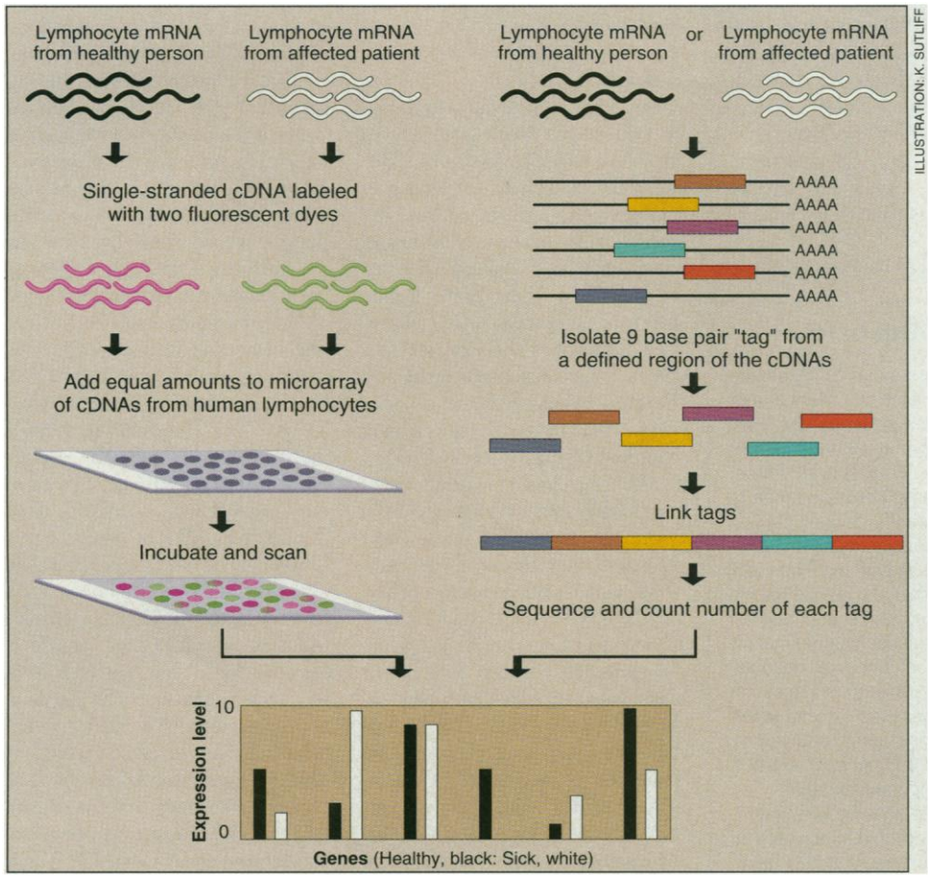
Entering the Postgenome Era

With the gene databases rapidly filling, the next step is to find out what the genes really do, either by measuring the activity of panels of genes—or by analyzing the cell's protein complement

Keeping an organism up and running is a feat of mind-boggling complexity. To transform a single fertilized egg cell into an adult human body and then keep that body alive and healthy, for instance, requires some 100,000 genes, each adjusting its activity to precise degrees and at precise times and locations. Thanks in part to the Human Genome Project, geneticists have done a remarkable job of assembling vast amounts of raw data about that intricate genetic machinery. Already unique sequences long enough to identify unequivocally well over half the genes are in the bag. Now, as some teams are gearing up for the final push to spell out all 3 billion base pairs of the human genome, others are poised to step into the postgenome era and find out how those genes act in concert to regulate the whole organism. Two techniques described in this week's issue of *Science* may help unveil the genes' multifarious roles (see pp. 467 and 484).

The best way to do that is to monitor the genes' fluctuating activities in different tissues at different stages of development, and in good health and in bad. But this is such a cumbersome job that, for the most part, until now it has only been possible to tackle one gene at a time. As a result, "we don't have a clue about the function of three fourths of the [identified genes]," says molecular biologist Mark Adams of The Institute for Genomic Research, a private research institution in Gaithersburg, Maryland, that has churned out huge numbers of partial gene sequences called expressed sequence tags (ESTs).

