been detected using only the DNA-binding data. Another advantage of the combined approach is that it can also predict directionality of the edges in the network; that is, it can be inferred whether the genes in a module are upregulated or downregulated by examining their expression correlations. An important benefit of having a complete genetic network of an organism is its potential to provide clues on a gene's role in, for example, signal transduction pathways and thereby identify its interaction partners.

It is accepted that genes in the same network module generally have similar cellular functions. This has also been observed among network modules generated by GRAM. Notably, the authors found that in most cases in which a gene module is regulated by more than one transcription factor, previous evidence could always be found suggesting potential physical or functional interactions between these transcription factors. All these observations prove that the regulatory networks produced by GRAM are biologically relevant and promise to serve as a blueprint to direct future experiments.

Like microarrays in the late 1990s, it is almost certain that the new ChIP-chip technology will quickly catch on with researchers worldwide, and before long, hundreds of genome-wide DNA-binding data sets will be available. Powerful and sophisticated computer algorithms, such as GRAM, will be needed to analyze these data.

Finally, many other research avenues can be pursued. For example, these tools can be applied to determine the degree of conservation of modular network structures or regulatory interactions among closely related species, such as *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. This type of comparative analysis can potentially shed light on the evolution of regulatory networks. Also, the current knowledge on genetic networks does not paint a truly dynamic picture of the processes taking place inside a cell. Existing technologies and algorithms, such as GRAM, are the first steps toward the development of tools capable of capturing the dynamics of genetic regulatory networks.

1. Bar-Joseph, Z. *et al. Nat. Biotechnol.* 21, 1337–1342 (2003).
2. Chu, S. *et al. Science* 282, 699–705 (1998).
3. Ren, B. *et al. Science* 290, 2306–2309 (2000).
4. Iyer, V.R. *et al. Nature* 409, 533–538 (2001).
5. Horak, C.E. *et al. Genes Dev.* 16, 3017–3033 (2002).
6. Lee, T.I. *et al. Science* 298, 799–804 (2002).
7. Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. & Gerstein, M. *J. Mol. Biol.* 314, 1053–1066 (2001).
8. Ihmels, J. *et al. Nat. Genet.* 31, 370–377 (2002).
9. Pilpel, Y., Sudarsanam, P. & Church, G.M. *Nat. Genet.* 29, 153–159 (2001).
10. Yu, H., Luscombe, N.M., Qian, J. & Gerstein, M. *Trends Genet.* 19, 422–427 (2003).

Playing tag with the yeast proteome

Brenda J Andrews, Gary D Bader & Charles Boone

Two tagged proteome studies offer the most intimate and detailed view into the inner works of yeast cells to date.

Proteomics—the study of the complement of expressed cellular proteins (or proteome)—has catapulted to the forefront of biological research. This advance is due to the development of enabling technologies for producing large-scale data sets of protein activities and to the increasing number of annotated genome sequences that can serve as prerequisite proteome 'blueprints'. Pioneering methods for analysis of the proteome have been developed in yeast and have relied on the systematic cloning of open reading frames (ORFs) for subsequent expression or generation of genomic sets of strains expressing tagged proteins suitable for a variety of array-based manipulations. In two recent *Nature* papers, the Weissman and O'Shea groups[1,2] report two notable additions to the arsenal of tools available for the comprehensive analysis of gene and protein function in yeast. The authors describe two collections of yeast strains in which each ORF is fused with affinity or fluorescence tags, thereby providing the most comprehensive and sensitive view yet of the expressed proteome and its subcellular location in a eukaryotic cell.

In the past few years, myriad genetic and biochemical methods have been used to query genomic sets of proteins for biochemical activity and protein-protein interactions. Notable landmarks on the road to the functional description of the yeast proteome include large-scale two-hybrid screens, immunoprecipitation–mass spectrometric analysis of protein complexes and the generation of tagged sets of proteins for production of functional protein chips (reviewed in ref. 3). The generation of protein complex interaction maps and functional surveys of proteins for DNA binding and other activities are providing a rich, but relatively static, view of the yeast 'interactome'. A more complete 'cell biological' view of the proteome will emerge from integration of proteomics information with functional genomics data derived from transcriptional profiling and gene disruption projects, as well as a picture of the subcellular distribution of proteins and their relative abundance.

