

# Visualization and Integration of Protein-Protein Interactions

Peter Uetz<sup>1,4</sup>, Trey Ideker<sup>2</sup> and Benno Schwikowski<sup>2,3</sup>

<sup>1</sup>Departments of Genetics, University of Washington, Box 357360, Seattle, WA 98195

<sup>2</sup>The Institute for Systems Biology, 4225 Roosevelt Way NE, Suite 200, Seattle, WA 98105, e-mail: [benno@systemsbiology.org](mailto:benno@systemsbiology.org), [tideker@systemsbiology.org](mailto:tideker@systemsbiology.org)

<sup>3</sup>Department of Computer Science and Engineering, University of Washington, Box 352350, Seattle, WA 98195

<sup>4</sup>present address: Institut für Genetik, Forschungszentrum Karlsruhe, Postfach 3640, D-76021 Karlsruhe, Germany, e-mail: [peter.uetz@itg.fzk.de](mailto:peter.uetz@itg.fzk.de) or [peter@uetz.de](mailto:peter@uetz.de)

## TABLE OF CONTENTS

<b>INTRODUCTION .....</b>	<b>2</b>
WHY DO WE NEED VISUALIZATION? .....	3
PROTEIN INTERACTION MAPS VERSUS METABOLIC PATHWAYS .....	3
PROTEIN NETWORKS, PROTEIN COMPLEXES, AND DYNAMIC PROTEIN INTERACTIONS .....	4
PROTEIN-PROTEIN INTERACTIONS AND ASSOCIATED INFORMATION .....	5
<b>VISUALIZATION.....</b>	<b>5</b>
RELATIONAL VISUALIZATION .....	5
SYMBOLS AND CONVENTIONS .....	10
GRAPHS AND GRAPH DRAWING .....	10
SPRING EMBEDDER ALGORITHM .....	13
LIMITATIONS FOR LARGE GRAPHS .....	13
EXTENSION TO THREE DIMENSIONS .....	13
TECHNIQUES FOR VISUALIZATION .....	14
AVAILABLE TOOLS FOR VISUALIZATION .....	15
<b>INTEGRATING PROTEIN-INTERACTION NETWORKS WITH SUPPLEMENTAL DATA.....</b>	<b>17</b>
SIMULTANEOUS DISPLAY OF COMPLEMENTARY DATA TYPES .....	17
AN EXAMPLE: INTEGRATED NETWORKS TO STUDY GALACTOSE METABOLISM.....	17
SUPERIMPOSITION OF mRNA- AND PROTEIN-EXPRESSION CHANGES ON THE NETWORK .....	20
INTEGRATED NETWORKS ENABLE NEW BIOLOGICAL INSIGHTS .....	20
CHOICE OF GRAPHICAL REPRESENTATION .....	22
FROM VISUAL REPRESENTATION TO A PREDICTIVE MODEL: EARLY STEPS TOWARDS GENE-EXPRESSION .....	23
MODELING USING PHYSICAL-INTERACTION NETWORKS.....	23
PREDICTION OF EXPRESSION CHANGES RESULTING FROM PARTICULAR PERTURBATIONS TO THE GALACTOSE-UTILIZATION NETWORK .....	24
<b>FUTURE DIRECTIONS .....</b>	<b>25</b>
<b>REFERENCES .....</b>	<b>26</b>

# INTRODUCTION

Eukaryotic cells are remarkably well understood in their chemical composition: we know the DNA sequences of many organisms more or less completely and can deduce many of their RNA and protein products. In the past, proteins have received most attention from biochemists because they are the most complicated and among the most important molecules in a cell. Accordingly, proteins and their interactions have been studied in great detail resulting in the identification of thousands of protein-protein interactions over the past 30-50 years. In addition to these classical studies, more recent large-scale proteomics projects are contributing huge amounts of systematic data relevant to the understanding of protein interactions. These large data sets also harbor information that is not immediately obvious without integrative analysis. Although integration and analysis have traditionally been carried out by humans, the sheer amount of data now calls for computer assistance.

Of course, cells not only consist of proteins but also a significant number of other molecules, ranging from small ions to high molecular weight carbohydrates and nucleic acids. Most of these non-proteinaceous compounds interact with at least one protein, as many represent the product of enzymatic reactions, and by default associate with the enzyme that generated them (e.g. pyruvate with pyruvate kinase; see Figure 1).

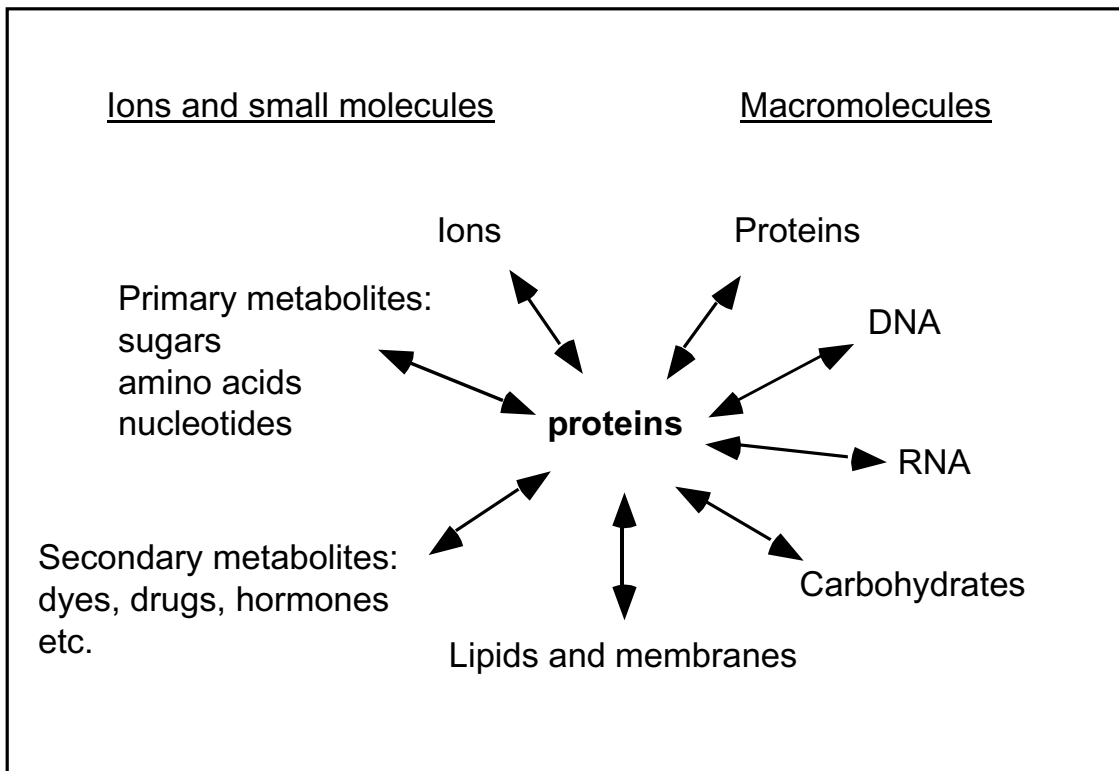


Figure 1. Physical interactions of proteins. Note that there are also interactions between the non-protein classes, e.g. between ions and other small molecules (such as  $Fe^{++}$  and heme). However, not many such interactions are reported. Additional interactions can be imagined with artificial molecules like drugs or synthetic ligands. In addition to physical interactions, genetic interactions (like synthetic lethality or suppression) can hint at potential physical interactions.

Although knowledge of these interactions is essential for a complete understanding of a cell, the number of such interactions is thought to be small compared to the number of interactions among proteins or between proteins and nucleic acids. Moreover, the study of protein-DNA interactions has been a more popular starting point for interaction network assembly than the study of protein-protein interactions. DNA behaves more predictably than proteins and is subject to fewer variables such as chemical modification which may confer significant structural change. Furthermore, DNA-binding domains in many transcription factors are generally better characterized than protein-protein interaction domains, and extensive structural information for such modules exist. In contrast, interactions between proteins, or between proteins and RNA are often harder to characterize due to their perceived inherent variability.

This chapter will present some computerized means to visualize and integrate protein-protein interaction data with data from DNA chip experiments and other sources, to create a starting point for future methods and discoveries.

### **Why do we need visualization?**

For most people, graphical representations of facts are much easier to understand than raw data. This is especially true for large datasets or complex situations. A long list of interacting proteins or a table of protein pairs falls short of capturing what happens in a cell, which is a dynamic process that occurs in at least 4 dimensions (including time). Instead, the use of graphics suits human preference for visual perception over every other sensory system. Like a road atlas, visual maps of protein interactions provide orientation for both novices and specialists. In order to make these maps useful for both audiences it is desirable to generate dynamic maps that allow concealment of detail when only a rough overview is needed. Finally, protein-interaction maps stimulate the formulation of hypotheses that can be tested experimentally. For example, if a membrane protein is found to interact with a transcription factor this might look like a false positive. However, such apparent incongruities have led to unexpected new insights into signal transduction, as in the cases of *notch* and *Suppressor of hairless*, Su(H), (Artavanis-Tsakonas et al., 1999) or with the SREBPs, transcription factors that are localized to the ER membrane (Edwards et al. 2000). Development of an appropriate map can aid in the identification of such informative anomalies.

### **Protein interaction maps versus metabolic pathways**

Considerable effort has been invested into the visualization of metabolic pathways (e.g. Michal 1993, 1998), with more recent efforts using computerized systems (e.g. Kffner et al. 2000). Although the structures of metabolic pathways and protein interaction maps are similar, there are a number of significant differences. While *metabolic pathways* focus on the conversion of small molecules and the enzymes responsible for these conversions, *protein interaction maps* (and signal transduction maps) concentrate mainly on physical contacts without obvious chemical conversions. Physical interactions are certainly of great utility when one studies single proteins or defined biological processes, but themselves do not reflect the huge amount of knowledge that has been accumulated in the biological literature. For example, Figure 5 implies that the PEX proteins in yeast form a complex, but it does not provide any information about the assembly of the complex, its biological function or its regulation. Although these shortcomings can be partly relieved by building in hyperlinks to protein databases, few such databases collate biological information in an easily accessible format.

In addition to physical and metabolic networks, several groups have presented models and systems for *genetic networks* (Kolpakov et al. 1998, Serov et al. 1998, von Dassow et al. 2000). Such networks do not require physical interactions and in fact can suggest factors affecting

control of a biological process that remain to be identified (von Dassow et al. 2000). Nevertheless, the ultimate goal of all such network models is a physical network integrated with genetic and metabolic information that is predictive of phenotype.

