BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S233–S240

Discovering regulatory and signalling circuits in molecular interaction networks

Trey Ideker[1,*], Owen Ozier[1], Benno Schwikowski[2] and Andrew F. Siegel[2, 3]

[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA, [2]Institute for Systems Biology, Seattle, WA 98103, USA and [3]Departments of Management Science, Finance, Statistics, and Genome Sciences, University of Washington, Seattle, WA 98195, USA

Received on January 24, 2002; revised and accepted on April 1, 2002

ABSTRACT

Motivation: In model organisms such as yeast, large databases of protein–protein and protein-DNA interactions have become an extremely important resource for the study of protein function, evolution, and gene regulatory dynamics. In this paper we demonstrate that by integrating these interactions with widely-available mRNA expression data, it is possible to generate concrete hypotheses for the underlying mechanisms governing the observed changes in gene expression. To perform this integration systematically and at large scale, we introduce an approach for screening a molecular interaction network to identify active subnetworks, i.e., connected regions of the network that show significant changes in expression over particular subsets of conditions. The method we present here combines a rigorous statistical measure for scoring subnetworks with a search algorithm for identifying subnetworks with high score.

Results: We evaluated our procedure on a small network of 332 genes and 362 interactions and a large network of 4160 genes containing all 7462 protein–protein and protein-DNA interactions in the yeast public databases. In the case of the small network, we identified five significant subnetworks that covered 41 out of 77 (53%) of all significant changes in expression. Both network analyses returned several top-scoring subnetworks with good correspondence to known regulatory mechanisms in the literature. These results demonstrate how large-scale genomic approaches may be used to uncover signalling and regulatory pathways in a systematic, integrative fashion.

Availability: The methods presented in this paper are implemented in the *Cytoscape* software package which is available to the academic community at http://www.cytoscape.org.

Contact: trey@wi.mit.edu

Keywords: molecular interactions; gene expression; data integration; simulated annealing; Monte carlo methods.

INTRODUCTION

Expression profiling and large-scale proteomics have revolutionized biology by generating vast amounts of data about cell state. Genes with significant changes in expression have immediate and widespread interest as markers for diseases, stages of development, and a variety of other cellular phenotypes (Altman and Raychaudhuri, 2001). Genes with correlated expression changes over many conditions are likely to be involved in similar functions or cellular processes; these genes often also share DNA sequence elements, providing evidence that they are regulated by common transcription factors. Analytical methods such as gene expression clustering (Eisen *et al.*, 1998; Tamayo *et al.*, 1999), significance testing (Kerr and Churchill, 2001; Rocke and Durbin, 2001; Ideker *et al.*, 2000), and sequence motif identification (Pilpel *et al.*, 2001) have been indispensable for enabling these discoveries and summarizing the data at each step. By performing these analyses, we hope ultimately to answer questions about the underlying molecular mechanism: *What are the signalling and regulatory interactions in control of the observed gene expression changes? How is this control exerted?*

For model organisms such as yeast, new technologies and data sets are making it possible to address these questions more directly than ever before. For example, systematic two-hybrid screens and co-immunoprecipitation experiments are populating the public databases with thousands of protein–protein interactions and complexes (Uetz *et al.*, 2000; Gavin *et al.*, 2002). Other ongoing projects are defining large numbers of protein→DNA interactions (Ren *et al.*, 2000), and protein microarrays are making it possible to map interactions between proteins and drugs, hormones, and other small molecules

[*]To whom correspondence should be addressed.

(Zhu *et al.*, 2001). These molecular interactions provide a convenient framework for understanding changes in gene expression and for integrating a wide variety of global state measurements.

Along these lines, in recent work we used a molecular interaction network to analyse changes in expression observed over 20 perturbations to the yeast galactose-utilization (GAL) pathway (Ideker *et al.*, 2001). To construct the network (shown in Figure 1), we screened a database of publicly available protein–protein and protein→DNA interactions to select 362 interactions linking genes that were differentially-expressed under one or more perturbations. We found that pairs of genes linked by molecular interactions in this network were more likely to have correlated expression profiles than genes chosen at random, and we reported particular interactions for which this pairwise correlation was strong. However, the general task of associating gene expression changes with higher-order groups of interactions, such as make up signalling and regulatory pathways in the cell, was not discussed.

To address this problem, we now introduce a general method for searching the network to find 'active subnetworks', i.e., connected sets of genes with unexpectedly high levels of differential expression[†]. When expression levels have been observed over multiple conditions, we also wish to determine which conditions significantly affect gene expression in each active subnetwork. In order to achieve these goals, we implement a statistical scoring system which captures the amount of gene expression change in a given subnetwork. We then describe a search algorithm, based on simulated annealing, for identifying the highest scoring subnetworks. We explore the performance of our method in two cases: a small interaction network with expression data from a single condition, and a large interaction network observed over multiple conditions.

