From genomics to proteomics

Mike Tyers* & Matthias Mann†

**Samuel Lunenfeld Research Institute, Mount Sinai Hospital, and Department of Medical Genetics and Microbiology, University of Toronto, Toronto, Canada M5G 1X5 (e-mail: tyers@mshri.on.ca)*
†Center for Experimental BioInformatics, Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark (e-mail: mann@bmb.sdu.dk)

Proteomics is the study of the function of all expressed proteins. Tremendous progress has been made in the past few years in generating large-scale data sets for protein–protein interactions, organelle composition, protein activity patterns and protein profiles in cancer patients. But further technological improvements, organization of international proteomics projects and open access to results are needed for proteomics to fulfil its potential.

The term proteome was first coined to describe the set of proteins encoded by the genome[1]. The study of the proteome, called proteomics, now evokes not only all the proteins in any given cell, but also the set of all protein isoforms and modifications, the interactions between them, the structural description of proteins and their higher-order complexes, and for that matter almost everything 'post-genomic'. In this overview we will use proteomics in an overall sense to mean protein biochemistry on an unprecedented, high-throughput scale. The hope, now being realized, is that this high-throughput biochemistry will contribute at a direct level to a full description of cellular function.

Proteomics complements other functional genomics approaches, including microarray-based expression profiles[2], systematic phenotypic profiles at the cell and organism level[3,4], systematic genetics[5,6] and small-molecule-based arrays[7] (Fig. 1). Integration of these data sets through bioinformatics will yield a comprehensive database of gene function that will serve as a powerful reference of protein properties and functions, and a useful tool for the individual researcher to both build and test hypotheses. Moreover, large-scale data sets will be crucial for the emerging field of systems biology[8].

Challenges and approaches in proteomics

Proteomics would not be possible without the previous achievements of genomics, which provided the 'blueprint' of possible gene products that are the focal point of proteomics studies. Although almost trite, the tasks of proteomics can usefully be contrasted with the huge but straightforward challenges initially facing the genome projects. Unlike the scalable exercise of DNA sequencing, with its attendant enabling technologies such as the polymerase chain reaction and automated sequencing, proteomics must deal with unavoidable problems of limited and variable sample material, sample degradation, vast dynamic range (more than 10^6-fold for protein abundance alone), a plethora of post-translational modifications, almost boundless tissue, developmental and temporal specificity, and disease and drug perturbations. While proteomics is by definition expected to yield direct biological insights, all of these difficulties render any comprehensive proteomics project an inherently intimidating and often humbling exercise.

In this *Nature* Insight, five central pillars of proteomics research are discussed with an emphasis on technological developments and applications. These areas are mass spectrometry-based proteomics, proteome-wide biochemical assays, systematic structural biology and imaging techniques, proteome informatics, and clinical applications of proteomics. As is apparent from the reviews, the divisions between these areas are somewhat arbitrary, not least because technological breakthroughs often find immediate application on several fronts. More important, biologically useful insights into protein function often emerge from the combination of different proteomic approaches.

Mass spectrometry-based proteomics

The ability of mass spectrometry to identify ever smaller amounts of protein from increasingly complex mixtures is a primary driving force in proteomics, as described in the review on page 198 by Aebersold and Mann. Initial proteomics efforts relied on protein separation by two-dimensional gel electrophoresis, with subsequent mass spectrometric identification of protein spots. An inherent limitation of this approach is the depth of coverage, which is necessarily constrained to the most abundant proteins in the sample. The rapid developments in mass spectrometry have shifted the balance to direct mass spectrometric analysis, and further developments will increase sensitivity, robustness and data handling.

