In science, the important thing is to modify and change one's ideas as science advances.
Herbert Spencer (1820–1903)

GENOMICS

Encyclopedia of DNA

The completion of the Human Genome Project in April 2003 was a landmark accomplishment, but much remains to be learned before scientists fully understand the true functionality of the DNA sequences in our genetic matter. To that end, the National Human Genome Research Institute (NHGRI) has initiated a variety of research projects to better understand the sequence. One of the most intriguing and potentially far-reaching of these efforts is the Encyclopedia of DNA Elements, or ENCODE, project, which aspires to create a complete catalog of all the functional elements of the human genome.

"The ultimate goal of the ENCODE project is to create a reference work that will help researchers fully utilize the human sequence to gain a deeper understanding of human biology, as well as to develop new strategies for preventing and treating disease," says Elise A. Feingold, one of the NHGRI program directors in charge of the ENCODE project.

The goal of the Human Genome Project was simply to sequence the human genome; no distinctions were made between protein coding and noncoding regions. The ENCODE project is intended to pick up where the Human Genome Project left off, by providing answers about the roles that are played by the different genetic elements in the sequence. In addition to studying the human genome, the ENCODE project is also looking at genomic sequences from a variety of animals to provide multispecies comparisons. This will help to identify conserved sequences, which are thought to be strong indicators of functionally important regions in the human genome.

Encyclopedia genomica. The goal of the ENCODE project of the National Human Genome Research Institute is to create a complete catalog of all of the functional elements of the human genome.

NHGRI launched ENCODE last year with the first round of a total $36 million in grants that will be awarded over a three-year period. The first round of awards went to 14 recipients in the United States and abroad. In addition to the grantees, several other academic and scientific groups are providing specific technical expertise, such as database coordination, to assist the project.

According to Feingold, the grantees and other contributors are working as a consortium to analyze about 1% of the genome. Their goal is to determine the most effective set of methodologies, which will then be applied to the remaining 99%.

One of the grantees is Anindya Dutta, a professor of biochemistry and molecular genetics at the University of Virginia. He and his colleagues are studying ways to map replication elements on human chromosomes. The completion of the Human Genome Project created what Dutta considers an obvious opportunity to embark on such a replication study.

"There are very few origins of replication mapped in human cells—five to ten if you're generous, but I would say three or four," Dutta says, referring to genetic elements that are necessary to initiate DNA synthesis. "It was pretty clear when the sequence of the human genome came out that this is a great tool for us to find hundreds of origins and how they are controlled by chromatin structure, gene density, promoter activity, and, of course, sequence."

Following the model established by the Human Genome Project, NHGRI is calling for the data generated by the ENCODE project to be stored in databases and made freely available to the scientific community. The Center for Biomolecular Science and Engineering at the University of California, Santa Cruz—which is one of the institutions involved in providing support work for the ENCODE project and which also developed the computer programs that ran the sequencing of the human genome—is in charge of maintaining the database for sequence-related ENCODE data. In June 2004 the center added an ENCODE page (**http://genome.ucsc.edu/encode/**) to its existing genome browser, which gets 5,000 visits a day.

It's not yet clear what might happen further with the data after the initial pilot project ends in 2006. Feingold says that when the first period ends, "we will evaluate what we have learned and determine the best path for moving forward." **–Richard Dahl**

Medio Images, Matt Ray/EHP

SYSTEMS BIOLOGY

BAC to the Future

In a step forward into the future of gene expression research, molecular biologists and neurobiologists have joined forces to map the genes that control brain structure and neural circuits. The project, called the Gene Expression Nervous System Atlas, or GENSAT, maps mouse genes that are also present in the human genome as expressed in the central nervous system. According to project director Nathaniel Heintz, head of the Laboratory of Molecular Biology at The Rockefeller University, New York, GENSAT means that researchers studying degenerative conditions such as Parkinson disease can now have access to gene expression within the brain without having to do their own molecular genetics from scratch. Some unexpected insights have already come to light, giving neuroscientists new places to search for the roots of cognitive impairment.

