IDG.net

Resolving Bottlenecks: Target Elimination **P.52** Strategic Insights: Regulatory Compliance **P.61**

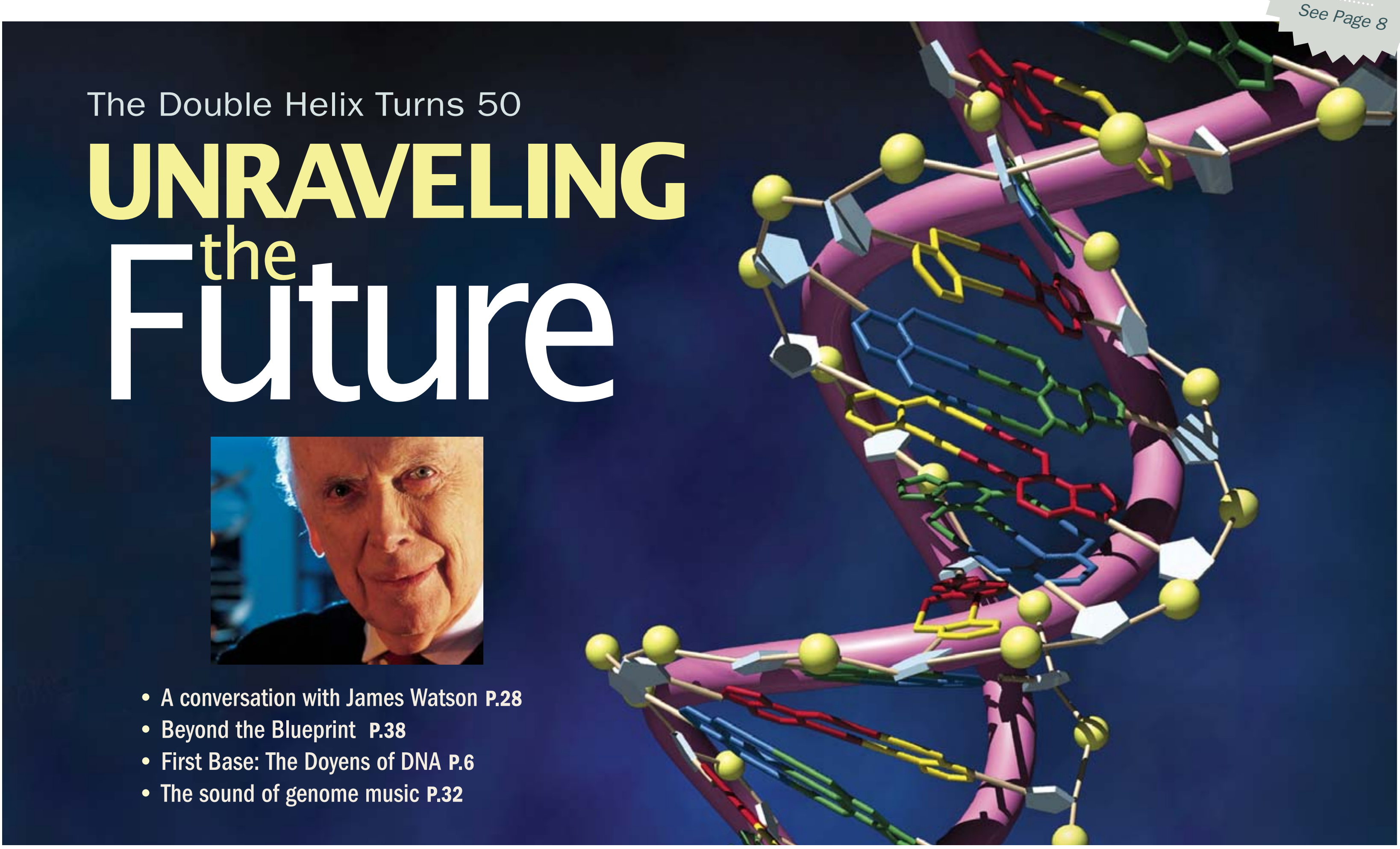
www.bio-itworld.com

Bio·IT World

TECHNOLOGY FOR THE LIFE SCIENCES

APRIL 2003 • VOL. 2, NO. 4

Winner of Two 2003 Neal Awards
See Page 8



The Double Helix Turns 50

UNRAVELING the Future

- A conversation with James Watson **P.28**
- Beyond the Blueprint **P.38**
- First Base: The Doyens of DNA **P.6**
- The sound of genome music **P.32**