The past year has seen partial analysis of the yeast interactome, the malaria proteome, bacterial proteomes and various organellar proteomes (see review by Aebersold and Mann, page 198). These vast data sets represent but the tip of the iceberg for biological discovery and drug development. An enormous challenge resides in the obvious fact that the proteome is a dynamic, not a static, entity. Initial efforts to gauge proteome-wide regulatory events in single experiments have been directed at the yeast phosphoproteome[9] and the ubiquitin-mediated 'degradome' (S. P. Gygi, personal communication). Much higher throughput and sensitivity will be needed to enable true proteome dynamics and moment-by-moment snap shots of cellular responses. Nascent methods for gel-free analysis of complex mixtures hold great promise in this regard[10]. Further needs will include more complete sequence coverage of each individual protein, robust and varied methods for sample preparation, and sophisticated algorithms for automated protein identification and detection of post-translational modifications. The ambitious goals of systems biology, which aims to comprehensively model cellular behaviour at the whole-system level[8,11], will also require reliable quantitative methods.

Array-based proteomics

A number of established and emergent proteome-wide platforms complement mass spectrometric methods, as

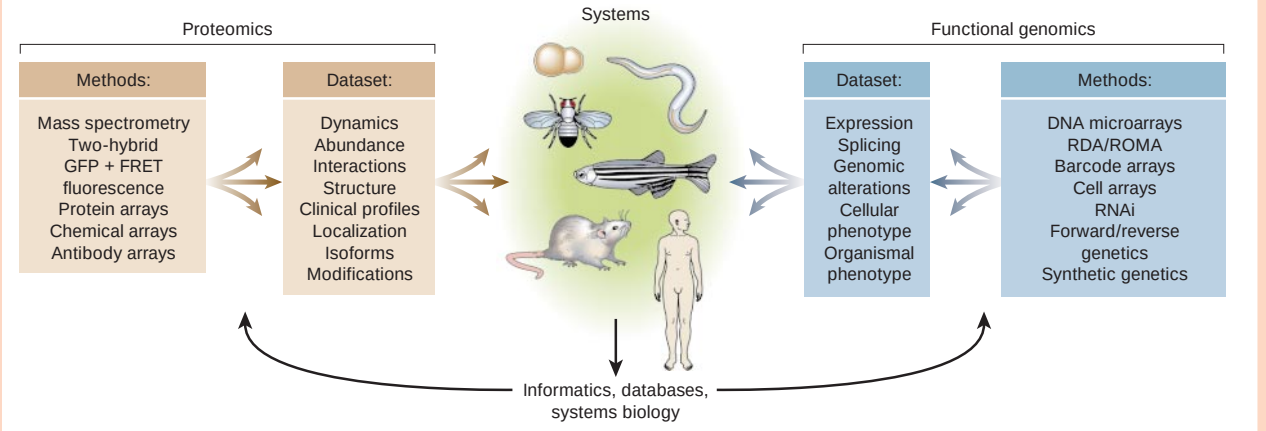

Figure 1 Platforms for proteomics and functional genomics. Methodology is shown in the outer columns, resultant data sets in the middle columns, and model systems in the centre.

reviewed on page 208 of this issue by Stan Fields and co-workers. The forerunner amongst these efforts is the systematic two-hybrid screen developed by Fields[12]. Unlike direct biochemical methods that are constrained by protein abundance, two-hybrid methods can often detect weak interactions between low-abundance proteins, albeit at the expense of false positives.

More recently, various protein-array formats promise to allow rapid interrogation of protein activity on a proteomic scale. These arrays may be based on either recombinant proteins or, conversely, reagents that interact specifically with proteins, including antibodies, peptides and small molecules[13]. Readouts for protein-based arrays can derive from protein interactions, protein modifications or enzymatic activities. A current challenge is to effectively couple high-end mass spectrometry to array formats. Array-based approaches can also use *in vivo* readouts, for example in the systematic analysis of protein localization in the cell through green fluorescent protein (GFP) signals or protein association through fluorescence resonance energy transfer (FRET) between protein fusions to different wavelength variants of GFP. Finally, cell- and tissue-based arrays enable yet another layer of functional interrogation.