GENSAT is sponsored by the National Institute of Neurological Disorders and Stroke (NINDS) and is based at The Rockefeller University, although prescreening of candidate genes is conducted by Tom Curran, chair of developmental neurobiology at St. Jude Children's Research Hospital in Memphis, Tennessee. *In situ* hybridization is used to screen thousands of candidates to find genes that are active in the central nervous system. Of these, an advisory committee selects 250 genes each year for in-depth analysis by the Rockefeller group. Says Heintz, "Having an advisory committee means this research is done with consensus from many parts of the neuroscience community."

Information gathered through the project is posted in a public database at **http://www.gensat.org/**. Started in 2003, the GENSAT database contains detailed information for 300 genes and is updated regularly. With the goal of analyzing 250 genes yearly, the project is planned to run for at least several more years, according to Heintz.

The main tools of GENSAT are bacterial artificial chromosomes (BACs), which are simple loops of bacterial DNA that reproduce outside the cell. BACs adeptly incorporate chunks of introduced DNA from other species, which are preserved and duplicated along with the BACs. The Human Genome Project relied on BACs to help map the human genome.

To measure gene activity and patterns of gene expression in the brain, the GENSAT team inserts a reporter gene for enhanced green fluorescent protein into each BAC. When genes are active, the enhanced green fluorescent protein glows bright green. Each BAC is then inserted into eggs harvested from mice, and the eggs are implanted into foster mothers.

The resulting offspring carry the BAC throughout their bodies in all of the cells that express the corresponding gene. Groups of mice are sacrificed at three time points—two of which correspond to critical periods of human central nervous system development—and their brains and spinal cords are analyzed. Mapping gene activity at three different points reveals how the cells migrate and interact.

Mr. Greengenes. The GENSAT project uses enhanced green fluorescent protein to map mouse genes that are also present in humans and expressed in the central nervous system.

Photodisc, Matt Ray/EHP

The first samples are taken when the mouse embryos are 15 days old, which corresponds to the sixth to seventh month of human gestation. "During this period the cortex forms, and defects that lead to malformations occur," explains project codirector Mary Beth Hatten, head of the Laboratory of Developmental Neurobiology at Rockefeller. The second time point, at 7 days after birth, is equivalent to 6–8 months of age in humans. At this age, interconnections form in the cerebellum, which controls movement, and in the hippocampus, which controls short-term memory. The final observations are made on adult mouse brains at age 7 months, which are similar to those of 30-year-old humans.

Findings published in the 30 October 2003 issue of *Nature* reveal some of the surprising connections the GENSAT project is uncovering. For example, people with DiGeorge syndrome, a congenital condition marked by heart defects and learning disorders, lack a gene called *Gscl*. Heintz, Hatten, and other GENSAT researchers discovered that *Gscl* is produced by neurons in the interpeduncular nucleus, the brain region that also regulates rapid-eye-movement sleep. Another finding reported in this paper relates to the striatum, which degenerates in patients with Parkinson disease. In end-stage Parkinson disease, up to 95% of so-called spiny neurons are lost. Until recently, the striatum had been the only place where spiny neurons were found, says Hatten. Yet, the BAC method identified vectors that can be used to separately analyze spiny neurons that project to the substantia nigra and the globus pallidus.

The GENSAT methods can also monitor the effects of environmental toxicants, such as lead, on brain development. "You can expose the BAC mice to any environmental condition you want, to see how the migration and maturation of neurons changes," says Hatten.

"The tools and mouse lines provided by this project allow the neuroscience community to perform detailed studies of each gene," says Laura Mamounas, the GENSAT project officer at the NINDS. "GENSAT also may serve as a model for future gene expression projects."

