3. A SYSTEMS APPROACH TO DISCOVERING SIGNALING AND REGULATORY PATHWAYS

—or, how to digest large interaction networks into relevant pieces

Trey Ideker*

ABSTRACT

In the post-genomic era, the first step in any study of protein function is a homology search against the complete genome sequence of the organism of interest. By analogy, if we also wish to elucidate the cadre of signaling and regulatory pathways in the cell, an extremely powerful first step is to construct a complete network of protein-protein and transcriptional interactions and then search through this network to identify critical pathways in a top-down fashion. Like genomic sequence, the molecular interaction network provides a broad foundation for more directed studies to follow. We illustrate this strategy using a large network of 12,232 interactions in yeast. A variety of applications are discussed, including screening the network to identify pathways responsible for gene expression changes observed in galactose-induced cells, and identifying groups of interacting proteins that are essential (by phenotypic assay) for the cellular response to DNA damage.

3.1. INTRODUCTION

In today's post-genomic era, it practically goes without saying that any study of protein function depends on first having a relatively complete genome sequence map of the species of interest. By analogy, if we are interested not just in protein function, but also in how proteins are interconnected within a complex web of signaling and regulatory pathways in the cell, knowing the genome is not quite enough. In addition to the genome, we should also have as our base a comprehensive "interactome"—that is, the network of all protein-protein, protein-DNA, protein-small molecule, and other

* Trey Ideker, Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, MA 02142-1479.

interactions that drive cell function. Then, just as we might use BLAST to search the genome for particular proteins of interest, novel computational tools will allow us to filter through the interaction network to extract relevant signaling or regulatory pathways of interest.

There are two fundamental approaches for studying this interaction network: (1) directly observing the molecular interactions themselves; and (2) observing the molecular and cellular states induced by the interaction wiring. In terms of the first approach, recent systematic two-hybrid (Ito et al., 2001) and co-immunoprecipitation (Mann, Hendrickson, and Pandey, 2001) studies have resulted in a combined database of 15,000-20,000 protein-protein interactions in yeast. Similarly, a new technology known as ChIP-to-chip analysis allows us to measure protein-DNA interactions at large scale. In this analysis, the first "ChIP" stage uses Chromatin ImmunoPrecipitation to pull down transcription factors of interest and all of the promoters they bind, whereas the second stage identifies promoters bound by each transcription factor by labeling and hybridization against a microarray "chip." Lee et al. (2002) have now performed this procedure systematically for approximately 100 transcription factors in yeast, resulting in about 6000 known protein-DNA interactions. Of course, interactions between proteins or between proteins and DNA are not the only types of interactions mediating signaling and regulatory pathways. Other important interactions occur between proteins and hormones, proteins and drugs, or proteins and metabolites, but cannot yet be measured at large scale.

And as for the second fundamental approach, observing the molecular states induced by the interactions? Certainly, DNA microarrays are now widespread in molecular biology for measuring gene expression changes at large scales. In addition, mass-spectrometry-based approaches are now making it possible to interrogate the abundances and phosphorylation states of many proteins simultaneously. Other molecular states, such as abundance levels for the thousands of intracellular metabolites, cannot yet be measured systematically, although mass spectrometry and NMR promise to revolutionize this area as well.

3.2. INTEGRATING INTERACTIONS AND MOLECULAR STATES

Given databases of interactions and states, there is now a tremendous need for computational models and tools able to integrate these large-scale data within a common modeling framework. One goal of this integration is to search the interaction network to identify particular pathways of interactions that correlate with or explain changes in the molecular state.

For instance, consider the integrated network shown in Figure 1, representing a region of the known interaction network surrounding the process of galactose utilization (GAL) in yeast. A node in this network represents a gene and its protein, whereas a link between nodes (i.e., an edge) represents either a protein-DNA (yellow arrow) or protein-protein (blue line) interaction that has been previously determined by some experimental method. The protein-protein interactions shown here are from the BIND (Bader et al., 2001) or DIP (Xenarios and Eisenberg, 2001) databases, while the protein-DNA interactions are drawn from either TRANSFAC (Wingender et al., 2001) or taken from a recent publication by Lee et al. (2002).