One practical bottleneck to these approaches, and indeed to most systematic approaches, has been the limited availability of validated genome-wide complementary DNA for use in the capture of protein complexes with epitope tags. The FlexGene consortium between academic institutions and industry aims to develop complete cDNA collections in recombination-based cloning formats for the biomedical community (see http://www.hip.harvard.edu).

Structural proteomics

Beyond a description of protein primary structure, abundance and activities, the ambitious goal of systematically understanding the structural basis for protein interactions and function is reviewed by Baumeister *et al.* on page 216 of this issue. Through literary metaphor, the authors make a compelling argument that a full description of cell behaviour necessitates structural information at the level not only of all single proteins, but of all salient protein complexes and the organization of such complexes at a cellular scale. This all-encompassing structural endeavour spans several orders of magnitude in measurement scale and requires a battery of structural techniques, from X-ray crystallography and nuclear magnetic resonance (NMR) at the protein level, to electron microscopy of mega-complexes and electron tomography for high-resolution visualization of the entire cellular milieu. The recurrent proteomic theme of throughput and sensitivity runs through each of these structural methods, and Baumeister *et al.* suggest novel solutions, even including eliminating the crystals from crystallography! NMR and *in silico* docking will be necessary to build in dynamics of protein interactions, much of which may be controlled through largely unstructured regions[14].

Informatics

As with any data-rich enterprise, informatics issues loom large on several proteomics fronts. On page 233 of this issue, Boguski and McIntosh highlight the importance of sample documentation, the implementation of rigorous standards and proper annotation of gene function[15]. It is crucial that software development is linked at an early stage through agreed documentation, XML-based definitions and controlled vocabularies that allow different tools to exchange primary data sets. Considerable effort has already gone into interaction databases[16] and systems biology software infrastructure[17] that should be built upon by future proteomics initiatives. The development of statistically sound methods for assignment of protein identity from incomplete mass spectral data will be critical for automated deposition into databases, which is currently a painstaking manual and error-prone process. Lessons learned from analysis of DNA microarray data, including clustering, compendium and pattern-matching approaches, should be transportable to proteomic analysis[2], and it is encouraging that the European Bioinformatics Institute and the Human Proteome Organisation (HUPO) have together started an initiative on the exchange of protein–protein interaction and other proteomic data (see http://psidev.sourceforge.net/)

Clinical proteomics

Proteomics is set to have a profound impact on clinical diagnosis and drug discovery, as is fittingly reviewed by Sam Hanash on page 226, the inaugural president of HUPO. Because most drug targets are proteins, it is inescapable that proteomics will enable drug discovery, development and clinical practice. The form(s) in which proteomics will best fulfil this mandate is in a state of flux owing to a multitude of factors, not the least of which are the varied technological platforms in different stages of implementation.

The detection of protein profiles associated with disease states dates back to the very beginning of proteomics, when two-dimensional gel electrophoresis was first applied to clinical material. The advent of mass spectrometers now able to resolve many tens of thousands of protein and peptide species in body fluids is set to revolutionize protein-based diagnostics, as demonstrated in recent retrospective studies of cancer patients[18]. The robust and high-throughput nature of mass spectrometric instrumentation is imminently suited to clinical applications. Protein- and antibody-based arrays with validated diagnostic readouts may also become amenable to the clinical