METHODS

Basic z-score calculation

To rate the biological activity of a particular subnetwork, we begin by assessing the significance of differential expression for each gene. We use the error model provided by the program *VERA* (Ideker *et al.*, 2000) to obtain p-values p_i representing the significance of expression change[‡] for each gene i. Each p_i is then converted to a z-score $z_i = \Phi^{-1}(1 - p_i)$, where Φ^{-1} is the inverse normal CDF. Thus in random data, p-values are distributed uniformly from 0 to 1 and z-scores follow a standard normal, with smaller p-values corresponding to larger z-scores.

To produce an aggregate z-score z_A for an entire subnetwork A of k genes, we sum the z_i over all genes in the subnetwork:

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \qquad (1)$$

Subnetworks of all sizes are comparable under this scoring system: if the z_i are independently drawn from a standard normal distribution, z_A will also be distributed according to a standard normal[§], independent of k. A high z_A indicates a biologically active subnetwork.

Calibrating z against the background distribution

In order to properly capture the connection between expression and network topology, we must determine whether the score z_A of a subnetwork is higher than expected relative to a random set of genes (drawn from the same expression data but independently of the network). We randomly sample gene sets of size k using a Monte Carlo approach, compute their scores z_A, then use these to derive estimates for the score mean μ_k and standard deviation σ_k for each k. Because we expect the means and standard deviations to be a smooth function of k, we can reduce noise in the Monte Carlo estimates using a sliding window average. Using these estimates, the corrected subnet score s_A is:

$$s_A = \frac{(z_A - \mu_k)}{\sigma_k} \qquad (2)$$

Using this correction, the scores s_A of randomized subnets are guaranteed to have $\mu = 0$ and $\sigma = 1$. An overview of this scoring process is shown in Figure 2a.

Scoring over multiple conditions

Our scoring system may be extended to accommodate gene expression changes measured over multiple conditions as shown in Figure 2b. In this case, we start with a matrix of p-values (genes versus conditions) and corresponding z-scores. Given a subnetwork A, we use eqn. (1) to produce m different aggregate scores $(z_{A1}, z_{A2}, \ldots, z_{Am})$, one for each condition. These are then sorted from highest to lowest $(z_{A(1)}, \ldots, z_{A(j)}, \ldots, z_{A(m)})$; we compute the significance $r_{A(j)}$ of the jth highest score using a binomial

[†] If the network included all interactions (and interaction types) important for cell function, then there would exist a single subnetwork connecting *all* of the genes with significant expression changes. Since the network is clearly missing important interactions, we may potentially find many subnetworks of interest. A search is necessary to select just those interactions relevant to the expression data of interest while rejecting false-positive interactions.

[‡] Available at http://www.systemsbiology.org/VERAandSAM. Any gene expression analysis tool that robustly models error would suffice, so long as its output can be converted to p-values.

[§] This is true because the variance of a sum is the sum of the variances for independent random variables.

