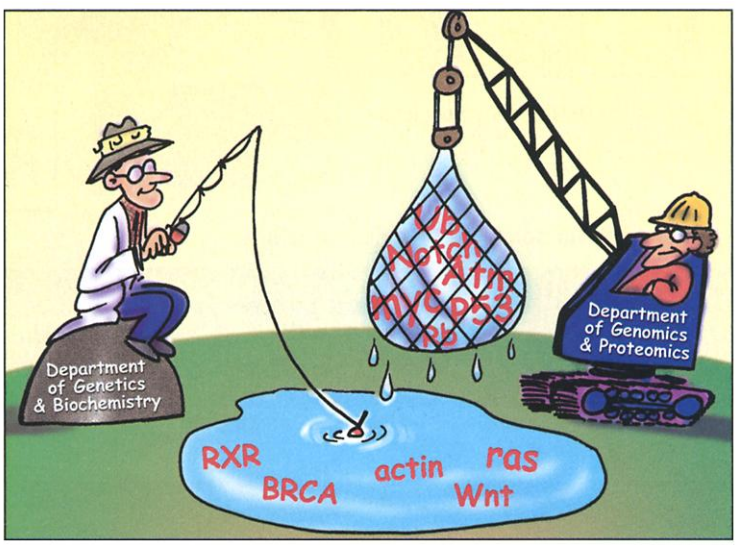
FUTURE DIRECTIONS: PROTEOMICS

Proteomics in Genomeland

Stanley Fields

If the architect you hired to design your home brought you a blueprint that solely consisted of a long list of parts that began "windowwabeborogovestaircasedoorjubjub...," you might start to wonder if and when you will see your new house. Some people have similar reservations about the recently "completed" human genome sequence, heralded as the "genetic blueprint" that will revolutionize biology and medicine. Deciphering how a mere 10^7 nucleotides result in a yeast cell—let alone how 3×10^9 nucleotides result in Tiger Woods or Britney Spears—cannot begin until the genes have been annotated. This step includes figuring out the proteins that these genes encode and what they do for a living. But understanding how all of these proteins collaborate to carry out cellular processes is the real enterprise at hand.



In the wonderland of complete sequences, there is much that genomics cannot do, and so the future belongs to proteomics: the analysis of complete complements of proteins. Proteomics includes not only the identification and quantification of proteins, but also the determination of their localization, modifications, interactions, activities, and, ultimately, their function. Initially encompassing just two-dimensional (2D) gel electrophoresis for protein separation and identification, proteomics now refers to any procedure that characterizes large sets of proteins. The explosive growth of this field is driven by multiple forces—genomics and its revelation of more and more new proteins; powerful protein technologies, such as newly developed mass spectrometry approaches, global two-hybrid techniques, and spin-offs from DNA arrays; and innovative computational tools and methods to process, analyze, and interpret prodigious amounts of data.

The shift in thinking from genomics to proteomics comes with an appreciation of the difficulty of the task: Proteins are much more complicated than nucleic acids. Unlike the decoratively challenged DNA, proteins get phosphorylated, glycosylated, acetylated, ubiquitinated, farnesylated, sulphated, linked to glycophosphatidylinositol anchors, and embellished in numerous other ways. A single gene can encode multiple different proteins—these can be produced by alternative splicing of the mRNA transcript, by varying translation start or stop sites, or by frameshifting during which a different set of triplet codons in the mRNA is translated. All of these possibilities result in a proteome estimated to be an order of magnitude more complex than the genome. (So it may be fortunate for proteomicists that humans might have as few as six times the number of genes that yeast have!) What is more, proteins respond to altered conditions by changing their location within the cell, getting cleaved into pieces, and adjusting their stability as well as changing what they bind to (other proteins, nucleic acids, lipids, small molecules, or other ligands). Protein levels often do not reflect mRNA levels (*1*), and even the presence of an open reading frame does not guarantee the existence of a protein. Lastly, a single protein may be involved in more than one process, and conversely, similar functions may be carried out by different proteins.