Indeed, BAC mice can be used to screen gene activity in other organs. The BAC mice are made available to other researchers who are interested in performing systematic studies of gene expression. Scientists in other specialties are "just starting to bootstrap our efforts to get their particular information," says Heintz. **–Carol Potera**

BIOINFORMATICS

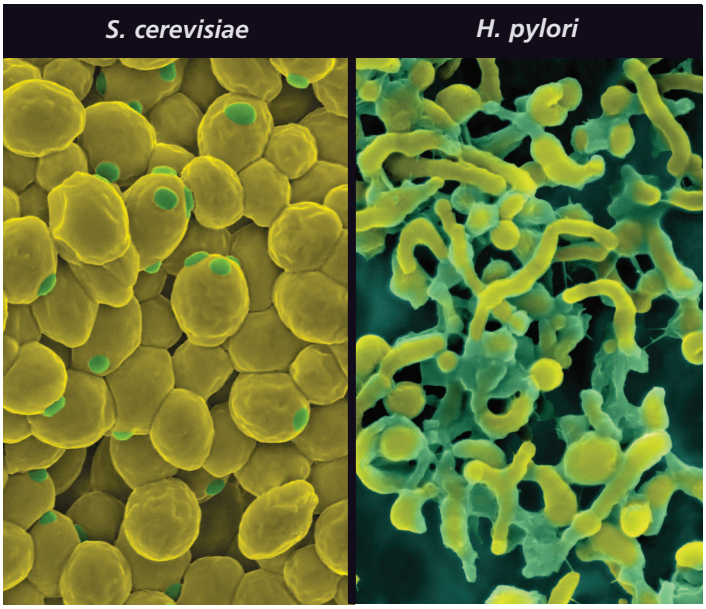
The Path to Species Comparison

Systems biology relies on integrating genetic, proteomics, and metabolic data, and on understanding interdependent cellular and intercellular events that are constantly in flux. To accomplish this feat, researchers have relied on DNA and protein sequence databases and high-throughput expression analysis techniques such as microarrays to produce ever-growing libraries of expression data. DNA and protein sequences can be quickly compared using software tools such as BLAST (Basic Local Alignment Search Tool), a program that identifies similar genes in different organisms. Now scientists are applying this computational approach to protein interaction networks, which are the means by which proteins communicate.

"As we move from a focus on sequences to one on networks, we need a tool similar to BLAST," says Trey Ideker, an assistant professor of bioengineering at the University of California, San Diego. The software program PathBLAST was developed to fill this need by a group consisting of researchers from Ideker's lab and the lab of Brent Stockwell, now an assistant professor of biological sciences at Columbia University. At the time, both Ideker and Stockwell were fellows at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, and worked on the program development with Richard Karp, a professor of bioengineering and mathematics at the University of California, Berkeley, known for his work in combinatorial algorithms and bioinformatics.

The PathBLAST program rapidly compares protein interaction networks across two different organisms using fast-executing algorithms. The program searches for high-scoring alignments involving one path from each network. The proteins of the first path are paired with putative homologs—or proteins presumed to have a common origin and function—from the other species and occurring in the same order in the second path. PathBLAST is built as a plug-in to Cytoscape, a widely used software platform. Scientists use Cytoscape to visualize molecular interaction networks and integrate these interactions with gene expression profiles and other data.

"The important stuff in biology is revealed by comparing things," says Ideker. "By comparing protein interaction networks of two different species or even within species, we can identify pathways and complexes that have been conserved over evolution." These evolutionarily conserved pathways allow interpretation of the network of a poorly understood organism based on its similarity to that of a well-known species. This comparison could provide a model of signaling and regulatory pathways that are related to a response to an environmental toxicant. It could also help target drugs to pathways that are present in a pathogenic organism but absent from its human host. Such a model could furthermore help identify drugs that would repair damaged pathways or even cause new ones to be formed.



Conservation comparison. PathBLAST software allows researchers to identify protein interaction networks that are conserved across multiple species.