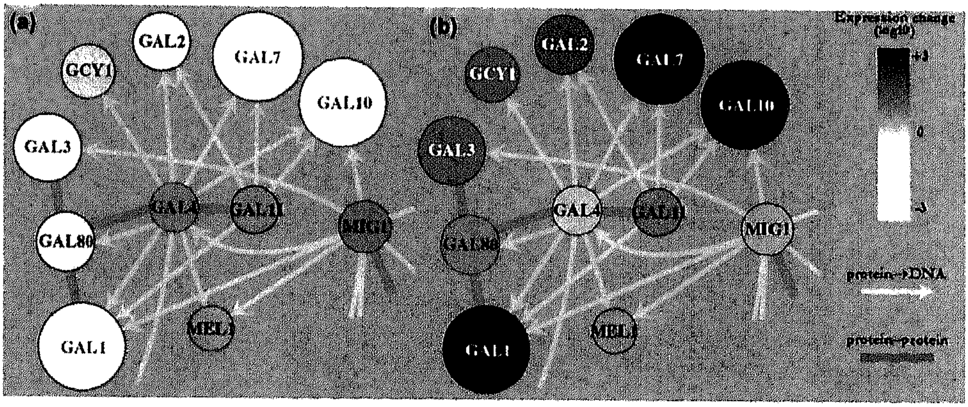


Figure 1. Integrated network representing a region of the known interaction network. Reprinted with permission from Ideker et al. *Science* 292, 929-934 (2001). American Association for the Advancement of Science.

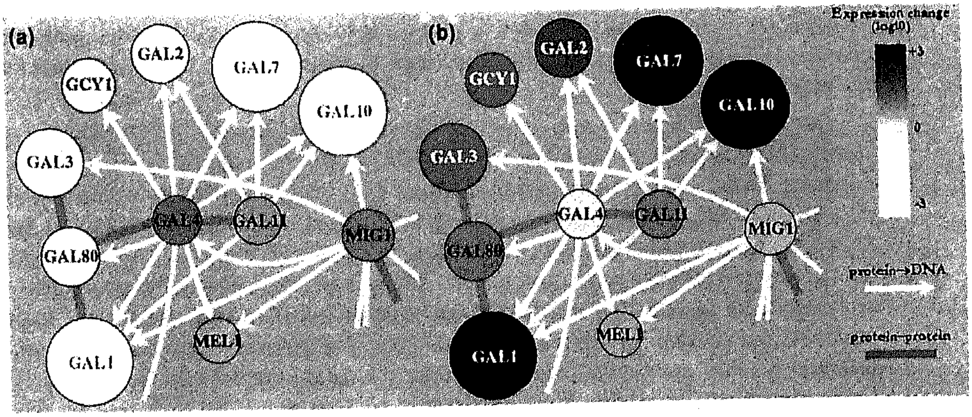


Figure 1. Integrated network representing a region of the known interaction network. Reprinted with permission from Ideker et al. *Science* 292, 929-934 (2001). American Association for the Advancement of Science.

The colors of the nodes represent the states being measured. Figure 1a shows changes in mRNA expression measured in response to a deletion of *GAL4*, whereas the intensities of the other nodes indicate their resulting change in mRNA concentration (Ideker et al., 2001). Background gray represents no change in expression; increasing shades of gray represent increasing levels of mRNA expression; and decreasing shades of gray represent decreasing levels of expression. When *GAL4* is deleted, we see strong decreases in expression of *GAL1*, *7*, and *10*. Importantly, we can begin to explain why we see these changes using interactions present in the underlying network. In this case, the explanation is quite simple: *GAL4* connects directly to *GAL1*, *7*, and *10* through protein-DNA interactions, and it is reasonable to suppose that this is the path by which a *GAL4* deletion evokes these downstream changes.

