strated SAGE's powers by using it to analyze the genes that are expressed in the human pancreas. They first extracted all the messenger RNA (mRNA), the products of active genes, from pancreatic tissue and copied it into complementary DNAs (cDNAs) that have the same sequences as the coding parts of the original pancreatic genes. During this step, the researchers also tagged the 3′- (far) ends of the cDNAs with a biotin molecule. After cutting the cDNAs into pieces with an enzyme called a restriction endonuclease, they were able to isolate the pieces containing each cDNAs' 3′-end with the aid of biotin-binding streptavidin beads.

Then, with a second restriction endonuclease, they clipped out a piece of DNA containing at least nine base pairs from those fragments. To complete the process, the researchers used the polymerase chain reaction (PCR) to create hundreds of copies of each short "SAGE tag," joined 30 to 50 different tags together in a single DNA molecule, and then cloned and sequenced these molecules. (The SAGE method includes a step that identifies any rogue tags that are preferentially amplified by PCR.) Because so many tags can be sequenced at one time, one technician, using a single state-of-the-art automated sequencer, can monitor the activity of 20,000 genes in as little as a month, says Kinzler. With the old methods, the same output would take years.

The technique shows not only which genes are active in a tissue, but also at what level. The Hopkins team reviewed the expression patterns of the pancreatic genes, by analyzing a total of 840 tags. (To improve sensitivity in actual experiments, thousands of tags would be screened.) Forty percent of these occurred as single copies, providing a baseline of the lowest level of detectable gene activity. But 77 tags occurred more than once, and, as predicted, the most abundant of those—one occurred 64 times and accounted for almost 8% of the RNA in the pancreas—encoded well-known pancreatic enzymes such as trypsinogen 2 and pancreatic lipase. "The number of times you see each tag is an index of the gene's expression," says Kinzler, who likens the SAGE tags to the ubiquitous supermarket bar code. The SAGE methodology, he says, "is the [genetic] equivalent of a cash register keeping track of the number of each item [a customer] buys."

"[SAGE] is very clever. You can get a large amount of information very rapidly," says Adams. "It offers the potential to small laboratories to do comparative studies and to take advantage of all the EST sequencing that's already been done." The pancreas experiment also demonstrated SAGE's power to help hunt down new genes. Several of the tags that occurred at high frequency in the

(continued on page 371)

MOLECULAR BIOLOGY

From Genome to Proteome: Looking at a Cell's Proteins

It sounds like Mission Impossible: Follow the changes taking place inside a cell—during embryonic development, for example—by identifying the thousands of different proteins the cell produces and watching how they ebb and flow over time. To a growing band of researchers, however, such a mission is becoming more realistic, thanks to the increasing volume of sequence data and to improved analytical techniques for proteins. Indeed, they believe such studies are a wave of the future for genome research and many areas of cellular biology, and have even coined a term for the emerging field: "proteome" research.

The growing interest in proteome projects comes as genome scientists are producing sequence data on more genes than they can put a function to. Some researchers are trying to find out what genes do by monitoring their expression patterns (see News story on p. 368 and Reports on pp. 467 and 484). But proteome researchers are approaching the task from the other end—looking at the proteins the genes produce. Although this approach is more complex and big obstacles remain, focusing directly on cellular proteins is "an important complement to studying DNA," says Jonathan Knowles, director of the Glaxo Institute for Molecular Biology in Geneva. "We will not [be able to] define disease mechanisms in molecular terms ... just using nucleic acids," he says.

"It's going to be critical to see what happens to proteins as they live out their lives in the cell."

—Leigh Anderson

That's partly because the level of gene expression is only one of the factors that determine how much of a protein is present in a cell. What's more, a gene sequence does not completely describe a protein's structure: After synthesis, proteins usually undergo "posttranslational modifications," such as addition of phosphate groups or removal of amino acids from the ends, and these changes can alter their activities.

But until recently, the advantages of focusing on proteins were overwhelmed by the difficulties. That is now changing fast, with the advent of powerful new methods of mass spectrometry that vastly simplify protein analysis, even on very small samples, and enable researchers to match them to their corresponding genes in the rapidly filling sequence databases. And, when protein studies "connect with what's known ... suddenly the whole approach has a lot more power," says yeast biologist Jim Garrels of Proteome Inc. in Beverly, Massachusetts.