In a tour-de-force of strain construction, Ghaemmaghami *et al.*[1] used a PCR-based homologous recombination strategy to insert a tandem affinity purification (TAP) tag at the C termini of all predicted yeast ORFs. They reasoned that an explanation of the biological properties of the proteome would require not only a description of macromolecular complexes and their subcellular location, but also an experimental description of the expressed proteome and a reasonable measure of the absolute levels of proteins in the cell. Two features of the strain collection allow both a survey of expressed proteins in a particular physiological circumstance and a measure of their cellular abundance. First, the tagged proteins are expressed from their native promoters in their endogenous chromosomal location and should be responsive to normal regulatory circuitry. Second, each ORF is marked with a common tag allowing measurement of the absolute abundance of each protein using quantitative western-blot analyses (see http://yeastgfp.ucsf.edu/). A sensible set of test cases suggests that the regulation and activity of most yeast proteins is unperturbed by the C-terminal tag, which bodes well for the utility of the strain set in future genetic and cell biological studies and is good news for the many other projects that have used convenient tags to study gene and protein function.

The authors were able to successfully TAP-tag 6,109 of the 6,243 predicted ORFs and observed a protein product for 4,251 or 70% of the tagged proteome in log-phase yeast cells grown in optimal laboratory conditions[1]. A

Brenda J. Andrews is at the Department of Medical Genetics & Microbiology, and Charles Boone is at the Department of Medical Genetics & Microbiology and the Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5S 1A8, Canada. Gary D. Bader is at the Computational Biology Center, Memorial Sloan-Kettering Cancer Center, Box 460, 1275 York Ave., New York, New York 10021, USA.
e-mail: charlie.boone@utoronto.ca; brenda.andrews@utoronto.ca