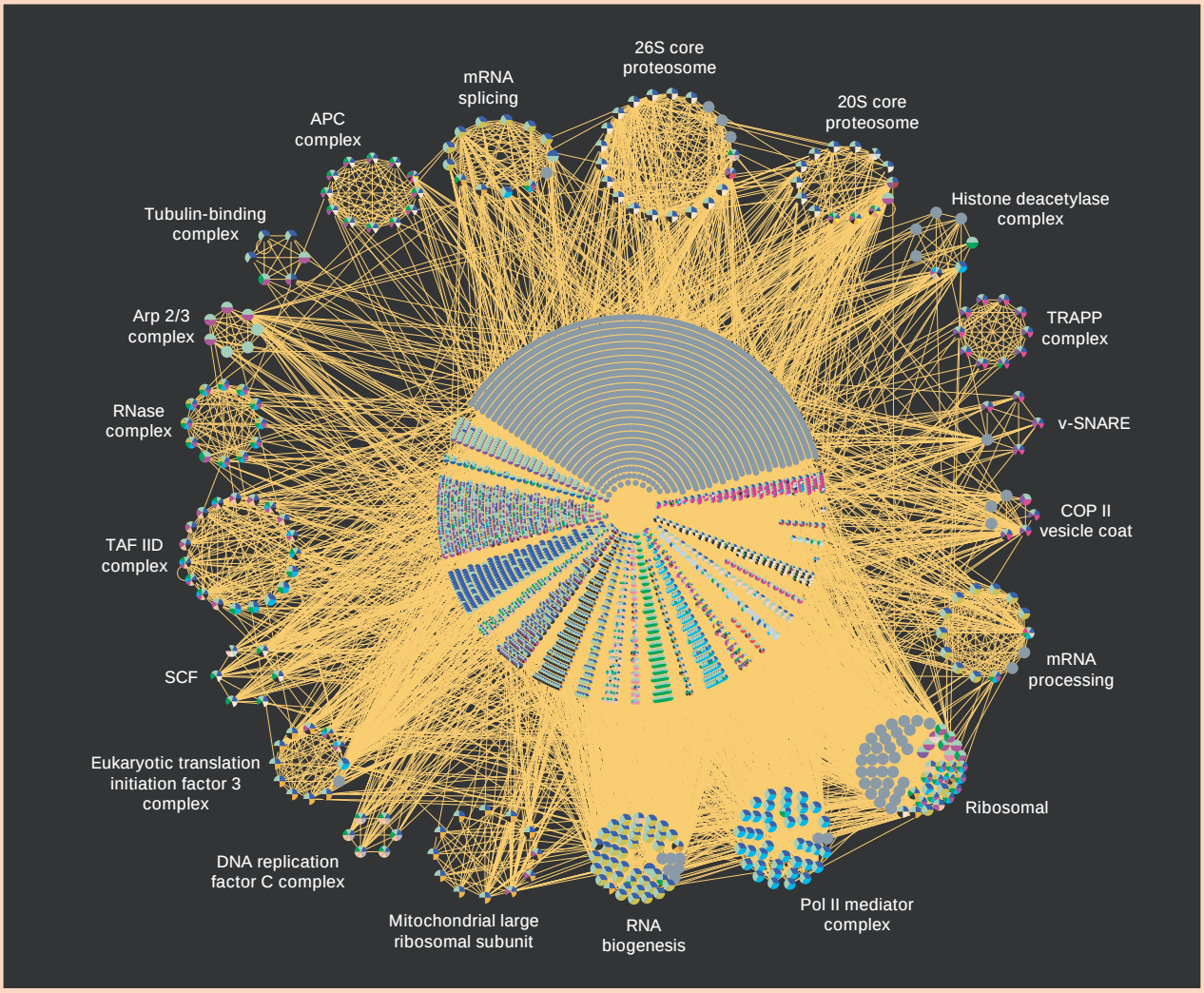

Figure 2 Visualization of combined, large-scale interaction data sets in yeast. A total of 14,000 physical interactions obtained from the GRID database were represented with the Osprey network visualization system (see http://biodata.mshri.on.ca/grid). Each edge in the graph represents an interaction between nodes, which are coloured according to Gene Ontology (GO) functional annotation. Highly connected complexes within the data set, shown at the perimeter of the central mass, are built from nodes that share at least three interactions within other complex members. The complete graph contains 4,543 nodes of ~6,000 proteins encoded by the yeast genome, 12,843 interactions and an average connectivity of 2.82 per node. The 20 highly connected complexes contain 340 genes, 1,835 connections and an average connectivity of 5.39.

setting. As with all clinical interfaces, issues of standardized sample preparation, storage and annotation must be addressed.