Although the term "proteome" made its first appearance in the scientific literature only this year, in papers by Marc Wilkins and Keith Williams at Macquarie University in Sydney, Australia, the idea of analyzing the proteins expressed in different cell types has been around for nearly 2 decades. That was when Patrick O'Farrell, then at the University of Colorado, Boulder, developed two-dimensional gel electrophoresis, which made such an analysis possible. In this method, cell extracts are put onto a gel and the individual proteins separated first by charge and then by size. The result is a characteristic picture of 1000 to 3000 spots, each usually a single protein. In principle, these gel patterns reveal not only the amounts of proteins but also many of the posttranslational modifications they have undergone.

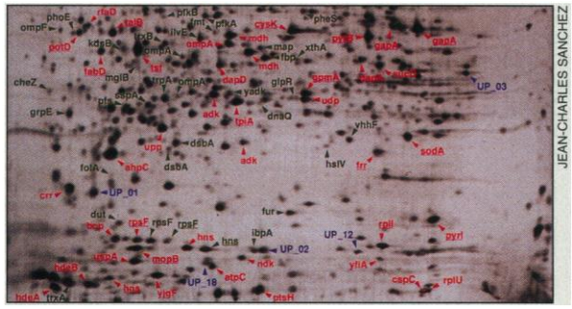
But by the mid-1980s, "people were getting discouraged," says Denis Hochstrasser, who heads the Clinical Chemistry Laboratory at the University of Geneva Hospital. Several problems had surfaced. Two-dimensional gels were not easily reproducible, making it nearly impossible to compare data from different labs. They revealed only the more abundant proteins in a cell—hardly a complete picture. Worst, says Hochstrasser, "it was very difficult to get information on the spots. It was like looking at the sky. You see the stars but don't know which ones they are."

Ironically, the worst problem is turning out to be the first one solved. The key to identifying a spot is to learn something about it that can be used to search protein databases. This first became possible in the 1980s after several groups developed methods to transfer spots from gels onto membranes so they could be analyzed. Amino acid composition, analysis of peptide fragments, and partial amino acid sequences can then often identify them as known proteins or products of predicted genes.

But it is the new mass spectrometry (MS)

methods—the "tool of the future," says protein expert Mike Waterfield of the Ludwig Institute for Cancer Research in London—that open the door to large-scale, automated spot identification. The two basic technologies—laser desorption MS, developed by Franz Hillenkamp at the University of Münster, Germany, and electrospray MS, devised by John Fenn at Yale University, both in the late 1980s—enable scientists to measure the molecular weights of large biomolecules easily and accurately, which in turn can help match proteins to those in the databases. Fenn's electrospray MS also generates partial sequences on purified peptides.

Still, it took several more years to bring the MS tools to the present stage, where they can begin tackling large-scale work. The crucial step—identifying proteins automatically and unambiguously—was just taken by Matthias Mann at the European Molecular Biology Laboratory in Heidelberg, Germany, and independently by Ruedi Aebersold and John Yates at the University of Washington, Seattle. Success was mostly a matter of getting more precise molecular weight measurements and developing more powerful database-scanning software, says Mann. His lab also devised methods for obtaining partial sequence data from unseparated peptides, making it possible to analyze mixtures of proteins at high sensitivity. Taken together, he says, all the improvements allow them to analyze some 10 proteins per day and identify the known ones with certainty.



MS marks the spots. Improved mass spectrometry helps identify the protein spots in 2D gels, like this one prepared from cells of the bacterium *E. coli.*

Besides being fast, MS methods have other advantages over conventional ways of analyzing proteins. They need far less protein—only one-tenth, or even one-hundredth as much, says Mann, and that allows identification of many of the less abundant spots on 2D gels. MS can also reveal many of the modifications to a protein's structure.

Researchers are now trying to scale up the MS methods so they can handle the thousands of proteins necessary for proteome work. For this step, the soon-to-be-completed yeast and *Escherichia coli* genome projects should provide the acid test. Once these genomes are fully sequenced, it should be possible, at least in theory, to link every 2D gel spot with the gene that encodes it, and with data on how the spot varies in different mutants and growth conditions. That, in turn, should help assign roles to many genes of unknown function—for example, by suggesting whether they belong to the same biochemical pathway or regulatory unit.

