Proteomics

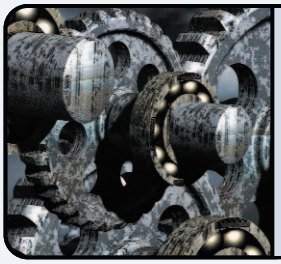
Characterizing the Cogs in the Machinery of Life

Now that the human genome sequence is complete, the quest to extract beneficial knowledge from it is on. One of the most promising, active areas of exploration lies in the human proteome—the global expression of proteins, those marvelous strings of amino acids responsible for all human biologic processes. Proteins are life, and the recently developed ability to study them on a large scale, quantitatively and qualitatively, is known as proteomics.

The human proteome may never be completely solved in the same way the genome was. The genome is relatively static, and presented a finite end point. The proteome is dynamic, changing constantly with time and conditions, with proteins interacting to form networks and pathways to respond to stimuli and carry on the endless business of cellular function. The challenge of completely mapping the proteome is widely considered to be several orders of magnitude greater than that of the genome. The picture is so complex and so dynamic that some proteomics experts

question the very concept of the existence of a measurable human proteome. Famed genomicist J. Craig Venter put this doubt succinctly when he told the 5 April 2001 *Wall Street Journal* that "there ain't nosuch thing as a proteome."

Nevertheless, there can be no debate that proteomics is poised to deliver vast amounts of useful information about physiologic function at the subcellular, cellular, organic, and systemic levels, yielding profound new insights into disease and drug mechanisms, the effects of environmental exposures, and much more. Although a comprehensive map of the entire human proteome may never be accomplished, protein maps of human organs, glands, and fluids and of entire less-complex organisms are within sight, and major efforts are under way to document many of those proteomes.



The proteome is constantly changing. Technologically, you're always trying to hit a moving target.

— B. Alex Merrick
NIEHS

Evolution of Proteomics

In large measure, proteomics has emerged in parallel fashion with the other "-omics" fields such as transcriptomics and metabonomics. The technologies, methodologies, and grand ambitions of the Human Genome Project have rapidly proliferated and now permeate virtually every area of the life sciences.

Just as the advent of genomics brought the ability to discover large numbers of genes quickly, proteomics was born when technologic advances allowed scientists to widen their focus from the painstaking isolation and identification of single proteins to a more comprehensive view of the entire protein complement expressed in a given cell line, tissue, or organism. However, proteomics researchers employ their own unique mix of tools, approaches, and skills to address the questions they seek to answer.

Although the term "proteomics" did not exist until 1994, when Australian postdoctoral student Marc Wilkins coined it, the practice of the science has been going on since the mid-1970s. Two milestone technologic breakthroughs facilitated the ability to look at multiplicities of proteins, both of which, although much refined, are still in wide use today in laboratories around the world.

At its core, proteomics is all about separation and identification—the process of taking a sample of interest, separating out all of the proteins therein, and then identifying them. The first major breakthrough, which was a great leap forward in separation, took place in 1975 with the introduction of two-dimensional gel electrophoresis (2DE).

With this method—still the first step in many proteomics experiments—proteins from a sample are separated on a polyacrylamide gel according to their mass and charge, which, along with intensity, are what provide the spectrum that makes up a protein's distinctive signature. The more abundant the protein, the larger and more intensely staining the spot on the gel.

The only problem is that 2DE, while it allows separation and visualization of the protein complement, does little or nothing to address identification. Regardless, the advent of 2DE was so exciting that in 1980 it spawned the proposal of a Human Protein Index project—an effort to catalog all human proteins and then use that knowledge to define the genome (although Congress considered the project, it was never funded, and advances in genomics soon bypassed the idea).

The second major breakthrough, which really brought proteomics into its own, was the arrival of two crucial techniques in the 1980s that made possible the use of mass spectrometry (MS) to identify proteins: matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI). These methods allow protein samples to be ionized for analysis in a mass spectrometer, producing a pattern called a mass spectrum. These mass spectra—which often number in the thousands for a given sample—can then be used to positively identify proteins or protein digests (strings of peptides or protein fragments produced when the proteins are ionized) through the automatic querying of protein databases. Unidentified or novel proteins can be analyzed through further MS runs or by other techniques.

"ESI and MALDI were a quantum leap," says William Pierce, a professor of pharmacology, toxicology, and chemistry at the University of Louisville School of Medicine. The subsequent development of time-of-flight (TOF) detection, which expanded the range of ionic molecular weights detectable by MS instruments, brought further analytic capabilities to MALDI. Today, MALDI-TOF is in widespread use in proteomics laboratories.