The biggest challenge for coherent display of physical networks remains the large amount of biological information that is available about many molecules and their interactions. Graphical networks can illustrate the complexity of biological interactions, but still fail to explain processes, mainly because important details are not visualized, such as spatial and temporal expression patterns or the conditions under which certain interactions occur.

### Protein networks, protein complexes, and dynamic protein interactions

Currently, the most practicable way to identify the components of a protein complex is mass spectrometric (MS) analysis (Yates 2000). Unfortunately, MS usually doesn't provide information about topology, so that additional methods are required to decipher which proteins bind to which. The two-hybrid system can provide complementary information about direct interactions. However, it remains a challenge to integrate data from different experimental approaches, especially when they have been collected under different biological conditions (e.g. when cells were grown in different media). Figure 2 illustrates such non-overlapping datasets. Protein interactions can occur in stable complexes or as transient, usually regulated interactions. Unfortunately, most interactions are described qualitatively and we don't know how strong the interaction really is. For example, the Database of Interacting Proteins (DIP) lists binding constants for less than 20 protein pairs (as of December, 2000, see also Xenarios et al. 2000). Because there are hardly any quantitative data about protein interactions, protein complexes are currently difficult to analyze quantitatively based on an informatics approach.

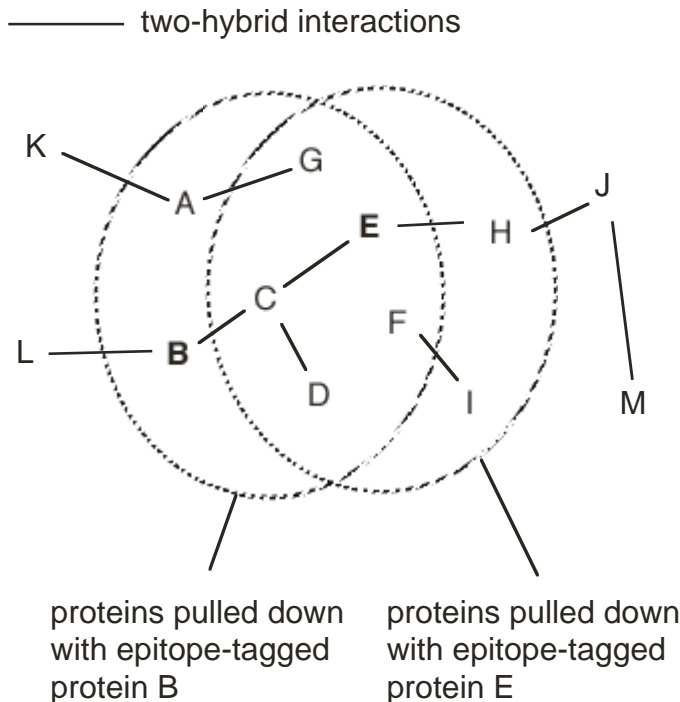


Figure 2. Protein complexes and networks. Mass spectrometry allows to identify proteins in a complex but the composition may be dependent on which protein was tagged in order to purify the complex. Two-hybrid interactions allow to reconstruct protein networks but not physical complexes. Most protein interaction diagrams ignore such contradictions.

Theoretically, computational analysis of protein structures should allow us to find fitting surfaces among known protein structures and thereby predict their interactions (Tsai et al. 1996, Palma et al. 2000). Of the ~6000 yeast proteins comprising the yeast proteome, about 600 have experimentally determined 3D structures, while for up to ~2600 (~44%) the structure can be modeled based on homology (<http://jura.ebi.ac.uk:8765/ext-genequiz//genomes/sc/index.html>, Sanchez et al. 2000, Vitkup et al. 2001, and U. Pieper, pers. comm., and M. Andrade, pers. comm.). However, limitations in experimental data and computing power still prohibit detailed and thereby successful predictions in most cases.

## Protein-protein interactions and associated information

Much information is required to describe molecular interactions, especially if they are dynamic and dependent on many different parameters. For instance, once interacting proteins have been identified, one wants to know which parts or *domains* of the two partners interact because this might lead immediately to further hypotheses about the specificity and nature of the interaction. The *strength* of interaction is also an important parameter, because it indicates whether the two proteins form a complex or interact only transiently. The strength of an interaction may be also *modulated* by phosphorylation or other types of modification: for example, SH2- and some WW-domain proteins only bind to phosphorylated target proteins.

Although many interactions among structural proteins are static and therefore relatively straightforward to describe, dynamic interactions are more difficult to study and visualize. For example, the activity of an enzyme might be regulated by several subunits, whose interaction is dependent on physical or biochemical parameters such as temperature, phosphorylation, or concentration. Likewise, protein interactions might be described in terms of actions they exert on other proteins, much like the interactions between enzymes and their substrates. Common examples are protein acetylases or proteases. Such interactions might necessarily be weak and transient to assure a high reaction rate. We have summarized parameters describing protein interactions and their conditions in Table 1. Table 2 lists databases and websites that have such data for large sets of proteins.

## VISUALIZATION

### Relational visualization

Protein-protein interactions are frequently represented as a linear list of protein pairs (such as in Uetz *et al.* 2000). In contrast, *relational visualization* seeks to represent entities and their relationships in a graphical form (Figure 3). The complexity of such representations ranges from simple (Figure 3a-c) to highly complex (Figures 3d-h and Figure 4). Figures 3f and 3g illustrate the usefulness of graphics on a small network of protein-protein interactions in yeast.

Although both representations reflect identical information, the graphical representation (frequently called layout) has fundamental advantages with respect to human perception.

1. *Localization: single versus multiple entries.* In a textual representation, the protein interactions involving a given protein *X* are usually spread out over different positions in the list, which requires an exhaustive search through the whole list to find all interactions involving *X*. In a graphical layout, *X* occurs exactly once.

2. *Context.* Once *X* has been identified in the layout, its immediate and indirect neighbors are easily identified, and their relation to *X* studied.

3. *Mental map.* A graphical representation facilitates to memorize proteins by position in a mental map (Eades *et al.* 1991). In positioning the nodes, secondary information can be employed to guide the layout; for example, proteins can be spatially grouped by localization or function. In this way, a particular arrangement of the proteins can increase the information content of the layout, and facilitate its comprehension at the same time.

Table 1. Parameters of molecular interactions and complementary sources of biological information

Parameter	Molecule					
	Prot.	DNA	RNA	Lipids	Carb.	Met.
Concentration	x	-	x	(x)	x	x
Localization	x	(x)	(x)	x	x	x
Covalent modifications	x	x	x	x	x	x
phosphorylation	x	-	?	x	(x)	?
acetylation	x	?	?	?	?	?
methylation	x	x	?	?	?	?
other modifications	x	?	x	x	?	?
cleavage (degradation)	x	(x)	x	(x)	x	x
Non-covalent modification <sup>a</sup>	x	(x)	x	?	?	?
Logical state (ON/OFF)	x	?	?	?	?	?
Binding sites	x	x	x	(x)	(x)	(x)

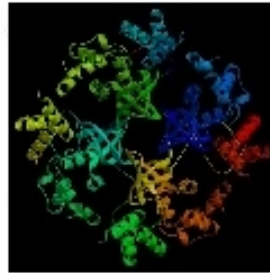
For types of molecules see also Figure 1. Carb. = carbohydrates, Met. = metabolites. "x" indicates whether this parameter is relevant for the given molecule. Information about molecules and their actions can be found to various degrees in the databases listed in Table 2. Please note that many modification states or the activity of molecules is dependent on the input from other molecules, e.g. when phosphorylation activates a protein. Such actions and their conditions are usually not recorded systematically or in a standardized nomenclature, and thus are difficult to use for automated map generation. Protein interaction maps therefore should have an option to enter some free-text annotation that can be accessed from the graphical output. (a) Non-covalent modifications may be conformational states such as allosteric isoforms.

a text: X-Y

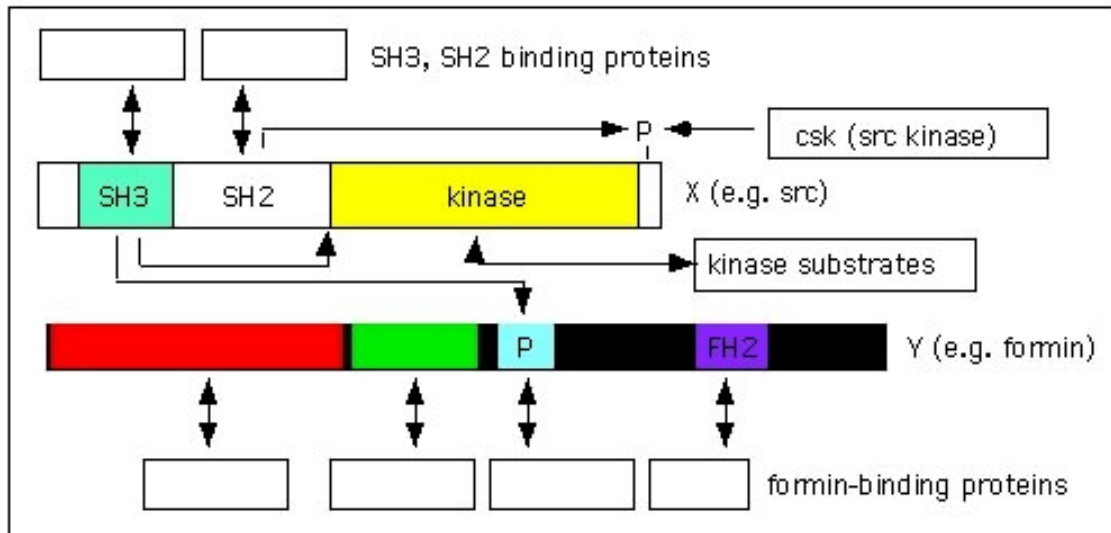
d "3D":

b schematic: 

c shapes: 



e domains, schematic:



f list:

Ypt1-YPL246C  
Akr2-YHR105W  
Yip1-YGL161W  
YPL246C-Vam7  
YGL161W-Pep12  
YPL246C-YHR105W  
YHR105W-YGL161W

g graphical display of list:

Ypt1	Akr2	Yip1
YPL246C	— YHR105W	— YGL161W
Vam7		Pep12

h dynamic, 3D representation: electronic display  
not possible on paper

Figure 3. Visualization of protein interactions. For computerized display, (a) and (b) are the most common ones. Shapes, as in (c) are more difficult to generate automatically by computer because topology has to be taken into account. (d) shows a computer-generated 3D structure of a protein complex but such data is only available for a small set of proteins. (e) takes into account the domain structure of proteins and functional interactions such as phosphorylation, but there are no systems available yet to generate such diagrams automatically, partly because the pertinent information is not available in databases.