When we examine different cellular perturbations or biological conditions, the node colors change to reflect these new states. For instance, if we now knock out the *GAL80* gene instead of *GAL4*, the colors reveal a marked increase in *GAL1*, *7*, and *10* (Figure 1b). In this case, a path of length 2 connects *GAL80* to these downstream expression changes: *GAL80* connects to *GAL4* through a protein-protein interaction, while *GAL4* connects to *GAL1*, *7*, and *10* through a series of protein-DNA interactions. In fact, this interaction path turns out to be biologically correct: (Lohr et al., 1995) *GAL80* encodes a repressor protein, which binds to *GAL4* through a protein-protein interaction and keeps it from activating *GAL1*, *7*, and *10*. When GAL80 is knocked out, this protein-protein interaction no longer occurs, and *GAL4* is free to transcribe the GAL genes at a high level.

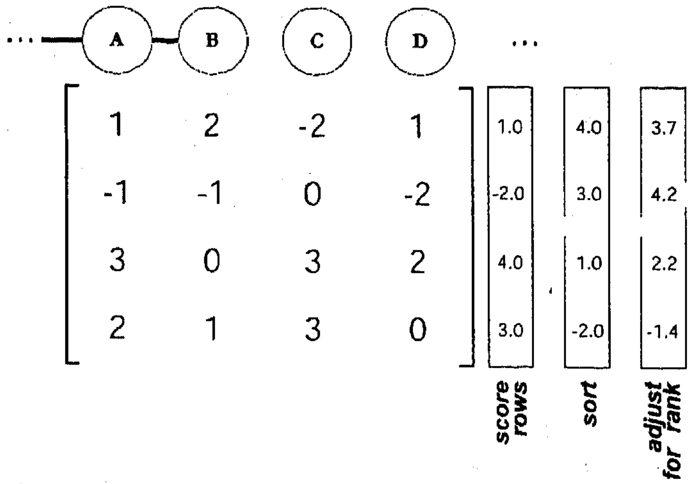
3.3. AUTOMATICALLY EXTRACTING INTERACTION PATHWAYS FROM THE NETWORK

The galactose-related genes and interactions account for just a small piece of the full yeast molecular interaction network. The full network is actually quite large: recall that the public databases currently contain approximately 20,000 protein-protein and protein-DNA interactions for yeast. In such a large network, we can no longer use a quick visual assessment to pull out putative pathways to explain superimposed gene expression changes. However, the basic ideas illustrated in the context of the GAL system extend to the general case.

In general, when some gene is deleted or otherwise perturbed, the resulting significant expression changes will be distributed about the molecular interaction network. Some of these expression changes may in fact be transmitted from the initial perturbation through a pathway or subnetwork of interactions contained within the network. At a high level, we would like to "connect the dots" by identifying paths connecting perturbed to affected genes. Because of the large number of false positives and negatives in both the interaction and expression data sets, we do not expect these paths to be present or relevant for all gene expression changes. However, for the interactions that are present and transmitting a signal, we should be able to find them. Once identified, we define these interaction pathways as "active"; that is, transmitting expression changes from one gene to another in a particular perturbation or condition. Of course, these "active pathway" hypotheses are only predictions—they must be verified or rejected by directed biochemical assays—but they can be generated automatically.

To search for these pathways and pull them out systematically, we first need a mathematical definition of what it means for a pathway to be active [details of this approach have been previously reported elsewhere (Ideker et al., 2002)]. Consider a network consisting of four proteins A, B, C, and D, as shown in Figure 2. Proteins A and B connect to each other through a protein-protein interaction; proteins B and D regulate C through protein-DNA interactions. Now assume that we have observed gene expression changes over four conditions (rows in Figure 2). We are interested not in the ratio of gene expression, but in the *significance of gene expression change*. Whether the expression ratio goes up or down is irrelevant for the purposes of finding pathways—we are simply looking for regions of change.

