In Pursuit of the Proteome

BY STEVEN PELECH, PhD

2001 not only marks the beginning of the new millennium, but also the onset of the post-genomics era. Roughly 30,000 genes uncovered with the sequencing of the human genome are now finally at hand. The discovery of all of the human genes is akin to amassing all of the pieces in an unassembled jigsaw puzzle with no guide as to their purpose and proper arrangement. The challenge now is to connect the proteins encoded by these genes to furnish a clear picture of how physiological processes are mediated and regulated. This endeavour has been termed "functional genomics," although "functional proteomics" would be more apt phraseology. The prize for this enterprise will be vastly improved ways of diagnosing and treating disease in a new era of personalized medicine.

Mapping the Generic Proteome

With only a few exceptions (e.g. germ-line cells, red blood cells and tumour cells), almost all of the 200 or so different specialized types of cells in the human body share the same genes. However, they differ profoundly with respect to which of these genes is actively turned on to produce proteins. The term "proteome" has been adopted to specifically describe the unique complement of proteins that reside in a cell. With the human genome largely deduced, several companies have begun to direct their attention to proteomics. Notable Canadian players in the proteomics arena include MDS Proteomics, Inc., Caprion Pharmaceuticals, Inc., Kinetek Pharmaceuticals, Inc. and Integrative Proteomics Inc. But they face stiff competition. For example, in April of this year, Myriad Pharmaceuticals, Hitachi and Oracle Corp. formally announced a collaboration to map the human proteome in less than three years.

Mapping the proteomes of humans and other organisms will be several orders of magnitude more difficult than sequencing their genomes. As much as 20 per cent of all genes may be actively expressed in a typical cell. Many genes can specify the synthesis of multiple proteins through alternative splicing during gene transcription. Furthermore, most proteins undergo extensive post-translational modifications, including phosphorylation, glycosylation, sulphation, methylation and acylation, among others. These covalent modifications can have profound effects on the functional activities and locations of proteins within cells.

It is likely that, on average, each gene may specify 10 or more protein variants that arise from alternative splicing and covalent modification. Therefore, the number of potential protein entities in the human genome is probably in the order of 300,000. In December 2000, Large Scale Proteomics Corp. and its parent, Large Scale Biology Corp., completed the first version of their Human Protein Index, which inventories protein identities and amounts in the major human tissues and subcellular fractions. The Human Protein Index is based on 157 different tissue samples primarily from one female, with male-specific tissues from another individual. Over 115,000 distinct protein entities have been catalogued, about 18,000 of which are identified.

Apart from the staggering multitude of different potential proteins species within any cell, another major issue is the very dynamic nature of the proteome. A cell's protein composition markedly varies with cell type, age, health and environmental conditions. To properly understand how the cell operates under normal physiological circumstances and in disease, there is a far ranging list of questions about each protein that will have to be addressed. Where and when is the protein produced? The answer must extend beyond the type of cell to the precise subcellular location of the protein and even whether it exists in stable complexes with other proteins. What is the exact biological function of the protein? Knowledge of the protein's primary structure predicted from its gene sequence can provide some clues. Is the protein in its active or inactive state? How is it regulated? The answers to these latter questions must encompass information about the other proteins, macromolecules or metabolites with which it interacts, and the physiological processes that the protein impacts.

BIND (Bimolecular Interaction Network Database) is an important Canadian initiative to track all of the known interactions between proteins and other biomolecules. With backing from MDS Proteomics and IBM, BIND is a publicly accessible bioinformatics database that allows scientists worldwide to submit and review the results of research about molecular interactions. Such data is presently being generated from a variety of approaches, including the employment of the yeast dihybrid system, affinity purification and co-immunoprecipitation of proteins. In the near future, through the development of microarrays with immobilized bait proteins and the detection of captured proteins by imaging surface plasma resonance, it may be possible to evaluate the interactions of thousands of proteins within minutes.

Over time it should become possible to define the "rules of engagement" between proteins. This will encompass whether there are any precedents or predictions for interaction between any two proteins, and the outcome of

such an interaction. Whether two proteins can actually interact in a proteome also depends on their co-localization within the same cell. The function of any protein is not only dictated by its primary structure, but also the context in which it must work. The same protein may have multiple functions, depending on which other target proteins are present. Thus it is of equal importance to track the wide-scale distribution of proteins to hone in on their functions.

Tracking the Proteome

Through microarray technology, the transcription of thousands of genes can be monitored simultaneously by assessing the levels of the mRNAs for these genes in tissue or cell lysates. Although the results of gene microarray analysis are not particularly quantitative, the scope, sensitivity and speed of this method is expected to spawn a $1 billion US industry within the next five years. Altered expressions of many genes serve as markers of disease, drug action and the potential sensitivities of patients to specific drugs. This form of analysis is particularly powerful when critical single nucleotide polymorphisms (SNPs) in genes important in disease or drug uptake and metabolism are tracked.

