> **Man is only man at the surface. Remove the skin, dissect, and immediately you come to machinery.**
> Paul Valéry (1871–1945)

eTag, You're It!

Researchers use microarrays to cull from the 30,000 genes identified by the Human Genome Project those that interact with environmental toxicants. Generally, anywhere from 50 to 200 genes stand out and are further scrutinized by time-consuming bioinformatics algorithms. Now a new tool called the eTag Assay System rapidly identifies not only genes but also related proteins without the need for complex sample preparation and follow-up bioinformatics. eTags further allow scientists to look at multiple concentrations of an environmental toxicant to determine at what dose it becomes toxic, says Sharat Singh, inventor of the eTag chemistry and senior vice president of technology and assay development at Aclara Biosciences in Mountain View, California.

Short for "electrophoretic tags," eTags are small fluorescent molecules linked to nucleic acids or antibodies and are designed to bind one specific messenger RNA or protein, respectively. After the eTag binds its target, a special proprietary enzyme cleaves the bound eTag from the target. The signal generated from the released eTag, called a "reporter," is proportional to the amount of target messenger RNA or protein in the sample.

The eTag reporters are identified by capillary electrophoresis using a capillary-based DNA sequencer. The unique charge-to-mass ratio of each eTag reporter—that is, its electrical charge divided by its molecular weight—makes it show up as a specific peak on the capillary electrophoresis readout.

Unlike microarrays, samples used with the eTag system do not require expensive preparation steps, such as amplification by polymerase chain reaction. Nanogram quantities of cells, tissue, blood, or bodily fluids are mixed directly with eTag reagents in standard 96-well laboratory microplates.

The ability to look at genes and proteins at the same time yields multiple advantages for researchers. Running only one assay means less sample is needed, and time and equipment costs are lower. This ability also increases the specificity of the assay by confirming that certain genes express certain proteins. Finally, because some genes are regulated at the transcriptional level and others at the translational level, measuring genes and proteins gives a better understanding of the regulation of gene expression.

Because of the costs related to running microarrays, researchers often obtain data for just one dose of a given compound. Yet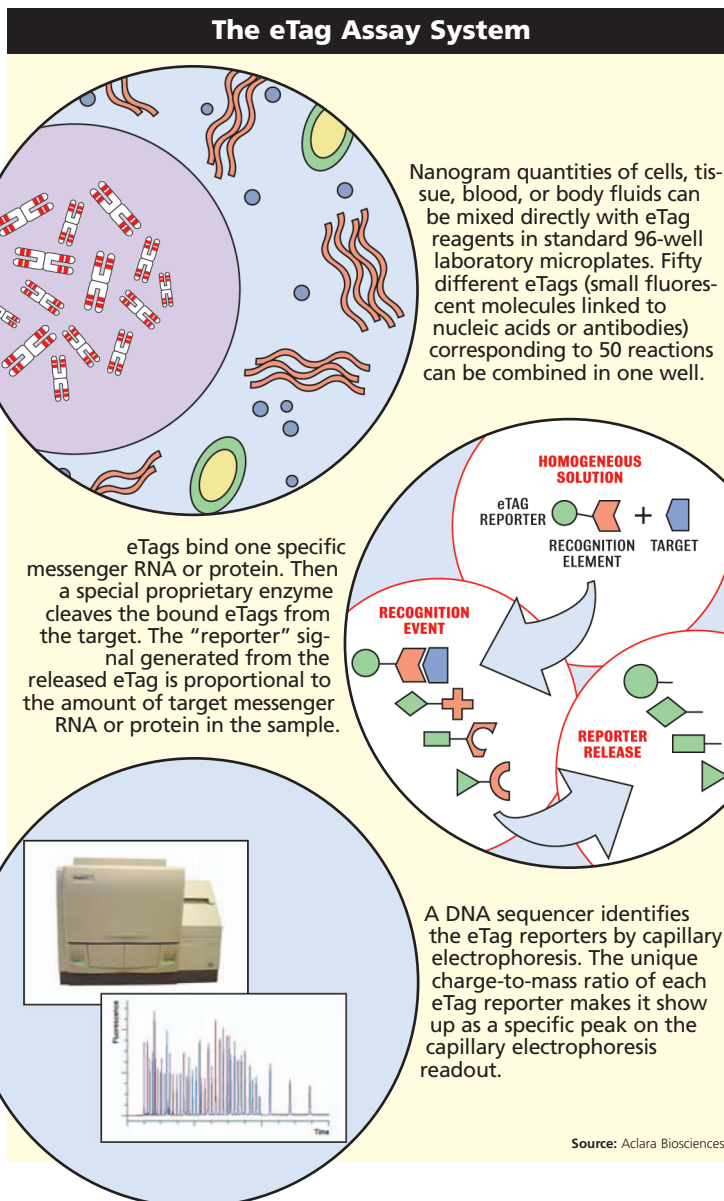, "it would be enlightening to look at one hundred compounds at ten concentrations and ten time points to fully understand toxic reactions," says Singh. Such a comprehensive experiment would require hundreds of microarrays (at about $300 apiece) and a few months to complete. In contrast, eTags could accomplish the task in days, at a cost of approximately $200.