The new techniques, one developed by a research team led by molecular oncologists Kenneth Kinzler and Bert Vogelstein of Johns Hopkins University in Baltimore and the other by biochemist Patrick Brown of Stanford University in California, and his colleagues, aim to break up this logjam. They allow researchers to assess the activity patterns of thousands of genes simultaneously, generating in a matter of weeks information that might otherwise have taken years to gather.



Measuring gene activities. At left is a diagram of the microarray assay for gene expression; the SAGE technique is illustrated at right. Here, the procedures assess how gene expression differs in lymphocytes from a healthy person and those from a person fighting off an infection.

And geneticists are purportedly hungry for that knowledge: "I guarantee that in a year from now dozens of laboratories will have tried to adopt these approaches," says molecular biologist Jeffrey Trent of the National Center for Human Genome Research at the National Institutes of Health. "The value, the uniqueness [of the new approaches] is that they give a broad look at patterns of gene expression," he notes. This ability "is very, very important," adds molecular biologist Leroy Hood of the University of Washington, Seattle. "The future of biology is in the analysis of complex systems. And you can't look at [the expression of] one gene and understand how the system works." Molecular geneticist Mel Simon of Caltech in Pasadena agrees: "It represents a new era in the kinds of analysis we can do. It's a fantastic breakthrough for the study of the [genetic] mechanisms involved in development and in the control of differentiation."

Gene hunters are also likely to employ the new techniques in their frenzied quest to capture the crippled human genes that cause disease. Identifying genes whose activity is altered in diseased tissue will help researchers home in on those most likely to contain disease-causing mutations. And all these possibilities will not be lost on the pharmaceutical industry. Indeed, in anticipation of commercial interest, both teams have applied for patents on their techniques, which Hopkins has licensed to a biotech company called PharmaGenics Inc. of Allendale, New Jersey, and Stanford to a new company—Synteni Inc. of Palo Alto—founded last year by team member Dari Shalon.

The Hopkins team calls its technique Serial Analysis of Gene Expression, or SAGE. It relies on the fact that a sequence as short as nine base pairs is all it takes to identify 95% of human genes, provided the sequence is picked from the same place in all the genes surveyed. The amount of that sequence in a particular tissue is the measure of the gene's activity.

The Johns Hopkins workers demon-

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)

(continued from page 369)

pancreatic tissue had no counterparts in the gene databases. Using the SAGE tags, the Kinzler-Vogelstein team identified the clones for those genes in a pancreatic gene library, sequenced the clones, and added the sequences to the database.

Brown and his colleagues reach the same endpoint as the Johns Hopkins team—a detailed description of gene activities in a given tissue or cell—but they get there not by sequencing gene fragments, but by using a miniaturized system that makes use of the fact that similar DNA strands bind or hybridize to complementary sequences. "Suppose you're [from] one of those many labs that have been madly sequencing cDNAs," says Brown. "You have sequences of tens of thousands of cDNAs, but little information about where they are expressed, and you want to find out very quickly." With their new "microarray" assay, he says, it's feasible to monitor the activity of thousands of genes per day.

For its proof-of-principle experiment, the Brown team turned to a weed called *Arabidopsis thaliana*, the fruit fly of plant genetics. Using a tiny computer-controlled two-pronged fork that they had designed specifically for the task, the researchers dropped onto a microscope slide spots of solutions, each containing a different double-stranded cDNA from an *Arabidopsis* gene library. After fixing this array of spots to the slide with heat and chemicals, the Brown team added pooled cDNA prepared from the protein-coding mRNA extracted from *Arabidopsis* leaves and labeled with a dye that glows red, and cDNA prepared from the protein-coding mRNA extracted from *Arabidopsis* roots and labeled with a dye that glows green. The spots where cDNA from the plant leaves or roots bound to the corresponding cDNA in the microarray fluoresce red and green.

The fluorescence patterns, measured by a computerized scanner, indicate the relative levels of expression of the genes in the two tissues, and the absolute activity of each gene can be determined by comparing its fluorescence to standards of known amounts of cDNA. Expression of some of the genes was 100-fold or greater in one tissue than the other, Brown says, "and when we sequenced them, it was exactly what you would have expected." For example, he says, the genes for photosynthetic enzymes were turned on in the leaves, but not the roots. In this initial test case, the microarray contained only 45 cDNAs, but since then the team has created microarrays with 1800 yeast DNA sequences, increasing the information gleaned from each experiment 40-fold.

