

Towards an understanding of complex protein networks

Chandra L. Tucker, Joseph F. Gera and Peter Uetz

Large-scale two-hybrid screens have generated a wealth of information describing potential protein–protein interactions. When compiled with data from systematic localizations of proteins, mutant screens and other functional tests, a network of interactions among proteins and between proteins and other components of eukaryotic cells can be deduced. These networks can be viewed as maps of the cell, depicting potential signaling pathways and interactive complexes. Most importantly, they provide potential clues to the function of previously uncharacterized proteins. Focusing on recent experiments, we explore these protein–interaction studies and the maps derived from such efforts.

The recent completion of many genome-sequencing projects has prompted a shift in the focus of large-scale biological science from DNA (genomics) to RNA (transcriptional profiling) and proteins (proteomics)^{1,2}. While sequence data provide a necessary framework of knowledge, they are in most cases insufficient for understanding the biological function of particular proteins or understanding the interplay of these proteins with other molecules in a cell. A complete understanding of protein functionality will require information on many levels: knowledge of transcriptional, translational and posttranslational regulation, binding constants, structures, protein interactions and cellular networking. However, answering basic questions, such as what partners does each protein have, should provide a framework onto which more complex regulatory information can be built. In this review, we discuss the efforts of several groups in the large-scale identification and display of protein interactions. The culminations of some of these studies are protein–interaction maps that represent a population of interacting proteins displayed as networks or circuits. Such networks show not only the potential binding partners of a specific protein but also the complexities of these interactions on a global level.