Fifty different eTags corresponding to 50 reactions can be combined in one well. Aclara offers eTag assay kits to detect certain pathways, such as cytokines and cytochrome P450s associated with toxicity, or they can customize sets of eTags to suit a researcher's needs.

At Stanford University School of Medicine, rheumatologist Paul J. Utz received funding from the National Heart, Lung, and Blood Institute to use eTags to identify cytokines and chemokines as biomarkers for systemic lupus erythematosus and rheumatoid arthritis. In mouse models of rheumatoid arthritis, environmental toxicants such as mercury trigger the production of characteristic autoantibodies, which may also be pathogenic. In past experiments, Utz has relied on homemade protein arrays to detect biomarkers such as autoantibodies and cytokines. However, these protein arrays "are a pain to print, they do not allow us to study all proteins, and the sensitivity is not as good as what we get with eTags," he says. Moreover, Utz adds, eTags "are easier to use and have great flexibility."

Because eTags are small molecules that do not interfere with biological processes, the opportunities for their use are broad, says Singh, ranging from detecting biowarfare agents such as anthrax and smallpox, to discovering biomarkers of cancer. Aclara scientists have created 600 eTags for genes and proteins relevant to toxicology and human illnesses, providing significant coverage of key targets. The ultimate goal is to create 1,000 eTags by the end of 2004, says Singh. **–Carol Potera**

The eTag Assay System

Nanogram quantities of cells, tissue, blood, or body fluids can be mixed directly with eTag reagents in standard 96-well laboratory microplates. Fifty different eTags (small fluorescent molecules linked to nucleic acids or antibodies) corresponding to 50 reactions can be combined in one well.

eTags bind one specific messenger RNA or protein. Then a special proprietary enzyme cleaves the bound eTags from the target. The "reporter" signal generated from the released eTag is proportional to the amount of target messenger RNA or protein in the sample.

A DNA sequencer identifies the eTag reporters by capillary electrophoresis. The unique charge-to-mass ratio of each eTag reporter makes it show up as a specific peak on the capillary electrophoresis readout.

Source: Aclara Biosciences

Silencing of Mutant Genes with RNAi

RNA works hard at the business of expressing genetic information. It carries instructions from DNA in the cell nucleus into the cytoplasm, where basic housekeeping functions are carried out and proteins manufactured. When messenger RNA arrives in the cytoplasm, it binds to the ribosomes and guides the assembly of amino acids into proteins. Now advances in genomics have led to the discovery that, in addition to its transport and manufacturing roles, RNA can silence

siRNA therapies for heritable diseases such as Machado-Joseph disease (MJD) and other dementias in which defective proteins clump together and impair brain and nervous system function. In MJD, for example, a mutation of the *MJD1* gene produces multiple copies of the amino acid glutamine, which makes a protein that is toxic to cells. And in frontotemporal dementia with parkinsonism, a mutation of the *tau* gene produces defective tau protein consisting of the tangled filaments that lead to cell death in dementia disorders such as Alzheimer disease.

Miller and colleagues conducted experiments using siRNAs to silence genes of these two diseases. In the experiments, siRNAs were produced in a test

down disease-causing proteins coded by dominant genes offers hope for new and effective treatment of diseases that other genetic engineering strategies—such as gene replacement therapy—cannot address.