A dazzling panoply of MS refinements and enhancements, as well as the development of other technologies, now facilitate the application of proteomic techniques to a highly diversified universe of research pursuits. Virtually every proteomics laboratory, whether it's connected with government, academia, or industry, seems to have its own favored technologic approach, and many have developed their own in-house methods, along with customized bioinformatics software, to help sort through and make sense of the massive amounts of data their systems generate.

"I think we need more than one platform to be able to adequately do service to measuring proteins in a global fashion," says B. Alex Merrick, head of the Proteomics Group at the NIEHS National Center for Toxicogenomics (NCT). "The proteome is constantly changing. Technologically, you're always trying to hit a moving target." Merrick cautions, however, that "because proteins have so many properties, or attributes, and we have the potential to measure them, it spawns many, many platforms, and technologically we haven't sorted out which platforms are the best ones."

The general feeling among proteomics researchers is that the field is on the verge of consolidating years of method development into a flood of knowledge. "I think we're about to transition out of the age of explorers in proteomics to the age of applications," says Daniel Liebler, a professor of biochemistry and director of proteomics at the Vanderbilt University School of

Getty Images, Matt Ray/EHP

Medicine. "Proteomics techniques are not going to be done just as demonstrations of powerful technology, but will really be integrated into studies in basic laboratory science, animal and nonanimal models of diseases and environmental exposures, and actually in human clinical studies as well."

Proteomics in Action: Cancer Detection

Thanks to progress in clinical proteomics, someday soon a simple blood test could hold the key to early diagnosis of certain cancers. That is one of the many goals of the Clinical Proteomics Program, a joint research effort that is codirected by biochemist Emanuel Petricoin of the U.S. Food and Drug Administration (FDA) and pathologist Lance Liotta of the National Cancer Institute (NCI).

The group has developed a method of identifying protein patterns in blood serum that is potentially indicative of the presence of a wide range of diseases. Their initial study, published 16 February 2002 in *The Lancet*, focused on ovarian cancer, which presently has both a poor late-stage survival rate and a poor early detection rate—a deadly combination fostering an urgent need for better diagnostic tests, especially for women at high risk for developing the disease.

The investigators used surface-enhanced laser desorption/ionization TOF (SELDI-TOF), a variation on MALDI-TOF that incorporates protein microarrays and is particularly well suited to detecting patterns of proteins in samples. First, a "training" set of known, unblinded samples was run through the instrument—in this case, serum samples from both healthy women and women with ovarian cancer. With the high throughput of the equipment, spectra for each sample—each one containing 15,200 data points, or individual pieces of information—were quickly generated.

Next, the raw data were processed by a unique bioinformatics system that incorporates a form of artificial intelligence called a genetic algorithm. The genetic algorithm compares the patterns of protein expression in the diseased samples to those in the healthy samples, looking for those patterns that optimally discriminate between the two. The algorithm learns as it goes in a process that involves hundreds of millions of pattern combinations and comparisons. The end product is a pattern of unidentified proteins—in this case, five—that precisely distinguishes healthy samples from diseased ones.

The next step was to run a set of known but blinded samples through the same process, and then compare the results to assess the predictive power of the patterns. In this study, the investigators achieved a sensitivity of 100%—that is, all of the cancerous samples were correctly identified, with no false negatives—and a specificity of 95%, meaning only 5% of the identifications were false positive. This was vastly superior to the 35% positive predictive value in the same samples of cancer antigen 125, the present gold standard clinical biomarker.

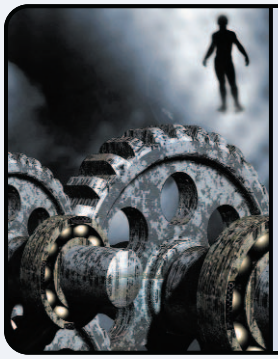
Subsequent technical refinements (which included a switch to a much higher-resolution and more stable mass spectrometer, and incorporation of advanced spectral quality control methods) have improved the system's sensitivity and specificity to 100% in a larger blinded set of ovarian cancer and high-risk samples. The team is currently enrolling participants in a clinical trial to test their methodology in detecting recurrence of ovarian cancer.

Although the proteins in the discriminatory pattern generated by this method are at least initially unidentified, Petricoin says that is beside the point. "We as scientists want to understand what the nature of the beast is, and we're hunting that down," he says. "We're already making great progress to that end. But we don't see identity as necessary for its use in diagnostics."