KAYOMI TUKIMOTO; MIRIAM CHUA/COLD SPRING HARBOR LABORATORY (INSET)

EDC SHREDS PAPER TRAILS

Electronic data capture finds increased acceptance

BY AMANDA FOX

As the pressure rises to compress the time it takes to conduct clinical trials, commercial and research institutions alike are adopting electronic data capture (EDC) to improve clinical trial management. Recently, Biogen and Dana Farber/Partners CancerCare (DF/PCC) have gotten in on the action, demonstrating that EDC applies to both settings.

With an increasing number of federal clinical trial and drug

(CONTINUED ON PAGE 24)

DEBATING DNA IN MONTEREY

At *Time* conference, top brains spin genomic future amid potshots at pharma

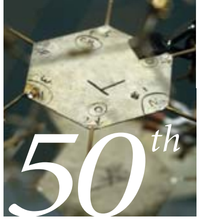
BY MARK D. UEHLING

"This whole meeting has been a mixture of exhilaration and terror," said Francis Collins, director of the National Human Genome Research Institute. If Collins were scared, having run the public effort to sequence the human genome, imagine the reeling minds of preachers, financiers, teachers, and other civilians attending "The Future of Life," a conference sponsored by *Time* magazine in Monterey, Calif.

The February meeting looked both backward at the discovery of the double helix and forward to the benefits and havoc DNA may wreak. Genomics, clearly, isn't for only scientists anymore. Ordinary citizens gawked at the speakers, begging autographs and snapshots of uber-nerds and scientists such as Leroy Hood, Ray Kurzweil, Richard Dawkins, Stewart Brand, and E.O. Wilson. The conference featured a baroness, an ambassador, and two Nobelists.

As Collins sketched a number of April events to celebrate the finished human genetic sequence, including a plan to dispatch 1,000

(CONTINUED ON PAGE 20)



50th Anniversary of the Double Helix

BEYOND the

As investigators celebrate the golden anniversary of the double helix, how will the wealth of data emanating from the human genome and allied technologies impact research on health and disease?

BY MALORYE BRANCA

Contemplating DNA's inner beauty from blurry X-ray images and cardboard cutouts, James Watson and Francis Crick could hardly have imagined that someday, scientists would be surfing the double helix from their desktops, making discoveries with the click of a mouse.

Since Watson and Crick's landmark April 1953 publication in *Nature*, molecular biology has become a mainstay of drug discovery and development, culminating in the sequencing, and public sharing, of the human genome sequence. The announcement of the "substantially complete" human genome sequence this month, coinciding with the 50th anniversary of the double helix, heralds a new phase in the application of genome science to improving human health.

"There has been a remarkable transformation in the way we think about biology," says Eric Lander, director of MIT's Whitehead Institute Center for Genome Research. "We think about biology now as information." Aravinda Chakravarti of Johns Hopkins University agrees: "Today, my students take it for granted they can browse the genome. Someday, we'll do all these studies — sequence analysis, proteomics, genotyping — from a single desktop."

That reality is gaining credence as the "completed" sequence of the human and other genomes, transcriptomes, and proteomes become publicly available, along with powerful new tools to investigate them.

To commemorate the 50th anniversary of the double helix, we asked 50 experts to share their views about the key developments, particularly since the release of the first draft genome sequence in 2000, as well as speculate on future advances. The staggering progress in genomic medicine in the past few years is just a prelude of the excitement in store.

