When we try to pick out anything by itself, we find it hitched to everything else in the universe.
John Muir
My First Summer in the Sierra (1911)

INNOVATIVE TECHNOLOGIES

Silent Advances

A growing body of research shows that gene silencing is a critical component of many diseases. In particular, scientists continue to learn more about how enzymes known as histone deacetylases, or HDACs, work to silence genes. Better understanding of how HDACs silence genes is particularly relevant to understanding, and perhaps better managing, diseases characterized by abnormal cell growth, such as cancer and neurological disorders.

Chromosomes contain DNA, and this genetic material is tightly packed into chromatin. The smallest unit of chromatin is the nucleosome, where proteins known as histones tightly bind DNA. All this wrapping protects genes from being decoded and expressed inappropriately. Histone acetylases switch genes on by freeing DNA from tightly packed chromatin. HDACs are counterpart enzymes that operate in reverse; they shut off genes.

Eleven types of human HDAC were already known to occur in complex mixtures with related proteins, such as gene repressors and hormone receptors. In the course of deciphering the components of one of these complexes, Ramin Shiekhattar, an associate professor in the Gene Expression Program at the University of Pennsylvania's Wistar Institute, discovered an entirely new family of complexes containing HDACs. All the members share a common core composed of HDAC linked to another protein called BHC110. A variety of other proteins are attached to this core unit, including one involved in X-linked mental retardation and another associated with breast cancer. These findings are described in the 28 February 2003 issue of the *Journal of Biological Chemistry*.

The HDAC section of the new complex binds to chromatin to shut off genes, just like all other HDACs; the challenge lies in uncovering what the BHC110 component does. Scientists have identified enzymes that acetylate, deacetylate, phosphorylate, dephosphorylate, and methylate histones. "What's missing is an enzyme that demethylates histones," says Shiekhattar. He speculates that histone demethylation may actually be the role played by BHC110. If this is indeed the case, "BHC110 is going to be a hot protein," says Shiekhattar.

Another mystery is why diverse proteins are attached to the HDAC/BHC110 core, in contrast to the other HDACs, which bind only one type of protein to their cores. Shiekhattar suspects that the different proteins direct the complex to specific tissues. For instance, one member of the new family contains the *ZNF217* gene that is amplified in breast cancer. The HDAC/BHC110 complex with this particular subunit attached may be involved in the regulation of breast cancer. "My gut feeling is that we found a set of complexes that repress different genes based on their unique subunit," says Shiekhattar. Experiments are currently under way to explore this theory.

Shiekhattar's findings add to "the collective work of other laboratories that study HDAC to impact our understanding of diseases," says Danny Reinberg, an investigator at the Howard Hughes Medical Institute and a distinguished professor of biochemistry at the University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School in Piscataway. The overall goal of HDAC research is to learn how HDAC complexes control cellular functions, then identify compounds to block undesirable actions.

For example, defects in the acetylation/deacetylation machinery occur in tumors and in Huntington disease. Scientists at the Memorial Sloan-Kettering Cancer Center (MSKCC) in New York City, led by MSKCC president emeritus Paul Marks, have discovered that the HDAC inhibitor suberoylanilide hydroxamic acid (SAHA) causes cancer cells to stop growing and die. Their findings are published in the 15 May 2003 issue of *Blood*. SAHA, which was first synthesized 15 years ago by MSKCC researchers to control the cell cycle, is undergoing clinical trials in cancer patients, who show early positive outcomes. By inhibiting HDAC, SAHA increases the level of histone acetylation, resulting in increased expression of genes and proteins (such as p27kipl and gelsolin) that are directly implicated in tumor suppression.

SAHA has also been shown to prevent movement disorders in a mouse version of Huntington disease, where the buildup of abnormal proteins in brain cells jams the acetylation–deacetylation regulatory system. A team from King's College London published findings in the 18 February 2003 issue of *Proceedings of the National Academy of Sciences* that mice with the disease that had drunk water laced with SAHA showed significantly less loss of movement than those that drank plain water.

"In years to come, we will learn that other diseases are affected by HDAC," predicts Reinberg. It all goes to show that, as Shiekhattar puts it, gene regulation is "like driving a car"—safe driving relies as much on braking as on accelerating. **–Carol Potera**

Hope may come from HDACs. More information on how histone deacetylase (HDAC) compounds control cellular function could eventually lead to treatment for conditions such as Huntington disease.