Two of the world's largest reference laboratory companies apparently agree. Quest Diagnostics and LabCorp have sublicensed the technology from Correlogic Systems (which developed the initial genetics algorithm and licensed the technology from the U.S. government), and plan to start offering the proteomic test as an ovarian cancer screening tool for women at high risk by the end of 2003. Initially, they will market the procedure under the FDA's "home brew" provision, which allows the companies to perform the service only in their own validated laboratories.

Petricoin is optimistic that proteomic pattern diagnostics could impact medical diagnostics in a big way. "This is a different type of diagnostic paradigm," he says. "[It] completely changes and turns on its head the normal tried-and-true route—which we would suggest is failing—of looking at discovery biomarkers. . . . I think if either our or LabCorp/Quest's efforts are successful, it's really going to throw a gauntlet down on a completely different type of diagnostic procedure being used in the clinic. That could have reverberations throughout disease detection, period."

The FDA/NCI group is applying this proteomic technique in similar studies of breast, lung, pancreatic, esophageal, brain, and prostate cancers, as well as efforts to detect cancer drug cardiovascular toxicity before symptoms occur, and to assess the effectiveness of molecularly targeted cancer drugs.



Getty Images, Matt Ray/EHP

> Proteomics techniques are not going to be done just as demonstrations of powerful technology, but will really be integrated into studies in basic laboratory science, animal and nonanimal models of diseases and environmental exposures, and actually in human clinical studies as well.
>
> —Daniel Liebler
> Vanderbilt University School of Medicine

Another Approach: Cancer Profiling

The Mass Spectrometry Research Center at the Vanderbilt University School of

Medicine, headed by Richard Caprioli, the Stanley Cohen Professor of Biochemistry, also is pursuing methods of distinguishing diseased tissue from healthy tissue, particularly in cancer. But Caprioli's group takes a very different approach, in which identification of the proteins in the affected tissues themselves, rather than in plasma, is central. It's called tissue proteome profiling, and it appears to be a powerful new tool for both diagnosis and prognosis.

Tissue proteome profiling has several advantages, including the ability to accurately detect factors such as life expectancy and tumor aggressiveness. Tissue proteome profiling could also directly identify potential targets for drug intervention, as well as contribute to understanding of mechanisms of the disease.

In their most complete study to date, published 9 August 2003 in *The Lancet*, Caprioli and his coworkers concentrated on lung tumors. They took hundreds of lung tumor biopsy samples and analyzed their protein complements via MALDI MS, looking at several spots on each sample, each of which generated thousands of signals in a specific pattern of proteins. Then, using a series of bioinformatics tools, they correlated the tissue proteome profiling information with known information about the individual patients, some of whom had already died of their disease.

"We found that at the first level, we could find unique sweeps of proteins that helped us actually classify the disease," says Caprioli. "So if you take the biggest set of lung tumors, non–small cell lung carcinomas, we could further classify them as adenocarcinomas, squamous cell carcinomas, and so on."

Of course, pathologists can do the same, but Caprioli says they didn't stop there. "We asked, 'Can we correlate these patterns with the life expectancy or the prognostic value of these diseases?' And it turned out to be of very high accuracy." He says the researchers could tell from the protein pattern which patients would go on to survive for long periods of time, and which patients would die of cancer—"so the aggressiveness of the disease was apparent in the protein profile."

They were further able to pick out with approximately 80% accuracy those patients whose tumors had metastasized, causing nodal involvement and the often inoperable development of secondary tumors. There is presently no other method of making such a crucial clinical prediction.

Although appropriately cautious to point out that these results were in just one type of tumor study, Caprioli is excited about the possibility of using protein patterns to identify types of tumors that have an aggressive posture for nodal involvement. "It begins to get you out of just diagnosing to actual patient care, so that the clinician can now identify a high-risk group and make the appropriate therapeutic decisions," he says.

Caprioli's group is nearing completion of a similar study of brain tumors with similar results in terms of the power of tissue protein profiling for prognostication. They're also looking at diabetes mellitus, cardiac and pulmonary diseases, and several other conditions. Caprioli asserts that this platform, along with other clinical proteomics work, ultimately constitutes an entry point into the field of individualized medicine, centered around the concept that each patient's disease is unique at the molecular level. "It's a whole new way of looking at things," he says. "There's no doubt in my mind that as we collectively learn more and more about the molecular ways of diagnosing disease, of predicting disease progression, that this individualized way of looking at diseases will become more and more common."

Toxicoproteomics: Mechanisms and Biomarkers

Liebler's main focus is toxicoproteomics. His group concentrates on understanding how reactive intermediates produce deleterious effects by modifying proteins. These unstable chemical species enter a cell as a result of environmental exposures and tend to bind to proteins or DNA, modifying their properties in an injurious way and forming new biomolecules known as adducts.