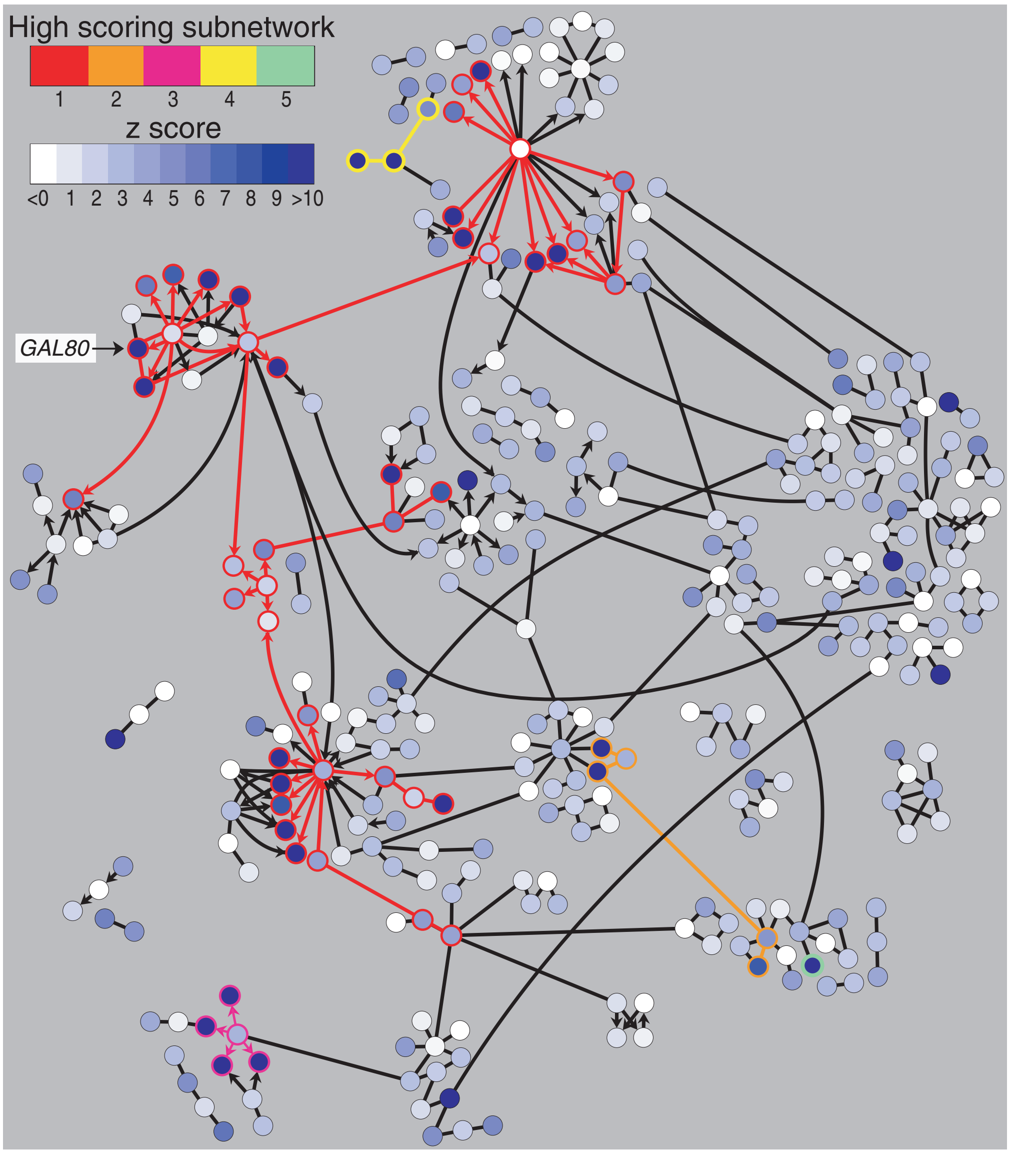

Fig. 1. Performance on a small molecular interaction network. Nodes represent genes, an edge directed from one node to another signifies that the protein encoded by the first gene can influence the transcription of the second by DNA binding (protein→DNA), and an undirected edge between nodes signifies that the corresponding proteins can physically interact. Z-scores (blue scale) indicate the likelihood of differential expression of each gene in a *GAL80* knockout experiment. Z-scores were used to search for active subnetworks using our simulated annealing method; the five top-scoring subnets are shown. For further information on this network, including gene labels, see Figure 4 in (Ideker *et al.*, 2001).

order statistic (Kendall *et al.*, 1987) as follows: Let $P_z = 1 - \Phi(z_{A(j)})$ represent the probability that any single condition has a z-score above $z_{A(j)}$. Then:

$$p_{A(j)} = \sum_{h=j}^{m} \binom{m}{h} (P_z)^h (1 - P_z)^{m-h} \qquad (3)$$

This summation gives the probability that at least j of the m conditions had scores above $z_{A(j)}$, which is equivalent to the p-value for $z_{A(j)}$ as the jth largest of a standard normal sample. We use the inverse CDF $r_{A(j)} = \Phi^{-1}(1 - p_{A(j)})$ to convert back into a standard normal z-score, which is now adjusted for rank; the maximum of these is the subnet's new score, $r_A^{\max}$:

$$r_A^{\max} = \max_j (r_{A(j)}) \qquad (4)$$

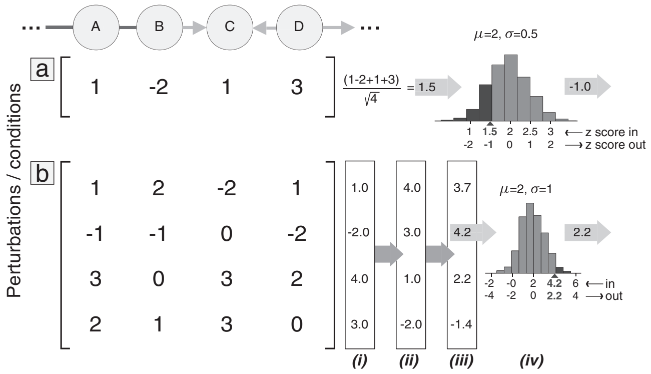

Fig. 2. Scoring an example subnetwork. In panel **[a]**, individual scores z_i are used to compute a single aggregate score $z_A = 1.5$ for the subnet; this value is then compared to the distribution of scores for random gene sets of size $k = 4$ (depicted by the histogram) producing the score of -1.0 by eqn. (2). Panel **[b]** shows the extended procedure for scoring a subnet under multiple conditions. Aggregate z scores for each condition (i) are sorted (ii) and adjusted for rank (iii). As in **[a]**, the maximum score of 4.2 is then corrected to 2.2 using the background score distribution (iv).