According to Hui Zhang, an associate professor of genetics at the Yale University School of Medicine, the paper is highly significant in that it provides a conceptual as well as potential way of treating diseases that contain point mutations through interfering with normal cellular function. "The authors provided detailed analysis for allele-specific silencing of the disease genes using the siRNA approach, which may provide a therapeutic answer to many mutation-based diseases," says

Cross-species application. Because RNA interference is evolutionarily conserved, it may be useful for silencing genes in various animal models of disease.

gene expression by a process called RNA interference (RNAi). RNAi provides a new tool for investigating gene function that also has potential for developing novel clinical treatments for certain previously untreatable diseases.

RNAi is an evolutionarily conserved cellular mechanism in worms, plants, and animals. In the RNAi pathway, long pieces of double-stranded RNA are cut into smaller pieces by the "dicer" enzyme to form small interfering RNAs (siRNAs) that are about 21 nucleotides long. These siRNAs bind with other molecules to form the RNA-induced silencing complex, which allows the siRNAs to target specific messenger RNAs to block production of protein.

In the 10 June 2003 issue of *Proceedings of the National Academy of Sciences*, graduate research assistant Victor M. Miller, associate neurology professor Henry L. Paulson, and colleagues at the University of Iowa report results that will faciliate development of

tube, then added to cells to see if they inhibited or suppressed expression of the targeted gene. Sequences for siRNAs that worked were inserted into a plasmid for production of short hairpin RNA, which the cell converts into siRNA, using the specific sequence for each different siRNA the team wanted to clone.

Targeting a single-nucleotide difference between the mutant and healthy *MJD1* gene enabled the scientists to almost completely eliminate production of the defective protein in a human cell model system. The experiments using siRNAs to knock down expression of the *tau* gene also succeeded in reducing production of the protein that causes disease.

Paulson says, "Because the human genome is full of polymorphisms, including countless single-nucleotide polymorphisms, it is conceivable that some of these might be associated with diseases, or traits, that allow them to be the 'hook' by which selective targeting [of a gene] can occur." The ability of siRNAs to knock

Zhang. "Their conclusion . . . is consistent with our understanding of the siRNA targeting mechanism reported by many others."

The next big hurdle will be to test siRNAs in an animal model. In a collaboration between Paulson and Beverly Davidson, the Roy J. Carver Associate Professor of Internal Medicine at the University of Iowa, the team is now employing a viral vector to introduce siRNAs into mouse models of human neurodegenerative disease.

Although human therapy is the ultimate goal, there are a number of challenges ahead before this new technology will be available. These include possible rapid degradation of siRNA in the cell, nonspecific effects on gene expression, and the need for high specificity to prevent unwanted side effects of treatment, such as possible interference with other proteins or biological pathways, or elicitation of immune system responses.
–Mary W. Eubanks

ADDLs: A New Explanation for Alzheimer Disease?

Clumps of large, sticky proteins forming senile plaques have been observed in the brains of people with Alzheimer disease ever since neurologist Alois Alzheimer first described the disorder nearly a century ago. However, only in the last few years have researchers begun to understand how the primary component of these proteins, a compound known as β-amyloid (Aβ), disables and kills brain cells. Recent work shows that other forms of Aβ known as Aβ-derived diffusible ligands, or ADDLs, may, along with senile plaques, play a key role in the pathogenicity of Alzheimer disease. True to the promise of environmental medicine, these findings could contribute to better methods for diagnosing Alzheimer disease, as well as new therapies to halt its progress.

Recent research by postdoctoral fellow Yuesong Gong, neurobiology and physiology professor William Klein, and other researchers at Northwestern University, published in the 2 September 2003 *Proceedings of the National Academy of Sciences*, shows that ADDLs bind to synapses, connection points that allow the exchange of signals between neurons. There they disrupt the signaling needed to form memories, says Klein, a member of the Northwestern Cognitive Neurology and Alzheimer's Disease Center.

Aβ is formed by the breakdown of the amyloid precursor protein, a molecule of unknown function that is embedded in the membrane of some cells. It is not unusual for cells to make Aβ, but in persons with Alzheimer disease, for reasons that are not yet known, production of Aβ either dramatically increases or its breakdown decreases. The Northwestern team found up to 70 times as many ADDLs in brain tissue from persons with Alzheimer disease compared to brain tissue from persons without the disease.