Using a form of MS called tandem MS, or MS/MS, along with a novel algorithm and proprietary bioinformatics software called Scoring Algorithm for Spectral Analysis, Liebler and his group are able to analyze the mass spectra of peptides to establish their sequences, the positions of any modifications, and, by mapping that information back onto the entire protein sequence, the sites of modification in the protein itself. Ultimately, two major questions are addressed: What are the protein targets of reactive intermediates? And what are the cellular responses to protein modification?

Answers to these questions will shed light on some of the most important avenues of contemporary research in toxicology. Liebler sees the biggest near-term payoff of this type of work as coming in two general areas. One area is the understanding of mechanisms of toxicity. The other is the identification of biomarkers of exposure. "If we can figure out what the targets of some of these environmental compounds are or what reactive intermediates come from environmental stimuli or stresses, by understanding mechanisms we then know what components of the cell or tissues might be amenable to some kind of protective intervention," he says.

As proof of principle, Liebler's team published a study in the June 2002 issue of *Chemical Research in Toxicology* documenting their system's ability to map hemoglobin adducts of the aliphatic epoxides, a group of common industrial chemicals. The team is most interested in investigating biomarkers of oxidative stress, the damaging phenomenon implicated in many disease processes and often the result of environmental exposure. "What we would like to do is identify some of the most abundant of these reactive intermediates that are formed under representative conditions either *in vitro* or in animal models *in vivo*, where we can manipulate oxidative stress," says Liebler.

Pierce's biomolecular MS lab at Louisville is involved in similar functional proteomics work. His group looks at subsets, or small clusters, of functionally interactive proteins. They isolate post-translational modifications, any of more than 100 different types of changes that can be made to proteins by a variety of factors after their original creation. (This partially accounts for the vastly larger number of proteins than genes.) Pierce's group also works to develop or validate biomarkers in cases of specific environmental or xenobiotic exposures and those agents' interactions with nucleic acids and proteins.

In a collaboration with Louisville professor of medicine Aruni Bhatnagar, Pierce and colleagues are looking at a very large, ubiquitous set of chemicals, the aldehydes, which form adducts with proteins, potentially contributing to cardiovascular disease. Aldehydes are not just environmental contaminants, but are also naturally present in food and are intermediates in human metabolism. By identifying aldehyde-induced adducts and elucidating how they might influence protein function, the team hopes to characterize a novel mechanism involved in hypertension, stroke, and other forms of cardiovascular disease.

Pierce is also working on a project with university associate professor of medicine James Summersgill, investigating the

Getty Images, Matt Ray/EHP

Proteomics Resources

Groups and Initiatives

Human Proteome Organisation (HUPO)
http://www.hupo.org/
An international research consortium intended to encourage large-scale analysis of the human proteome

Human Proteomics Initiative
http://www.expasy.org/sprot/hpi/
Joint effort of the Swiss Institute of Bioinformatics and the European Bioinformatics Institute that seeks to comprehensively annotate all known human proteins

National Heart, Lung, and Blood Institute Proteomics Initiative
A seven-year, $157 million program to accelerate the development of innovative technologies to characterize healthy and diseased heart, lung, blood, and sleep processes in 10 special centers of proteomics research across the country

Protein Structure Initiative
http://www.structuralgenomics.org/
A 10-year project funded by the National Institute of General Medical Sciences to determine the three-dimensional structures of 10,000 unique proteins, while dramatically reducing the time and costs involved in the process

Databases

Biomolecular Interaction Network Database
http://www.blueprint.org/
A comprehensive, publicly accessible repository administered by blueprint WORLDWIDE for data and software tools related to critical biomolecular functions

Chemical Effects in Biological Systems Knowledge Base
http://www.niehs.nih.gov/nct/cebs.htm
National Center for Toxicogenomics database that will exhaustively document the toxic effects of chemicals in the environment and be fully searchable by compound, structure, toxicity, pathology, gene, gene group, single-nucleotide polymorphism, pathway, and network

Human Protein Reference Database
http://www.hprd.org/
Joint project of The Johns Hopkins University and the Institute of Bioinformatics that is expected to eventually contain comprehensive entries on 10,000 human proteins, including domain architecture, post-translational modifications, interaction networks, and disease associations

Protein Sequence Database
http://pir.georgetown.edu/pirwww/dbinfo/pirpsd.html
A comprehensive annotated protein sequence database in the public domain, maintained by the Protein Information Resource, that contained more than 283,000 entries as of November 2003

Swiss-Prot
http://us.expasy.org/sprot/
A curated protein sequence database developed by the Swiss Institute of Bioinformatics and the European Bioinformatics Institute that strives to provide a high level of annotation, a minimal level of redundancy, and high level of integration with other databases

TrEMBL
http://www.ebi.ac.uk/trembl/
A database maintained by the European Bioinformatics Institute that contains the translations of all coding sequences present in the European Molecular Biology Laboratory's Nucleotide Sequence Database that are not yet integrated into Swiss-Prot

United Protein Database
http://www.uniprot.org/
With $15 million in funding from six NIH institutes and centers, will combine the resources of Swiss-Prot, TrEMBL, and the Protein Sequence Database.