From the proteomics perspective, however, there are serious drawbacks to gene microarrays. mRNAs are translated by the cell's protein synthesis machinery to produce proteins. But, quantitation of the mRNA level for a protein provides only an indirect measure of the amount of that protein in a tissue or cell sample. There seems to be only a 50 per cent correlation on average between the levels of a given mRNA and its translated protein product. This is a consequence of regulation of mRNA translation and protein degradation following its synthesis. Other shortcomings of gene microarrays include lack of data about the subcellular locations of proteins, their states of covalent modification and their possible interactions with other proteins.

Two-dimensional (2-D) gel electrophoresis has been a robust tool during the last two decades for the resolution and simultaneous detection of several thousand proteins in cell lysates. This method involves the separation of proteins initially by isoelectric focusing in the first dimension, followed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) in the second dimension. In the first dimension, proteins migrate within a pH gradient until their net charge is zero (isoelectric pH). In the second dimension, the proteins are further resolved on the basis of their molecular size. Proteins are subsequently visualized as constellations of spots by silver staining. The greater amount of a given protein in a cell-derived sample, the larger and darker its specific spot displays. Stained proteins can be identified following their isolation from the 2-D gels and analysis by MALDI-TOF (Matrix-Assisted Laser Desorption Ionization Time-of-Flight) mass spectrometry. The direct identification of these proteins has become possible by matching the charge to mass ratio values of fragments of these proteins with those predicted from the sequencing of the human genome.

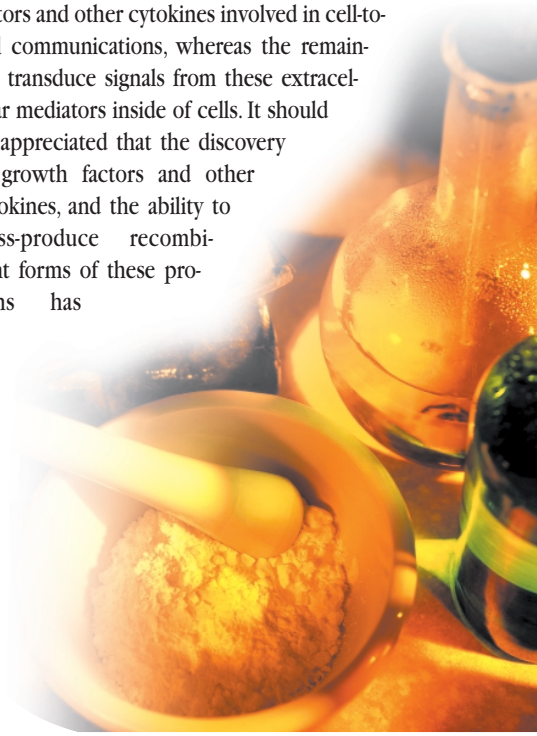
One of the serious limitations of 2-D gel electrophoresis is that this technique can at best resolve 6,000 to 7,000 of the potential 60,000 protein species in a typical cell. Furthermore, the recovery of proteins from the isoelectric focusing step into the SDS-PAGE gel can be less than 10 per cent. This can occur because proteins are generally least soluble at their isoelectric pH, and may precipitate during electrophoresis in the first dimension. However, the most pressing issue with 2-D gel electrophoresis is that the most interesting proteins to the pharmaceutical industry are usually present at such minute levels that they are not readily detected by silver staining. Those proteins that are visualized in 2-D gels tend to be metabolic pathway enzymes and structural proteins that are expressed in abundance. Therefore, it is often necessary to incorporate selective enrichment techniques, such as subcellular fractionation or partial purification, to enrich low abundance proteins as a prelude to 2-D gel electrophoresis.

The use of specific antibody probes can permit the detection and quantitation of even rarely expressed proteins in tissue and cell lysates by immunoblotting and immunohistochemistry approaches. Over 20,000 antibodies are already commercially available for detection of several thousand diverse proteins. With knowledge of the primary structures of all of the proteins in the human genome, antibodies can be developed that target any protein of interest. Companies such as Becton-Dickinson Transduction Laboratories and Kinexus Bioinformatics Corp. already offer antibody-based screening services to track hundreds of proteins involved in cell regulation. The immobilization of specific antibodies into a microarray format would provide a powerful platform for future proteomics studies.