Once formed, Aβ molecules can fold on themselves and bind to each other. Early in this process, they link in globules of 12–24 molecules to form ADDLs. Eventually, in a process described in the 18 April 2003 issue of *Science* by a team including Carl Cotman,

director of the University of California, Irvine, Institute for Brain Aging and Dementia, ADDLs can bind to cells until they begin to appear as diffuse plaques.

In the early 1990s, researchers thought that plaques were the most toxic form of Aβ. But recent research indicates that ADDLs may be more damaging. Studies in mice have shown that memory loss correlates more strongly with the presence of ADDLs than with the presence of senile plaques, and that treatments to reduce ADDL levels can actually reverse memory loss. "It's very likely that [senile plaques] can be bioactive; they've been seen attached to the sides of nerve cells, but the damage they cause is probably limited," says Klein, who

Addled synapses? In a newly discovered Alzheimer disease process, ligands known as ADDLs bind to synapses and disrupt the formation of memories.

further speculates, "The ADDLs are much more insidious because they diffuse between cells until they find just the right target."

Cotman's team and others have also found that ADDLs are similar in size and shape to prions, molecules that have been linked to the transmission of bovine spongiform encephalopathy ("mad cow disease") and other neurodegenerative diseases. They may therefore share a common mechanism of toxicity.

ADDLs have been measured in cerebrospinal fluid. Measurement of ADDL levels could eventually serve as a diagnosis

of Alzheimer disease, especially in its early stages, and could be more accurate than the cognitive evaluations currently used, says Klein. His team is working to see if ADDL levels that correlate to the disease can be detected in less-invasive blood tests.

Recent ADDL discoveries could also contribute to research into potential anti-amyloid therapies, which currently fall into three groups: immunotherapies that prompt the body's immune system to destroy Aβ, antiaggregants that keep the molecules from clumping, and enzyme (secretase) modulators that prevent the creation of Aβ or hasten its destruction. "All three approaches show some efficacy in transgenic animal models," says Samuel Gandy, vice chair of the medical and scientific advisory council of the Alzheimer's Association and director of the Farber Institute for Neurosciences at Thomas Jefferson University.

An immunotherapy approach using vaccine has shown mixed results so far. Although some patients in phase II trials showed cognitive benefits, some 5% also developed acute allergic encephalitis for reasons yet unknown. Klein speculates that a vaccine that focused on Aβ in ADDL form, as opposed to plaques, could be more effective, cause less inflammation, and require less vaccine. Further, he says, there are no reasons *a priori* that should prevent development of ADDL-specific therapeutic antibodies, which would constitute a "passive" vaccine. "It's a challenging and exciting approach, but there's a lot of work to go yet," says Cotman.

Several drugs to inhibit Aβ clumping are in clinical trials, according to Gandy. Enzyme-based therapies could use existing compounds—recent studies indicate that some antioxidants, including vitamin E and curcumin (the main ingredient in curry), may inhibit Aβ accumulation. Other research has shown that statins (cholesterol-lowering drugs) and estrogen activate a form of secretase that destroys Aβ, says Gandy. There are indications that other lifestyle and environmental factors may also decrease Aβ levels. For example, some epidemiology studies link exercise with delayed onset of Alzheimer disease, says Cotman. "Ultimately," he says, "we may be looking at a combination of behavior, new drugs, and nutrients to treat Alzheimer's disease." **–Kris Freeman**

txgnet

Harvard Institute of Proteomics

The discovery, study, and characterization of the vast number of proteins in the human body is an enormous challenge. One research center that has been formed to take on this challenge is the Harvard Institute of Proteomics (HIP). The HIP's website, located at **http://www.hip.harvard.edu/**, outlines the institute's research program. Founded within the Harvard Medical School in 1999, the HIP is at present laying the groundwork for determining the function of every protein encoded by the human genome in order to help understand how protein malformations contribute to disease.

building of an expression-ready gene repository for these gene sets are described on the Pseudomonas page of the Research section of the HIP site. Currently HIP investigators are focusing on developing gene sets for *Pseudomonas aeruginosa* PAO1 (the leading cause of death in cystic fibrosis patients), the malaria parasite *Plasmodium falciparum* 3D7, and strains of arboviruses including West Nile virus and dengue virus strains 2 and 3.