Pieces of the Proteomics Puzzle

Proteomics encompasses several different subdisciplines, each with its own unique approach and its own contribution to the overall quest to glean knowledge from the proteome.

Expression Proteomics

In expression (or profiling) proteomics, researchers seek to discover and quantify significant differences in the totality of expressed proteins between known samples—often diseased versus nondiseased or exposed versus unexposed. These differences appear as patterns that can have a very high degree of predictive power, whether the proteins in the pattern are identified (as some experts contend is necessary) or remain unidentified (which others argue is sufficient). Expression proteomics studies yield hypotheses that are then confirmed or refuted by other methods. Clinical proteomics investigations, which seek to apply proteomics knowledge directly to medical practice, typically employ expression proteomics methods.

Functional Proteomics

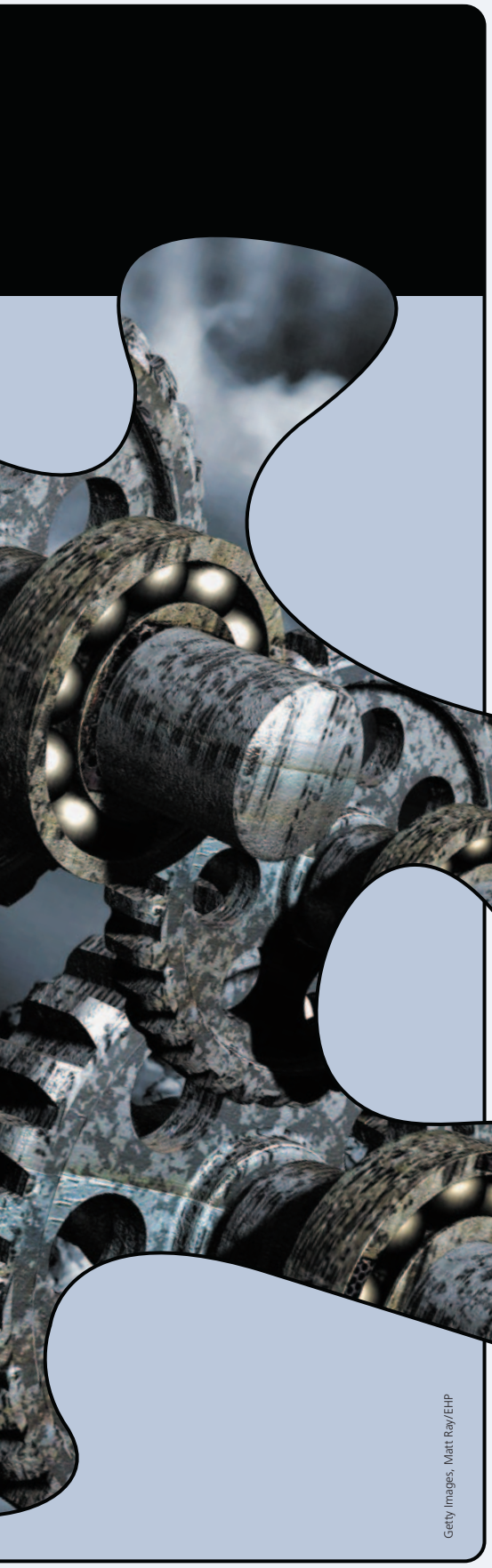
Functional proteomics encompasses a wide variety of studies involving subsets of proteins. These studies seek to analyze and characterize specific functions, including signaling pathways, interactions, disease mechanisms, and biomarkers of disease or environmental exposures. In this field, hypotheses are tested rather than developed, and protein identification is vital to success.

Structural Proteomics

Structural proteomics concentrates on mapping the structure of protein complexes or those proteins present in a specific cellular organelle. Such information can provide valuable insights into cellular architecture, which greatly influences cellular function. X-ray crystallography and structural modeling by computational biology are the main methods utilized to unravel these extremely complicated systems.