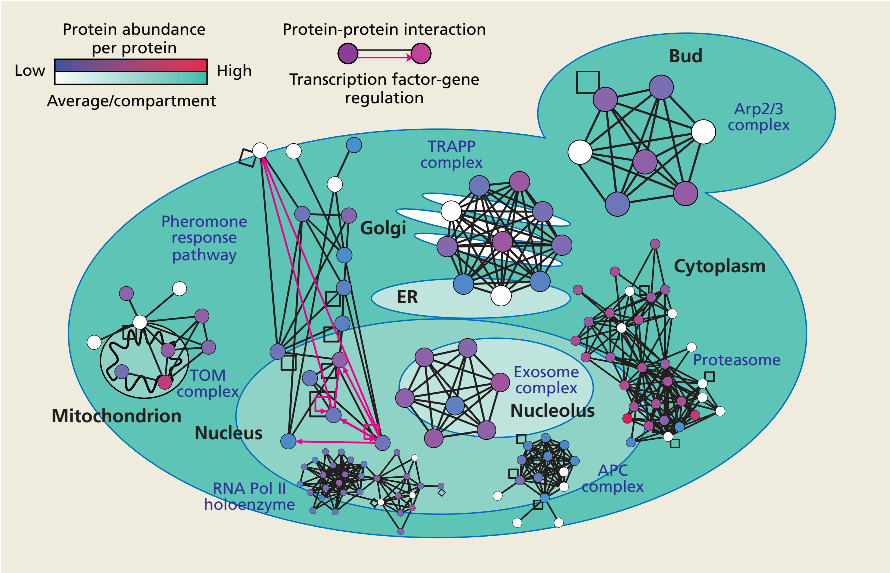

Figure 1 Integrated view of localization and protein abundance data[1,2] with protein-protein and gene regulation interactions in the context of selected complexes, pathways and the cell. Cellular compartments are colored by average protein abundance, with light colors representing compartments with low protein abundance (and dark colors those with high abundance). Black lines represent protein-protein interactions and red arrows point from transcription factors to the genes they regulate. Proteins are represented by circles colored by protein abundance continuously from blue to red indicating low to high abundance. White proteins have no abundance information. The TRAPP complex localizes to the 'endoplasmic reticulum (ER) to Golgi'. The transport outer membrane (TOM) complex localizes to the mitochondrion. The Arp2/3 complex localizes to cortical actin patches and the cytoplasm. The proteasome, anaphase promoting complex (APC) and RNA polymerase II (Pol II) complexes mostly contain subunits that localize to the nucleus, but some subunits localize to the cytoplasm (depicted as spanning both compartments). The pheromone-response mitogen-activated protein kinase cascade spans the cell from the surface to the nucleus. It contains a cell-surface receptor and other cortical components, cytoplasmic signaling molecules and nuclear transcription factor effectors, which control the expression of genes encoding components of the pathway as part of a regulatory circuit (red arrows).

subset of proteins were only seen with a second GFP-tagged strain set (see below), and the combined data show that ~80% of the proteome is expressed in happily growing yeast cells. This experiment considerably augments efforts to view the proteome using mass spectrometry and two-dimensional gel electrophoresis[4] and provides the most comprehensive and sensitive view, so far, of the expressed proteome in a eukaryotic cell. The analysis also confirms that the range of protein expression in the cell is massive, from 50 to well over 1,000,000 molecules per cell, although an even more sensitive assay may find lower abundance proteins. The observed protein set can be used as an experimental validation of the existence of hypothetical genes and, in combination with comparative genomics information, provides a powerful means of correcting errors in gene annotation (see http://www.yeastgenome.org/chromosomeupdates/start_changes.shtml).

In a parallel proteomics project, Huh *et al.*[2] used an identical strategy to generate a green fluorescent protein (GFP)-tagged yeast strain collection that provides both a second powerful experimental resource for the yeast community and the first view of the native yeast proteome in living cells. The first proteome-scale analysis of protein localization involved a description of the cellular location of almost half of yeast proteins using plasmid-based overexpression of epitope-tagged proteins and genome-wide transposon mutagenesis for high-throughput immunolocalization of tagged gene products[5]. This study affirmed the correlation between protein function and subcellular environment and highlighted the importance of generating a high-resolution and comprehensive view of protein localization.

Huh *et al.* analyzed 6,029 strains with GFP-tagged ORFs and found that three-quarters of the proteome and over two-thirds of the previously unlocalized proteins had a detectable GFP fluorescence signal in log-phase cells. In a first pass, the GFP patterns were classified as having one or more of 12 rather broad subcellular localization patterns, such as cell periphery, nucleus, mitochondrion and cytoskeleton. A second round of colocalization experiments, using monomeric red fluorescent protein fusions to reference proteins of known localization, distinguished another 11 localization categories, including the nucleolus and spindle pole body.

Several criteria suggest that this impressive two-stage binning of GFP patterns produced a high-quality view of the GFP-proteome that approximates the real situation in the cell. First, the results agree substantially with Saccharomyces Genome Database (http://www.yeastgenome.org/) annotations for the localization of ~2,500 yeast proteins and with the previous large-scale study that examined the proteome using immunofluorescence[5]. Second, over 90% of proteins identified were also found within the set detected by western-blot analysis of the TAP tag collection, indicating that GFP fluorescence can be used to detect a broad range of protein expression levels. And third, this comprehensive GFP data set encompasses a set of organellar proteomics projects that aim to identify subsets of proteins in various organelles. For example, 164 nucleolar proteins were identified, 82 of which overlapped with the 127 proteins catalogued in Saccharomyces Genome Database and 82 of which were newly defined. Because many of the characterized nucleolar proteins are involved in 'ribosomal RNA expression and processing' and 'ribosomal biogenesis', most of the newly localized proteins would be expected to participate in some aspect of these processes. Given the apparent high quality of the localization data, researchers interested in the function and regulation of the nucleolus ought to get busy. The authors note that proteins with crucial C-terminal targeting signals are often mislocalized in this study and new fusions will have to be constructed to get an accurate view of the subcellular location of this group of proteins. All GFP localization information has been admirably recorded in a public database (http://yeastgfp.ucsf.edu), making the data easily accessible for further analysis.