Institute researchers are also pushing forward to clone the entire range of kinases encoded within the human genome and to transfer them to a
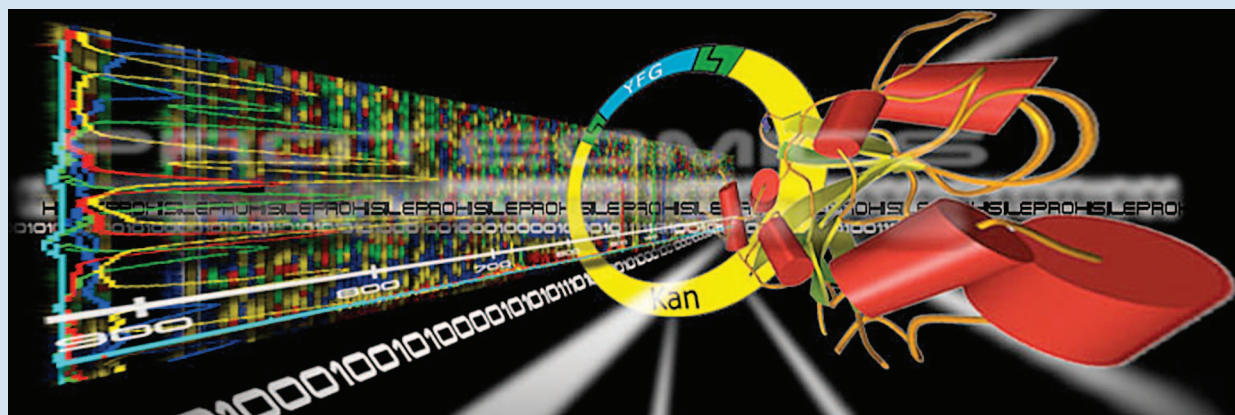
Toward this goal, HIP scientists are developing a novel robotics-based resource known as the FLEXGene repository, which will contain and distribute cloned copies of 20,000 human genes. Once this repository is complete, its developers say it will allow researchers to look at protein expression in all experimental formats and at any chosen scale. Whereas traditional DNA subcloning is time-consuming and labor-intensive, using clones from FLEXGene is quick, inexpensive, and efficient.

The research section of the site contains contact information for and descriptions of the eight areas of ongoing HIP research. One of these, the Breast Cancer 1,000 Project, is focused on discovering and understanding the biological functions of proteins related to breast development and breast cancer. Eventually, this research group will develop a repository of clones for 1,000 full-length complementary DNAs for genes that may contribute to the onset of breast cancer. The group is also working to convert their findings into technologies to support a broad range of functional experiments. The products generated through these efforts should also prove useful to research on other cancers that have been linked to this group of genes.

HIP projects devoted to the sequencing of the genomes of human pathogen organisms and the

database that supports cell-based assays relevant to processes including growth factor/cytokine signaling, apoptosis, and immunosuppression. Information on this project is available in the Research section, as are snapshots of projects to develop high-throughput methods for protein expression and purification, and to devise a method for producing protein microarrays that is more streamlined than current processes and that reduces the need for direct manipulation of the proteins.

The Informatics section of the HIP site contains links to databases and web-based programs used for proteomics applications. For example, the MedGene database mines text citations from PubMed to create a "co-occurrence network" that normalizes and ranks reported human gene–disease relationships. MedGene can generate input for disease-specific microarrays, sort gene profiling data, and compile lists of genes for use in screening experiments. The programs available on this page include PCR Oligo Calculator, which lets researchers design batch polymerase chain reaction primers for amplifying the open reading frames for a given set of genes, and Batch Gene Retriever, which allows downloads of full-length sequence information from National Center for Biotechnology Information databases in batch mode. –**Erin E. Dooley**

Crunching the Bio-numbers

The quest is on to extract useful information from the growing mountain of data from today's high-throughput, high-tech biology, and there is a constant demand for new data-mining techniques that are faster and smarter. The work presented at a recent symposium session on microarray and gene expression analysis shows how bio-number crunchers are contributing many ingenious new approaches to finding the scientific needles in the raw data haystacks.