Figure 3, cont d. Boxes denote other proteins. (f) list of interactions as text. (g) graphical representation of list in (f). (h) Ultimately we wish an integrated display of the aforementioned options which allows a user to look at the big picture of many proteins but also to zoom in to atomic detail. Such visualization tools are not available yet and will be possible only as electronic systems (i.e. not on paper). 3D structure in (d) reproduced by permission of the Protein Structure Database (<http://www.rcsb.org>); based on a figure in Hargreaves et al. (1998).

Table 2 - Databases and web sites

<b>Protein interactions</b>	
MIPS	<a href="http://www.mips.biochem.mpg.de/proj/yeast/tables/interaction/index.html">http://www.mips.biochem.mpg.de/proj/yeast/tables/interaction/index.html</a> , (Mewes et al. 2000)
YPD (Proteome);	<a href="http://www.proteome.com/">http://www.proteome.com/</a> (single interactions)
DIP	<a href="http://dip.doe-mbi.ucla.edu/">http://dip.doe-mbi.ucla.edu/</a>
SGD (Function Junction)	<a href="http://genome-www.stanford.edu/cgi-bin/SGD/functionJunction">http://genome-www.stanford.edu/cgi-bin/SGD/functionJunction</a>
Myriad-Pronet	<a href="http://www.myriad-pronet.com/">http://www.myriad-pronet.com/</a>
Curagen	<a href="http://portal.curagen.com">http://portal.curagen.com</a>
BIND	<a href="http://www.biond.org">http://www.biond.org</a>
BRITE	<a href="http://www.genome.ad.jp/brite/">http://www.genome.ad.jp/brite/</a>
<b>Protein networks</b>	
Biocarta	<a href="http://www.Biocarta.com">http://www.Biocarta.com</a>
Genmapp	<a href="http://gladstone-genome.ucsf.edu/introduction.asp">http://gladstone-genome.ucsf.edu/introduction.asp</a>
Kohn (1999)	<a href="http://discover.nci.nih.gov/kohnk/links.html">http://discover.nci.nih.gov/kohnk/links.html</a>
Signal Transduction Knowledge Environment	<a href="http://www.stke.org">http://www.stke.org</a>
Schwikowski et al.	<a href="http://depts.washington.edu/sfields/projects/YPLM/data/index.html">http://depts.washington.edu/sfields/projects/YPLM/data/index.html</a>

Eilbeck et al. (1999) described another protein-protein interaction database, which was not accessible at the time of this writing.



The magnitude of experimental data from large-scale experimental methods makes it seem impossible to visualize all protein-protein interactions in a single layout, even for relatively simple organisms such as yeast.

For example, Figure 4 shows a layout that contains the largest component of an experimentally determined protein-protein interaction map in yeast (as of April, 2000). Specific functions (according to the YPD classification, Costanzo *et al.* 2000) are highlighted by color, and it becomes clear from this map that certain proteins with similar function cluster together. However, the detail in Figure 5 illustrates that, while proteins and their interactions appear arranged well in the peripheral regions, in central regions of the layout edges and name labels are drawn on top of each other, making it impossible to discern individual interactions. Finally, computer-generated interaction maps have not been designed to contain as much information as hand-drawn maps.

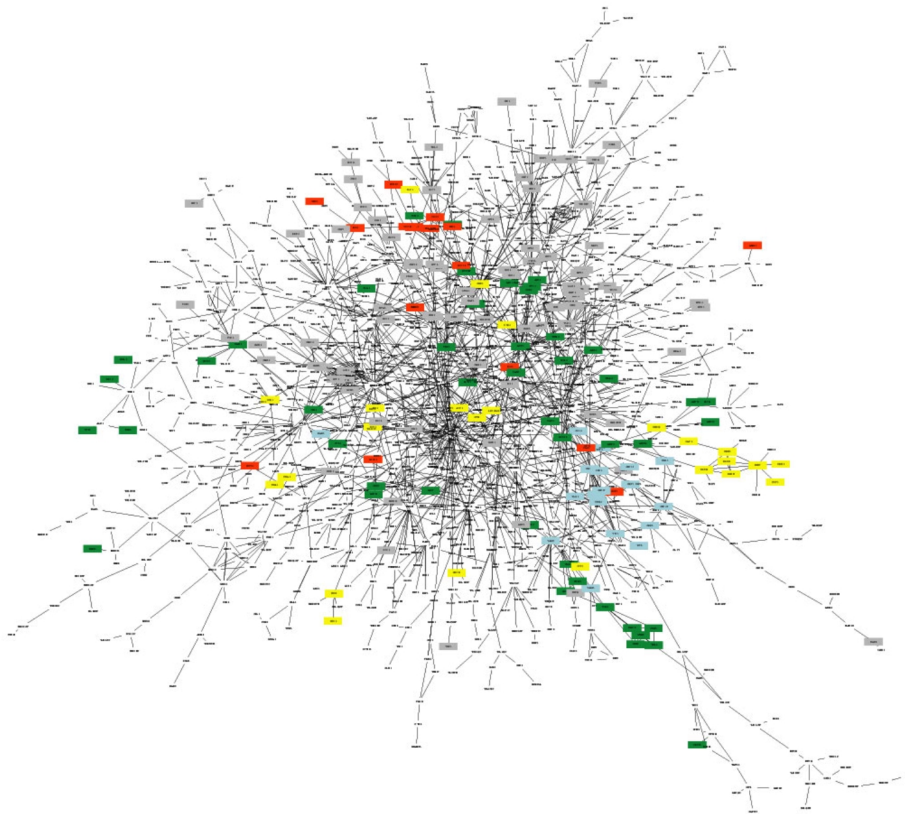


Figure 4. 2358 protein-protein interactions in yeast (from Schwikowski *et al.* 2000). Alternative wiring schemes for complex networks are shown in Kohn (1999) and Strogatz (2001).

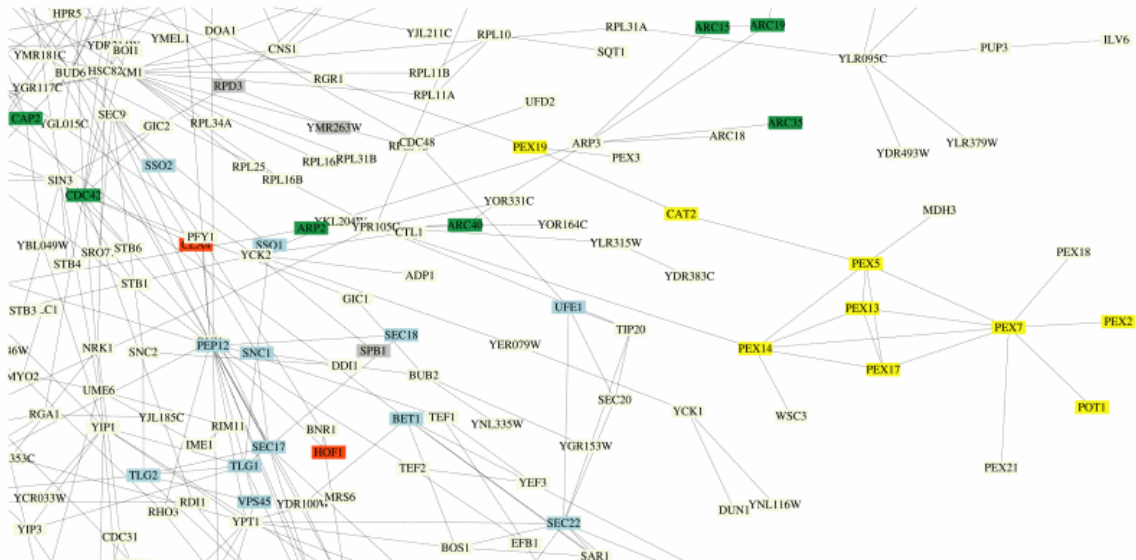


Figure 5. Detail of Figure 4.

In summary, hand-formatted maps (such as those in Michal 1993; Michal 1998; Kohn 1999) are usually of a higher quality, but — due to the large amount of work involved to construct them — available for very limited datasets. Finally, with the greater complexity of datasets arising from more complicated genomes, even hand-formatted maps are likely to be inadequate.

## Symbols and conventions

Several authors have suggested symbols for describing protein-protein interactions. Notably, Kohn (1999) suggested some conventions for building sophisticated models of protein interactions involved in cell-cycle control and DNA repair (Figures 6-9). Although Kohn's wiring diagrams are well worked out, they are not generated by an automated system and therefore have to be redrawn manually when larger changes need to be included. However, his symbols and conventions might also be used by a computerized system, and therefore are reproduced here. More recently, Cook et al. (2001) suggested another system for describing complex biological systems including protein interactions. Other projects are under way and we would like to refer readers to our web site for updates (<http://www.systemsbiology.org/pubs/vizprotein>).

## Graphs and Graph Drawing

The field of *graph drawing* deals with the automatic computation of maps from graphs. The abstract nature of a set of protein-protein interactions can be captured by the mathematical notion of a *graph*. Formally, a graph consists of nodes that represent the proteins, and edges between pairs of nodes that represent protein-protein interactions. Graphs arise in many other fields, such as sociology, project management, and software engineering, as well as in other areas of biology, such as taxonomy and biochemistry. Because of this ubiquity, there is an extensive body of *graph theory* that deals with mathematical properties of graphs (for an introductory text, see Bollobás 1998).


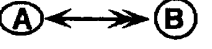

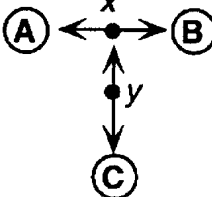

Symbol	MEANING
	Non-covalent binding, e.g. between proteins. A node represents the A-B complex itself.
	Asymmetric binding where protein A contributes a peptide that binds to a receptor site or pocket on protein B.
	Z is the combination of states defined by x and y.
	Multimolecular complex: x is A-B; y is (A-B)-C.
	Formation of a homodimer. Filled circle on the right represents another copy of A. The filled circle on the binding line represents the homodimer A-A.

Figure 6. Kohn's symbols for describing protein-protein interactions (after Kohn 1999). Please note that Kohn also suggested additional symbols for the stoichiometric conversion of A into B, degradation products, transport etc. A complete list can be found in Kohn (1999) or online at <http://discover.nci.nih.gov/kohnk/symbols.html>. Additional symbols and conventions have been proposed by other authors such as Cook et al. (2001).