Where We Are

It's worth noting that in the pre-proteomic era, thousands of proteins were exquisitely characterized—those in metabolic and signaling pathways; in the replication, transcription, and translation machinery; in secretory and cytoskeletal networks; and in a host of other cellular complexes. These functional assignments came from efforts to understand specific cellular processes, with three major factors fueling progress in the last two decades. First, a confluence of geneticists, cell biologists, biochemists, and structural biologists arrived to tackle the same problems from different directions. Second, the extraordinary conservation of fundamental cellular mechanisms enabled insights garnered from the study of one organism to be immediately applied to all others—what's true for the walrus is true for the carpenter. Third, developments in technology—the now-standard tools of molecular biology such as DNA sequencing, recombinant DNA, and the polymerase chain reaction—sparked new experimental strategies.

With the emergence of proteomics, additional proteins are now pouring in to join those already implicated in some process or other. This extra information, however, derives not from continued small-scale analysis of a biological activity, but from much larger and more systematic studies. Proteomics, like its precursor genomics, thus represents the emergence of a new way of doing research that is not dependent on the testing of specific models of cellular behavior. This style of science obviously does not replace, but rather will increasingly operate in tandem with, traditional biological research methods.

One general principle is that proteins prefer to hang out in the cell with others that they work with; thus, the identity of new proteins in the complexes left intact after cell lysis often provides clues to function. A big boost has come from recent advances in mass spectrometry that allow the rapid identification of proteins separated in a 2D gel or by chromatography. The mass spectrometer measures the masses of peptides (typically derived from a trypsin digestion), which are then compared to the predicted masses of peptides from in silico digestions of sequences in genomic databases (*2*). Although unambiguous identification of a protein cannot always be derived from the masses of a few of its peptides, in the tandem mass spectrometer, peptide ions from the first mass spectrometer run are fragmented and identified in a second run to yield the more valuable commodity of a peptide sequence. A single peptide sequence usually identifies a protein. Advances in automation, increased sensitivity, and higher throughput, combined with improved biochemical fractionations

The author is at the Howard Hughes Medical Institute, Departments of Genetics and Medicine, University of Washington, Seattle, WA 98195, USA. E-mail: fields@u.washington.edu

CREDIT: JOE SUTUFF

and the availability of vastly expanded databases, have extended the application of mass spectrometry to ever bigger jobs. For example, megadalton protein complexes can be purified, often with a single tagged component, and their constituents can be identified after gel electrophoresis. Such analyses have been performed on, among other complexes, the human spliceosome (*3*), the yeast nuclear pore complex (*4*), and the pea chloroplast (*5*). Bypassing even gel separation, the direct analysis of protein complexes identifies components of heterogeneous protein mixtures, often using 1D or 2D chromatography for fractionation before analysis by mass spectrometry. Application of this procedure to a whole-cell yeast lysate identified 189 proteins (*6*) and more recently 1484 proteins (*7*), including integral membrane proteins and those of low-abundance in the cell.

The complement to mass spectrometry, the yeast two-hybrid system, has been increasingly "genomicized." From its initial application to finding protein partners that interact with just one protein, the assay has been scaled up to handle, for example—15 proteins implicated in yeast mRNA splicing (*8*), 29 proteins involved in *Caenorhabditis elegans* development (*9*), the ~55 proteins of bacteriophage T7 (*10*), 266 proteins of vaccinia virus (*11*), and even 5345 proteins of *Saccharomyces cerevisiae* (*12*). For yeast, more than 2700 putative interactions involving at least 2000 different proteins have been identified, mostly through two-hybrid experiments. This set of interaction data can be visualized as protein networks, with one analysis yielding a network that encompasses over 2300 links (*13*). The validity of many of the links in this network is supported by database annotations. More than 70% of characterized proteins with partners that have also been described could be assigned a correct functional category according to the properties of these partners (compared with only 12% if the proteins in the network are kept constant and the links are scrambled). Thus, a protein of unknown function that binds to one of known function can be tentatively assigned to the same cellular category as its partner.