We consider the subnetwork to be 'active' under conditions ranked 1 through j. As in the single condition case, we must also calibrate the score against the background distribution: first, a Monte Carlo technique is used to estimate means μ_k and standard deviations σ_k for $r_A^{\max}$ computed from random gene sets of size k; using these estimates, the score $r_A^{\max}$ is corrected to produce the final score s_A.

Searching for high-scoring subnetworks via simulated annealing

The above methods allow us to score a given subnetwork, but we must also find the highest-scoring subnetwork(s) in a full network of molecular interactions. Because the problem of finding the maximal-scoring connected subgraph is NP-hard[¶], we implement an approach based on simulated annealing (Kirkpatrick *et al.*, 1983). In practice, this approach is not guaranteed to find the maximal score overall; however, all high-scoring subnetworks are of strong biological interest regardless of whether they are strictly maximal[‖].

Throughout the following algorithm, we associate an 'active/inactive' state with each node. G_w denotes the 'working' subgraph of G induced by the active nodes. At each iteration i, we define s_i as the score (s_A from above) of the highest-scoring component in G_w.

```
Input: A graph $G = (V, E)$ of molecular
interactions, a number $N$ of iterations,
and a temperature function $T_i$ which
decreases geometrically from $T_{start}$ to $T_{end}$
Output: A subgraph $G_w$ of $G$

(1) Initialize $G_w$ by setting each $v \in V$
    to active/inactive with probability $\frac{1}{2}$;
(2) FOR $i = 1$ to N DO
(3)     Randomly pick a node $v \in V$ and
        toggle its state;
(4)     Compute the score $s_i$ for the
        working subgraph $G_w$;
(5)     IF ($s_i > s_{i-1}$), keep $v$ toggled;
(6)     ELSE keep $v$ toggled with
        probability $p = e^{(s_i - s_{i-1})/T_i}$
(7) Output $G_w$ and its highest-scoring
    component $A$.
```

Finally, we 'quench' at temperature $T_i = 0$ until all adjoining possibilities have been explored and the score has reached a local maximum. Upon termination of annealing, the subnetwork A represents a signalling or regulatory circuit of high biological interest[**].

Heuristics for improved annealing

We have extended our annealing approach to search for M subnetworks simultaneously (where M is a user-definable parameter). Maintaining multiple components can dramatically improve the efficiency of annealing, because toggling the state of a node may cause a number of components with low score to merge into a single high-scoring component[††].

An additional heuristic increases the efficiency of annealing in networks with many 'hubs', i.e., nodes of high degree. Simulated annealing tends to perform poorly in such networks, because adding a hub to G_w immediately creates a large component incorporating all nodes adjacent to the hub. Unless all adjacent nodes with low z-score are inactive, the resulting component will also have low score regardless of the contribution of the hub node itself. We have addressed this problem by a straightforward modification to step (3) of the algorithm: when adding a node of degree greater than a user-definable parameter $d_{\min}$, simultaneously remove all neighbours that are not in the top-scoring component.

[¶] R. Karp, *personal communication*. A proof is shown in supplementary materials at http://www.cytoscape.org/ISMB2002/.

[‖] Given suitable values for annealing parameters T_i and N, the score of the final solution is guaranteed to be the global maximum (Lundy and Mess, 1986). However, these values are generally unknown and can be impractically large.

[**] This search is robust with respect to false-positive interactions, because adding an interaction never disrupts an existing subnetwork. With respect to false negatives, it is possible for a missing interaction to prevent the formation of an otherwise high-scoring subnetwork; however, the remaining pieces of the subnetwork will still score well.

[††] Full details are provided in the supplementary web materials.