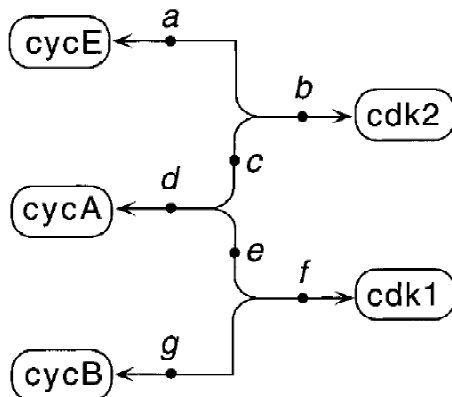


Figure 7. Kohn's representation of alternative binding modes. Example: heterodimers formed by Cyclins E, A, and B binding to Cdk1 or 2.; Note that a, c, e, and g each lie on a unique connector line, and each represents a unique heterodimer, namely (a) CycE:Cdk2, (c) CycA:Cdk2 (e) CycA:Cdk1, (g) cycB:Cdk1. Nodes b, d, and f, on the other hand, represent dimer combinations, namely (b) Cdk2 complexed with either CycE or CycA; (d) CycA complexed with either Cdk2 or Cdk1; (f) Cdk1 complexed with either CycA or CycB. This notation simplifies the representation of multiple alternative interactions: for example, the interactions of p21, p27, or p57 with various Cyclin: Cdk dimers. A formal rule, required to avoid ambiguity, is that lines representing alternative interactions must join at an acute angle. Reprinted from *Molecular Biology of the Cell*, (1999, volume 10, 2703-2734) with permission by the American Society for Cell Biology.

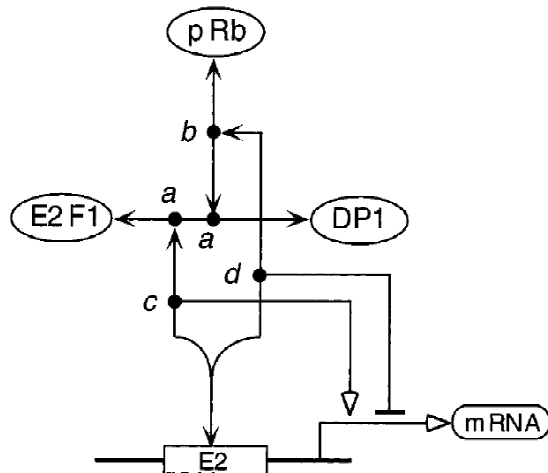


Figure 8. Kohn's representation of multimolecular complexes and integration of information on transcriptional regulation: Stimulatory and inhibitory complexes of E2F1, DP1, and pRb. Note that the promoter element E2 can be occupied either by E2F1:DP1 or by E2F1:DP1: pRb (alternative binding represented by interaction lines joined at an acute angle). Individual complexes are (a) E2F1:DP1 dimer; (b) E2F1:DP1:pRb trimer; (c) E2F1:DP1 bound to promoter element E2 (transcriptional activation shown); (d) E2F1:DP1:pRb bound to E2 (transcriptional inhibition shown). From Kohn 1999. Reprinted from *Molecular Biology of the Cell*, (1999, volume 10, 2703-2734) with permission by the American Society for Cell Biology.

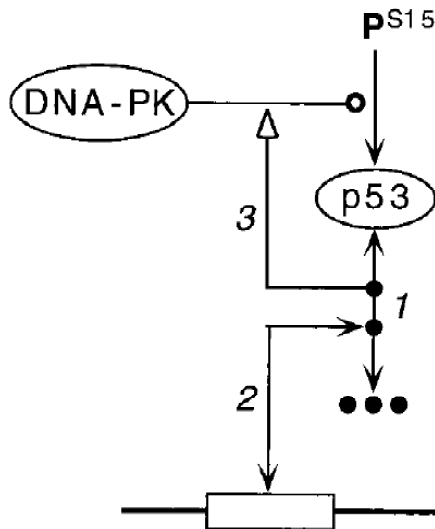


Figure 9. Kohn's representation of homopolymers: formation and effects of p53 homotetramer.  $\circ$  represent additional copies of p53. (1) The three additional copies of p53 monomer required to make up the tetramer are represented by three nodes placed side by side and linked to the identified p53 monomer; a node placed internally on this line represents the homotetramer itself. (2) p53 tetramer can bind to promoter element. (3) Tetramerization stimulates (or is required for) phosphorylation of p53 Ser15. Reprinted from *Molecular Biology of the Cell*, (1999, volume 10, 2703-2734) with permission by the American Society for Cell Biology.

A graph is specified completely by a set of nodes and a set of node pairs as edges, but graph theory does not stipulate where its nodes and edges are to be drawn. To obtain a drawing or layout of a graph, as in Figure 3 (h), one needs to further associate (two- or three-dimensional) coordinates with each node, and specify how the edges are drawn. Performing this task computationally is the object of *graph drawing*, a relatively young subfield of computer science (for an overview, see Battista *et al.* 1999). Various attempts have been made to quantify the quality of a two-dimensional graph layout. According to common definitions of quality, good layouts should have evenly spaced nodes: edges should be straight lines, identical or isomorphic subgraphs should be drawn identically, etc. One of the most prominent criteria, the number of edge intersections in a graph drawing, has been correlated empirically with human ability to solve simple problems using that drawing (Purchase 1997).

*Planar graphs* are graphs that are optimal in this respect, i.e., graphs that can, in some way, be drawn in two dimensions without any edge crossings. Planar graphs are important in applications such as the layout of electronic circuits, where different conducting paths are not allowed to cross each other. Although planar graphs usually permit many layouts without edge crossings, even efficiently testing whether a given graph is planar is not straightforward (Hopcroft *et al.* 1974). However, most graphs, such as those that represent protein-protein interactions, are not planar.

### **Spring embedder algorithm**

The most widely used algorithm for general larger two- and three-dimensional graphs is the spring embedder algorithm (Eades 1984). The layout of a graph is computed by modeling a mechanical system in which the edges of the graph correspond to springs, and nodes correspond to rings. The springs create an attracting force between the rings when they are far apart, and a repulsive force repels close rings. One searches for a placement of rings that minimizes the total energy present in the system, commonly by simulating the behavior of the mechanical system over a certain period of time. Figure 4 was created using a spring embedder algorithm.

### **Limitations for large graphs**

Working with layouts for very large graphs of 100 or more nodes presents certain technical limitations. First, the computer time required to execute most practical layout algorithms does not scale linearly, but rather with the square of the graph size (at least). Many layout objectives, such as the minimization of edge crossings, translate into NP-hard problems (Garey *et al.* 1979; Garey *et al.* 1983) and take even more time to achieve. Second, in an interactive system with thousands of nodes or more, just drawing a dynamically changing graph, even though this is a linear operation, can take unacceptably long.

While faster computers may eventually mitigate the above restrictions, even the best drawings of large graphs under any of the above quality criteria may not be aesthetically pleasing or practically usable.

### **Extension to three dimensions**

The increase in available computing power and the advancement of graphical displays and software standards has inspired work on three-dimensional graph layout. For graphs in three dimensions, the edge-crossing criterion is no longer helpful in selection of good drawings. Every graph can be drawn in three dimensions without edge crossings in many ways (Fary 1948). For displaying protein interactions in three dimensions, a variation of a spring embedder algorithm has been suggested (Basalaj and Eilbeck 1999).

## Techniques for visualization

There are several techniques to relieve the above described problems with the visualization of large graphs.

### Zoom and pan

A common approach is zoom and pan. This is the same technique that is used by Web browsers: Instead of showing a Web page from beginning to end, only part of it is shown at every given moment, and the user can continuously scroll through the Web page by means of a scroll bar.

### Focus and context techniques

Zoom and pan has the disadvantage that zooming makes certain regions of the layout invisible: one creates a focus, but the context is lost. Focus and context techniques avoid loss of context by compressing a layout towards the edges of the window, instead of hiding part of it. One example of a such a technique is the well-known fisheye effect. Note that focus and context techniques are complementary to zoom and pan they can be used together.

### Collapsing protein classes

A third, complementary, technique to simplify a layout collapses groups of proteins (*classes*) into single nodes. Figure 10 was generated from 2709 protein-protein interactions in yeast (Schwikowski *et al.* 2000) on the basis of the functional classification of the involved proteins according to YPD (Costanzo *et al.* 2000). Each node represents a functional class. Proteins that have been assigned multiple functions thus contribute to multiple classes. The aggregated information is summarized in several ways: The number on each edge *A-B* indicates the number of protein interactions between proteins of function *A* and proteins of function *B*. This number is also reflected in the thickness of the edge. The numbers in parentheses indicate the number of intra-class interactions, and the number of proteins in the class, respectively.

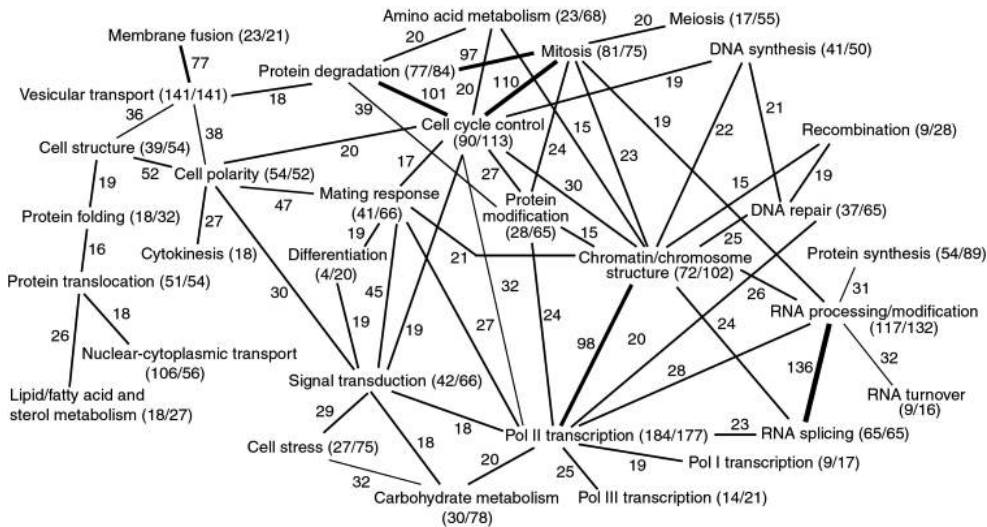


Figure 10. Protein classes by functional classification (after Schwikowski *et al.* 2000).

## Available tools for visualization

Table 3 lists some currently available software packages that visualize arbitrary protein interactions or can be customized for that task. The databases listed in Table 4 visualize predefined, limited sets of protein interactions.