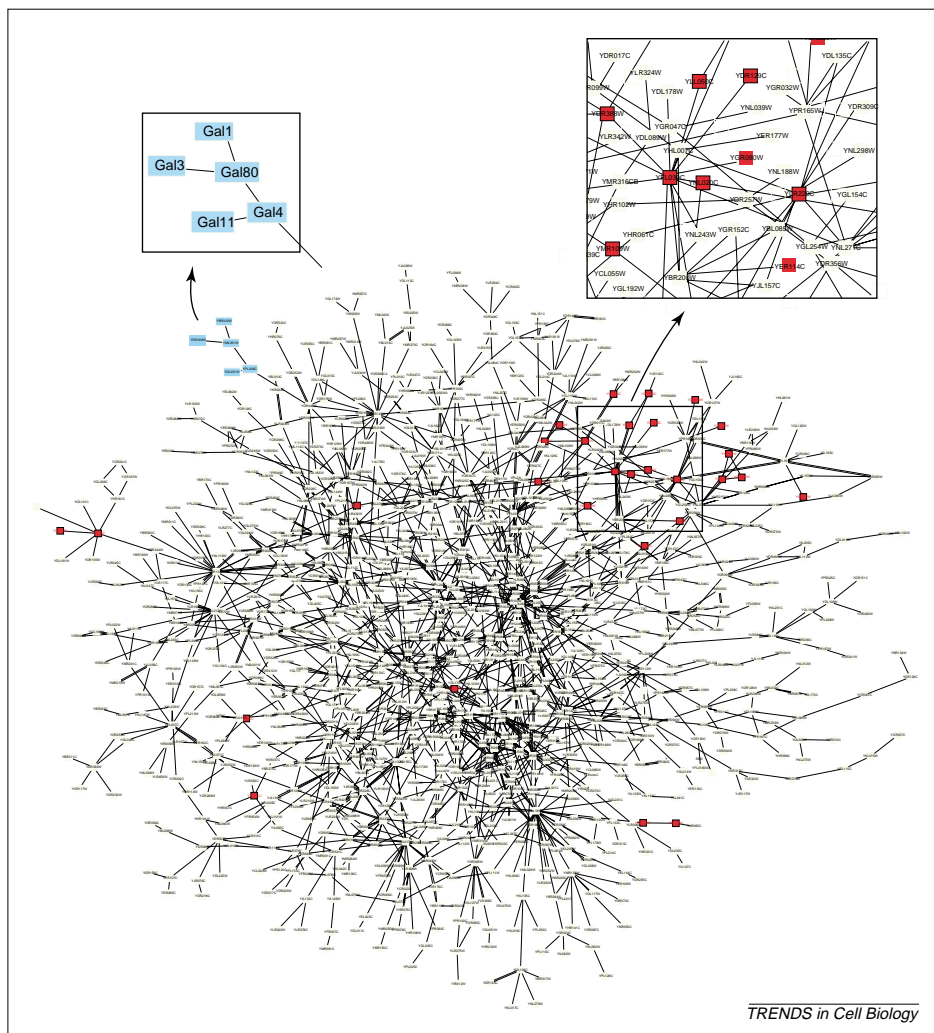


Fig. 1. Yeast protein interaction network of ~1200 interacting proteins based on published interactions (modified from Ref. 10). Inset shows close-up of region highlighted in box. Highlighted in red are cell structure proteins (a single functional class). Proteins in this category can be observed to cluster primarily in one region. Although interacting proteins are not depicted in a way that is consistent with their known cellular location (i.e. those proteins known to be present in the nucleus in the center of the interaction map and those present in plasma membranes in the periphery), signal-transduction pathways (or at least protein contact paths) can be inferred from this diagram. Also highlighted (in blue) are proteins involved in galactose regulation. These proteins are put in a functional context in Fig. 4. To experience the full detail of this figure, readers are advised to view the online version (<http://journals.bmn.com>).

Protein–interaction maps

One of the primary methodologies that allowed for the large-scale analysis of protein interactions was the development of the yeast two-hybrid system³. Through this method, proteins could be assayed for interaction simply by measurement of the growth of yeast colonies on a plate. The first large-scale two-hybrid data were generated for the T7 bacteriophage⁴, whereas other

analyses have focused on selected proteins from *Caenorhabditis elegans*⁵ and *Saccharomyces cerevisiae*^{6–8}. Recently, a two-hybrid protein–interaction analysis was undertaken of the entire *S. cerevisiae* proteome⁹. The results from this project were combined with other *S. cerevisiae* interaction data annotated in the Yeast Proteome Database (YPD; <http://www.proteome.com>) and the MIPS database (Munich Information

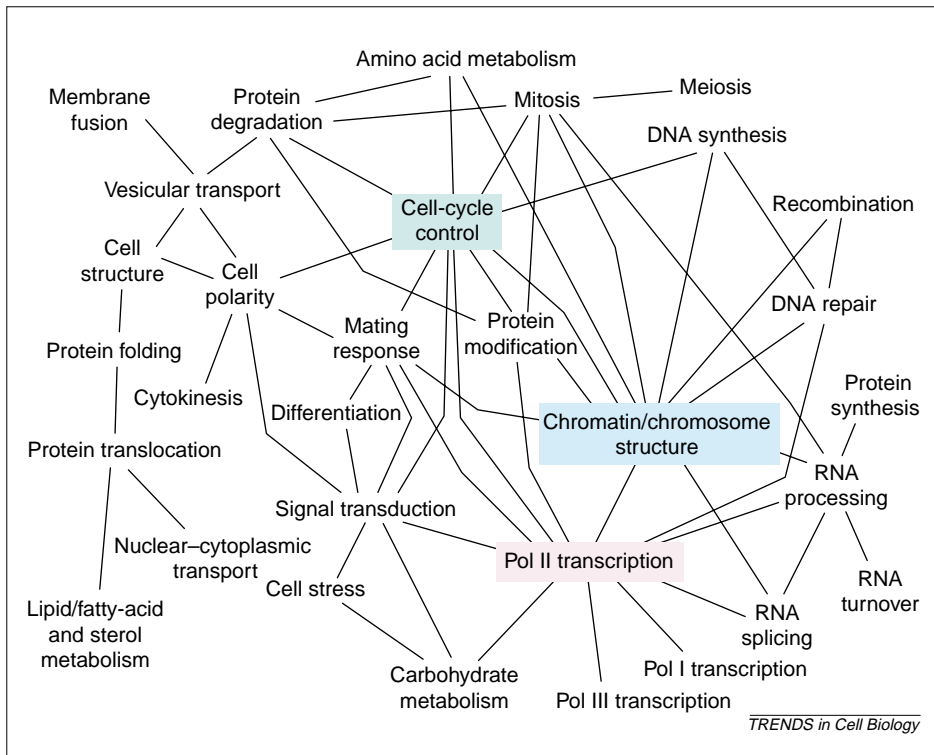


Fig. 2. Functional group interaction map based on Fig. 1 (modified from Ref. 10). Shown are interactions between functional groups of yeast proteins. Each line indicates that there are 15 or more interactions between proteins of the connected groups. Connections with fewer than 15 interactions are not shown because one or a few interactions occur between almost all groups and often tend to be spurious – that is, based on false positives in two-hybrid screens or other assays. Note that only proteins with known function are included and that about one-third of all yeast proteins belong to several classes.