Photodisc

POLICY

FDA Eyes Pharmacogenomics Data

The U.S. Food and Drug Administration (FDA) is looking at how microarray and toxicogenomics data may be incorporated into its drug review process. Field insiders expect microarray data will eventually be a standard component of submissions for both investigational new drug applications (for use in clinical tests) and new drug applications (for marketing new drugs in the United States). However, in the short run, the FDA's capacity to manage a deluge of these additional data is limited. And key questions remain as to exactly how and when the FDA will accept microarray data.

Only about 1 out of every 10 drugs makes it to the first phase of clinical trials, according to industry estimates. Current estimates of the cost to develop a drug run $800 million. Experts contend that if the use of microarray data could even double the efficiency of drug development—for example, by increasing the number of drug candidates that make it to clinical trials—the savings would be substantial. And the potential to increase the efficiency is greater than that, says Leslie Browne, chief operating officer of Iconix Pharmaceuticals, based in Mountain View, California.

Microarray data could also improve drug quality. Research has shown that gene expression data can catch changes early on that traditionally are seen only in pathology. "A tool like this provides an opportunity to weed out compounds early that will have problems," explains Browne. In a rat study, for example, lesions were caught at day 5 compared to day 28 for classical histologic methods. Other retrospective studies have demonstrated the strength of the microarray as a predictive tool across species.

The FDA released draft guidance on pharmacogenomics data submissions in November 2003. "The draft guidance is a great start to this process, and the developing debate will enhance the field," says Browne. By embracing the technology early on, "the FDA in this case has been helping to push this forward," he says.

Norris Alderson, the FDA senior associate commissioner for science, intends to develop one set of standards for use throughout the agency, including all FDA centers. "We're striving to achieve—as much as possible—harmony within the agency as we move forward to apply genomics in a regulatory setting," explains John Leighton, supervisory pharmacologist in the FDA Division of Oncology Drug Products. "Our thinking is evolving as we see more and more submissions containing genomic data and gain a better understanding of what is useful and what isn't from a regulatory standpoint."

To help develop the guidance and learn how to address microarray data, the FDA

Artville, Chris Reuther/EHP

Office of Testing and Research has launched two gene expression database projects. The first, a collaboration with Iconix, will familiarize FDA reviewers with microarray basics using Iconix's DrugMatrix toxicogenomics database. So far, DrugMatrix contains findings on 600 compounds at multiple dosage amounts and times. Gene expression data are linked to information on pharmacology, histopathology, clinical chemistry, and toxicology, providing a reference for FDA reviewers to compare findings with known results. Iconix is also training FDA reviewers on quality control for microarray data generation, as well as how to analyze data across multiple microarray product platforms and validate biomarkers from integrated chemogenic data sets.

The second project, in partnership with Schering-Plough and Affymetrix services provider Expression Analysis, based in Durham, North Carolina, is building a database for mock gene expression data submissions. According to the 23 June 2003 edition of the online news source *Bioinform*, the planned internal gene expression database will help educate FDA reviewers about the format, content, and context of microarray data submissions.

Most experts agree that the FDA has been legitimately conservative in its use of toxicogenomics data so far, because there are real risks in adapting microarrays and similar technologies before they are mature. "The idiosyncratic response of individuals to drugs is still quite unknown, and just because we could measure forty thousand genes at a time doesn't make this problem any easier to solve," says Atul Butte, a physician and instructor of endocrinology and informatics at Children's Hospital Boston and Harvard Medical School.

Initially, Leighton sees microarray technology as an adjunct to traditional drug evaluation tools that will help researchers better understand the underlying mechanisms of toxicity, especially for long-term studies. Moreover, he believes such data will play a greater role, at least initially, as a tool for enhancing an understanding of a compound's pharmacology rather than its toxicologic properties.

But much work needs to be done before the FDA can determine how microarray data should be used in regulation, and standards need to be established before the agency can decide how to use such information in risk assessment. For example, there are no known valid biomarkers to date, as called for in the guidance. Among other technical challenges, a process needs to be established for how a biomarker progresses from "experimental" to "probable" status, and then to being a known biomarker. The FDA and many other research groups are striving to correlate content and format of gene expression microarray data with standard toxicology and pharmacology data.