Table 3: Visualization tools for general networks

---

The LEDA library/GraphWin	<a href="http://www.mpi-sb.mpg.de/LEDA">http://www.mpi-sb.mpg.de/LEDA</a>
C++ library for efficient data structures and algorithms; contains graph drawing demo application <b>Platforms:</b> Linux-PC, Sun, Silicon Graphics, HP, Windows 95/NT (commercial) <b>Availability:</b> commercial, free license for academic users. <b>Ref:</b> Mehlhorn <i>et al.</i> 1999	
Y-Files	<a href="http://www-pr.informatik.uni-tuebingen.de/yfiles">http://www-pr.informatik.uni-tuebingen.de/yfiles</a>
Extensible, programmable graph editor, with graph algorithms. Extensions to generation of biochemical pathway diagrams are underway <b>Platforms:</b> PC, Macintosh <b>Availability:</b> free binaries/source code for academic purposes. <b>Ref:</b> Himsolt 1997	
Graphlet	<a href="http://www.infosun.fmi.uni-passau.de/Graphlet/">http://www.infosun.fmi.uni-passau.de/Graphlet/</a>
Extensible, programmable graph editor, with graph algorithms. Extensions to generation of biochemical pathway diagrams are underway <b>Platforms:</b> Windows NT/98 or higher, Solaris, Linux <b>Availability:</b> free binaries/source code for academic purposes. <b>Ref:</b> Himsolt 1997	
XGvis	<a href="http://www.research.att.com/areas/stat/xgobi/index.html">http://www.research.att.com/areas/stat/xgobi/index.html</a>
Interactive visualization system for proximity data, graphs and networks <b>Platforms:</b> Linux, Solaris, other UNIX systems <b>Availability:</b> Free, incl. source code. <b>Ref:</b> Buja <i>et al.</i> 1998	
Tom Sawyer Software	<a href="http://www.tomsawyer.com/">http://www.tomsawyer.com/</a>
Cross-platform software library and tools for drawing general graphs. <b>Availability:</b> commercial <b>Platform:</b> Macintosh, many UNIX versions, standard WWW browsers <b>Platforms:</b> Most major operating systems, incl. Windows, Apple Macintosh	
CUtenet	<a href="http://genome6.cpmc.columbia.edu/~tkoike/cutenet/">http://genome6.cpmc.columbia.edu/~tkoike/cutenet/</a>
Interactive graphic editor for signal-transduction pathways and protein interactions. <b>Availability:</b> ? <b>Platform:</b> Most major operating systems (Java application). <b>Ref:</b> Koike & Rzhetsky 2000	

---

Table 4: Visualization of specific datasets.

---

DIP (Xenarios <i>et al.</i> 2001)	<a href="http://dip.doe-mbi.ucla.edu">http://dip.doe-mbi.ucla.edu</a>	Visualization of protein-protein interactions in DIP database. Static images depict neighborhoods 2 and 3 steps away from center protein. Availability: free Platforms: Web browser
ProNet	<a href="http://pronet.doubletwist.com/">http://pronet.doubletwist.com/</a>	Interactive visualization of protein-protein interactions in the ProNet database Availability: free Platforms: Standard WWW browser
GeneNet (Kolpakov <i>et al.</i> 1998)	<a href="http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/">http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/</a>	Interactive visualization of protein interactions in GeneNet database, based on a number of predefined diagrams. GeneNet includes entries for DNA, RNA, protein, and cellular interactions. Availability: free Platforms: Standard Web browser
PIMRider (Rain <i>et al.</i> 2001)	<a href="http://pim.hybrigenics.com">http://pim.hybrigenics.com</a>	Interactive visualization of protein-protein interactions with different viewers Availability: commercial, free license for academic users Platform: Standard Web browser
BindDB (Bader <i>et al.</i> 2001)	<a href="http://www.binddb.org">http://www.binddb.org</a>	Interactive visualization of protein interactions in BindDB database. BindDB contains general biomolecular interactions. Platform: Standard Web browser Availability: free

---



# INTEGRATING PROTEIN-INTERACTION NETWORKS WITH SUPPLEMENTAL DATA

## Simultaneous display of complementary data types

Proteins are not the only molecules of interest to biologists; nor do protein-interaction networks function in isolation to govern cellular processes or to influence phenotypes. On the contrary, successful visualization of protein-interaction networks leads almost immediately to questions such as: How do these proteins interact with DNA, substrates, and other cellular components? What impact does protein interaction have on the properties and behaviors of each interacting protein?

In fact, it is not hard to envision incorporating a number of supplemental data types into the basic network display. Sources of supplemental data generally fall into one of four categories: new types of interactions, new types of molecules, new information on existing interactions, and new information on existing molecules (see also Table 1). For instance, we may wish to visualize not only interactions between pairs of proteins, but also interactions between these proteins and ligands or other small molecules. Similarly, proteins that function as transcription factors may form complexes that bind particular DNA sequences, as well as interact with a variety of protein co-factors and annotation or organization taking these properties into account may be useful. Moreover, we may find it useful to place the protein-interaction network in the context of indirect evidence such as genetic interactions, or we may wish ultimately to annotate each protein with its instantaneous expression pattern or each interaction with its relative strength of binding as these points become known.

To date, numerous articles and textbooks have included figures displaying different types of molecules and interactions between them. However, these figures usually invoke a limited number of components to describe an isolated biochemical process or signaling pathway, are carefully tailored to illustrate a predetermined concept, and rely heavily on accompanying textual descriptions (Pirson et al. 2000). In contrast, there is a pressing need for visual representations that can *systematically present and organize* the extremely large amounts of protein-interaction and expression data rapidly accumulating in the wake of two-hybrid screens, DNA microarray technology, and high-throughput proteomics. Such displays are not hand-tailored to illustrate a foregone conclusion, but should ideally stimulate the discovery of new protein functions and biological relationships. As the raw data become increasingly complex with each type of supplemental information, tools that are both visual and interactive become increasingly important for emphasizing and extracting the key features.

## An example: integrated networks to study galactose metabolism

We now illustrate one method to systematically create and display an integrated interaction network, as described in Ideker *et al.* (2001). Suppose that we are interested in viewing the molecular interactions that govern a particular cellular process: that of galactose utilization in yeast. Our specific goals might be to assess the impact of these interactions on expression of the galactose-utilization (GAL) genes and to understand how the process of galactose utilization interacts with other metabolic processes in yeast.

To begin, we construct a database representing all known protein-protein and protein-DNA interactions in the yeast *Saccharomyces cerevisiae*. Protein interactions may be drawn from any of the sources listed in Table 2 and 4: in this example, we utilize the 2709 protein-protein interactions compiled by Schwikowski *et al.* (2000). Similarly, we obtain all of the 317 protein-DNA interactions present in either of two publicly-accessible on-line databases (as of July, 2000): TRANSFAC (Wingender et al. 2000) or the *Saccharomyces cerevisiae* Promoter Database (SCPD) (Zhu & Zhang 1999). These sources link known transcription factors (proteins) to the genes they regulate (DNA).

Next, we use a program based on *GraphWin* (Mehlhorn & Nher 1999) to display these physical interactions as a graph structure or *network*, as discussed above. Because several types of interactions are now involved, but all of them reflect physical binding events, we refer to this network as a *physical-interaction network*. As shown in Figures 11 through 13, each node represents a gene and is labeled with its corresponding gene name. An arrow, or *directed edge*, from one node to another signifies that the protein encoded by the first gene can influence the transcription of the second by DNA binding (a protein→DNA interaction), while a line, or *undirected edge*, between two nodes signifies that the proteins encoded by each gene can physically interact (a protein–protein interaction). Network layout is performed using the spring-embedder algorithm included with *GraphWin*, so that proteins with related functions or that are involved in the same molecular pathway often end up in the same region of the display.

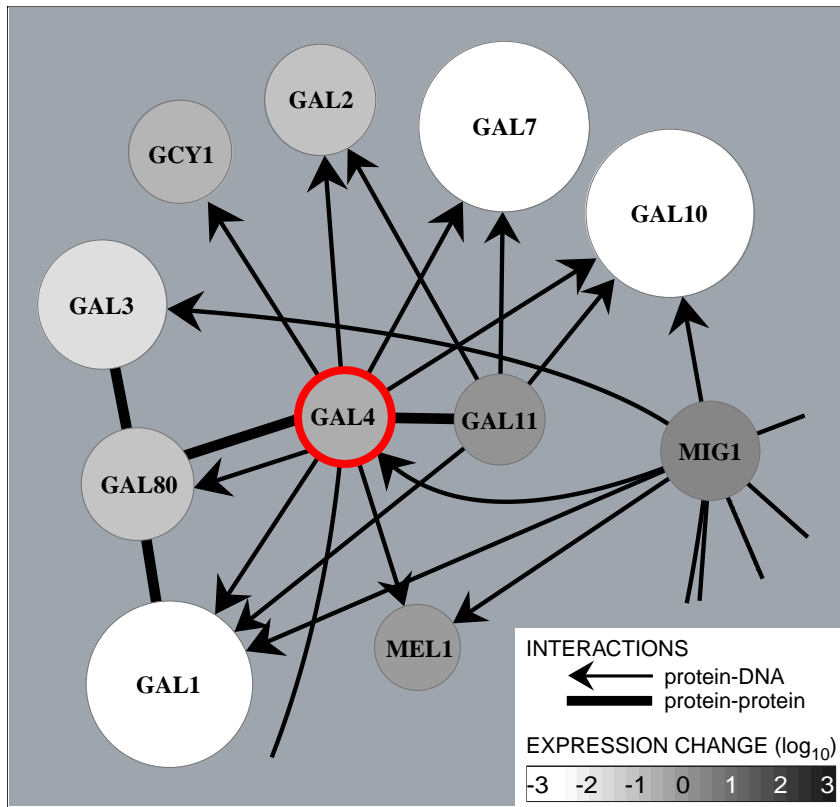


Figure 11. Sample region of an integrated physical-interaction network containing genes involved in galactose metabolism. Each node represents a gene, an arrow directed from one node to another represents a protein-DNA interaction, and an undirected line between nodes represents a protein-protein interaction. The grayscale intensity of each node indicates the change in mRNA expression of its corresponding gene, with medium-gray representing no change and darker or lighter spots representing an increase or decrease in expression, respectively (to draw attention to genes with large expression changes, node diameter also scales with the magnitude of change). To signify that the expression level of *GAL4* has been perturbed by external means, it is highlighted with a red border. According to the interactions in the network, Gal4 regulates expression of many other GAL genes through protein-DNA interactions, while Gal4's activity is impacted by protein-protein interactions with Gal80 and Gal11 (see text for details).

Thus, the region shown in Figure 11 corresponds to the process of galactose utilization, while the regions shown in Figures 12 and 13 correspond to amino-acid biosynthesis and glycogen synthesis, respectively.