Center for Protein Sequences; <http://www.mips.biochem.mpg.de>) to generate a global yeast protein-interaction map¹⁰. Several public databases also contain protein-interaction maps, including Myriad Corporation's Pronet[®] database (<http://www.myriad-pronet.com>), which is generated solely from previously reported interactions of human proteins, and Curagen Corporation's PathCalling[®] database (<http://portal.curagen.com>). Protein interactions have additionally been deduced by purely computational methods¹¹.

Protein network organization

The graphical representations of protein-interaction maps provide a rough outline of the complexity of protein associations. Schwikowski *et al.*¹⁰ compiled a list of about 2700 published protein interactions from the *S. cerevisiae* literature and found that 1548 yeast proteins could be depicted in a single large network (Fig. 1). When proteins of specific functional categories are highlighted in this network, as shown in Fig. 1 for cell structure proteins

(in red), proteins of like function tend to cluster together. By classifying proteins into these types of functional categories, Schwikowski *et al.* also generated a functional linkage map from this data¹⁰ (Fig. 2). Proteins in certain functional classes, such as cell-cycle regulation, transcription and chromatin regulation, have interactions with proteins of many other classes, consistent with their central roles in the cell. Other processes such as membrane fusion are more isolated, with proteins interacting mainly within this group or with a related group, vesicular transport.

The functional classification of proteins also allowed Schwikowski *et al.* to evaluate the plausibility of the network¹⁰. They found that 72% of all interactions between experimentally characterized proteins in this network are between two partners of the same functional class. When the interactions are randomized among the same set of proteins, only 12% of all interactions belong to the same class¹⁰. It is harder to evaluate the plausibility of interactions among proteins of different functional classes. These might be false-positive

interactions, but they also might be crosstalk interactions or interactions in related pathways. Even if an interaction appears implausible, they might be related pathways in which a link has not been uncovered. Two examples of seemingly implausible interactions are those between membrane-associated proteins and transcription factors as in β -catenin and TCF/LEF1¹² or the Notch receptor and Su(H)¹³.

Regulatory networks similar to that shown in Fig. 2 for yeast proteins can also be drawn for vertebrates. Figure 3 shows an interaction map of human proteins generated by the Myriad Pronet[®] database (www.myriad-pronet.com) based on the signaling pathway of the tumor-suppressor BRCA1. The original map was greatly simplified by removing all single interactors (proteins that interact with only one other protein) and single diverging branches (pathways that branch off and eventually dead-end), leaving a set of 'core proteins' that contain interaction pathways that connect to each other. In simplifying this map, we obtain a picture of the cell not unlike that shown with yeast. The core interactors fall into the same three central functional categories displayed in Fig. 2 – 21 out of 24 core proteins are involved in growth control, chromosome structure/chromatin remodeling or transcription. It is interesting, but perhaps not surprising, that the majority of these core proteins are associated with disease, given their central roles in growth control and gene regulation. Disease proteins are typically the most studied and thus have the most complete sets of interaction data. When examining maps generated from interactions reported in the literature, this bias must obviously be considered. As the number of large-scale interaction studies increases, it will be interesting to see how maps generated from less biased data sets differ from the literature-based maps. It is likely that the most studied proteins have the most, or the most central, interactions because these are typically the proteins involved in crucial functions in the cell. This concept is supported by data from large-scale yeast two-hybrid screens, where known proteins tend to have twice as many interactions as uncharacterized ones (P. Uetz *et al.* unpublished).