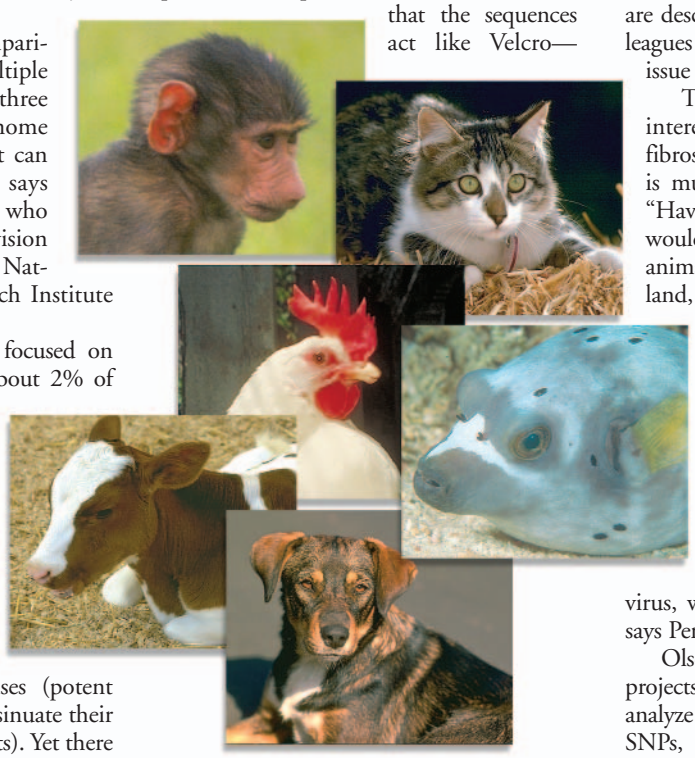
Industry in general has been slower than the FDA to promote the use of microarrays in the development of new drugs. "Drug companies have been reluctant to embrace it because they have realized the disadvantages," Browne says. But the FDA is working to demonstrate that voluntary submission won't come with penalties. "We hope to overcome the fear by some in industry that the agency won't know how to use the data or make inappropriate use of the data," Leighton says. **–Julie Wakefield**

GENOMICS

Sequencing a Zoo

Recent comparative sequencing and analysis of 10 genes in 13 vertebrate species has found hundreds of identical and potentially functional sequences in stretches of the genome that scientists once referred to as "junk" DNA. A recent report claims that these sequences have been conserved through hundreds of millions of years of evolution, a fact that suggests they may perform important roles and are worthy of future study.

"This kind of focused comparison of a few genes across multiple species can filter down the vast three billion letters of the human genome to a more manageable set that can be explored for function," says report coauthor Eric D. Green, who is scientific director of the Division of Intramural Research at the National Human Genome Research Institute (NHGRI).



Most genetic research has focused on exons, sequences making up about 2% of the total human genome that code for the creation of proteins. Until recently, it was believed that only exonic sequences were functional, and that the rest of the DNA was a sort of genetic detritus, consisting of useless code such as defective copies of genes, nonsensical repeats, and the remains of disabled retroviruses (potent viruses such as HIV that can insinuate their code into the DNA of their hosts). Yet there is increasing evidence that certain sequences outside exons that make up another 2–3% of the human genome also play critical roles. However, finding these nonexonic sequences has been difficult, in part because many are smaller than exons and lack the "start" and "stop" signals that mark protein-coding regions.

In the report, published in the 14 August 2003 issue of *Nature*, the team describes sequencing the genomic region corresponding to a section on human chromosome 7 that contains 10 genes. The 10 genes were sequenced in the human and 12 other species: chimpanzee, baboon, cat, dog, cow, pig, rat, mouse, chicken, zebrafish, and two species of pufferfish. The sequences were then compared using two different statistical methods in a hunt for "multispecies conserved sequences," or MCSs.

A total of 1,194 MCSs were identified between the two techniques. Of these, the vast majority were nonexonic—only 244 overlapped exons. A little over half (648) were found in introns, sequences that are transcribed to messenger RNA but removed before the RNA is transcribed into proteins. The rest of the MCSs (302) were found in areas between genes. The nonexonic sequences may possibly regulate protein transcription or perform other functions.

Comparing two species usually isn't enough to find these conserved sequences, says Green. "For example, if you take the human genome and mouse genome, at forty percent of the places, the DNA is so similar that the sequences act like Velcro—they stick together, or align." The vast majority of these sequences are identical only because the two species had a common ancestor not all that long ago in evolutionary terms, and there hasn't been enough time for any nonfunctional sequences to diverge from each other. To find the small stretches of sequences that are critical to gene and organism function, it's necessary to compare multiple genomes, says Green.