The PathBLAST development group published a paper in the 30 September 2003 issue of *Proceedings of the National Academy of Sciences* in which they identified the conserved pathways within the yeast *Saccharomyces cerevisiae* and the bacterium *Helicobacter pylori*. For example, the authors found that one pathway that was critical in catalyzing DNA replication and another in protein degradation were conserved in both organisms as a single network. Within seconds, the program had determined that the bacterium contained 1,465 interactions among 732 proteins, and the yeast contained 14,489 interactions among 4,688 proteins.

This report proved that the method works for matching conserved networks from among all the networks in two species, according to software engineer Brian Kelley, a member of Stockwell's lab. Kelley says, "The next step is to prove the software in a novel application where you start with a given disease network and see if it is conserved in other species. Once you prove this utility, then the use of PathBLAST will skyrocket." Kelley adds that research into the mTOR cell growth–triggering protein pathway may prove to be that application. This pathway is composed of a complex of proteins that respond to nutrient cues; understanding it will clarify the role that nutrients and metabolism play in disease.

Other researchers have taken a complementary approach by comparing what's known about a disease to a known network. At Beyond Genomics in Waltham, Massachusetts, researchers measure quantitative differences between transcripts, proteins, and metabolites across a given disease model, determine correlations within the data set, and then compare the experimentally derived network with a known biological network or pathway.

"As the protein interaction databases become more heavily populated with interactions among higher eukaryotes, PathBLAST and related approaches will start to shine as they can help elucidate the set of core biological networks for a given genome," says Tom Plasterer, the principal scientist for bioinformatics at Beyond Genomics. "These networks—when coupled with a tightly defined experimental context—will be invaluable in understanding mechanisms of disease, where one expects compensatory and subtly differing biological networks to emerge."

The PathBLAST website is hosted by the Whitehead Institute and available at **http://www.pathblast.org/**; it will soon be mirrored at the San Diego Supercomputer Center at the University of California, San Diego. And as for whether industry will embrace PathBLAST, Ideker says, "It's still early. Speculating too far about these technologies is like asking industry in 1980, 'Is genome sequencing going to revolutionize your drug discovery pipeline?' Even in 2004, the verdict is still out on that one!" **–W. Conard Holton**

Dennis Kunkel Microscopy

PHARMACOGENOMICS

Activating Cancer Drug Discovery

Normal cells that transform into cancer cells undergo various metabolic changes, including shifts in activities of enzymes that mediate macromolecule synthesis and growth-signaling pathways. Proteomics technology now provides an elegant way to identify the enzymes that are active in processes linked with tumor progression. As demonstrated by a recent study conducted at The Burnham Institute's Cancer Center in La Jolla, California, this approach is beginning to unveil some novel high-efficacy targets for cancer control and treatment.

In work published 15 March 2004 in *Cancer Research*, the Burnham team used a novel proteomics screen based on probes that bind to the active site of the enzyme target. By competing with such probes for the active site, one can simultaneously identify protein targets and screen for their inhibitors. Activity-based proteomics screening is fast emerging as "the wave of the future," says coauthor Steven J. Kridel, a postdoctoral fellow at the time of the research and now an assistant professor of cancer biology at Wake Forest University of Winston-Salem, North Carolina—it enables the generation of hypotheses that can lead to meaningful clinical applications. The chemical strategy for activity-based proteomics was pioneered in the laboratories of cell biologist Ben Cravatt of The Scripps Research Institute and pathologist Matthew Bogyo of Stanford University. Kridel and colleague Jeffrey Smith, associate scientific director for technology at The Burnham Institute, are among the first to use the approach to identify a therapeutic lead.

The activity-based strategy may mark a major improvement over the usual proteomics approaches, which are based on the relative abundance of a particular protein target. "Measuring the abundance of a protein only provides a static picture of a potential target enzyme," says Kridel. "There are several levels of regulation between protein abundance and protein activity. With activity-based proteomics, you also can tell whether there is a specific physiologic state that turns off the enzyme's activity and whether an inhibitor of that particular enzyme exists."