According to the network, Gal4p is a transcription factor that binds to the promoters of many other GAL genes, thereby regulating their transcription through protein-DNA interactions (Figure 11). The network also shows clearly that the activity of Gal4p may be influenced by protein-protein interactions with Gal80p and Gal11p. Note that the network specifies only that a particular protein-DNA interaction takes place: it does not dictate whether the interaction activates or represses transcription, whether the effect on transcription is rapid or gradual, or in the case that multiple interactions affect a gene, how these interactions should be combined to produce an overall level of transcription. Since these levels of information are not encoded in the protein-DNA databases, they are also absent from the network display. Similarly, the protein-protein databases do not specify whether the Gal80p-Gal4p protein interaction, as shown in the figure, results in these proteins forming a functional complex or whether one protein modifies another. All of this information is known outside of the databases: classic genetic and biochemical experiments (Johnston & Carlson 1992, Lohr et al. 1995) have determined that Gal4p is a strong transcriptional activator, and that Gal80p can bind to Gal4p to repress this function.

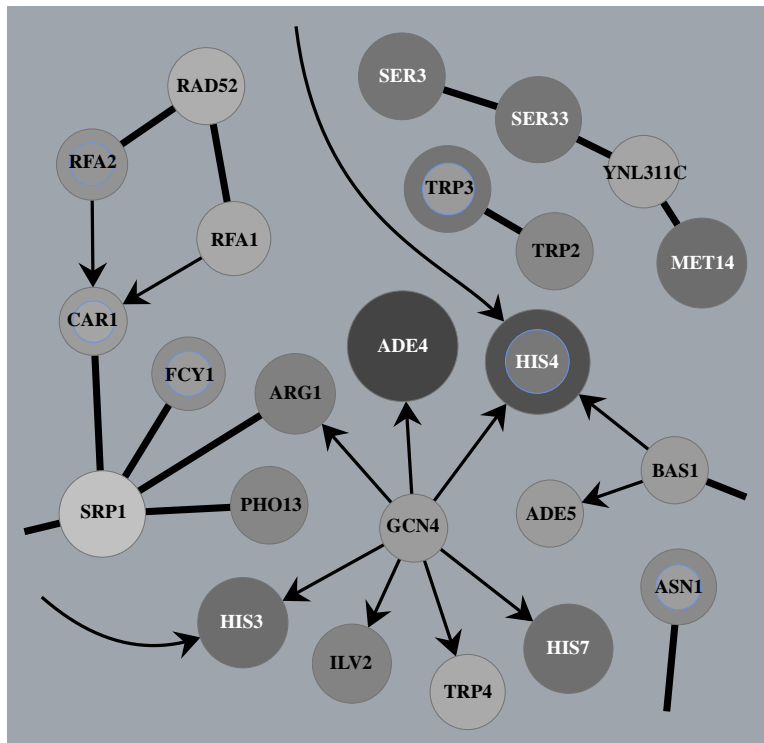


Figure 12. Integration of protein-expression response in the region corresponding to amino-acid biosynthesis. Nodes and interactions appear as in Fig. 11, with a solid grayscale intensity representing the change in mRNA expression. Nodes for which protein data are also available contain two distinct regions: an outer circle, or ring, representing the change in mRNA expression, and an inner circle representing the change in protein expression. Both mRNA and protein intensity scales are identical to that used in Fig. 11.

## Superimposition of mRNA- and protein-expression changes on the network

To better understand how the physical-interaction network regulates genes, it can be extremely effective to augment the network with information about gene expression. As described previously (Ideker et al. 2001), we [T. Ideker] measured global changes in gene expression over 20 genetic and environmental perturbations to the GAL pathway. Wild type (*wt*) and nine genetically-altered yeast strains were examined, each with a complete deletion of a different GAL gene (*gal1Δ*, *2Δ,3Δ*, *4Δ*, *5Δ*, *6Δ*, *7Δ*, *10Δ*, or *80Δ*). These ten strains were perturbed environmentally by growth to steady state in the presence (+gal) or absence (—gal) of 2% galactose. Since many of the deletion strains cannot grow in galactose, 2% raffinose was also provided in both media as an alternate supply of sugar. In each of these 20 perturbation conditions, we monitored changes in mRNA expression over the approximately 6200 nuclear yeast genes using a whole-yeast-genome microarray.

For any particular perturbation, we can integrate, *i.e.* graphically superimpose, the resulting changes in mRNA expression on the network. Although a number of visual representations are possible, an obvious choice is to use node color to represent a change in expression of the corresponding gene. For example, Figure 11 shows the expression changes resulting from the perturbation *gal4Δ*+gal, a deletion of the *GAL4* gene in the presence of galactose.

When protein-expression data are available, they too can be superimposed on the network display. For example, we measured changes in both mRNA *and* protein levels in wild-type cells grown in the presence *vs.* absence of galactose (Ideker et al. 2001). Using a procedure based on isotope-coded-affinity-tags (ICAT) and tandem mass spectrometry (Gygi et al. 1999), we detected a total of 289 proteins and quantified their expression-level changes between these two conditions. Figure 12 illustrates the addition of this information to the visual display, focusing on the region of the network corresponding to amino-acid biosynthesis. By comparing the mRNA- and protein-expression responses displayed on each node, one can visually assess whether the mRNA and protein data are correlated and quickly spot genes for which they are remarkably discordant.

## Integrated networks enable new biological insights

In the previous example, we have integrated at least four types of data into the same graphical display: protein-protein interactions, protein-DNA interactions, mRNA-expression changes, and protein-expression changes. In addition, because interconnected groups of genes tend to have related functions, the display also confers information about cellular function or process. What exactly is gained from this level of integration? Is superimposing data from multiple, complex sources on the same graphical display really worth all of the added clutter?

Most generally, the integrated network display is useful because it provides a lucid means of summarizing existing biological knowledge about molecular behavior. Although individual researchers may amass a great deal of knowledge about the molecular interactions underlying one particular pathway, no single biologist can be familiar with the extremely large and complex number of interactions in an entire cell. A computer database, however, tracks all of these, provided the proper representation is available to allow a biologist to access, display, and interpret the information. Moreover, since changes to the database are automatically reflected in the graphical display, the integrated network is continually up-to-date. In short, the physical-interaction databases and the graphical display together constitute *an expert system*, providing knowledge about the molecular makeup of the cell which can be queried and viewed by a biologist.

In our case study of galactose utilization, the network display engenders at least three types of biological insights. First, it provides plausible cause-and-effect explanations for numerous changes in gene expression observed in response to each of the twenty perturbations to the GAL

pathway. Note that when *GAL4* is deleted in Figure 11, expression levels of *GAL1*, 7, 10, and several other genes decrease dramatically, consistent with *GAL4*'s known role as a transcriptional activator. Similarly, in Figure 12 we see that the increase in expression of *HIS3*, *HIS4*, *HIS7*, *ADE4*, *ARG1* and *ILV2* could be controlled by the *GCN4* transcription factor. Interestingly, *GCN4* itself does not change perceptibly in mRNA expression. However, because seven of the eight genes it regulates *do* change, we feel intuitively that *GCN4* is somehow involved. A subsequent literature search on *GCN4* reveals that this gene is in fact regulated translationally, not transcriptionally (McCarthy 1998), a detail not represented by the network display because we have not yet measured protein-expression changes for *GCN4*. Thus, through a rapid visual scan, we can determine which gene-expression changes could be caused by known protein-DNA interactions, and which changes require further research to identify the particular transcription factors involved. Going a step further, one can then seek explanations for how each transcription factor is itself controlled, either by protein-DNA interactions with still other transcription factors, or through protein-protein interactions with cofactors or signaling proteins.

Second, the network graph highlights groups of physically-interacting proteins that display joint increases or decreases in expression level across many experimental conditions. These coordinate changes suggest that the proteins are controlled by one or more common transcription factors. For instance, the genes *GAC1*, *GIP1*, *PIG2*, and *GSY2*, shown in Figure 13, are not only involved in protein-protein interactions with each other, but *display concomitant increases in expression*. Moreover, the expression levels of these genes are highly correlated over the twenty perturbation conditions. Examples of *inverse* regulation are also abundant among physically interacting proteins, where often one protein is known to inactivate the other. For example, we observe an increase in expression of Gsy2p, a glycogen synthase, and a corresponding decrease in expression of Pcl10p, a protein that interacts with and inactivates Gsy2p (Wilson et al. 1999). Thus, the integrated network suggests that glycogen synthesis is controlled by upregulating an enzyme *and* downregulating the enzyme's inhibitor.

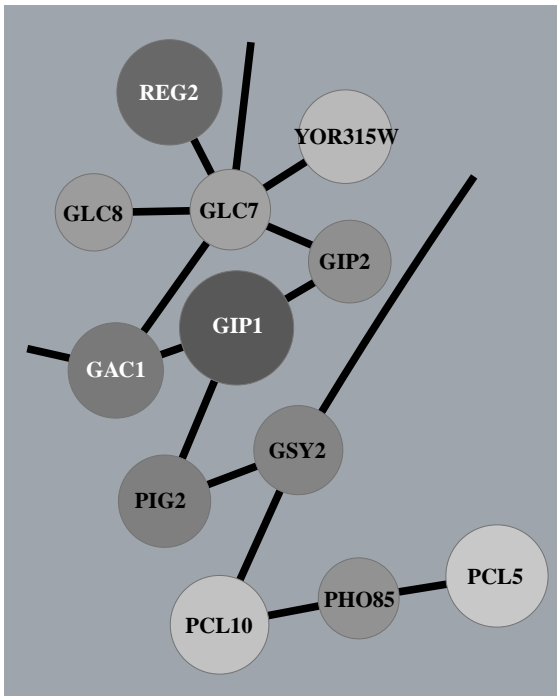


Figure 13. Co- and inverse-regulation of interacting proteins. The integrated physical-interaction network is shown as in Fig. 11, for the region corresponding to glycogen synthesis. Changes in mRNA-expression are due to deletion of the *GAL4* gene (a *gal4Δ* vs. wild-type strain, galactose present). In this and many other perturbation conditions, the Gac1-Gip1-Pig2-Gsy2 enzyme complex increases in mRNA expression, while Pcl10 (which functions to inactivate Gsy2) shows a corresponding expression decrease.

Finally, the network graph may be used to confirm newly discovered or controversial interactions. Consistent co- or inverse-regulation between two physically-interacting proteins, over many perturbation conditions, provides strong evidence that the interaction occurs *in vivo* and is not an artifact of the particular assay originally used to determine the interaction. This confirmation is especially useful given that some experimental techniques for establishing a physical interaction, such as the two-hybrid screen, may return a substantial number of false-positive interactions.