"This work has had two immediate consequences," says Maynard V. Olson, director of the University of Washington Human Genome Center. "First, bioinformaticians are using these data extensively to fine-tune their methods for finding conserved sequences. The second immediate consequence is that this work is already guiding choices for whole-genome [sequencing] projects." One of the ideas of sequencing this whole "zoo," says Green, is to get a better idea of which genomes would be most cost-effective to sequence completely in terms of sorting out critical areas and looking for elements such as MCSs.

NHGRI researchers are continuing their analyses to determine which genome comparisons are most effective at finding conserved sequences, and how many comparisons are necessary to find the largest possible number of MCSs. For example, the team found that eliminating chimpanzee and baboon sequences from their 13-species analysis didn't reduce the number of MCSs found, but removing the nonmammals reduced the total by 17%. Their methods are described in a paper by Green and colleagues published in the December 2003 issue of *Genome Research*.

The *Nature* study may be of special interest to researchers studying cystic fibrosis, because one of the genes analyzed is mutated in people with that disease. "Having the data from so many species would certainly aid in the construction of animal models," says Christopher Penland, director of research for the Cystic Fibrosis Foundation. The results could also be useful in studies of gene therapy, in which viruses are engineered to transport potentially therapeutic genes. "You could use this research to look for regions in and around the gene that nature has deemed valuable and omit other areas to reduce the overall load to be carried by the virus, when the virus capacity is limited," says Penland.

Olson says such multispecies sequencing projects may also help develop better tools to analyze single-nucleotide polymorphisms, or SNPs, in humans—in effect, comparing numerous humans as opposed to numerous species. "A major issue in human genetics right now is to improve our ability to look at very large SNP databases and develop better quantitative models for determining which ones might affect function, as opposed to being background noise," he says.

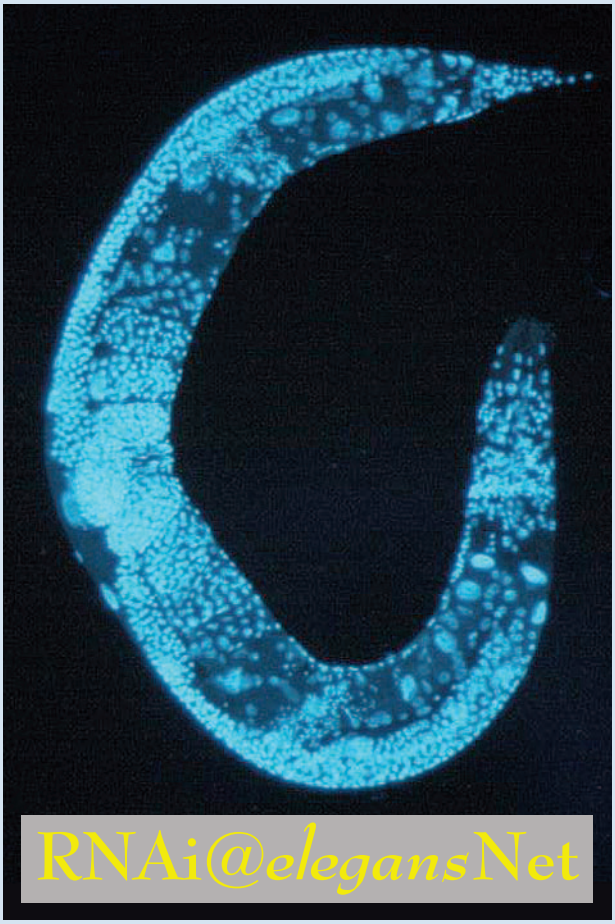
The study also addresses a controversy in evolutionary genetics regarding the pace of genetic mutation. A theory called the "molecular clock" states that mutations occur at a steady pace across time, regardless of species. However, the NHGRI studies indicate that the genomes of rodents are mutating faster than those of primates, carnivores, or artiodactyls (a type of ungulate). The NHGRI study also confirms previous work indicating that primates are more closely related to rodents than they are to carnivores (such as cats and dogs) or to the hoofed artiodactyls (such as cows and pigs).