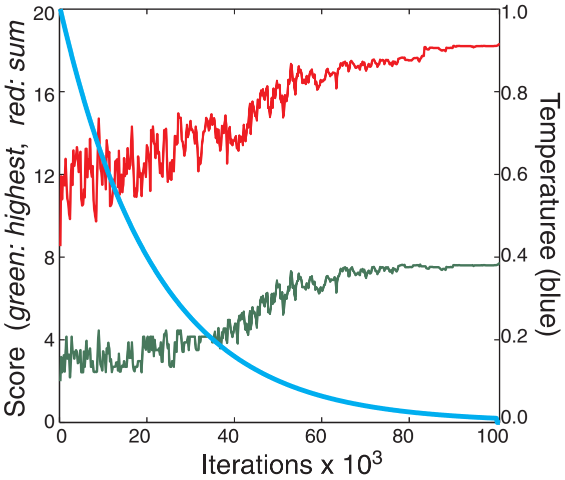

Fig. 3. Score and temperature versus number of iterations. Simulated annealing was performed for the network and expression profile shown in Fig. 1, with parameters $N = 1 \times 10^5$, $T_{\text{start}} = 1$, $M = 5$, $d_{\text{min}} = 100$. Annealing temperature (right vertical axis; solid grey trace) decreases geometrically over consecutive iterations according to the set schedule. By the end of the run, scores for each of the five top scoring subnetworks have increased to a local maximum (left vertical axis; solid black trace = top score, dashed = sum of all five).

RESULTS

Small network with a single perturbation

We searched for active subnetworks in the network of 362 protein–protein and protein→DNA interactions from our previous galactose utilization study (Ideker *et al.*, 2001). This search was performed against gene expression changes measured in response to a single perturbation: a strain with a complete deletion of the *GAL80* gene versus wild type yeast (as provided in the same publication). The significance of each gene expression ratio in this data set has already been assessed according to a statistical error model implemented by the program *VERA* (Ideker *et al.*, 2000). We converted these significance values to z-scores, shown superimposed on the network in Figure 1.

We ran our annealing algorithm with parameters ($N = 100\,000$; $T_{\text{start}} = 1$; $T_{\text{end}} = 0.01$; $M = 5$; $d_{\text{min}} = 100$). Figure 1 shows the nodes involved in each of the five top-scoring subnetworks; Figure 3 tracks the increases in subnetwork score over the course of the run. Scores of the five subnetworks were (7.7, 3.1, 2.8, 2.5, 2.3); their corresponding sizes were (43, 5, 5, 3, 1). To verify that these subnetworks contained higher-than-expected levels of differential expression, we compared their scores to annealing runs performed on randomized expression data sets. As shown in Figure 4, the top score was greater than and non-overlapping with top scores from randomized data.

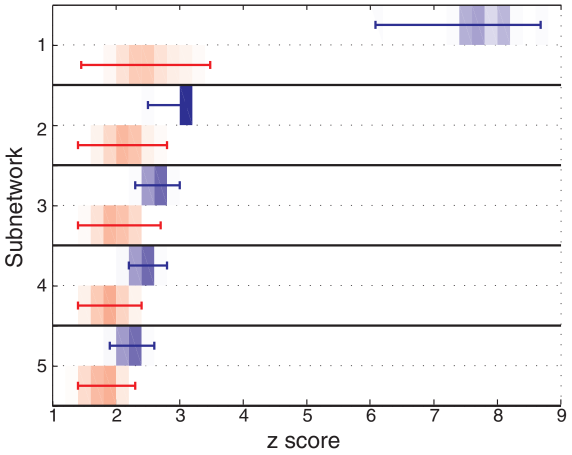

Fig. 4. Distribution of subnetwork scores in actual and randomized data. Score distributions are shown for 200 simulated-annealing runs on actual (a) versus scrambled (b) data, for each of the top-five scoring subnetworks (vertical axis). Different initial conditions were used for each run, with identical network, expression data, and annealing parameters to those described for Figure 1. Greyscale intensity is proportional to the number of runs achieving a particular score; the bounding lines indicate the maximum and minimum values of each distribution. Scrambled data were generated by shuffling the mapping between genes and z-scores, thus preserving the score distribution over all genes but removing any correlation between gene expression and network location.

Taken together, the five subnetworks contained 41 out of 77 genes in the network with significant changes in expression ($p < 10^{-5}$; $z > 4.27$). In each case, the subnetwork gives clues as to how gene expression is transmitted from one gene to another. For instance, the top-scoring subnetwork included *GAL80*, the gene that was knocked out to produce the observed expression changes. It neighbours *GAL4* (immediately right of *GAL80* in Figure 1), a hub with protein→DNA interactions to seven other genes in the subnetwork. Therefore, one hypothesis is that *GAL80* influences expression of these genes by a path through *GAL4*[‡‡].

The subnetworks contain many examples of genes with low z-score that were required to connect together several high-scoring genes. For example, subnetwork 3 (lower left corner of Figure 1) consists of four genes connected to a central transcription factor through protein→DNA interactions. As is typical for regulatory networks, the transcription factor shows only modest expression change compared to the four genes it regulates. Our search identifies this subnetwork because its total level of significance remains relatively high.

[‡‡] As described below in the Discussion, this hypothesis is well supported by the literature (Lohr *et al.*, 1995).