To indicate significance of expression, we use an error model and an associated statistical test that assigns z-scores to each expression change in each condition (Ideker et al., 2000). Briefly, this method works as follows: if there is no significant expression



$$\begin{bmatrix} 1 & 2 & -2 & 1 \\ -1 & -1 & 0 & -2 \\ 3 & 0 & 3 & 2 \\ 2 & 1 & 3 & 0 \end{bmatrix} \quad \begin{array}{|c|} \hline 1.0 \\ -2.0 \\ 4.0 \\ 3.0 \\ \hline \end{array} \quad \begin{array}{|c|} \hline 4.0 \\ 3.0 \\ 1.0 \\ -2.0 \\ \hline \end{array} \quad \begin{array}{|c|} \hline 3.7 \\ 4.2 \\ 2.2 \\ -1.4 \\ \hline \end{array}$$

Figure 2. Example interaction path with expression data over four conditions.

change for a gene in a condition, then the z-score follows a standard normal distribution (with mean 0 and standard deviation 1). If there is significant expression change for a gene in a condition, its z-score should be significantly higher than expected by this standard normal distribution. The higher the z-score, the more surprising the gene expression change, whether the gene is induced or repressed. For example, out of all four genes shown in Figure 2, we are most confident that gene B has changed in expression in condition 1. We are somewhat less confident that the levels for A or D have changed in this condition, and we are fairly sure that the level of C has not changed.

Once we have computed z-scores using the error model, it is straightforward to score pathway activity by adding together the z-scores of all nodes in the pathway. If no genes are differentially expressed, this sum will itself follow a standard normal. Otherwise, the sum is significantly higher. For example, to score the pathway ABCD in condition 1, we compute the sum 1+2-2+1 = 2 and then divide by the square root of the number of nodes (to normalize the z-score back down to standard deviation 1), resulting in an aggregate "pathway activity" score of $2/\sqrt{4} = 1$. Scoring a pathway over multiple conditions is more complex and is explained in full in Ideker et al. (2002).

Scoring a pathway is only half the problem. Given this scoring system, how do we find the absolute highest scoring pathways in the entire network of 20,000 protein-protein and protein-DNA interactions? This problem can be shown to be NP complete, which means that an exact solution is not obtainable in polynomial time. Instead of solving it exactly, we use an approximation algorithm based on simulated annealing. This algorithm finds, if not the single highest-scoring pathway, a collection of several relatively high-scoring "active" pathways. To search for active pathways using simulated annealing, we take the full molecular interaction network (of ~25,000 interactions among 6000 nodes in the case of yeast) and randomly choose several pathways as initial seeds. Then, over a number of iterations, we add/subtract nodes to each pathway in an attempt to improve its score. If the score increases, we keep the change, whereas if the score decreases, we discard the change with a certain probability dictated by annealing temperature. Given enough iterations, the score starts out low and gradually improves until it stabilizes. In this way, the annealing algorithm is guaranteed to produce pathways that have at least a local optimum in score.

3.4. SCREENING FOR ACTIVE PATHWAYS RESPONDING TO GALACTOSE-GENE PERTURBATIONS

Now let's use this automated pathway search procedure to investigate a specific biological problem of interest. In a proof-of-principle application, we recently screened the yeast interaction network to find pathways active under different perturbations to the galactose utilization network in yeast (Ideker et al., 2002). Seven perturbations were performed, by first generating gene knock-outs of *GAL1*, *2*, *5*, *6*, *7*, *10*, and *80* in separate strains, then measuring the corresponding cellular responses with a whole genome mRNA expression profile.

We ran the automated pathway search procedure to identify which pathways from the yeast interaction network were most activated by these perturbations. Five high-scoring pathways were identified and are shown in Figure 3a. As in Figure 1, a line