Indeed, a pilot project to test this approach is already under way, coordinated by yeast geneticist Piotr Slonimski at the Center of Molecular Genetics in Gif-sur-Yvette, France. The idea is to probe the function of newly discovered yeast genes and work out their interdependencies by disrupting them and then looking for changes on 2D gels.

So far, network members have analyzed 25 yeast strains, each with a disruption in a different gene. About half show changes in up to a few dozen protein spots, says Helian Boucherie at the Institute of Biochemistry and Cellular Genetics in Bordeaux, France, who is doing the 2D gel analysis together with Peter Mose Larsen at the University of Århus in Denmark. When these spots are identified—which should go quickly now that MS expert Mann is on board—they hope for a rich payoff in understanding the genes' functions. If the strategy works, the team will go on to a large study of 1000 genes.

Yeast biologists have high hopes for these efforts. "This is what people will want to do," says Garrels—"knock out their favorite gene and see its effect. It tells you a lot about the function of any gene you care to study."

But even though it's getting easier to identify protein spots on the 2D gels, not all the problems have been solved. First and foremost, says Jean-Charles Sanchez, who heads the 2D gel lab at the University of Geneva, is the difficulty in comparing gels from different labs—a frustrating problem now that there are over a dozen 2D gel databases on the World Wide Web. The difficulty is largely due to the gels themselves, Sanchez says, although he adds that things are improving as more labs adopt a method of fixing a pH gradient to the gel, eliminating some irreproducibility. Better software for comparing gel images also helps. Using the MELANIE II package developed by Geneva's Ron Appel, Sanchez could match spots on his *E. coli* gels to those in a database with valuable data on different mutant strains, which is maintained by Ruth VanBogelen of Parke-Davis Research and Frederick Neidhardt of the University of Michigan, both in Ann Arbor.

In parallel with these efforts, bioinformaticist Amos Bairoch and Appel have developed a system called ExPASy—the Expert Protein Analysis System—which links the Geneva 2D gel databases to the SWISS-PROT protein sequence data bank, and from there to many other databases and software tools. ExPASy, which receives some 300,000 queries monthly on the World Wide Web, also helps researchers guess a protein's function from its sequence. Bairoch is now working on predicting which sites in a protein are likely targets for posttranslational modifications—information that will help in interpreting 2D gels and understanding protein function. Last month, this work won Bairoch the SFr2 million ($1.7 million) Helmut Horten Foundation Award in biomedical research.

But for the Geneva team, the yeast and *E. coli* projects are only steppingstones to their real goal: using 2D gels as "clinical molecular scanners," says Hochstrasser. They pin their hopes on a growing body of data—collected in some half-dozen 2D gel databases worldwide—that link changes in gel patterns to specific diseases. Hochstrasser's unit is focusing on posttranslational modifications, which he calls "a major factor in disease." A case in point: Several cancer-causing oncogenes arise as a result of mutations in genes encoding enzymes that add phosphate groups to proteins. By studying such modifications closely, Hochstrasser predicts that researchers can develop diagnostic tools such as "marker" proteins for specific diseases or for assessing how advanced cancers are.

Scientists who like to think even bigger can look to the Large Scale Biology Corp. in Rockville, Maryland—which didn't get its name for nothing. Its mission, says President Leigh Anderson, is to harness the power of proteome studies for discovering new drugs and analyzing how they work. The idea is that drugs induce changes in protein patterns in the tissues of treated animals, and these changes give important clues about both the drug's desirable effects and its toxicity. For example, says Anderson, "if a drug inhibits cholesterol synthesis, we should see responses that identify it as active on cholesterol metabolism, and we do."

Anderson acknowledges there are still obstacles. Slow, laborious 2D gel methods remain a big problem, he notes, especially compared to the simple kits and procedures of DNA research. But he is optimistic that ongoing improvements and automation of 2D technology will make a big difference.

If these developments pan out, they will provide better ways to look at thousands of proteins on one gel, a feat even DNA researchers might envy. Then the key will be making sense of it all—a difficult task, but one Anderson says has to be faced. "It's going to be critical to see what happens to proteins as they live out their lives in the cell," he says. "There is no cheap, easy way around these problems. People are realizing that complexity has to be confronted."

–Patricia Kahn