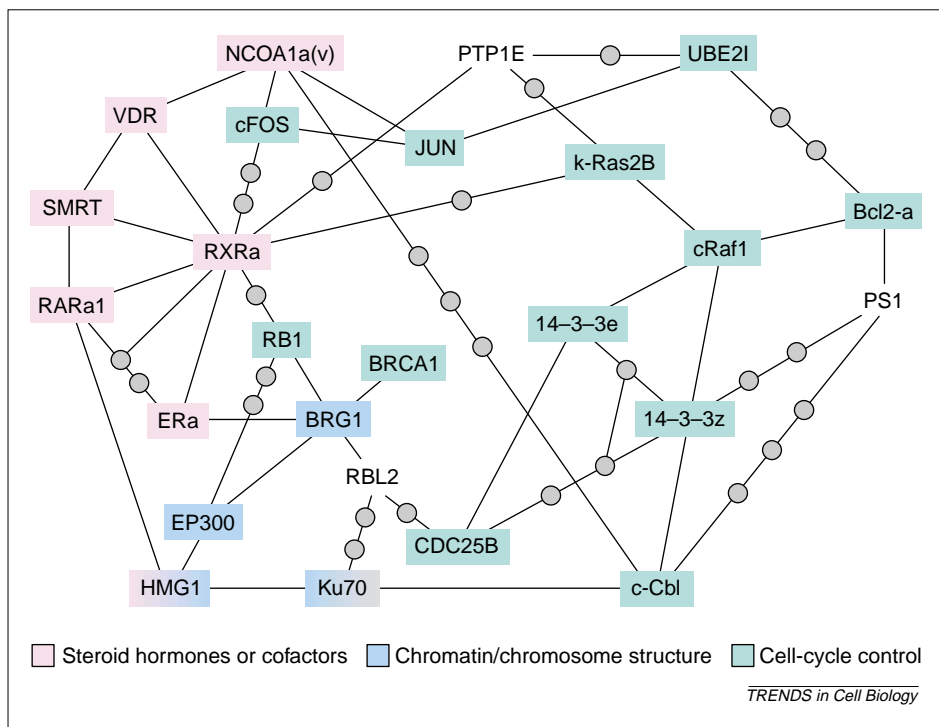


Fig. 3. Simplified protein-interaction map generated using Myriad Pronet® software (www.myriad-pronet.com), showing core proteins involved in the BRCA1 signaling network. This map was generated by removing all single-interacting proteins as well as any single pathways that branch off from the core. Highlighted in pink are proteins involved in transcription, in green are proteins involved in growth control, and in blue, chromatin/chromosome structure proteins. Some proteins, such as Ku70, have multiple functions. In general, no more than seven interactions are required to link any one core protein with another, while the degrees of separation between proteins outside of the core are much greater.

From protein interactions to functional assignments

While protein-interaction studies are useful for describing protein networks, the main goal of most interaction studies today is to learn about individual proteins: their potential partners, functions and interactive complexes. By identifying known partners of an unknown protein, a putative functional category for the unknown protein can often be assigned. High-throughput screens can have an advantage in this sense over individual studies both in the number of novel proteins that can be assigned potential functions and in the accuracy of the assignments. If protein X (uncharacterized) is found to interact with protein Y and protein Z, and both Y and Z are components of the RNA-processing machinery, then it is quite likely that protein X is also involved in RNA processing, perhaps as part of a complex with Y and Z. Based on their protein-interaction map of yeast, Schwikowski *et al.* compiled a list of about 370 proteins of unknown function that interact with at least one protein of known function¹⁰. They found that about

10% of these 370 had at least two protein partners of the same function. Such uncharacterized proteins are likely to have a role similar or related to that of their known interactors. Although identifying just one or two interactors for an uncharacterized protein might be insufficient to predict its function reliably, it does suggest an activity that can be tested experimentally. However, it is important to keep in mind the false-positive and false-negative interactions in these studies, one of the main limitations of comprehensive protein-interaction mapping^{14,15}.