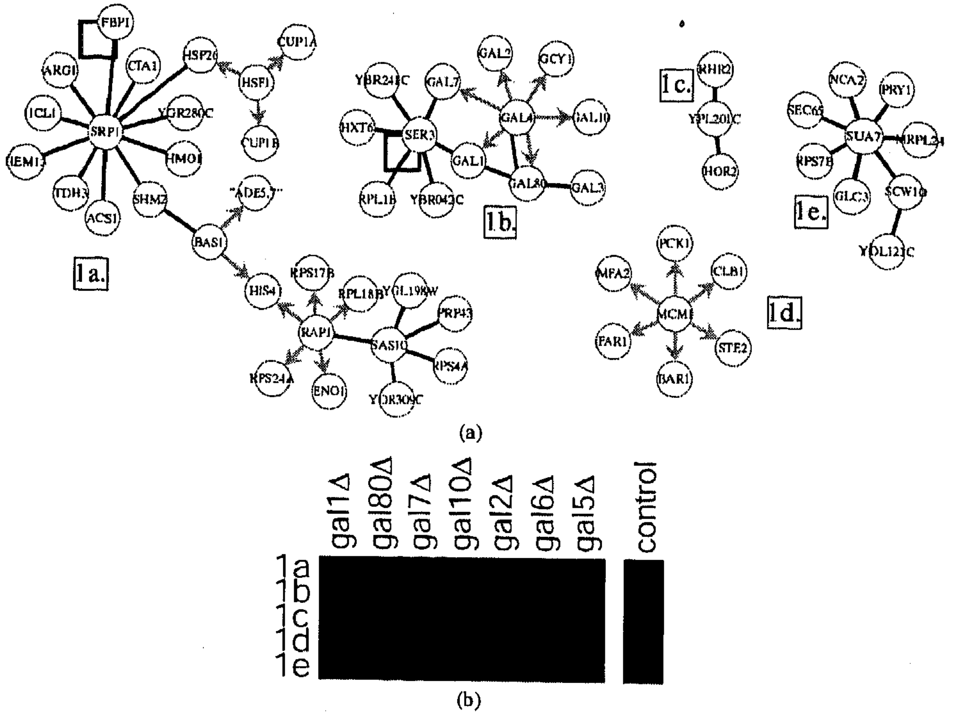


Figure 3. High-scoring pathways. Reprinted with permission from Ideker et al. (2001).

represents a protein-protein interaction and an arrow represents a protein-DNA interaction: all of these interactions are derived and filtered from the whole molecular interaction network.

Figure 3b indicates the particular conditions (columns) activating each of the five pathways (rows). For instance, pathway 1a is active under the *GAL1*, *8*, *7*, *10*, and *2* perturbation experiments, but not under the *GAL7* or *5* perturbation experiments. Likewise, pathway 1b is activated by a *GAL80* perturbation only. Using Figure 3b, we can compare different pathways on the basis of the expression experiments which activate them. For instance, note that pathway 1a and 1c have an identical perturbation profile, which is very different from that of pathway 1b.

The five active pathways represent a combination of known and unknown regulatory processes in yeast. As a "positive control," pathway 1b contains much of the GAL module shown in Figure 1, including the *GAL4* central transcriptional activator and the *GAL80* transcriptional repressor. Given that we are directly perturbing many of the genes in this pathway, we expect it to be active.

Other active pathways represent new discoveries. These provide testable hypotheses for the underlying regulatory and signaling interactions responsible for the observed expression changes. It was not known, for instance, that MCM1 and its downstream regulated genes were involved in the galactose response.

We are currently in the process of applying this approach to a variety of other pathways and expression data sets. One exciting implication for this technology is in the area of drug development. Many drugs are well characterized in terms of what proteins and pathways are being targeted, but not in terms of their possible toxicological side effects. The problem, therefore, is not to discover new drug targets, but to reveal additional pathways that may be affected by the drugs. Here, the limiting factor is obtaining a molecular interaction network relevant to humans. As large interaction networks are determined for key human cell lines—for example, hepatocytes and cancer cells—such an analysis will become possible.

3.5. PATHWAYS RESPONDING TO DNA DAMAGE AS REVEALED BY HIGH-THROUGHPUT PHENOTYPIC ASSAYS

Another method of filtering the molecular interaction network to identify biologically relevant pathways is to use deletion phenotypes. In recent work performed in collaboration with Leona Samson's laboratory (Begley et al., 2002), such an approach was used to map genes and pathways required for the cellular response to DNA damage. For each gene-knockout strain in yeast (libraries of all single gene-knockout strains are now publicly available), we tested whether the strain was able to grow in the presence of MMS, a powerful DNA-damaging agent. Wild-type cells can, in fact, grow under a moderate concentration of MMS, but many gene-knockout mutants either grow slowly or not at all under these conditions.