–Kris Freeman

txgnet

RNAi@*elegans*Net

Caenorhabditis elegans was the first multicellular organism to be completely sequenced. With that 1998 achievement came a fresh appreciation for this popular and valuable research model, which has become a staple of genomics studies. Today, researchers around the world take advantage of the Internet to share genomics data not just on *C. elegans* but also on other organisms. Accordingly, the nonprofit *elegans*Net website is a hub to a universe of information not just on the titular nematode but also on several other species, as well as genomics-relevant technologies.



The goal of *elegans*Net is to simplify navigation to resources on the World Wide Web, to enhance interaction among researchers in molecular, cellular, systems, and organism biology, and to promote science appreciation among the public. The site is vast, with approximately 29,000 links indexed, and it draws on all kinds of resources, from pharmaceutical company–produced educational materials to peer-reviewed journals. One of the technologies highlighted by *elegans*Net is gene silencing through RNA interference (RNAi), through a resource page located at **http://c.elegans.tripod.com/RNAi.htm**.

Under the Hot header on the homepage, visitors will find links to RNAi research published in the past month. This section also includes a history and overview of RNAi provided by biology products company Ambion, access to Ambion's *Silencer* newsletter on gene silencing research and technologies, and online news sources related to RNAi. The Literature Searches section expands on these offerings by taking visitors to the National Library of Medicine PubMed homepage and running preselected searches on pertinent topics, including RNAi therapy, transcriptional gene silencing, and posttranscriptional gene silencing. This gives visitors the power to access the most up-to-date citations literally at the click of a button. Visitors are have ready access to published research on high-throughput screens and reviews of RNAi as a gene therapy approach. Papers on these topics are available under the High-Throughput Screens (HTS) and RNAi Therapeutic Models Reviews headers on the homepage.

The links under the Players header take visitors to online journal articles describing key elements of the world of RNAi, including Dicer, microRNA, and short hairpin RNAs. The Animations & Images section directs visitors to websites housing time-lapse films of *C. elegans*.

The Resources & Services section lists links to a number of research centers, databases, and search engines. For example, the RNA World Databases site of the Institut für Molekulare Biotechnologie in Jena, Germany, provides access to a vast wealth of databases, web-based tools, and software. The RNAi.net page lists educational, career, and business opportunities for scientists. The RNAi Phenotype Search, part of the WormBase consortium of *C. elegans* researchers, allows visitors to search for genes with positive or wild-type RNAi assays by any of a number of maternal, embryonic, and/or postembryonic phenotypes. And the Harvard Medical School *Drosophila* RNAi Screen Center makes available a library of double-stranded RNAs that can be used by researchers to conduct high-throughput cell-based RNAi screens to identify genes involved in various assays. **–Susan M. Booker**

BIOINFORMATICS

Cluster Busters

Bioinformatics experts are always working to design better statistical algorithms to comprehend the expression patterns of tens of thousands of genes. Different algorithms may better serve diverse scientific goals, such as screening for potential tumor markers or obtaining a comprehensive window into the state of a cell as it reacts to an environmental toxicant. Now, in a new use for a preexisting methodology, Raj Acharya and Jyotsna Kasturi, two computer scientists at The Pennsylvania State University, have applied a mathematical approach called Kullback-Leibler (KL) clustering to the identification of patterns in microarray data.

Microarrays shed light the effects of environmental toxicants on genes by measuring the expression of thousands of messenger RNAs simultaneously. Microarray experiments generate vast amounts of data, which bioinformatics experts examine using statistical algorithms designed to detect patterns. Similar genes are sorted into groups, or clusters, that provide insights into gene interactions and thus help to explain underlying biological processes.

Traditionally, mathematicians and engineers have used KL methods to explore theoretical concepts. But KL clustering has proven to be a powerful method for looking at gene expression over time in response to drugs or environmental toxicants, says Murali Ramanathan, an associate professor of pharmaceutical sciences at the University at Buffalo–The State University of New York, who collaborated with the Penn State scientists in proof-of-concept studies on the approach.

In general, clustering procedures find similarities among data set items that form the basis for sorting them into a series of groups. "Clustering is like sorting different-color balls into bins, each containing one color," explains Acharya, who is director of the Penn State Advanced Laboratory for Information Systems and Analysis. Each ball is sorted by how closely it matches the color of the other balls already assigned to bins. The algorithm calculates a similarity score for all pairs of genes and assigns them to a cluster. Genes within clusters carry out similar tasks, such as cholesterol synthesis or wound healing. Any genes of unknown function are "guilty by association"—they are suspected to have a function similar to those of known genes in the same cluster.