The Prime Real Estate of the Human Genome

Presently, 90 per cent of the drugs sold in the $300 billion annual global pharmaceutical drug market treat the symptoms of disease and not the underlying causes. Less than 500 of the proteins encoded by the human genome were targeted by the biopharmaceutical industry for the discovery and development of these drugs. With the completion of the Human Genome Project, there are now about 5,000 interesting potential drug targets. However, the industry does not have the capacity to fully exploit all of these targets. It typically costs $500 million US over a 10-year period to identify a drug against a single target and successfully bring it to the marketplace. This high cost mostly reflects the enormous failure rate at the various stages of the drug discovery, development and testing processes. The most common and expensive setback associated with experimental drugs is the unforeseen side effects in late-stage human clinical trials. It has been estimated that the pharmaceutical industry misdirects approximately $5 billion per year alone on screening for compounds against targets that are inappropriate. Substantial reductions in these costs could be achieved with the identification of the more promising drug targets and markers of toxicity. This recognition of potential cost savings by the pharmaceutical industry will drive the proteomics market to $5.8 billion by 2005, according to Frost and Sullivan.

Most of the diseases associated with aging involve defects in cellular regulation proteins. Not surprisingly, most pharmacologically active compounds target signalling proteins. At least 21 per cent of the genes in the human genome encode regulatory proteins. About 5 per cent of these signalling proteins are hormones, growth factors and other cytokines involved in cell-to-cell communications, whereas the remainder transduce signals from these extracellular mediators inside of cells. It should be appreciated that the discovery of growth factors and other cytokines, and the ability to mass-produce recombinant forms of these proteins has



spurred the growth of the biotechnology industry for the last two decades. There remains high interest in novel genes that encode secreted proteins for their therapeutic potential; however, the intracellular signal transduction proteins have also captured the attention of the biopharmaceutical industry in recent years. The importance of signal transduction is underscored by the more than 100,000 researchers in university, industrial, hospital and government laboratories worldwide who are actively investigating cell regulation.

One of the most exciting groups of cell signalling proteins are enzymes called protein kinases. Two to three per cent of human genes encode at least 868 different protein kinases. These kinases catalyse the reversible phosphorylation of about a third of all proteins in cells, and in this manner, govern all cellular activities. Protein kinases play a major role in the precise orchestration of metabolism and other cellular processes in response to environmental cues. They are the principal components in the communication and control pathways that link receptors for extracellular mediators to appropriate effector responses throughout the cell. Cascades of sequentially activating kinases are further integrated into networks termed "kineomes," which operate as the functional equivalent of the nervous systems of cells. Deduction of the composition and architecture of kineomes in cells will be much more tractable than dealing with their larger proteomes.

Through protein kinases and other signalling proteins, cells monitor their environment and react appropriately to the benefit of the organism, even if this means that individual cells must commit suicide. Real havoc ensues when these proteins malfunction as a consequence of mutations in the genes that encode them. More than 400 human diseases have been linked to defects in protein kinase signalling pathways. For example, about half of the approximately 100 known cancer genes (i.e. oncogenes) encode protein kinases. Most of the remaining oncogenes specify upstream regulators or downstream targets of kinases. Often, inappropriately overproduced or hyperactivated kinases play pivotal roles in the proliferation and metastasis of tumour cells. Although protein kinases are present at 100- to 1,000-fold lower concentrations than metabolic pathway enzymes and structural proteins, they can be successfully tracked with specific antibody probes in tissue biopsy samples from patients. Therefore, it is feasible to identify those protein kinases that contribute to each cancer on a case-by-case basis. Armed with this knowledge, the cancer patient could be successfully treated with a specific inhibitor for the culprit protein kinase.

The pharmaceutical industry has also been drawn to protein kinases, because they represent the largest family of cell signalling enzymes. Due to their catalytic active sites, enzymes are in general well-suited for drug discovery. Their active sites lie within crevices on the surface of enzymes, and act to specifically dock substrates and catalyse their chemical transformations. They also represent the Achilles heel of the enzymes, since small molecule compounds that bind within this region can be disruptive. Such inhibitors can be readily uncovered in natural product and combinatorial chemistry libraries by high-throughput, *in vitro* kinase activity screens with robotic systems. Since the 3-D X-ray crystallographic structures of many kinases are continually being elucidated, it is also possible to use molecular modelling and rational drug design to develop specific protein kinase inhibitors. Such inhibitors will become valuable additions to our pharmaceutical armament in the war on disease.

Twenty-first century medicine will ultimately be based on molecular analyses to ensure that therapy is specifically tailored for each patient's unique situation. The expression and activity patterns for protein kinases and other signalling proteins could provide valuable information for disease diagnosis and prognosis. Moreover, they will be important targets for therapeutic intervention.

An old adage is that a doctor can learn much about a patient's ailments by paying careful attention to what the patient has to say about their condition. The new paradigm is that we can learn more about how to silence disease by listening directly to the cells of the body and eavesdropping on their molecular communication systems.

Dr. Steven Pelech is the founder and former president of Kinetek Pharmaceuticals, Inc. He is currently a full professor in the De-partment of Medicine at the University of British Columbia, and the founder and president of Kinexus Bioinformatics Corp.