **15,** 607–611