Protein localization within the cell can now be addressed at a genomic level. In a tour de force of transposon tagging and analysis (*14*), over 11,000 yeast strains were generated with more than 2000 *S. cerevisiae* genes affected; indirect immunofluorescence was then used to determine subcellular localizations for over 1300 of the tagged proteins. Biochemistry, too, is feeling the impact of complete sequence information. The entire set of predicted yeast proteins has been fused to the "purification hook" of glutathione S-transferase (*15*). This set enables a biochemical genomics strategy in which the fusions are purified as 64 pools of 96 proteins each. The pools can be assayed for any biochemical activity, and the protein responsible for the activity in a pool can be quickly identified. Because the pools are derived from an array of yeast strains harboring a single gene, the gene encoding the activity is immediately known.

For the expanding number of genome sequences available, clever algorithms have been developed that assign functions to previously unknown proteins that do not rely on amino acid similarity. One approach scores the presence or absence of a given protein in all sequenced genomes, revealing sets of proteins that have co-evolved (that is, all members of a set are either present or absent in an organism), and are therefore likely to act in the same cellular process (*16*). A second approach is based on the observation that many proteins consist of two domains in one organism, whereas the domains are two separate proteins in another organism (*17*, *18*). The existence of the fusion, in which the two domains clearly interact, suggests that in the second organism the two separate proteins also interact. A third approach identifies cases in which multiple genomes harbor the same set of neighboring genes (*19*, *20*), a situation implying that each set encodes proteins of related function. Such operons in prokaryotes typically specify functionally linked proteins, but some examples are also found in eukaryotes.

Although strictly speaking not a proteomics technique, DNA arrays often provide insight into the functions of sizable collections of proteins. Genes that are transcriptionally co-regulated generally code for proteins that act in the same process, as demonstrated by yeast genes that operate in the cell division cycle (*21*, *22*), sporulation (*23*), and the diauxic shift (*24*). Expression profiles reveal up- or down-regulated mRNAs (and thus, presumptively, their protein products) in disease processes such as cancer, and consequently can be used to classify tumors (*25*). Microarray technology can identify classes of proteins—for example, membrane-bound and secreted proteins have been identified through the localization of their mRNAs (*26*), and proteins that bind to a DNA sequence have been identified by their interaction with a double-stranded DNA array (*27*). Microarray-based assays can also be used to detect polymorphisms (variations in the DNA), thereby associating protein variants with a disease state. An early application of this approach correctly identified 14 of 15 patients carrying known mutations in the hereditary breast and ovarian cancer gene *BRCA1* (*28*).

Given the current genomic and proteomic commotion, we should keep in mind that a protein found to be "in the spliceosome complex," "interacting with actin," "co-evolving with a prion protein," or "up-regulated in leukemia" has not been functionally characterized in the traditional sense to which biologists are accustomed. Instead, these types of results often serve only to place a protein in the appropriate bailiwick for follow-up analysis.

Where We're Heading

So far, most proteomic measurements have been performed in a cataloging mode, but the future will see more studies that address the dynamics of cellular processes. The protein composition of a cell is not static, therefore, it is crucial to obtain quantitative comparisons after a cell's environment changes. Proteomic strategies increasingly allow such quantitative analyses to be carried out. For example, stable isotopes enable two protein populations to be labeled with either a heavy or a light affinity tag, then mixed, trypsinized, and fractionated to enrich for subsets of proteins (*29*). Because the peptides in the two populations are identical except for the defined mass difference of the two tags, quantitation by mass spectrometry is possible. These studies are in their early stages and their potential is tremendous. Increasingly, proteins will undergo wholesale analyses to probe for their various modifications. Affinity purification approaches using specific antibodies, metals, lectins, or other reagents allow enrichment for modified proteins, which then can be detected by mass spectrometry (*30*). These types of strategies should make it feasible to follow, at the level of the proteome, a series of complicated cellular events such as those that ensue after a T cell encounters an antigen. Advances in direct analysis by mass spectrometry of peptide mixtures generated by the digestion of complex protein samples will lead to an escalating number of protein identifications in one experiment. This procedure may allow human tissues to be used as the protein source and renders feasible the discovery of early disease markers (through the comparison of the protein content of pathogenic cells with that of their normal counterparts).