The biggest shock in the human genome is the paucity of genes. Even the earlier estimates of 30,000 may prove too high. This begs the troubling question, "Where is the human complexity coming from?" The answer shifts the spotlight from the genome to the proteome (and beyond).

50/50: Reflections on the Double Helix

In honor of the 50th anniversary of Crick and Watson's *Nature* paper describing the double helix, *Bio•IT World* presents an exclusive online series of interviews with 50 experts in genomic medicine.

bio-itworld.com/news/reflections_index.html

SCIENCE MUSEUM/SCIENCE & SOCIETY PICTURE LIBRARY (TOP)

Illustration by Kayomi Tukimoto www.tukimoto.com

Blueprint

> *"This structure has novel features which are of considerable biological interest."*
>
> — **J. WATSON & F. CRICK,** *Nature*, 25 April, 1953

ERIC LANDER

director, Whitehead Institute Center for Genome Research

> *"Watson and Crick's discovery represented the high point of the molecular biology revolution, in that they reduced biology to molecules. But that contained the seed of the next revolution, to reduce biology beyond molecules, to pure information."*

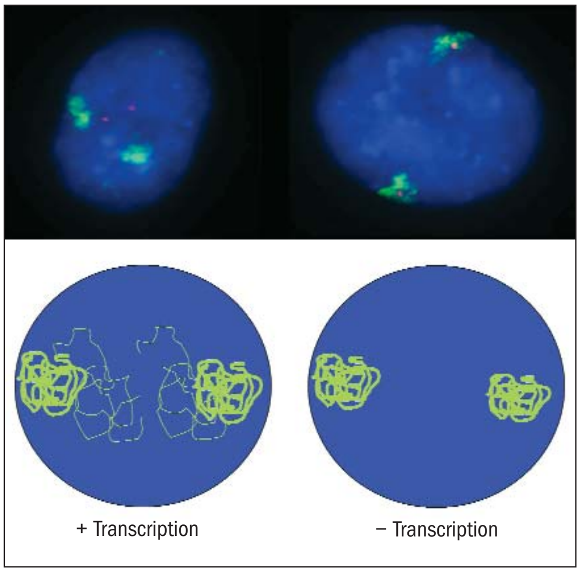
The proteome — the sum of all proteins — is far larger than the genome. For example, alternative gene splicing produces multiple proteins from a single gene, which are then chemically decorated with various moieties, producing a bewildering array of protein forms.

Even RNA — the molecular messenger between genes and proteins — may play more roles than previously thought. The recent discovery that RNA interference (RNAi), an exciting new tool for gene manipulation first studied in plants, also works in mammalian cells has not only opened new avenues of inquiry but also may yield a new class of drugs (see "The RNAi Revolution").

And last year, a study by Affymetrix found much more RNA than expected in a survey of all genes expressed from chromosomes 21 and 22. "We found that 30 to 35 percent of the nonrepetitive DNA is being expressed," says Affymetrix President Steve Fodor. "This is remarkable, because the lore would be that only 1.5 to 2 percent of the genome would be expressed."

Either the gene tally is wrong, or RNA serves many more functions than a simple messenger. "Besides the standard genes that code for protein, there are also genes that code for small RNAs," says Adrian Krainer at Cold Spring Harbor Laboratory. "There is this whole new machinery associated with RNAi and related processes. There appear to be hundreds of genes coding for small RNAs that are part of this."

With so much attention lavished on gene count, the critical issue of gene regulation — the role of non-coding DNA sequences, the mystery of accessing tightly bundled chromosomal DNA to switch on gene sequences — will be a fascinating research vista in the coming years.



CHUBB, J.R. AND BICKMORE, W.A. *CELL* 112. WITH PERMISSION ELSEVIER SCIENCE

Reading the genome: *Chromosomes containing active genes (red dots) loop out (left), whereas inactive chromosomes adopt a more condensed state (right).*

The RNAi Revolution

The discovery of RNA interference (RNAi) could not have been more timely. "Genomics generated a much larger universe of targets," says Bristol-Myers Squibb's Nicholas Dracopoli. "The newer targets, which we don't have much experience with, have slowed down the industry's success rate."