Integration of different functional genomics data sets enables biological hypotheses to be formulated with increasing levels of confidence. In an effort to leverage the biological information in their data set, Huh *et al.* analyzed their subcellular localization data in light of transcriptional coregulation information and combined physical and genetic interaction data sets. Both types of integrative analysis provided useful insight. In the first bioinformatics exercise, the authors investigated whether transcriptional coregulation is related to subcellular localization. To do this, they calculated the relative representation of proteins with a given localization for 33 general transcriptional modules, defined previously from an analysis of over 1,000 microarray data sets[6]. Notably, statistically significant enrichment was seen for 19 of the 22 most highly expressed modules, indicating that colocalization is highly correlated with transcriptional coexpression. Deeper analysis of the data can provide information on biological function that could not be gleaned from analysis of either data set alone.

In a second computational analysis, the authors examined the relationship between colocalization and physical or genetic interaction. The relative enrichment for colocalization was assessed for the combination of protein-protein or genetic interactions in the GRID database (http://biodata.mshri.on.ca/grid/), a collection of information derived from existing databases and large-scale data sets. As expected, because both genetic and physical interactions are indicative of a functional relationship, they were highly enriched between proteins that colocalize. An enrichment of interactions was also observed for protein pairs showing distinct localization categories, such as microtubule and spindle pole body. This illustrates the potential of this approach for identifying the network of functional relationships between subcellular localizations, a network that may reflect a dynamic interchange of proteins between compartments or genetic buffering of compartments.

With the unveiling of these two new tools, researchers are in the privileged position of having a comprehensive description of yeast that includes positive identification of nearly all of its genes, many proteins categorized by their interactions in complexes, and, of course, data on the abundance and location of most known proteins. The generation of a global *in vivo* view of the yeast proteome means that we can start to assemble diagrams of its cellular pathways and complexes with unprecedented detail. For example, using colocalization and abundance data together with existing interaction data, we can overlay the architecture of complexes and signaling pathways with specific cellular compartments and environments (**Fig. 1**).

Next, these strain sets can be used to move from a relatively static view of the proteome to an analysis of the dynamic abundance and localization of proteins during developmental programs or in response to environmental and genetic insults. For example, the GFP-tagged strains could be grown as high-density arrays in solid medium and assayed for colony fluorescence to monitor global changes in protein levels in response to drug treatments[7]. High-throughput strain construction methods[8] will also allow introduction of the entire GFP- or TAP-tagged proteome into any genetic background. In this way, the genetic requirements for protein localization and protein complex formation can be systematically assessed for pathways and proteins of interest. The combination of the GFP- and TAP-tagged strain set in a matrix format would create a doubly tagged set that should allow proteome-wide coimmunoprecipitation tests as a sensitive means for assessing protein-protein interactions globally and in defined genetic backgrounds. These tools open the door for numerous other focused and large-scale analyses of the yeast proteome.

1. Ghaemmaghami, S. *et al. Nature* **425**, 737–741 (2003).
2. Huh, W.-K. *et al. Nature* **425**, 686–691 (2003).
3. Phizicky, E.M., Bastiaens, P.I.H., Zhu, H., Snyder, M. & Fields, S. *Nature* **422**, 208–215 (2003).
4. Washburn, M.P., Wolters, D. & Yates, J.R *Nat. Biotechnol.* **19**, 242–247 (2001).
5. Kumar, A. *et al. Genes Dev.* **16**, 707–719 (2002).
6. Ihmels, J. *et al. Nat. Genet.* **31**, 370–377 (2002).
7. Dimster-Denk, D. *et al. J. Lipid Res.* **40**, 850–860 (1999).
8. Tong, A.H.Y. *et al. Science* **294**, 2364–2368 (2001).