Two-hybrid screens and their role in protein-interaction studies

Owing to the limited number and variety of large-scale systematic studies, there is at the present time no good estimate of how many protein interactions there are in a cell. Ideally, an accurate estimate of the total number of protein interactions would be based on experimentally determined interactions between proteins that are known to be present at the same time and in the same general location (i.e. nucleus or cytoplasm). Most

of the data so far generated from large protein-interaction studies have come from two-hybrid screening, a technique known to generate false positives and requiring experimental confirmation of any potential hit^{14,15}. Additionally, two-hybrid screening might miss interactors because of requirements for the potential interacting protein to be stably expressed as a fusion protein in the nucleus or owing to requirements for posttranslational modifications. Until the accuracy of these studies has been evaluated on a larger scale, these issues must be considered in any prediction of the number of protein interactions.

How many protein interactions are there in a cell?

With the above caveats in mind, we analyzed one of the larger reported protein interaction studies⁹ to make a crude estimate of the number of protein interactions in a cell. In the two-hybrid screens carried out by Uetz *et al.*⁹, reproducible positives were found for approximately half of the proteins screened against the entire set of yeast ORFs, with an average of 3.2 non-redundant interactions each. While some of these might indeed be false positives, approximately 70% of the known interactors were proteins of the same functional category (listed in Fig. 2), suggesting that at least ~70% of the interactions are plausible. Since half of the screened proteins ('baits') were unable to interact with the array of yeast ORFs in these screens ('preys'), we estimate that potentially half of the 'preys' could also be unable to participate in a two-hybrid interaction. Considering a potential false-negative rate of 50%, and a potential false-positive rate of 10–30%, the average number of interactions would be around 4.5 to 5.8. A similar two-hybrid false-negative rate of 45% was estimated by Walhout *et al.* in their *C. elegans* studies⁵. An average of ~5 interactions per protein is similar to a previous estimate of six interactions per protein that was compiled from two-hybrid screens reported in the literature¹⁵. When we compared the array two-hybrid results for eight of the most studied proteins in the yeast literature with published interactions, only 10% of previously known interactions were detected. If these screens indeed detect only 10% of all