The session was part of the Atlantic Symposium on Computational Biology and Genome Informatics, which was one of 11 conferences, symposia, and workshops convened under the umbrella of the 7th Joint Conference on Information Sciences, held 26–30 September 2003 in Research Triangle Park, North Carolina. The conference was sponsored by the NIEHS, the Association for Intelligent Machinery, Duke University, the journal *Information Sciences*, and the Harbin Institute of Technology in China.

According to session chair Björn Olsson, a lecturer in computer sciences at the University of Skövde, Sweden, the presentations showed that the field of bioinformatics is coming to terms with the capabilities it has to offer. "Going back a few years, when this type of data was completely new and everyone was excited about it, I think people were fumbling in the dark a bit about what to do with all of this data," he says. "It's becoming more clear now what directions we can go in."

Simon Lin, manager of the Duke Bioinformatics Shared Resource, led off the session by presenting the results of a study he and colleagues recently completed proposing an improved method of data classification in proteomics-based research using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry, an enhanced version of the original technology used for protein identification. Peaks from MALDI-TOF raw data (which exhibit the proteins extant in a biosample) must be brought into registration to correct for random fluctuations before they can be used to classify samples (for example, as diseased or nondiseased). Lin and his team used a new algorithmic approach to the registration problem,

employing a statistical model based on the chemical analysis of normal mixtures to achieve registration.

Applying this method to an existing data set of 11 tissue samples from cancerous lungs and 11 samples from healthy lungs achieved a classification rate of 90.9%, with a false-positive rate of 0%. Further, the researchers correctly identified two previously known lung cancer protein markers in the cancerous samples, and discovered seven novel markers worthy of further investigation for biologic relevance.

Improving the accurate classification of diseased versus nondiseased samples is one of the ongoing challenges in the world of bioinformatics. "Decision trees" help researchers classify samples by offering sequential tests of individual attributes. The result of each test determines which test, or branch, should be applied next, until a final classification is reached. Olsson presented work in which he and his colleagues applied the decision tree algorithm C4.5 to microarray-based gene expression data in order to induce decision trees for identification of breast cancer patients.

Using the expression values of the 108 genes identified in the literature as breast

More than a mouthful. A recent bioinformatics meeting highlighted the many data-crunching technologies being brought to bear on the "-omics."

cancer–related as input to the decision tree algorithm, the team analyzed gene expression data from 75 women, 53 of whom had been diagnosed with breast cancer. The decision tree method achieved 89% accuracy in classifying samples, based on their gene expression data. Olsson also described the potential utility of decision tree algorithms to study signaling pathways based on gene expression data, as well as to discover additional cancer-related genes.

From decision trees, graduate student Tao Shi of the Department of Human Genetics at the University of California, Los Angeles, took the audience into the woods as he described the use of "random forest" predictors to derive information from microarray data. The random forest approach, which uses a suite of decision trees, can be used to detect clusters in the data, a vital and informative but often difficult step in accurate classification. Shi showed that the random forest approach could help meet the challenge of using gene expression data to classify tumor types, which is increasingly important in molecular biology efforts to characterize cancer subtypes.

Rounding out the session, mathematician Takeharu Yamanaka and research fellow Fred Parham of the NIEHS Laboratory of Computational Biology and Risk Analysis presented a new method of analyzing gene expression to infer genetic interactions, which can help identify signal transduction pathways crucial to the sequence of biochemical events that control cellular function. The method uses a Bayesian network, a type of mathematical framework useful for representing known or hypothesized causal relationships. The network measures levels of messenger RNA from different genes and uses conditional assumptions to represent the influence one gene has on another. "By using the Bayesian network," says Yamanaka, "we can incorporate statistical thought into the analysis, unlike present methods." Parham adds, "Also, with the Bayesian networks, we can hopefully . . . say not just that this group of genes is related, but we can also see the causality."

As today's automated, high-throughput instruments routinely churn out masses of data that would have been unimaginable just a decade ago, innovations in bioinformatics such as those presented at the session will be required to bring method to the madness, and ultimately help deliver the improvements in human and environmental health promised by molecular biology. **–Ernie Hood**