Toxicoproteomics

In toxicoproteomics, the full range of proteomics methods and technologies are used in efforts to uncover the cellular and subcellular mechanisms at work in response to xenobiotic exposures. Researchers in this area are particularly interested in discovering biomarkers of exposure.



Getty Images, Matt Ray/EHP

interactions of the microorganism *Chlamydia pneumoniae* with the cardiovascular system. Just as *Helicobacter pylori* has been implicated in gastric ulcers, there is a theory that microorganisms in the cardiovascular system could cause systemic infection, leading to plaque development and atherosclerosis. "We study the chlamydial proteome and look at changes in it and how that might be reflected in the production of products that then stimulate atherosclerotic lesions," says Pierce.

Like all proteomics practitioners, he is enthusiastic about the possibilities that lie ahead in the field. "The infinite variety of states of proteins in the cell will give us the opportunity, more so than in genomics, to uncover new mechanisms in biology," he says. "In certain aspects you're looking at a dynamic system that is growing and changing, and we can actually 'catch biology happening.' And because of that, we'll find new mechanisms and be more likely to develop new ways to look at mechanisms or affect them."

NIEHS researchers are also delving into the field of proteomics. Merrick and the NCT Proteomics Group, working in partnership with Kenneth B. Tomer, who heads the NCT Mass Spectrometry Group, aid the center's efforts to discover more and better information about the adverse effects of chemicals and toxic compounds. Merrick's group works mainly in expression proteomics experiments with animals. "We want to be able to evaluate the effects of chemicals in experimental animals under the most controlled conditions possible," he says, "so that we can separate out the true effects of the chemicals from the nonspecific 'noise' effects that you always see with these types of technologies."

Among other projects, Merrick's group uses SELDI to examine the mechanisms of action of two key proteins involved in cell growth and cell death—p53 and NFκB. "p53 is often regarded as one of the 'master switches' of life and death and cell growth within the cell," says Merrick. "In the same sense, NFκB is a 'master switch' for inflammation and immune response. In these two proteins, we're looking for specific markers, specific states that would distinguish them in terms of their being activated or deactivated in association with a particular disease state or state of cellular function."

In the 3 April 2001 issue of *Biochemistry*, the Merrick and Tomer groups reported the results of their MS research on p53. They were able to isolate the entire protein from the cell for comprehensive MS analysis for the first time, in an effort to shed light on how the fine structure of the protein influences its ability to control cellular life and death. It has long been suspected that phosphorylation, a type of post-translational modification, may be involved in the process. The group discovered six specific phosphorylation sites on the protein, one of which, Ser(315), was particularly phosphorylated. Unraveling the mystery of how p53 exerts its "master switch" control over cellular mortality would be an important advance in biology, and this study constitutes a major step toward that discovery.

The group is also undertaking a number of clinical projects in neurodegenerative disease and cancer, taking advantage of access to blood and serum samples from NIEHS epidemiologic activities. Serum proteomic analysis has much to tell, according to Merrick. "The soluble proteins within serum or plasma can be reflective of a disease state, or of toxicity or injury to a particular disease site, whether it be in heart disease or liver disease," he explains. "So proteomics can shine in analysis of serum or plasma because of the nature of disease, in that you may either have release of a biomarker from a particular organ, or there may be indications of a repair process going on with serum or plasma that you can [detect in the bodily fluids]."

Toxicoproteomics can also be used to discover previously unknown sites within the cell, says Merrick. "When you're dealing with proteins, you're dealing with time and space," he says. "These proteins occupy a certain amount of space within tissues or cells, and to be able to isolate these subcellular portions that are important targets of either therapy or of toxicity is an area where proteomics can make a special contribution."

Calcium, Oxidation, and Aging

At the Pacific Northwest National Laboratory in Richland, Washington, researcher Thomas Squier practices proteomics as part of the laboratory's systems biology approach, which integrates information from all of the -omics disciplines to first determine how a cell functions and then develop predictive models. The lab's

Biomolecular Systems Initiative, which includes its proteomics work, is part of the U.S. Department of Energy's Genomes to Life program, which aims at uncovering biologic solutions for major environmental issues such as clean energy production, removal of excess carbon dioxide from the atmosphere, and remediation of contaminated environments.

Squier's group concentrates on analyzing calcium regulation in cells and how oxidative stress can trigger adaptive mechanisms, resulting in post-translational modifications to key calcium sensor proteins. Calcium maintains a 10,000-fold gradient in cellular systems and is the key player in the signaling that modulates energy metabolism. Changing calcium levels are responsible for much of the cell's sensing of the environment. By identifying post-translational modifications of the key calcium sensor proteins (changes such as methionine oxidation and protein nitration), potentially important new biomarkers of exposure can be isolated. For example, the lab has identified the calcium signaling protein calmodulin as a major target of oxidative stress, as described in the January 2003 issue of *Chemical Research in Toxicology*. This discovery could contribute significantly to understanding of adaptive cellular responses to environmental exposures, particularly in how repair and maintenance systems are triggered.