Proteomics will inevitably accelerate drug discovery, although the pace of progress in this area has been slower than was initially envisaged. Identification of new disease-specific targets, often those present on the cell surface, has been greatly enabled with current technology. An understanding of the biological networks that lie below the cell's exterior will provide a rational basis for preliminary decisions on target suitability.

Orthogonal omics

A caveat of all high-throughput approaches, including proteomics, is that the very scale of experimentation often precludes repetition and rigorous confirmation that is the essence of sound research. However, the intersection between proteomic data sets from different species or between proteomic and other genome-wide data sets often allows robust cross-validation (Fig. 1). This point is aptly illustrated by recent proteomic analysis of the yeast and human nucleolus, in which both directed and undirected efforts uncovered a vast network of protein interactions, many of which impinge on the conserved process of ribosome biogenesis[19]. Independent systematic analysis of yeast-cell size mutants (phenomics) and the gene set regulated by one of these size-control genes (transcriptomics) revealed an unanticipated regulatory relationship between ribosome biogenesis and commitment to cell division[20].

Similarly, the integration of interactome, phenome and transcriptome data sets has been used to deduce a new regulatory network in the nematode germline[21]. The combined use of physical, phenotypic and expression data sets can generate non-obvious hypotheses that would otherwise not arise from any individual approach. Even with limited data sets, educated guesses can made based on simple parameters. For example, an algorithm called ScanSite was used to identify tuberous sclerosis complex-1 as a physiologically relevant substrate of protein kinase B (PKB), based solely on the apparent mass by electrophoresis of the phosphorylated species and an abundance of PKB consensus site sequences[22]. Finally, new information can often be gained by re-investigating known complexes with new methods. For example, three new components of the heavily studied anaphase-promoting complex have recently been found by multidimensional mass spectrometry[23].

With the numerous initiatives to systematically correlate phenotype with loss of gene function in many model organisms including yeast, nematode, fruitfly, zebrafish, mouse and human, the insights gained from the combined use of large-scale cell biological, transcriptional and proteomic data sets should become synergistic as coverage increases. Most recently, the rapid acquisition of phenotypic data by RNA interference methods, with which it is now possible to systematically interrogate the human genome in tissue-culture cells[6], will greatly accelerate functional discovery when coupled to proteomic data sets.

Future developments and challenges

As the highly successful effort to sequence the human genome has illustrated, faster and cheaper is the inevitable mantra of any large-scale enterprise. This rhetoric applies doubly so to proteomics, although there is far more to proteomics than just throughput. In its absolute sense, the proteome will be as unreachable as the horizon; rather proteomics will coalesce with other technologies in as yet unimagined ways to converge on an accurate description of cellular properties.

By all criteria, current instrumentation is far from optimal, in part because manufacturers have not yet had the necessary lead time to build machines and associated hardware that are perfectly tailored to protein analysis. Mass spectrometry-based proteomics is nowhere near the physical limit of the few ions needed to register a peak and so a huge increase in performance can be expected in the coming years. As refinements are made in next-generation proteomic instruments, it will be possible to monitor many relevant post-translational modifications and protein interactions in ever more complex mixtures[24]. As one anticipated example of innovation, throughput and coverage could be greatly enabled by storing mass spectrometric signatures of every protein for real-time data-dependent analysis of highly complex mixtures.

At the level of the individual laboratory, there is undoubtedly a huge market for sensitive and affordable bench-top mass spectrometers for routine applications as analytical devices in all aspects of biological research. Developments in robotic sample preparation, alternative readouts for protein interactions, and microfluidics to minimize sample losses will all factor into achieving the goal of delivering high-powered proteomics to the masses. Equally important, availability of reasonably complete sets of expression and antibody reagents for all proteins would improve the speed and scope of both small- and large-scale proteomics.

With regard to the proteomes of even simple model organisms, all indications are that extant interaction maps are far from saturated. As the density of known interactions increases, testable hypotheses should emerge from the data set at an increasing rate, especially in combination with other genome-wide data sets, including predictions from structural data. Once sufficient dynamics data become available to build first-draft models of cellular behaviour, model refinement will require reiteration of proteomic analyses in numerous mutant and drug-treated conditions. If modelling of simple Boolean networks is a guide, the systems-level behaviour of bona fide protein interaction networks is sure to yield some surprises[25].