Large network with multiple perturbations

Having characterized our methods on a small network, we wished to explore their performance on larger networks containing nearly all catalogued molecular interactions in yeast. We also wished to investigate how searching for active subnetworks was impacted by multiple conditions. To construct a large molecular interaction network for yeast, we included all 7145 protein–protein interactions in the BIND database (Bader *et al.*, 2001) and all 317 protein→DNA interactions present in TRANSFAC (Wingender *et al.*, 2001) as of September 2001. To find active subnetworks, we screened this network against a complete data set of 20 mRNA expression profiles gathered in response to different perturbations to genes in the GAL pathway (Ideker *et al.*, 2001) (the *GAL80* deletion was one of these). As before, all measurements were converted into z-scores representing the likelihood of differential expression for each gene and perturbation.

For this larger network, we optimized the performance of our annealing algorithm by evaluating a range of annealing parameters over multiple runs. As shown in Table 1, increasing N from 10^3 to 10^7 led to a more than five-fold increase in top score [rows a]; tracking fewer than 10 subnetworks negatively impacted score [b]; the optimal value for $d_{\min}$ was approximately 100 [c]; and temperature did not have a dramatic effect for the parameter sets we surveyed [d]. Using the near-optimal parameters ($N = 1 \times 10^7$, $T_{\text{start}} = 2$, $T_{\text{end}} = 0.01$, $M = 20$, $d_{\min} = 100$), we identified seven subnetworks whose scores were significantly higher than expected in randomized data (by an analysis similar to that shown in Figure 4 for the small network). These subnetworks are shown in Figure 5, while Figure 6 shows the particular conditions used to maximize the score of each subnetwork, i.e., *the conditions under which each subnetwork was active*.

These results demonstrate that our approach can produce active subnetworks that are extremely large—340 genes in the case of the highest-scoring subnetwork! Although this entire structure is statistically significant, it is cumbersome to inspect visually or for use in formulating biological hypotheses. However, we may apply our search algorithm recursively to identify substructures that are particularly significant, i.e., scoring higher than would be expected at random within the 340-node subnetwork. By reducing the population of genes to those in the subnetwork, we change the distribution used to compute the correction factors μ_k and σ_k in eqn. (2) and thus the scores s_A. Here, a second simulated annealing run performed on subnetwork 1 against all 20 perturbation conditions ($N = 1 \times 10^6$, $T_{\text{start}} = 1$, $M = 5$, $d_{\min} = 1000$) identified five substructures of sizes (28, 13, 3, 7, 9); these are labelled #1a through #1e in Figures 5 and 6.

Table 1. Annealing parameter optimization. Each row summarizes results from ten annealing runs starting from different random states of G_w. N is the number of iterations, T_{start} is the starting annealing temperature ($T_{end} = 0.01$ for all runs), M is the number of subnetworks to find, and d_{min} is the minimum node degree required to invoke the hubfinding extension (see text). The shaded parameter set was used for further analysis; each annealing run with these parameters required approximately 3 hours of computation time on a Pentium IV

					Top score	
	N	T_{start}	M	d_{min}	Mean	Stdev
(a)	10^3	2	20	100	5.35	0.96
	10^4	2	20	100	8.79	2.67
	10^5	2	20	100	23.51	2.59
	10^6	2	20	100	25.76	0.77
	10^7	2	20	100	28.92	0.39
(b)	10^5	2	1	100	9.33	4.97
	10^5	2	5	100	17.22	3.61
	10^5	2	10	100	21.82	4.41
	10^5	2	100	100	21.50	3.30
(c)	10^5	1	20	5	17.15	1.15
	10^5	1	20	10	19.31	2.37
	10^5	1	20	100	20.88	2.20
	10^5	1	20	1000	8.92	0.92
(d)	10^5	0	20	100	23.34	0.93
	10^5	2	20	100	23.51	2.59
	10^5	10	20	100	21.87	1.83
	10^5	100	20	100	20.01	1.71

DISCUSSION

Subnetworks are consistent with known regulatory circuits

Many of the subnetworks resulting from our analyses have striking overlap with well-known regulatory and signalling pathways described in the yeast literature. For example, the path *GAL3—GAL80—GAL4→GAL1*,7,10 contained in subnetwork 1b (Figures 5 and 6b) forms the core of the known galactose-induction circuit (Lohr *et al.*, 1995). Briefly, GAL4p is a transcription factor that induces the expression of the enzymes *GAL1,7,10* though protein→DNA interactions, but in the absence of galactose, GAL80p inhibits this activity through a protein–protein interaction. In the presence of galactose, GAL3p associates with GAL80p via another protein–protein interaction, allowing GAL4p to transcribe the enzymes at a high level.