Protein expression and purification technologies will continue to improve. The biochemical genomics strategy of purifying pools of tagged proteins will be particularly suitable for the many bacteria that have had their genomes sequenced, but it can be applied to multicellular organisms as well. These and other procedures that make use of protein arrays will become commonplace. The arrays may be generated by in vivo expression of tagged proteins, in vitro translation, peptide synthesis, or protein capture by

antibodies or oligonucleotide aptamers. Their potential applications include: revealing interactions among proteins and between proteins and small molecules (drugs) or other ligands, identifying substrates for a modifying enzyme such as a protein kinase, and searching for enzymatic activities. A harbinger of the promise of this approach is the recent demonstration of proteins in nanoliter droplets immobilized by covalent attachment to glass slides; more than 10,000 samples could be spotted onto each slide with this technique (*31*). The few test proteins in this array format were assayed for interactions with another protein or a small molecule, and for their phosphorylation by a protein kinase. Targeted arrays will allow the identification of all of the enzymes in an organism that are able to carry out a specific modification of a substrate; for example, protein arrays have tested nearly the entire set of the predicted protein kinases in yeast for their activity on 17 substrates (*32*).

Protein databases will need to become much more sophisticated if they are to help scientists make sense of the staggering number of experimental measurements that will soon emerge. Demands range from tracking all of the ligands for each analyzed member of a protein family (such as the SH3 domain) to cataloging all of the known substrates of each protein kinase, protein phosphatase, or other modifying activity. In addition, protein data will need to be integrated with results from expression profiling, genome-wide mutation or antisense analyses, and polymorphism detection. As proteomic data accumulate, we will become better at triangulating from multiple disparate bits of information to gain a bearing on what a protein does in the cell. Proteomics will come of age when its revelations about formerly uncharacterized proteins directly drive imaginative hypotheses about their functions.

What We Need

For a field so laden with razzmatazz methods, it is striking that the number one need in proteomics may be new technology. There are simply not enough assays that are sufficiently streamlined to allow the automation necessary to perform them on a genome's worth of proteins. Those currently available barely scratch the surface of the thousands of specialized analyses biologists use every day on their favorite proteins. What we need are experimental strategies that could be termed cell biological genomics, biophysical genomics, physiological genomics, and so on, to provide clues to function. In addition, a protein contains so many types of information that each of its properties needs to be assayed on a proteome-wide scale, ideally in a quantitative manner.

As we have argued (*33*), existing technology—and more importantly, the reagents (sets of genes, plasmids, strains, proteins, and the like) and equipment to handle these reagents—must rapidly spread from the specialized genomic and proteomic centers to the rest of the community. Only when every laboratory is comfortable doing proteomics will its power be exploited fully. Moreover, the likelihood of new approaches increases in proportion to the number of investigators participating in the field.

An interdisciplinary spirit will come to guide those excited by the global analysis of protein function. Geneticists need to talk to chemists, physiologists to physicists, cell biologists to computer scientists. With questions so grand, the expertise to answer them requires the entire spectrum of science. This combination of new technology and its widespread dispersion together with broad-ranging collaborative projects will culminate in the frabjous day when the undertaking that began with genome sequencing reaches fruition.