All this information must obviously be presented in a form that can be processed by the human user. To this end, a great deal more effort must be placed on development of visualization tools, including automated integration with other genome-wide data sets (Fig. 2). There is much room here for novel approaches, many of which are likely to come from other fields that are also suffering from information overload. Examples include sophisticated tools for clustering DNA microarray data and multivariate graphical representations that use coloured readouts to highlight overall trends[26], as well as the sophisticated, three dimensional interfaces used in modern computer games.

On the clinical front, comprehensive proteomic analysis of small amounts of diseased tissue will facilitate diagnosis and therapeutic monitoring, particularly as patterns of disease prediction are recognized empirically from large clinical data sets. Application of phosphoproteomic methods to clinical samples promises what may be the most informative and discriminating readout of cellular status, which can then be used to advantage in diagnosis, drug discovery and elucidation of mechanisms of drug action. The proteomics of host–pathogen interactions should also be an area rich in new drug targets. Regardless of the exact format, robust mass spectrometry and protein-array platforms must be moved into clinical medicine to replace the more expensive and less reliable biochemical assays that are the basis of traditional clinical chemistry. Finally, the nascent area of chemiproteomics will not only allow mechanism of action to be discovered for many drugs, but also has the potential to resurrect innumerable failed small molecules that have dire off-target effects of unknown basis. Relatively little investment in well characterized leads hidden in the archives of pharmaceutical companies may leverage huge therapeutic returns.

Open-access proteomics

An all too common refrain of proteomics has been the limited or non-existent access for the individual biomedical researcher. Although virtually all academic centres have a mass spectrometry facility of some sort, lost samples, failed identifications and inadequate throughput are commonplace. In part, these problems represent the teething stages of a complex technology; additional factors are unaffordable equipment costs and a dearth of highly trained personnel to oversee facilities. As a consequence, most breakthroughs and the generation of raw data in proteomics derive from the work of only a handful of technically inclined laboratories. The burden of improving this circumstance falls on instrument manufacturers, proteomics leaders, funding agencies, academic institutions and the individual user alike. National proteome centres have also been proposed as a way to ensure availability of both expertise and equipment[27].

The common effort to map and understand the proteome in its various guises can benefit from lessons learned by genome-sequencing consortia. First and foremost, public access to on-line raw data is essential if there is to be sense of participation across the biomedical research community. Agreements similar to the Bermuda guidelines issued at a critical juncture of the genome projects[28] that mandate public accessibility and non-patenting of basic proteomic data would facilitate research in both the academic and industrial sectors. Such data should include the primary structure, post-translational modification, localization and protein–protein interaction pattern of all proteins.

It is important that large-scale proteomics efforts are co-coordinated, both to avoid duplication and to provide strong rational for funding agencies. These bodies are in principle willing to support proteomics as a way to reap the rewards of the genome projects, but they will have to be presented with clear goals and rationales of how proteomics will build an infrastructure to advance biomedical science. HUPO is one body that is positioned to play an important coordinating role. HUPO has proclaimed five initial goals for world-wide proteomics research: definition of the plasma proteome, proposals for an in-depth proteomics assault on specific cell types, formation of a consortium to generate antibodies to all human proteins, development of new technologies and formation of an informatics infrastructure. To this list we would add cataloguing the primary structure of all proteins, mapping all organelles that can be purified, and generating protein interaction maps of model organisms, for both comparative proteomics and integration with on-going functional genomics projects.