Currently, both of the new techniques are in the prototype stage, and "it remains to be seen which technique will be more amenable to widespread use," says Trent. Nonetheless, he says, either technique—or one of the similar techniques coming down the pipeline—will be instrumental to the success of efforts to study how coordinated changes in the activity of batteries of genes convert undifferentiated cells into cells with specific tasks and attributes, trigger the responses of differentiated cells to radiation, hormones, or other outside stimuli, and drive healthy cells through the abnormal changes that end in disease. Other teams are making progress in developing techniques that allow them to assess directly what proteins are present in cells, although this work is not quite as far along (see p. 369).

Indeed, the two gene-expression techniques are already being put through their paces in real-life research situations. Both groups are trying to use them to spot the differences between normal and cancer cells. "As soon as we knew that SAGE worked," says Kinzler, the Hopkins team started a project to compare the activity patterns of genes in normal cells lining the colon with those in colon cancers. Kinzler expects definitive results within 6 months. Meanwhile, Trent and his colleagues, in collaboration with the Brown team, are using the microarray technique to search for the tumor-suppressor genes that may prevent abnormal, but not yet cancerous, skin cells from taking the final steps to malignancy.

Brown and team member Ronald Davis, also of Stanford University, have even bigger plans afoot. Sometime in 1996, when the sequence of the whole genome of the yeast *Saccharomyces cerevisiae* is complete, they intend to mass-produce microarrays containing the organism's entire suite of about 6500 genes. By studying changes in gene expression under different conditions—for example, when the nutrient-starved yeast produces spores, says Brown, "we will be able to see when the cells call different genes into action, and from that information generate new hypotheses about what the genes do."

Although the potential of having all this new information at their fingertips promises to make a geneticist's life more interesting, it is likely to generate another information glut, warns Kinzler. "Instead of the Krebs cycle," he says, referring to the complicated graphic depiction of the cell's major energy-generating system that adorns many laboratory walls, "we are now going to have expression maps of 100,000 different genes. Good luck figuring that out!"

–Rachel Nowak

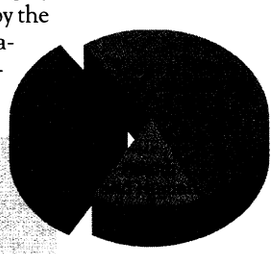
POLITICS

House Bundles 7 R&D Programs

The Senate is largely indifferent to it, the Administration is hostile, and it is unlikely to have any real effect on the 1996 budget. But last week the House passed a bill that, for the first time, lumps together spending authority for most nonmedical civilian science and technology programs. The 2-day debate leading up to the 248 to 161 vote provided a rare—and heated—discussion of federally sponsored research. In addition to putting science on center stage, it highlighted the widening gulf between the two parties on priorities for federal R&D.

Congressional action on civilian science and technology programs typically is scattered through the legislative calendar. This year, however, House Science Committee Chair Robert Walker (R–PA) championed a single bill that authorizes $21.5 billion—$3 billion less than current levels—for seven R&D agencies. "It's the first time we've focused attention on government R&D on the House floor," Walker told *Science*. "It makes more sense to look at science in a coordinated way." Congressional aides say the measure also demonstrates Walker's influence with the Republican leadership and advances his long-shot plan for a single Department of Science (*Science*, 31 March, p. 1900).

The funding figures in the omnibus authorization largely match the levels already approved by the House in a separate set of appropriations bills.



One piece. The House reauthorization of R&D programs covers almost a third of the federal science budget.

Those bills determine 1996 budgets for agencies including the National Science Foundation, the National Aeronautics and Space Administration (NASA), the Environmental Protection Agency (EPA), the Department of Energy (DOE), and parts of the Commerce Department. High-ranking Democrats including Vice President Al Gore and Representative George Brown (CA) used the debate to lambaste Republican plans to cancel industrial research programs like the Commerce Department's Advanced Technology