References and Notes

1. S. P. Gygi, Y. Rochon, B. R. Franza, R. Aebersold, *Mol. Cell. Biol.* **19**, 1720 (1999).
2. W. Blackstock, in *Proteomics: A Trends Guide*, W. Blackstock, M. Mann, Eds. (Elsevier Science, London, 2000), pp. 12–17.
3. G. Neubauer *et al.*, *Nature Genet.* **20**, 46 (1998).
4. M. P. Rout *et al.*, *J. Cell Biol.* **148**, 635 (2000).
5. J. B. Peltier *et al.*, *Plant Cell* **12**, 319 (2000).
6. A. J. Link *et al.*, *Nature Biotechnol.* **17**, 676 (1999).
7. M. P. Washburn, D. Wolters, J. R. I. Yates, *Nature Biotechnol.*, in press.
8. M. Fromont-Racine, J. C. Rain, P. Legrain, *Nature Genet.* **16**, 277 (1997).
9. A. J. Walhout *et al.*, *Science* **287**, 116 (2000).
10. P. L. Bartel, J. A. Roecklein, D. SenGupta, S. Fields, *Nature Genet.* **12**, 72 (1996).
11. S. McCraith, T. Holtzman, B. Moss, S. Fields, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 4879 (2000).
12. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
13. B. Schwikowski, P. Uetz, S. Fields, *Nature Biotechnol.*, in press.
14. P. Ross-Macdonald *et al.*, *Nature* **402**, 413 (1999).
15. M. R. Martzen *et al.*, *Science* **286**, 1153 (1999).
16. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4285 (1999).
17. E. M. Marcotte *et al.*, *Science* **285**, 751 (1999).
18. A. J. Enright, I. Iliopoulos, N. C. Kyrpides, C. A. Ouzounis, *Nature* **402**, 86 (1999).
19. T. Dandekar, B. Snel, M. Huynen, P. Bork, *Trends Biochem. Sci.* **23**, 324 (1998).
20. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, N. Maltsev, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2896 (1999).
21. R. J. Cho *et al.*, *Molecular Cell* **2**, 65 (1998).
22. P. T. Spellman *et al.*, *Mol. Biol. Cell* **9**, 3273 (1998).
23. S. Chu *et al.*, *Science* **282**, 699 (1998).
24. J. L. DeRisi, V. R. Iyer, P. O. Brown, *Science* **278**, 680 (1997).
25. C. M. Perou *et al.*, *Nature* **406**, 747 (2000).
26. M. Diehn, M. B. Eisen, D. Botstein, P. O. Brown, *Nature Genet.* **25**, 58 (2000).
27. M. L. Bulyk, E. Gentalen, D. J. Lockhart, G. M. Church, *Nature Biotechnol.* **17**, 573 (1999).
28. J. G. Hacia *et al.*, *Nature Genet.* **14**, 441 (1996).
29. S. P. Gygi *et al.*, *Nature Biotechnol.* **17**, 994 (1999).
30. O. N. Jensen, in *Proteomics: A Trends Guide*, W. Blackstock, M. Mann, Eds. (Elsevier Science, London, 2000), pp. 36–42.
31. G. MacBeath, S. L. Schreiber, *Science* **289**, 1760 (2000).
32. H. Zhu *et al.*, *Nature Genet.* **26**, 283 (2000).
33. M. Johnston, S. Fields, *Nature Genet.* **24**, 5 (2000).

34. I am especially grateful to E. Phizicky, M. Johnston, R. Aebersold, C. Manoil, and members of my laboratory for discussion and comments on the manuscript. Work from my laboratory has been supported by NIH grant P41 RR11823 from the National Center for Research Resources, and I am an investigator of the Howard Hughes Medical Institute.

FUTURE DIRECTIONS: GENOMICS AND MEDICINE

Dissecting Human Disease in the Postgenomic Era

Leena Peltonen and Victor A. McKusick

As overwhelmingly demonstrated by the sequencing papers in this issue, the complete anatomy of the human genome is now before us. In a very short time—within a decade—we have advanced from having very little information about the genetic details of biology to possessing an immense amount of structural information about individual genes. Currently, the complete genome sequences of more than 60 species are available in databases, and the prediction that there will be a total of 100 sequenced genomes in databases within the next few months seems realistic. This dramatic increase in the amount of genomic information will have a tremendous impact on biomedical research and on the way that medicine is practiced. When all the human genes are truly known, scientists will have produced a Periodic Table of Life, containing the complete list and structure of all genes and providing us with a collection of high-precision tools with which to study the details of human development and disease. New technologies will facilitate analyses of individual variations in the whole genome and the expression profiles of all genes in all cell types and tissues. The way will thus be paved for systems biology and for deciphering the genetic repertoires of many organisms. The complete genome sequence of humans and of many other species provides a new

L. Peltonen is in the Department of Human Genetics, University of California Los Angeles School of Medicine, Los Angeles, CA 90095–7088, USA. E-mail: lpeltonen@mednet.ucla.edu V. A. McKusick is in the Department of Medical Genetics, Johns Hopkins Medical Institutions, Baltimore, MD 21205, USA.