### **Choice of graphical representation**

Of course, it is not necessary to implement the same structural or color conventions used in the example. Nonetheless, due to the overwhelming amount of information to be loaded onto a single visual display, a clear, efficient, and consistent graphical representation remains extremely important. In the words of Edward Tufte (1983),

*Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.*

In constructing our integrated interaction graphs, we have relied upon a few short, common-sense guidelines:

1. *High visual density is to be desired, not avoided.* Often, clutter and confusion are failures of design, not complexity (Tufte 1983). Many different attributes can be varied on the same graph, each conveying a different type of information: node size and color; node border width and color; node-label font, font size, and font color; edge directionality, width, and color. We have varied only a small subset of these attributes in the visual displays of the previous example.
2. When including many different types of data in the same display, *use the fewest number of colors required to represent each type.* The key to managing visual complexity is to make each type of information use the fewest graphical resources possible. Because the eye is *particularly drawn to changes in color*, color should be used judiciously to emphasize only the most important features of the display. For example, although changes in gene-expression are often displayed with two color scales (*e.g.*, red for increases in expression and green for decreases in expression), these data are fundamentally one-dimensional; if necessary, they can be displayed with a single color or grayscale (*e.g.*, Figure 11), freeing up a larger range of colors to encode other types of data. Moreover, contrasts of red and green are particularly ill suited to convey information to the ~8% of men who are fully or partially red-green color blind (Passarge, 1995). In short, careful use of color is the key to representing highly-dimensional data sets.
3. *Display new types of data only if there is a clear biological goal that relies on these data.* The choice of which information to display, out of all information accessible from the databases, must always be driven by biological inquiry. For instance, if our goal is to understand which protein-DNA interactions can cause a particular gene expression pattern, information such as the amino-acid sequences of each protein or the genomic location of the corresponding gene would be of low interest. Alternatively, each of these data types may be assigned to a distinct visual *layer*, which can be temporarily hidden when the data in the layer are not directly relevant.

**From visual representation to a predictive model: early steps towards gene-expression modeling using physical-interaction networks**

Beyond their use as graphical displays, physical-interaction networks can function as *predictive models* of the cell. For instance, with a few added formalisms that we shall discuss shortly, the physical-interaction network developed in our example can predict all of the changes in gene expression that could be caused by a particular perturbation. Such predictions are highly informative when compared to their true, observed values measured in laboratory experiments: for any gene expression level that changes in experiment but not in simulation, we may conclude that one or several physical interactions are unknown and/or absent from the network model.

Predictions are obtained by running simulations. To see how, recall that the network represents two types of interactions, protein-DNA and protein-protein. These interactions can produce very different effects: a protein-DNA interaction can affect the expression level of a gene, while a protein-protein interaction can cause a protein to become active or inactive with regard to its biological function. However, a protein-protein interaction by itself cannot elicit a change in gene expression (without the aid of an associated protein-DNA interaction), just as a protein-DNA interaction usually affects protein activity only indirectly, by influencing whether the protein is expressed.

These interaction types imply at least two types of information associated with each node X: a gene expression level  $X_e$  and a protein activity  $X_a$ . A perturbation to the network may elicit a change in  $X_e$  through an incoming protein-DNA interaction from a node Y ( $Y \rightarrow X$ ), if Y also undergoes a change in activity  $Y_a$ . In contrast,  $X_a$  may change either if the perturbation causes a change in  $X_e$ , or if X is involved in a protein-protein interaction with a node Z which undergoes a corresponding change in activity  $Z_a$ . Figure 14 summarizes these rules.

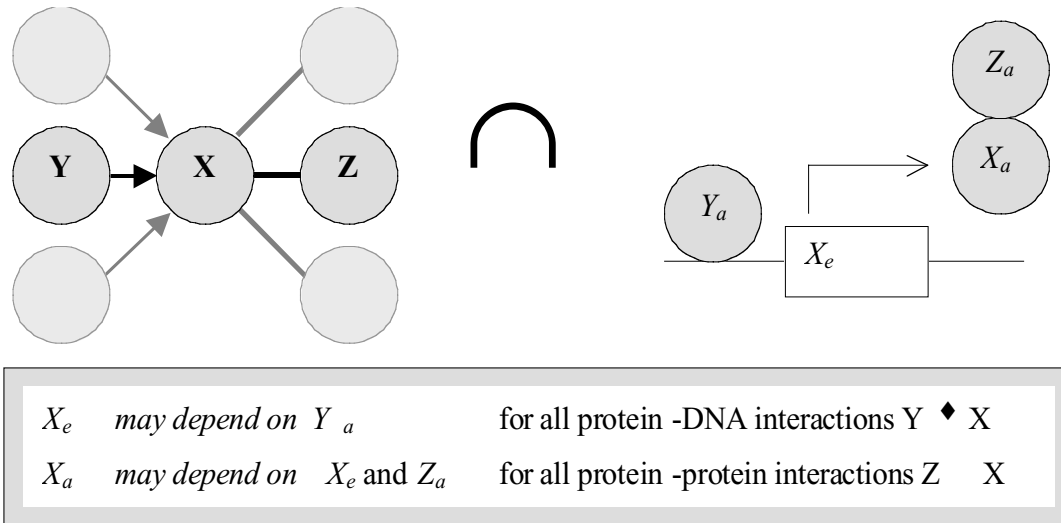


Figure 14. Rules governing the effects of physical interactions on gene expression and protein activity. Each node X in the physical-interaction network (left) has an associated expression level  $X_e$  and activity level  $X_a$  (right). These levels may change in response to (*i.e., may depend on*) changes at adjacent nodes Y or Z according to rules highlighted in the box. Rules are qualified with the word *may* because the network does not specify in what conditions the interactions occur or how multiple interactions should be combined to produce an overall expression or activity level.

Along with the network model, these rules are sufficient to predict possible changes in gene expression resulting from perturbation of any particular node in the network. It then becomes straightforward to perform these simulations automatically, by implementing these rules directly in software. Since the network model does not specify precisely how a node combines the relevant input interactions to determine its expression level or protein activity, it is not possible to state definitively whether a change actually occurs: we only know if a change *may* occur. However, it *is* possible to predict which nodes are not, under any circumstances, affected by a particular perturbation to the network.

### Prediction of expression changes resulting from particular perturbations to the galactose-utilization network

As an example, consider Figure 15a, which once again displays the region of the physical-interaction network corresponding to galactose utilization. In this case, the network has been perturbed by deletion of the *GAL3* gene in the presence of galactose. The resulting changes in gene expression, as observed by microarray experiment, are superimposed on this graph. To perform the corresponding simulation, we reason that deletion of *GAL3* is likely to affect *GAL3<sub>e</sub>* and *GAL3<sub>a</sub>*. In turn, a change in *GAL3<sub>a</sub>* may affect *GAL80<sub>a</sub>*, which then may affect *GAL1<sub>a</sub>* or *GAL4<sub>a</sub>*, as mediated through protein-protein interactions.

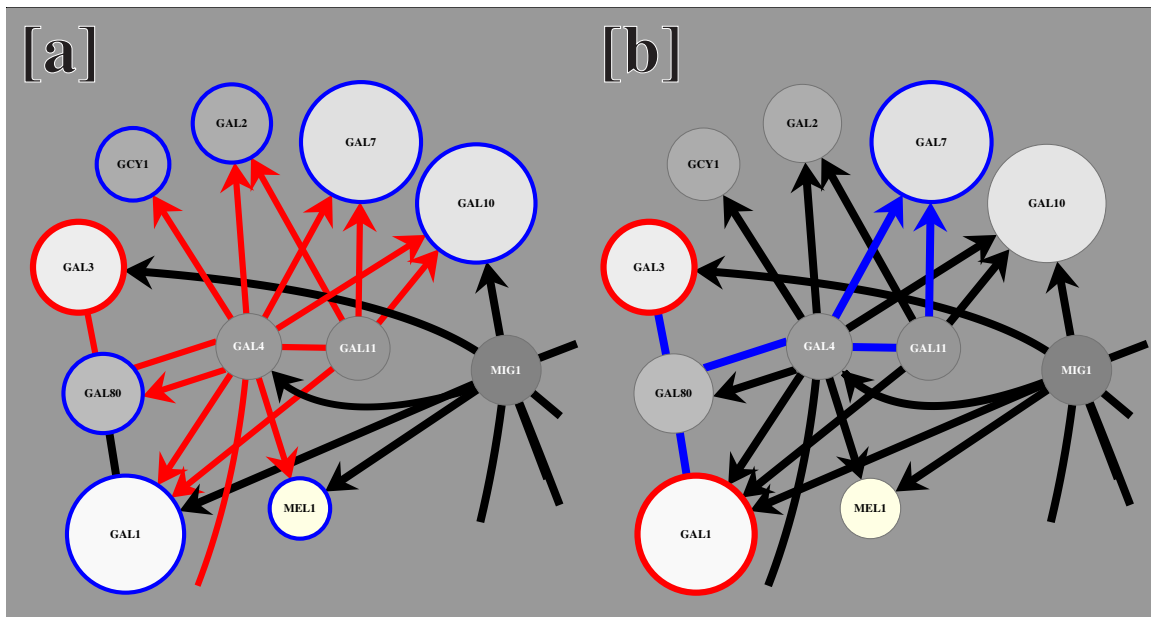


Figure 15. Predicting changes in gene expression in the region of the integrated network corresponding to galactose metabolism. The grayscale intensity of each node reflects the experimentally-observed change in mRNA expression for a *gal3Δ* vs. wild-type strain in galactose. **(a)** Forward simulation, starting from the perturbed gene *GAL3* (highlighted in red). Red edges denote interactions that may transmit a change, either in expression or activity, from one node to another (according to the rules described in Fig. 14). Nodes highlighted in blue denote genes whose expression levels may change as a result. Experimentally-observed expression changes in these blue genes are consistent with the simulation. **(b)** Reverse simulation, tracing backwards from *GAL7* (highlighted in blue), whose expression state has changed in response to perturbation of the network. Here, blue edges denote interactions which may transmit a change, leading to nodes that are highlighted in red if their corresponding genes were observed to change significantly in expression. These red nodes are possible causes of the change observed at *GAL7*.



Although a change in  $GAL1_a$  has no further impacts on the network, a change in  $GAL4_a$  can affect  $GAL7_e$ ,  $GAL10_e$ , and many other expression levels through protein-DNA interactions. Thus, experimentally-observed expression changes in  $GAL7$  and  $GAL10$  are consistent with those predicted by the network.