Kridel and Smith applied the activity-based strategy to identify proteins that exhibit different activities in cancer cells as compared to normal cells. They screened a group of enzymes known as serine hydrolases by measuring the activity levels of these enzymes in normal prostate epithelial cells and in three standard prostate cancer cell lines. They found that serine hydrolase expression was generally similar among all cell lines, with two key exceptions: one of the hydrolases was active in normal prostate cells but virtually inactive in all the tumor cells, while another was expressed in all of the tumor lines but absent in the normal cells. The latter enzyme was shown to be fatty acid synthase (FAS), which had earlier been strongly linked to tumor progression, making it an attractive therapeutic target.

Dual-purpose drug? A novel activity-based proteomics screen of the weight-loss drug orlistat revealed its surprising potential as a cancer treatment.

Having identified their molecular target of choice, the investigators then screened possible inhibitor drugs, hoping to find unforeseen side benefits in drugs already approved for human use. "Our goal from the outset was to find an anticancer drug that might not have been considered before," says Kridel. "We wanted a drug that inhibits a protein that is only expressed in cancer cells, not in normal cells, in part because we believed this would minimize toxic side effects." Among the many agents reviewed was the anti-obesity drug orlistat (trade name Xenical). Kridel says orlistat had not previously been shown to inhibit FAS, and FAS inhibition is not believed to be relevant to orlistat's mode of action in weight loss.

In cell culture studies, the Burnham team found that orlistat inhibited proliferation and induced apoptosis in at least two lines of prostate cancer cells. The antiproliferative effects were reversed by the addition of palmitate, the precursor for the majority of nonessential fatty acids, which cancer cells use primarily for energy and growth. This strongly implicated FAS inhibition, as FAS is the only eukaryotic enzyme capable of synthesizing palmitate. In rodent experiments, orlistat blocked tumor growth significantly, and the animals showed no outward signs of toxicity or adverse changes in blood chemistry.

By revealing some of the unanticipated effects of a drug, activity-based proteomics could markedly reduce the cost of drug development. "Orlistat just happens to be an approved drug with relatively minor toxicity that could be utilized quickly once its effectiveness in human prostate cancer is validated," says Massimo Loda, an associate professor of pathology at Harvard Medical School and the Dana Farber Cancer Institute in Boston, Massachusetts. "The implications of this study are dual: this activity-based proteomics approach can now be applied to the screening of diverse families of enzymes that sustain tumor survival, and it may reveal unsuspected activity of known drugs utilized in diseases other than cancer."

Such research may eventually pave the way for construction of a proteomics profile of susceptibility to cancer progression. "If a man presents with prostate cancer and has a biopsy, it is entirely possible that the proteomics screening approach can be used to assess whether his tumor has upregulated FAS," Smith says. "If it does, you can then prescribe a specific treatment regimen: to reduce dietary fat and block FAS activity using orlistat. This is moving toward personalized medicine."