Aging is an important factor in cellular adaptive ability as well. Aging is a major risk factor for most diseases and for sensitivity to environmental exposures, says Squier. "In aging," he says, "the key regulatory proteins get oxidized, and their oxidation slows metabolism down. We speculate that this is an adaptive mechanism to maintain this balance between reactive species and cell function."

Ultimately, this work on the detection of post-translational modifications of key sensor proteins could lead to the development of microarray-based assays that would rapidly analyze a person's antioxidant status. In terms of applications, Squier says, being able to quickly identify changes in protein expression and discern what post-translational modifications happen is going to provide a very high level of information about the health of an individual, which in turn could lead to greatly enhanced medical treatment.

Proteomics Initiatives and Databases

Considering the enormous challenges and opportunities posed by proteomics, it's unsurprising that there are several collaborative proteomics initiatives under way at the national and international levels. Perhaps the best known of these campaigns is the Human Proteome Organisation (HUPO), an international body intended to encourage large-scale analysis of the human proteome. HUPO seeks to consolidate national and regional proteome organizations into a worldwide research consortium. "Proteomics cannot be fully grasped and developed without a major international organized effort, which HUPO intends to facilitate," says Samir Hanash, the organization's president and a professor of pediatrics at the University of Michigan.

HUPO has established a goal of mapping 5,000 human proteins, and is coordinating and standardizing research in a variety of pertinent areas. Its major projects include analysis of specific regions of the body—the Human Plasma Proteome Project, the Human Liver Proteome Project, and the Human Brain Proteome Project. Another major project is the Proteomics Standards Initiative, which aims to define community standards for presentation of proteomics data. Still other projects include initiatives involving new proteomics technologies, cell models and tissue, bioinformatics, and the development of a collection of standardized, high-quality antibodies for every human protein. HUPO's shared resources, data, and establishment of standardized protocols and reporting guidelines should contribute substantially to the understanding of disease processes and chemical exposures.

The Human Proteomics Initiative seeks to comprehensively annotate all known human proteins, which means parsing out each protein's function, domain structure, subcellular location, post-translational modifications, variants, similarities to other proteins, and protein sequence polymorphisms. This ambitious project is sponsored by the Swiss Institute of Bioinformatics and the European Bioinformatics Institute, the keepers of one of the most widely used protein sequence databases, Swiss-Prot.

In certain aspects you're looking at a dynamic system that is growing and changing, and we can actually 'catch biology happening.' And because of that, we'll find new mechanisms and be more likely to develop new ways to look at mechanisms or affect them.

—William Pierce
University of Louisville School of Medicine

Virtually all proteomics experiments involve accessing and querying protein databases as an integral step in the process, allowing the identification and characterization of detected proteins and peptides. That vital link between data and knowledge should be greatly enhanced by the establishment of the United Protein Database, or UniProt. Funded in October 2002 by a three-year, $15 million grant subsidized primarily by the National Human Genome Research Institute along with five other institutes and centers of the NIH, UniProt will combine the resources of Swiss-Prot and two other major annotated protein databases, the European Bioinformatics Institute's TrEMBL and the Protein Information Resource's Protein Sequence Database. By January 2005, UniProt will be fully operational and available to all users free of charge.

In the world of proteomics, the main action often centers around interactions, molecular complexes, and pathways. The

Getty Images, Matt Ray/EHP

Biomolecular Interaction Network Database serves as a comprehensive, publicly accessible repository for data and software tools related to those critical biomolecular functions. The database is administered by blueprint WORLDWIDE, a nonprofit organization cofounded for that purpose by IBM and MDS Proteomics of Toronto, Canada.

Another important resource in the protein database arena has recently been launched as a joint project between researchers at The Johns Hopkins University in Baltimore, Maryland, and the Institute of Bioinformatics in Bangalore, India. By the end of 2003, the Human Protein Reference Database is expected to contain comprehensive entries on 10,000 human proteins, including domain architecture, post-translational modifications, interaction networks, and disease associations. The information in this database has been manually extracted from the literature by biologists who read, interpret, and analyze the published data.