For observed changes in gene expression that are not predicted by the network model, a legal path between the perturbed and affected gene does not exist. However, it is possible that a *segment* of this path is present in the network, offering at least a *partial* explanation for the observed change. One approach to finding this partial path is to start at an affected gene and work backwards towards the perturbed one. Figure 15b gives an example of this type of simulation. Here, we start at  $GAL7$ , which exhibits a clear decrease in gene expression under this perturbation condition. Working backwards, we see that a change in  $GAL7_e$  could be explained by incoming protein-DNA interactions implicating  $GAL4_a$  or  $GAL11_a$ .  $GAL4_a$  could likewise be affected by  $GAL80_a$ , and  $GAL80_a$  affected by  $GAL1_a$  or  $GAL3_a$ . Since dramatic changes in gene expression were observed for  $GAL1$  and  $GAL3$ , they become possible causes of the change observed at  $GAL7$ .

Although our simulation implicates  $GAL3$  as a possible cause of the expression change at  $GAL7$ , the perturbed node may not always be reachable through a backwards path. However, in performing the simulation, we hope to identify upstream nodes that are one or several steps closer to it.

## FUTURE DIRECTIONS

In the future, the available methods for data integration and network visualization should be extended in a number of important directions. First, there is a need for more complex integration schemes than have been presented here. For instance, although small molecules such as metabolites, drugs, or hormones are known to directly influence the expression of many genes and proteins, they do not appear in the network graph. One could represent these compounds as nodes in the graph and define a new type of physical interaction to represent the enzymatic transformation of one metabolite to another.

Second, we also need better algorithms for automated layout. Although the spring embedder and similar algorithms draw the graph so that strongly-connected subsets of nodes are grouped together in two dimensions, an improved layout algorithm would not only group together interacting proteins, but could attempt to explicitly group together nodes that are of similar biological function, subcellular localization, or that have similar gene-expression responses to perturbation.

Finally, it would also be extremely useful to display a much wider range of information about existing nodes and interactions. For example, the improved display might supply information about the structural or functional implications of a protein-protein interaction. Are the interacting proteins subunits of a larger complex, or does the interaction instead result in covalent modification of one of the proteins? Alternatively, when using the network graph as a guide to explain experimental observations, one might like to know how much *confidence* to place in each interaction. For example, was the interaction predicted computationally or determined experimentally, and is it supported by corroborating evidence? We anticipate that these types of data will become increasingly available as annotation of the public databases becomes more systematic and complete. The challenge will then be to integrate the new data in such a way as to increase our understanding of the underlying biological processes, not obscure them in convoluted figures or excessive detail. Ultimately, these added layers of information will make the network even more powerful as a model on which simulations may be performed to predict experimental outcomes.

## REFERENCES

- Artavanis-Tsakonas S., Rand M.D., and Lake R.J. 1999. Notch signaling: cell fate control and signal integration in development. *Science* 284, 770-776.
- Bader G.,D., Donaldson I., Woting,C., Ouellette B.F.F., Pawson T. & Hogue C.W.V. 2001. BIND- The biomolecular interaction network database. *Nucl. Acids Res.* 29 (1): 242-245.
- Basalaj W. and Eilbeck K.. 1999. Straight-Line Drawings of Protein Interactions. In *Graph Drawing 1999* (ed. J. Kratochvil), pp. 259-266. Springer-Verlag, Heidelberg-New York.
- Bollob s B. 1998. Modern graph theory. Springer, New York.
- Buja A., Swayne D.F., Littman M. and Dean N. 1998. XGvis: Interactive Data Visualization with Multidimensional Scaling. To appear in the *Journal of Computational and Graphical Statistics*, available from <http://www.research.att.com/~andreas/xgobi/>.
- Cook D.L., Farley J.F., and Tapscott, S.J. 2001. A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. *Genome Biology* 2001, 2 (4): research0012.1—0012.10 (online at <http://genomebiology.com/2001/2/4/research/0012>)
- Costanzo M.C., Hogan J.D., Cusick M.E., Davis B.P., Fancher A.M., Hodges P.E., Kondu P., Lengieza C., Lew-Smith J.E., Lingner C., Roberg-Perez K.J., Tillberg M., Brooks J.E. and Garrels J.I. 2000. The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): Comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res* **28**: 73-76.
- Eilbeck K., Brass A., Paton N. and Hodgman, C. 1999. INTERACT: an object oriented protein-protein interaction database. *Ismb.* 1999;:87-94.
- Eades P. 1984. A Heuristic for Graph Drawing. *Congressus Numerantium* **42**: 149-160.
- Eades P., Lai W., Misue K. and Sugiyama K. 1991. Preserving the mental map of a diagram. In *Compugraphics '91*, pp. 34-43.
- Edwards P.A., Tabor D., Kast H.R. and Venkateswaran, A. 2000. Regulation of gene expression by SREBP and SCAP. *Biochim Biophys Acta.* **15**;1529 (1-3):103-113.
- Fary I. 1948. On Straight Lines Representation of Planar Graphs. *Acta Sci. Math. Szeged* **11**: 229-233.
- Garey, M.R. and D.S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company.
- Garey M.R. and D.S. Johnson. 1983. Crossing number is NP-complete. *SIAM Journal of Algebraic and Discrete Methods* **4**: 312-316.
- Gygi S.P., Rist B., Gerber S.A., Turecek F., Gelb M.H. and Aebersold R. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994-999.
- Hargreaves, D., Rice, D. W., Sedelnikova, S. E., Artymiuk, P. J., Lloyd, R. G., Rafferty, J. B. 1998. Crystal structure of *E.coli* RuvA with bound DNA Holliday junction at 6 Å resolution. *Nat Struct Biol* **5**: 441.

- Himsolt M. 1997. The Graphlet System (system demonstration). In *Proc. of Graph Drawing '96*, pp. 233-240.
- Hopcroft J. and Tarjan R.E. 1974. Efficient Planarity Testing. *Journal of the ACM* **21**: 549-568.
- 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.
- Johnston M. and Carlson M. 1992. *Regulation of Carbon and Phosphate Utilization* (eds. Jones, E., Pringle, J. & Broach, J.) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor).
- Kolpakov F.A., Ananko E.A., Kolesov G.B. and Kolchanov N.A. 1998. GeneNet: a gene network database and its automated visualization. *Bioinformatics* **14** (6): 529-537.
- Kohn K.W. 1999. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* **10**, 2703-2734.
- Koike T. and Rzhetsky A. 2000. A graphic editor for analyzing signal-transduction pathways. *Gene* **259**: 235-244.
- Kffner R., Zimmer R. & Lengauer T. 2000. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics* **16** (9): 825-836.
- 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.
- McCarthy J. E. G. 1998. Posttranscriptional control of gene expression in yeast. *Microbiol Mol Biol Rev* **62**, 1492-1553.
- Mehlhorn K. and Nher S.. 1999. *LEDA. A platform for combinatorial and geometric computing*. Cambridge University Press, Cambridge.
- Mewes H.W., Frishman D., Gruber C., Geier B., Haase D., Kaps A., Lemcke K., Mannhaupt G., Pfeiffer F., Schuller C., Stocker S. and Weil B. 2000. MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **28**, 37-40.
- Michal G. 1993. *Biochemical Pathways* [Poster]. Boehringer Mannheim GmbH.
- Michal G. 1998. On Representation of Metabolic Pathways. *Biosystems* **47**: 1—7.
- Palma P.N., Krippahl L., Wampler J.E., and Moura J.J. 2000. BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins*. **39** (4):372-384.
- Passarge E. 1995. *Color Atlas of Genetics*. Thieme, Stuttgart.
- Pirson I., Fortemaison N., Jacobs C., Dremier S., Dumont J.E. and Maenhaut C. 2000. The visual display of regulatory information and networks. *Trends Cell Biol.* **10**, 404-408.
- Purchase H.C. 1997. Which Aesthetic Has the Greatest Effect on Human Understanding? In *Proceedings of Symposium on Graph Drawing GD '97*, pp. 248-261, Berlin.
- Rain J.-C., Selig, L., De°Reuse, H., Battaglia V., Reverdy C., Simon S., Lenzen G., Petel F., Wojcik J., Schchter V., Chemama Y., Labigne A. and Legrain P. 2001. The protein—protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211 - 215.
- Sanchez R., Pieper U., Mirkovic N., de Bakker P.I., Wittenstein E. and Sali A. 2000. MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* **28** (1):250-253.

- Schwikowski B., Uetz P. and Fields S. 2000. A network of protein-protein interactions in yeast. *Nature Biotechnology* **18**: 1257-1261.
- Serov V.N., Spirov A.V. and Samsonova M.G. 1998. Graphical interface to the genetic network database GeNet. *Bioinformatics* **14** (6): 546-547.
- Strogatz S.H. 2001. Exploring complex networks. *Nature* **410**: 268-276
- Tollis I.G., Battista G.D., Eades P., and Tamassia R.. 1999. *Graph Drawing - Algorithms for the Visualization of Graphs*. Prentice Hall, Upper Saddle River, New Jersey.
- Tsai C.J., Lin S.L., Wolfson H.J. and Nussinov R. 1996. Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. *Crit Rev Biochem Mol Biol.* **31**(2):127-152.
- Tufte E. R. 1983. *The visual display of quantitative information* (Graphics Press, Cheshire, Conn. (Box 430, Cheshire 06410),
- 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., Vijayadmodar 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.
- Vitkup D., Melamud E., Moulton J. and Sander C. 2001. Completeness in structural genomics. *Nature Struct. Biol.* **8** (6): 559-566
- von Dassow G., Meir E., Munro E.M. and Odell G.M. 2000. The segment polarity network is a robust developmental module *Nature*. **406** (6792):188-192.
- Wilson W. A., Mahrenholz A. M. and Roach P. J. 1999. Substrate targeting of the yeast cyclin-dependent kinase Pho85p by the cyclin Pcl10p. *Mol Cell Biol* **19**, 7020-7030.
- Wingender E., Chen X., Hehl R., Karas H., Liebich I., Matys V., Meinhardt T., Pruss M., Reuter I. and Schacherer F. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**, 316-319.
- Xenarios I., Fernandez E., Salwinski L., Duan X.J. Thompson M.J., Marcotte E.M., and Eisenberg D. 2001. DIP: the database of interacting proteins: 2001 update. *Nucleic Acids Res.* **29** (1): 239-241.
- Yates J.R. 3rd. 2000. Mass spectrometry. From genomics to proteomics. *Trends Genet.* **16** (1):5-8.
- Zhu J. and Zhang M. Q. 1999. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* **15**, 607-611.

Updates and corrections to this chapter will be posted at <http://www.systemsbiology.org/pubs/vizprotein>