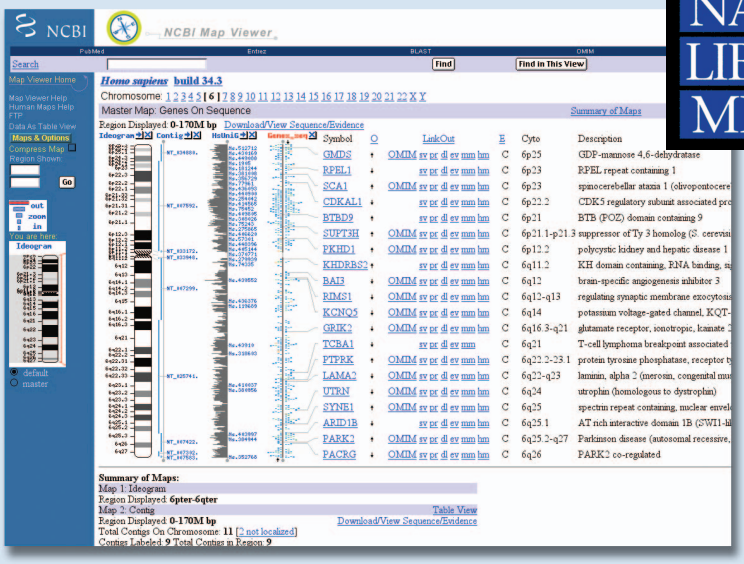
Smith believes a low-fat diet could reinforce orlistat's cancer-fighting effects in humans. "We know that tumor cells have a unique requirement for fat," he says. "If you restrict dietary fat and knock out the tumor's ability to synthesize its own fat from carbohydrates, then the antitumor effect should be even greater." **–M. Nathaniel Mead**

F. Hoffmann-La Roche

txgnet National Center for Biotechnology Information

The National Center for Biotechnology Information (NCBI) of the National Library of Medicine was created as a means of developing new information technologies to help advance the field of molecular biology. Composed of a group of scientists from disciplines including mathematics, research medicine, and structural biology, the center's programs and activities are centered on both basic and applied research in computational molecular biology. Since the center's inception in 1988, NCBI scientists have developed important novel algorithms and research approaches in the fields of computational biology and gene sequencing, among others. Today, researchers and others can find an impressive array of "omics" resources and tools collected into one central resource on the NCBI website at http://www.ncbi.nlm.nih.gov/.

General information on the center's work can be found within the top center section of the homepage as well as by clicking the About NCBI link on the left side of the homepage. A large portion of the NCBI website is devoted to the number of free, publicly accessible databases and software programs that the center hosts. Included among the databases on offer are GenBank, the Online Mendelian Inheritance in Man, the Cancer Genome Anatomy Project, and numerous organism-specific genome databases. The site classifies these databases as literature search databases, molecular databases, or genomic biology databases to make finding them easier; GenBank, NIH's annotated genetic sequence database, stands under its own heading. The Molecular Databases page—which is based on Entrez, the integrated search and retrieval system developed by the NCBI—has more than 25 additional database classifications such as nucleotide, protein, and expression. Entrez can also be quickly accessed through a pull-down menu at the top of the homepage.



The Genomic Biology section of the site provides a brief overview of this relatively new area of science. Visitors to this section will find a wealth of human genome resources, including a map viewer that can be used to browse the human genome. By selecting the Human Genome Resources page, users can access PDF background documents on the databases available on the site, two of which provide instructions on how to use the map viewer to explore genomes.

The main portion of this Genomic Biology area is divided into sections on genes and human health and contains links to Online Mendelian Inheritance in Man, RefSeq, dbSNP, and Gene Database. Visitors can also access BLAST for comparing genomic sequences and gene products, and a centralized registry of genomic clones, end sequences, mapping data, and distributor information. Other tools are available as well, and are classified under Maps and Markers, Transcribed Sequences, Cytogenetics, and Comparative Genomics.

The right-hand toolbar for the Genomic Biology section contains resources for 20 specific organisms. Selecting any one of these organisms takes the user to a guide outlining the many different search tools and other resources offered by the NCBI and other groups. These resources include gene maps, sequences, and annotation projects.

Educators and others using the site can access an online guide to GenBank and NCBI resources through the Education link on the homepage. Also on the Education page are tutorials for BLAST, Entrez, and other tools available on the site, as well as access to NCBI newsletters and map viewer exercises. Also within this section is an online science primer developed by the NCBI as a way to familiarize nonspecialists with terms and concepts such as "genome mapping," "expressed sequence tags," and "phylogenetics." **–Erin E. Dooley**

Top to bottom: National Library of Medicine; NCBI