As a second example, subnetwork 1d contains interactions known to regulate genes involved in both the mating response (e.g., *STE2*, *MFA2*, *BAR1*) and cell cycle arrest (*CLB1*, *FAR1*). However, even without expert knowledge from yeast biology, visual inspection of these subnetworks provides us with ready hypotheses for why their genes are differentially expressed. On the strength of these known cases, it will be extremely interesting to examine subnet-

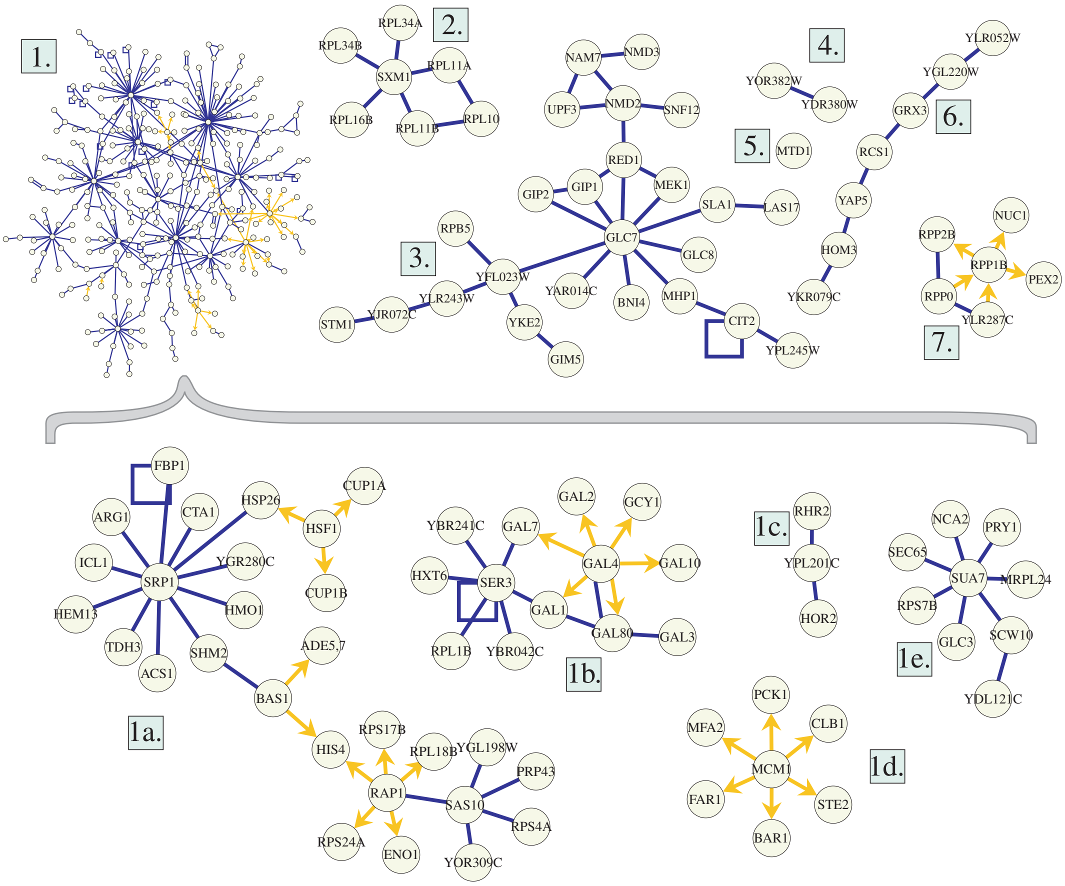

Fig. 5. Performance on a large interaction network using multiple conditions. The seven highest-scoring subnetworks screened from the large network are shown in decreasing order of score. Subnet 1, containing 340 nodes, was subjected to a second application of simulated annealing to produce five smaller subnets 1a–1e. Subnetworks are represented as in Figure 1, but with protein→DNA interactions denoted by yellow edges and protein–protein interactions denoted by blue edges.

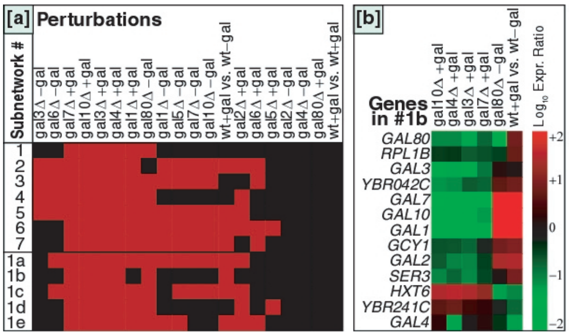

Fig. 6. Perturbations affecting each subnetwork. Panel **[a]**: Red blocks indicate the particular perturbations (columns) maximizing the score of each subnetwork (rows) shown in Fig. 5. Panel **[b]**: Changes in expression level are shown using a red/green colorimetric scale for the genes and perturbations associated with subnet #1b. Note that inversely correlated genes (e.g., *GAL1* and *HXT6*) may appear in the same subnet, and that genes without large changes in expression (e.g., *GAL4*) are nevertheless included because they connect several genes with dramatic changes (e.g., *GAL10* and *GAL7*).

works containing genes whose functions and/or regulatory mechanisms are not already well understood.