How do these "MMS-sensitive genes" map onto the protein-protein and protein-DNA interaction network? Figure 4 shows a sampling of interaction pathways containing significant numbers of MMS-sensitive proteins, as determined by the automated pathway screen described in Section 3. In the figure, a node is colored green if deletion of that gene results in slow growth or death in the presence of MMS; red if the deletion has no effect for growth in MMS; and gray if the node has not yet been tested by phenotypic growth assay. Of the gene knockouts tested so far, approximately 400 of them were MMS-sensitive. Using the automated screen for pathways, we were able to associate 100 of these with an "active pathway" having many other MMS-sensitive nodes in close proximity (75 of these appear as green nodes in Figure 4, while the remaining 25 were organized into several pathways not shown in the figure). One interesting observation is that MMS-sensitive nodes may be grouped in a single connected pathway even if several non-sensitive (or non-tested) nodes are required to do so. For instance, to include *MKC7*, *RRP6*, *GIS3*, and *CIN8* in the pathway shown in the upper-left-hand corner of Figure 4, it was necessary to also include YLR453C, which was not tested by phenotypic assay but is included because of its "MMS-sensitive" network neighborhood.

3.6. SUMMARY

A good metaphor for the pathway screening approaches discussed here is that of an information processor, or "black box," as shown in Figure 5. We pour into this black box, on the one hand, all of the molecular interactions previously determined for our organism of interest. On the other hand, we pour in molecular states measured in response to perturbations of a cellular process or biological response of interest. Here, we have used a network of approximately 25,000 protein-protein and protein-DNA interactions in yeast, with state changes measured either at the level of gene expression (Section 4) or growth phenotype (Section 5). After running the "active pathways" algorithm, the black box

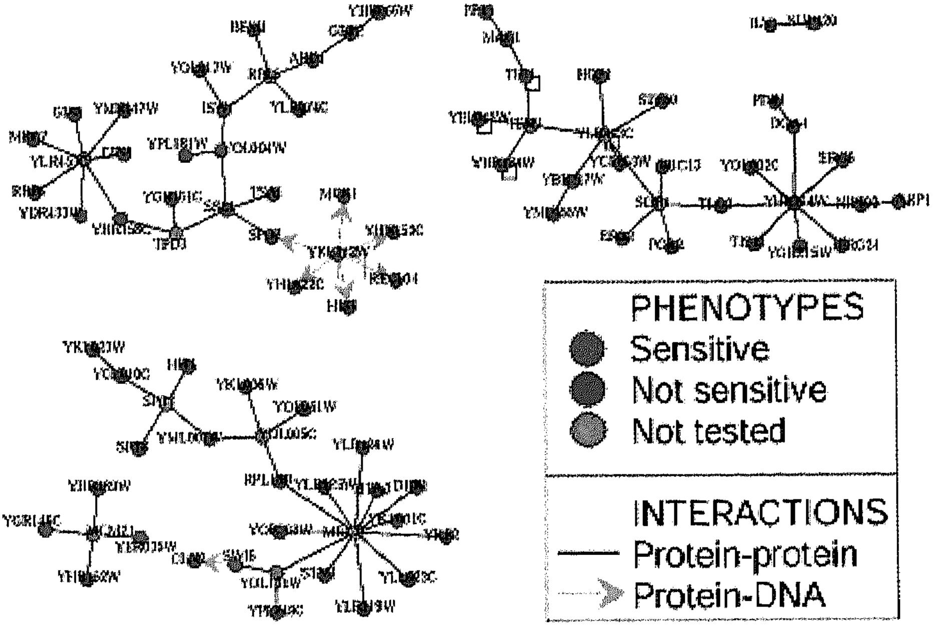


Figure 4. Interaction pathways containing significant numbers of MMS-sensitive proteins. Reprinted with permission from Begley et al. (2002).