To meet these laudable goals, it seems that a dedicated funding pool must be established for proteomics research, analogous to that created for the human and model-organism genome sequencing projects, or ongoing funding for these projects should be made available to proteomics. Given the cost of proteomic-scale projects, it

benefits academia and industry to collaborate as much as possible on method development, data acquisition and project coordination. Finally, a way must be established to integrate proteome-scale experiments with efforts of the many individual biology laboratories to develop and test biological models, the final key step in the discovery process that may always defy automation. Whatever the future holds, proteomics will yield great returns for all in what promises to be a knowledge watershed in biology and medicine. □

doi:10.1038/nature01510

1. Wilkins, M. R. *et al.* From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology* **14,** 61–65 (1996).
2. Shoemaker, D. D. & Linsley, P. S. Recent developments in DNA microarrays. *Curr. Opin. Microbiol.* **5,** 334–337 (2002).
3. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418,** 387–391 (2002).
4. Gerlai, R. Phenomics: fiction or the future? *Trends Neurosci.* **25,** 506–509 (2002).
5. Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294,** 2364–2368 (2001).
6. Hannon, G. J. RNA interference. *Nature* **418,** 244–251 (2002).
7. Kuruvilla, F. G., Shamji, A. F., Sternson, S. M., Hergenrother, P. J. & Schreiber, S. L. Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature* **416,** 653–657 (2002).
8. Csete, M. E. & Doyle, J. C. Reverse engineering of biological complexity. *Science* **295,** 1664–1669 (2002).
9. Ficarro, S. B. *et al.* Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nature Biotechnol.* **20,** 301–305 (2002).
10. Liu, H., Lin, D. & Yates, J. R. III Multidimensional separations for protein/peptide analysis in the post-genomic era. *Biotechniques* **32,** 898–911 (2002).
11. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934 (2001).
12. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340,** 245–246 (1989).
13. MacBeath, G. Protein microarrays and proteomics. *Nature Genet.* **32**(Suppl.), 526–532 (2002).
14. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293,** 321–331 (1999).
15. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25,** 25–29 (2000).
16. Bader, G. D. & Hogue, W. V. C. in *Genomics and Bioinformatics* (ed. Sensen, C. W.) 399–413 (Wiley-VCH, Weinheim, 2001).
17. Kitano, H. Systems biology: a brief overview. *Science* **295,** 1662–1664 (2002).
18. Petricoin, E. F., Zoon, K. C., Kohn, E. C., Barrett, J. C. & Liotta, L. A. Clinical proteomics: translating benchside promise into bedside reality. *Nature Rev. Drug Discov.* **1,** 683–695 (2002).
19. Andersen, J. S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12,** 1–11 (2002).
20. Jorgensen, P., Nishikawa, J. L., Breitkreutz, B. J. & Tyers, M. Systematic identification of pathways that couple cell growth and division in yeast. *Science* **297,** 395–400 (2002).
21. Walhout, A. J. *et al.* Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* **12,** 1952–1958 (2002).
22. Manning, B. D., Tee, A. R., Logsdon, M. N., Blenis, J. & Cantley, L. C. Identification of the tuberous sclerosis complex-2 tumor suppressor gene product tuberin as a target of the phosphoinositide 3-kinase/akt pathway. *Mol. Cell* **10,** 151–162 (2002).
23. Yoon, H. J. *et al.* Proteomics analysis identifies new components of the fission and budding yeast anaphase-promoting complexes. *Curr. Biol.* **12,** 2048–2054 (2002).
24. Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nature Biotechnol.* (in the press).
25. Huang, S. & Ingber, D. E. Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Exp. Cell Res.* **261,** 91–103 (2000).
26. Ball, P. Data visualization: picture this. *Nature* **418,** 11–13 (2002).
27. Aebersold, R. & Watts, J. D. The need for national centers for proteomics. *Nature Biotechnol.* **20,** 651 (2002).
28. Marshall, E. Bermuda rules: community spirit, with teeth. *Science* **291,** 1192 (2001).

Acknowledgements We thank B.-J. Breitkreutz for preparing Fig. 2, D. Figeys and members of the Center for Experimental BioInformatics (CEBI) for critical reading of the manuscript. CEBI is supported by a grant from the Danish Natural Research Foundation.