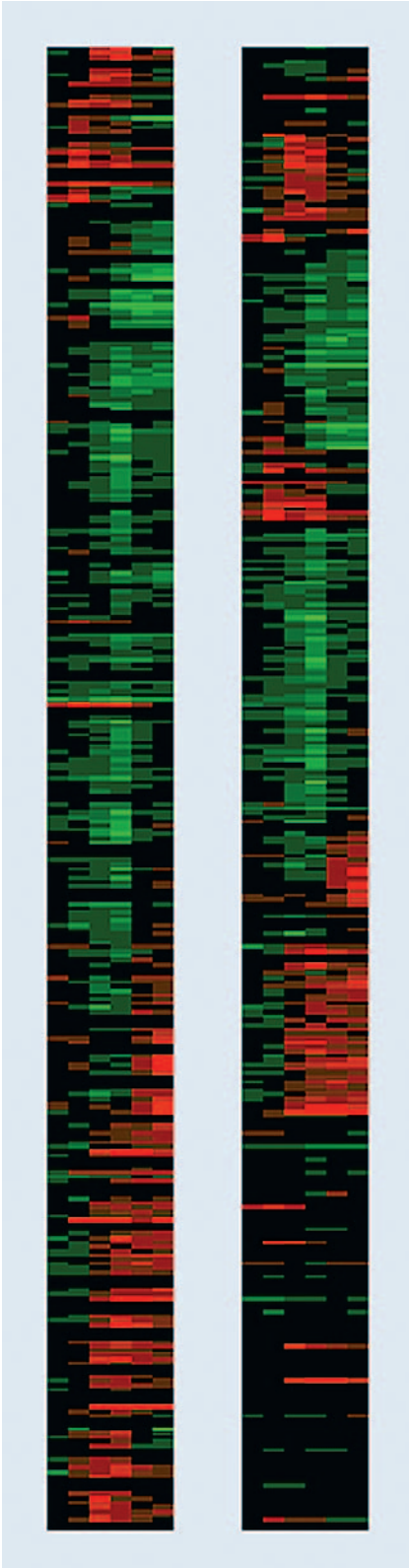
The main aim of cluster analysis of microarray data is to expose regulatory networks and assign function to sequences of no known function. So researchers desire small numbers of clusters, each densely packed with interrelated genes that reflect common pathways or biological functions. In contrast, techniques that generate many clusters containing just one or a few genes are undesirable.

In a proof-of-concept experiment, the researchers applied KL clustering to genetic data from the Onto-Express database of the Wayne State University Intelligent Systems and Bioinformatics Laboratory. Graduate student Kasturi had written a computer program to test KL clustering, which was applied to 517 genes from human fibroblasts treated with serum, representing 12 time points. In addition, a larger set of 4,579 yeast genes containing 18 time points in the cell cycle was analyzed. This test run, reported in the March 2003 issue of *Bioinformatics*, showed that KL clustering performed better at sorting microarray data than the standard method of hierarchical clustering, which uses a different algorithm to measure the similarity of genes.

For the fibroblast data, KL clustering produced about half as many clusters as did hierarchical clustering, and KL clusters were densely packed with similar genes. In contrast, many of the hierarchical clusters contained just one or two genes. A similar pattern emerged for the yeast data. "Using a small and large data set shows that the program is scalable to large sets of genes," says Kasturi.

Some KL clusters were compared with known genes from Onto-Express, which confirmed that clusters shared similar gene functions. For example, one cluster held genes related to cell–cell communication, whereas another cluster controlled cellular development. The Penn State team's KL clustering program is available by request by contacting Kasturi at jkasturi@cse.psu.edu.

Toxicologists are always searching for better ways to extract data from high-throughput screens, and new approaches to mining large data sets appear regularly in the bioinformatics literature. But Christopher Bradfield, a professor of oncology at the University of Wisconsin, Madison, and CEO of the Madison-based toxicogenomics service company Functional Biosciences, points out that new algorithms are only as good as their translation into clearer biologic understanding. "This may be a better mouse trap," he says, "but the real proof will be in how many mice it catches." **–Carol Potera**



Defining clusters. Comparison of red and green cluster plots for KL clustering (left) and hierarchical clustering (right) shows that the former creates fewer clusters with more genes—an important feature if scientists are to learn more about the function of genes based on the company they keep.

Jyotsna Kasturi