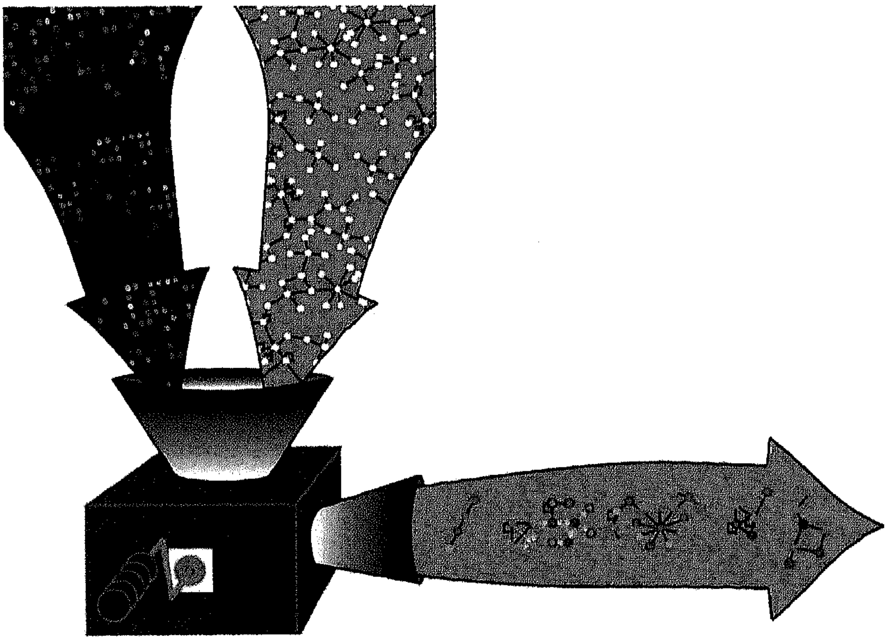


Figure 5. Information processor, or "black box." Reprinted with permission from Ideker and Lauffenburger, *Trends in Microbiology* (in press).

Figures 1, 3, and 4 are based on screen shots taken from a software package called Cytoscape, available as Open Source software from http://www.cytoscape.org as a platform-independent Java application. Cytoscape involves two main components: (1) a core platform for visualizing and manipulating large molecular interaction networks, and (2) an extensible plug-in architecture for writing algorithms and analyses that compute on these networks. The core contains all the routine graphical manipulation, visualization, and information management tasks for large networks: for instance, "How do we lay out these networks in two and three dimensions? Can we link these networks to underlying databases providing annotations for each gene, protein, and interaction"? Plug-ins further extend the basic functionality provided by the core—one such example is the Active Pathway finder discussed in Section 3. Cytoscape is a joint-development project with the Institute for Systems Biology in Seattle.

3.7. ACKNOWLEDGMENTS

This work would not have been possible without the aid of Owen Ozier, Scott McCuine, and other members of the Ideker laboratory; Rick Young (for advance access to data on protein-DNA interactions); and Leona Samson (for her work with the DNA-damage response). Many thanks are also owed to Paul Shannon, Andrew Markiel, and Benno Schwikowski, all at the Institute for Systems Biology, who are co-developers of the Cytoscape software. This work was supported by a grant from Pfizer.

3.8. REFERENCES

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F. F., Pawson, T., and Hogue, C. W. V., 2001, BIND-The biomolecular interaction network database, *Nucleic Acids Res.* **29**:242-245.

Begley, T. J., Rosenbach, A. S., Ideker, T., and Samson, L. D., 2002, Damage Recovery Pathways In Saccharomyces cerevisiae Revealed by Genomic Phenotyping and Interactome Mapping, *Mol. Cancer Res.* **1**:103-112.

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.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F., 2002, Discovering regulatory and signaling circuits in molecular interaction networks, *Bioinformatics* **18 Suppl 1**:S233-240.

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

Ito, T., Chiba, T., and Yoshida, M., 2001, Exploring the protein interactome using comprehensive two-hybrid projects, *Trends Biotechnol.* **19**:S23-S27.

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 Saccharomyces cerevisiae, *Science* **298**:799-804.

Lohr, D., Venkov, P., and Zlatanova, J., 1995, Transcriptional regulation in the yeast GAL gene family: a complex genetic network, *FASEB* J **9**:777-787.

Mann, M., Hendrickson, R., and Pandey, A., 2001, Analysis of proteins and proteomes by mass spectrometry, *Annu. Rev. Biochem.* **70**:437-473.

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.

Xenarios, I., and Eisenberg, D., 2001, Protein interaction databases, *Curr. Opin. Biotechnol.* **12**:334-339.