Subnetworks versus gene-expression clusters

It is instructive to compare existing methods for clustering genes by expression, e.g., (Eisen *et al.*, 1998; Tamayo *et al.*, 1999), to those introduced here. Although all of these methods form groups of genes (clusters versus subnets), our method differs from established clustering techniques in at least four major ways. First, and most importantly, our approach groups genes subject to the constraints of the molecular interaction network. A striking consequence of this constraint is that subnetworks may contain genes without large expression changes so long as they are required to connect other, differentially expressed genes. Second, subnetworks are scored over only a subset of conditions; thus, genes are not required to be co-regulated over all conditions in order to group together. Third, while most clustering methods group genes by both magnitude and direction of change, we consider only the significance of change. Thus, we may connect strongly repressed *and* induced genes to build subnets that represent complete signalling or regulatory pathways (see Figure 6b). Finally, our method leaves some genes unaffiliated with any subnetwork, unlike typical clustering methods which assign every gene to a distinct cluster.

Future work

In the near future, the most pressing task is to investigate our identified subnetworks in the laboratory. Because large interaction networks are suspected to contain many false-positives, an initial experiment would be to verify that the interactions in each subnetwork are reproducible and present under the subnet's particular set of conditions. We also wish to investigate a number of extensions to our approach, including:

- annotating each interaction with its directionality and/or the specific conditions and compartments in which it has been observed, constraining subnetworks to plausible causal chains.
- accommodating new types of interaction data such as interactions between proteins and small molecules, and new types of biological state measurements such as changes in protein abundances, protein modifications, or concentrations of intracellular metabolites.
- potentially correcting for local network topology. In random expression data, some subnets are much more likely than others to have the highest score; it may be desirable to correct for this bias by computing a prior probability over the population of subnetworks.

As the core algorithms are developed further, we expect this approach to have immense impact in elucidating the underlying molecular mechanisms of a variety of organisms. Ultimately, we envision that biologists will perform routine network screens to define novel modes of regulation, to identify evolutionarily conserved pathways, or to interrogate regulatory circuits responding to the entire spectrum of drugs and human diseases.

ACKNOWLEDGEMENTS

We are indebted to Richard Karp and Vesteinn Thorsson for their valuable ideas and technical feedback. We also wish to thank Paul Shannon, Andrew Markiel, Michael Johnson, Nada Amin, and Dan Ramage for their support of the *Cytoscape* software package. T.Ideker is funded by a Pfizer Fellowship in Computational Biology. A.F.Siegel holds the Grant I.Butterbaugh Professorship at the University of Washington.

REFERENCES

Altman,R.B. and Raychaudhuri,S. (2001) Whole-genome expression analysis:challenges beyond clustering. *Curr. Opin. Struct. Biol.*, **11**, 340–347.

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.

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

Gavin,A.-C., Bösche,M., Krause,R., Grandi,P. and Marzioch,M. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Ideker,T., Thorsson,V., Ranish,J.A., Christmas,R., Buhler,J., Bumgarner,R., Aebersold,R. and Hood,L. (2001) Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Ideker,T., Thorsson,V., Siegel,A. and Hood,L. (2000) Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.

Kendall,S.M., Stuart,A. and Ord,J.K. (1987) *Kendall's Advanced Theory of Statistics*, 5th edition, Oxford University Press, NY, pp. 446.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Kirkpatrick,S., Gelatt,C.D. and Vecchi,M.P. (1983) Optimization by Simulated Annealing. *Science*, **220**, 671–680.

Lohr,D., Venkov,P. and Zlatanova,J. (1995) Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *Faseb Journal*, **9**, 777–787.

Lundy,M. and Mess,A. (1986) Convergence of an Annealing Algorithm. *Mathematical Programming*, **34**, 111–124.

Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.*, (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.*, (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al.*, (2001) The TRANSFAC system on geneexpression regulation. *Nucleic Acids Res.*, **29**, 281–283.

Zhu,H., Bilgin,M., Bangham,R., Hall,D., Casamayor,A., Bertone,P., Lan,N., Jansen,R., Bidlingmaier,S., Houfek,T. *et al.*, (2001) Global analysis of protein activities using proteome chips. *Science*, **293**, 2101–2105.