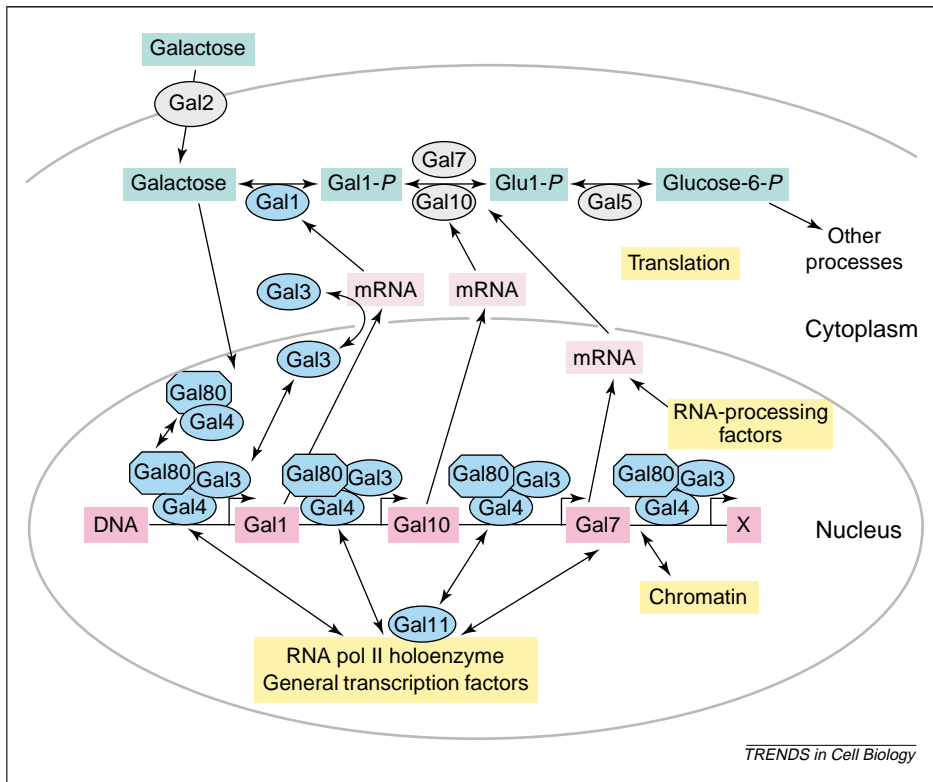


Fig. 4. Five proteins of the galactose pathway (represented in blue in Fig. 1) are diagrammed here in a functional context. The pathway shows the import of galactose through its eventual conversion to glucose-1-phosphate, proteins involved in regulating this process, and transcriptional regulation of proteins at the gene level. The individual Gal proteins interact with other proteins, as implicated in the protein-linkage maps, but also nucleic acids (pink), metabolites (green) and ions. Shown are a transporter (Gal2), metabolic enzymes (Gal1, Gal7, Gal10, Gal5), transcription factors (Gal4, Gal11) and regulatory proteins (Gal3, Gal80). In addition, hundreds of other proteins are required for processes such as transcription, RNA processing, RNA transport, translation and degradation.

interactions, this would give about 16 interactions per protein. However, the most studied proteins are also likely to have higher than average numbers of interactions, and 16 interactions per protein seems unreasonably high for the entire genome. Extrapolating to all yeast proteins and considering a potential false-positive rate of 50% and false-negative rate of 10–30%, this analysis would predict approximately 13 440 to 17 280 total interactions in yeast $[(6000 \times 4.5-5.8) / 2]$; note that the number is halved to rule out redundant interactions, i.e. A–B is the same interaction as B–A]. If we consider that (a) potentially 1000 yeast ORFs are questionable¹⁶ and (b) unknown ORFs tend to have only half as many interactions as characterized ones¹⁴ (the majority of baits in the Uetz *et al.* screens have been known proteins), the number would be reduced by (a) $1000 \times 5 / 2 = 2500$ and (b) $2000 \times 2.5 / 2 = 2500$, we arrive at a figure of roughly 8000 to 12 000 protein interactions.

Beyond interaction data: regulatory networks and cellular modeling
 Although comprehensive protein interaction maps of cells will eventually be generated, the intricate behavior of a cell is much more complex than can be displayed by a two-dimensional map such as that shown in Fig. 1. Transcriptional and translational regulation, posttranslational modifications, and spatial and temporal expression patterns must all be considered in cellular modeling scenarios. While current databases are sufficient for an investigator studying a particular protein or small network, they are not as useful in integrating complex information about cellular regulation, pathways, networks and cellular roles, and they lack coordination and the ability to exchange information between multiple data sources. Large studies that incorporate biochemical, physiological, morphological and temporal information will require the development of more integrative

databases as well as novel interfaces to display such information. Protein linkage maps that incorporate active linking (Myriad's ProNet[®]) or color-coding¹⁰ are a step in the direction of attempting to find a simple way to display a complex system to the user. However, even these maps are unable to provide information about the strength or kinetics of an interaction. Current protein interaction maps are also unable to differentiate between individual protein-interactions versus complexes. More complicated methods, such as described in Pirson *et al.*¹⁷ or Kohn *et al.*¹⁸, will be required to document regulatory pathways in greater detail, but they might eventually need to be automated to allow dynamic visualization and integration of biological information.

A study by Ideker *et al.*¹⁹ provides an example of the type of integrative study that will likely be more common in the post-genomic world. The authors used DNA microarrays, mass spectrometry and protein-interaction studies to analyze the galactose metabolism pathway in yeast (represented in Fig. 4). By studying RNA expression profiles in deletion mutants of yeast *GAL* genes involved in galactose metabolism, the authors identified genes that are transcriptionally affected by certain Gal proteins, revealing both known and unknown regulatory connections. They additionally used mass spectrometry to measure the levels of several hundred proteins, showing that, for many genes, there is a correlation between RNA and protein levels, but that for others there is surprisingly little correspondence. They integrated both previously published protein–protein interactions and metabolic pathways with these data to generated an integrated picture of this pathway.

Looking to the future

It is clear from the studies thus far that the complexity that emerges from large-scale protein interaction maps is daunting. We require new tools for visualizing complex information, better databases for exchanging this information, and new approaches to integrative science. The experience gained from studies of smaller genomes, such as that of *S. cerevisiae* (see also Box 1), will provide a basis for future exploration of higher eukaryotes.

Acknowledgements

We thank Benno Schwikowski for generously providing the map in Fig. 1. We also thank Stanley Fields and the reviewers for useful comments on the manuscript.