To spur the progress of clinical proteomics, in 2002 the National Heart, Lung, and Blood Institute launched a major initiative that created 10 special centers of proteomics research at academic institutions across the country. The seven-year, $157 million program is designed to accelerate the development of innovative technologies to characterize healthy and diseased heart, lung, blood, and sleep processes. Says the institute's proteomic program administrator Susan Old, "This should speed the delivery of potential new clinical applications from research into practice." The centers will investigate protein profiling, interactions, and post-translational modifications as they relate to a variety of conditions, including cardiovascular disease, autoimmune disease, airway inflammation, and cystic fibrosis.

The development of better tools and better knowledge of structural proteomics is the goal of the Protein Structure Initiative, a 10-year project funded by the National Institute of General Medical Sciences and launched in 2000 with an open-ended budget. Currently in its pilot phase, the initiative aims to determine the three-dimensional structure of 10,000 unique proteins, while dramatically reducing the time and costs involved in the process. By 2005, each of nine centers is expected to be able solve the structure of 100–200 proteins annually. By grouping proteins into structural families, "the initiative will develop a catalog of all the protein structures that exist in nature," said Marvin Cassman, then director of the National Institute of General Medical Sciences, at the time the initiative was launched. "We expect that it will yield major biological findings that will improve our understanding of health and disease."

Proteomics data are also expected to play a large role in the Chemical Effects in Biological Systems (CEBS) knowledge base being developed by the NCT. CEBS is designed to exhaustively document the toxic effects of chemicals in the environment and will be fully searchable by compound, structure, toxicity, pathology, gene, gene group, single-nucleotide polymorphism, pathway, and network. The knowledge base will be accessible by the public, and will be a major contributor to progress in the fields of toxicoproteomics and toxicogenomics.

Proteomics Prognostications

Scratch your average proteomics investigator and you will reveal an optimist just under the sober scientific surface. The excitement is palpable; the visions are grand. But all in the field agree that for proteomics to fulfill its lofty promise, certain key developments must take place, several of which are well on their way to fruition.

Technologic progress must continue and accelerate. Many in the field are anxious to see the replacement of the notoriously laborious 2DE method of protein separation with a more automated, high-throughput approach, such as antibody microarrays or isotope-coded affinity tags. All wish for further improvements and refinements in MS equipment and bioinformatics, as well as development of other technologies that could contribute to progress in the field.

"I think MS is becoming increasingly powerful, and we haven't yet realized the full power of these tools," says Liebler. "On the other hand, I think MS-based proteome analyses will give way to other kinds of less high-tech approaches, using perhaps some variants of array technologies: arrays of antibodies or aptamers [basic nucleic acid equivalents of antibodies] and perhaps small molecules that recognize proteins—little lab-on-a-chip devices that would be suitable for analysis of some components of proteins." There are a lot of competing technologies, he says, and it's hard to say what's going to work—"but if the last ten years have taught us anything, it's that we should be prepared to be surprised, regularly."

Philosophically, practitioners are confident that the knowledge gleaned from proteomics will ultimately converge and integrate with advances in the other -omics fields to evolve into a more holistic systems biology discipline with the ability to understand the processes and mechanisms of life in a truly global fashion. "We tend to think of ourselves as proteomics people, or genomics people, or lipids people, and in fact the cell and tissue only exist because all of these things are integrated," says Caprioli. "How these things all relate to one another is what's going to give us the key [to a more comprehensive understanding of systems biology]."

Researchers believe that proteomics will begin to make a tangible difference in medicine and environmental health quite soon. Merrick, for example, believes that within five years, there will be perhaps two or three key public databases that will offer access to gene and protein expression experiments that are done in a standardized way, and researchers will be able to query those databases for use in predicting human health responses to various environmental interactions. In 10 years, he says, "I believe that proteomics will be able to go right into the clinic, in terms of diagnostics and evaluation of blood and serum in a way that clinical chemistry can't approach or compete with. Typically, when you get blood drawn, you get maybe twenty or thirty analyses. . . . I think in the future this will just be dwarfed by the amount of useful information that will be derived at the proteomic level."

Petricoin is even less guardedly optimistic. He predicts that in five years "a patient will be able to have a pathophysiological portrait performed by high-throughput protein-based technologies that can read out hundreds of thousands of end points at once, and be able to provide the clinician a snapshot of what's going on in that organism." Within a decade, he says, will come the development of high-throughput proteomics coupled with artificial intelligence–type systems, with nanotechnology, and even with nano-intelligence systems, allowing clinicians to harvest information and deliver tailored therapeutics based on what's happening in the serum, plasma, and tissue of any given patient who visits the doctor. "It's really going to revolutionize the way in which molecular medicine is performed," says Petricoin. "It's going to happen."

Ernie Hood