References

- Eisenberg, D. *et al.* (2000) Protein function in the post-genomic era. *Nature* 405, 823–826
- Pandey, A. and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* 405, 837–846
- Fields, S. and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340, 245–246
- Bartel, P. *et al.* (1996) A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat. Genet.* 12, 72–77
- Walhout, A. *et al.* (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 287, 116–122
- Flores, A. *et al.* (1999) A protein–protein interaction map of yeast RNA polymerase III. *Proc. Natl. Acad. Sci. U. S. A.* 96, 7815–7820
- Ito, T. *et al.* (2000) Toward a protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1143–1147
- Fromont-Racine, M. *et al.* (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* 17, 95–110
- Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
- Schwikowski, B. *et al.* (2000) A network of interacting proteins in yeast. *Nat. Biotechnol.* 18, 1257–1261
- Marcotte, E. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753

- Barker, N. *et al.* (2000) The yin–yang of TCF/beta-catenin signaling. *Adv. Cancer Res.* 77, 1–24
- Artavanis-Tsakonas, S. *et al.* (1999) Notch signaling: cell fate control and signal integration in development. *Science* 284, 770–776
- Uetz, P. and Hughes, R. (2000) Systematic and large-scale two-hybrid screens. *Curr. Opin. Microbiol.* 3, 303–308
- Walhout, A. *et al.* (2000) Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* 17, 88–94
- Kowalczyk, M. *et al.* (1999) Total number of coding open reading frames in the yeast genome. *Yeast* 15, 1031–1034
- Pirson, I. *et al.* (2000) The visual display of regulatory information and networks. *Trends Cell Biol.* 10, 404–408
- Kohn, K.W. (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair system. *Mol. Biol. Cell* 10, 2703–2734
- Ideker, T. *et al.* Systematic perturbation and global analysis of galactose utilization in yeast. *Science* (in press)

Chandra L. Tucker*

Peter Uetz

Depts of Genetics and Medicine,
University of Washington Box 357360,
Seattle, WA 98195, USA.

*e-mails: ctucker@u.washington.edu

uetz@u.washington.edu

Joseph F. Gera

Depts of Genetics and Medicine and
Howard Hughes Medical Institute,
University of Washington Box 357360,
Seattle, WA 98195, USA.

e-mail: gera@u.washington.edu

Box 1. Note added in proof:

Helicobacter pylori protein–protein interaction map

Rain *et al.* recently screened 261 proteins of the human gastric pathogen *Helicobacter pylori* by using the two-hybrid system and found more than 1200 interactions^a. The interactions reported involve ~740 out of 1590 open reading frames – that is, 46% of the whole genome. Only about 2% of these interactions have been found among homologous proteins in *Escherichia coli*, showing how little we know even about the ‘best-studied’ organisms. Although they do not show a graphic interaction map, the authors provide software on their website (www.hybrigenics.com) to visualize subsets of interactions (free for academic users after registration). This technology is applicable to higher eukaryotes for which highly complex random-primed cDNA libraries can be screened for interacting domains.

Reference

- ^a Rain, J-C. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature* 409, 211–215

A guide to the new sections of *Trends in Cell Biology*

Research Update

features succinct mini-reviews that offer expert and balanced interpretation of important recent advances. Updates on new techniques and news from conferences are also provided in this section.

News & Comment

contains the Journal Club – short commentaries highlighting the most important recent papers in cell biology written by a diverse panel of active scientists – and also In Brief news, offering a glimpse of interesting developments in policy, funding, legal and ethical issues and the like, of relevance for cell biologists. Letters about articles in the journal or any issue of general interest to cell biologists are welcomed for publication in this section.

Opinion

provides a forum for personal viewpoints on research topics, including discussion of controversial or developing themes, and models or hypotheses. Articles for *Trends in Cell Biology* are generally commissioned, but suggestions for Opinion articles in particular are encouraged (a short synopsis with key references should be sent).

Review

articles form the core of the journal, providing authoritative, timely and objective coverage of important developments in cell-biological research. A peer-review process is an important part of the evolution of these articles. (Research News and Opinion articles are also peer-reviewed.)

Forum

contains lighter reading such as book and software reviews, and Pioneer, Career Perspective and Profile articles.