The standard tools to probe gene function were either cumbersome, such as knockout mice, or poorly informative, such as gene expression.

RNAi was described in 1998 by Andy Fire and Craig Mello at the Carnegie Institute in Washington, D.C. In plants and lower organisms such as the nematode, RNAi — a process in which double-stranded RNA fragments target and eliminate specific messenger RNA molecules — probably helps defend against viruses and other foreign molecules.

Two years ago, a paper in *Nature* dramatically showed that RNAi also worked in mammalian cells (Elbashir, S.M. *et al. Nature* 411, 494-498: 2001). The role of this process in humans remains vague, but, regardless, it possesses immense experimental — and therapeutic — potential.

Fast, easy, and inexpensive — that's what an experimentalist likes to hear. "To my mind, the most important new advance in biology is the RNAi approach," says Tom Cech, president of the Howard Hughes Medical Institute and Nobel laureate for the discovery of RNA enzymes. "RNAi is vastly more powerful than anything we have had. Even people who don't know how to spell RNA can use this successfully in diverse biological systems."

RNAi is already making an impact. Genome-wide knockdowns have been carried out in organisms including nematodes. Small interfering RNAs (siRNAs), which silence genes in mammalian cells, are now being designed against as many genes as possible.

One promising approach is to spot cells expressing defined genes on microchips for the analysis (see Ziauddin, J. and Sabatini, D.M. *Nature* 411, 107-110: 2001). "The hard part is not printing the chips or doing the experiment," says David Sabatini, Whitehead Institute Center for Genome Research Fellow and co-founder of Akceli. "It's picking the right sequences."

Each siRNA contains 21 nucleotides, but some sequences stick better than others. For once, researchers welcome the intense competition. "Hopefully, we will cover different genes, and get to them all more quickly," Sabatini says.

Early adopters of RNAi sound a cautionary note. "We've been using it for about five years," says Geoffrey Duyk, president of R&D at Exelixis. "The dirty little secret of RNAi is that you are knocking down messenger RNA to knock out protein. Because proteins have different turnover rates, you have to have a good way to measure protein level and activity."

Even skittish venture capitalists are falling for RNAi and its potential therapeutic value, with backing for companies such as Cenix and Alnylam Pharmaceuticals.

"The great thing about recombinant DNA and monoclonal antibodies was that they gave actual drugs right from the start," says Christoph Westphal, of Polaris Venture Partners. "With genomics, it just wasn't clear when it would develop a drug."

Firms such as Polaris, which is funding Alnylam, hope RNAi can fuel the next wave of biotech breakthroughs. "If we are very fortunate, siRNAs will make good drugs," Westphal says. "If we are unlucky, we still have a whole natural cellular machinery that is open to small-molecule development." —M. B.

BILL HASELTINE

chairman and CEO, Human Genome Sciences

"The human genome sequence is not an end in itself — it is the creation of a set of tools."

(CONTINUED FROM PAGE 40)

The analysis of genome variations, chiefly single nucleotide polymorphisms (SNPs), performed in concert with the sequencing project, has also been a revelation. The SNP Consortium has documented some 2 million SNPs in the human genome, which can serve as valuable landmarks in the search for disease genes. "We never expected that SNPs would be present in such densities," says Dahlia Cohen, head of functional genomics at Novartis.

Interestingly, SNPs in noncoding genome regions may be more important than expected. "Instead of changing the nature of the protein, variations may subtly change the amount of protein produced or the timing of its production," says University of Chicago's Nancy Cox.

Understanding diversity is where some of the major challenges lie. "Producing any draft genome sequence is trivial compared to what it is going to take to understand variation," says Anthony Brookes of the Karolinska Institute. But many groups are making rapid progress in documenting that variation, identifying shortcuts to land valuable disease markers. Chromosomal DNA is inherited in blocks, such that the inheritance of one SNP may be diagnostic for an entire series. This suggests that researchers don't have to sift through a million or more SNPs to find markers of disease or drug response.

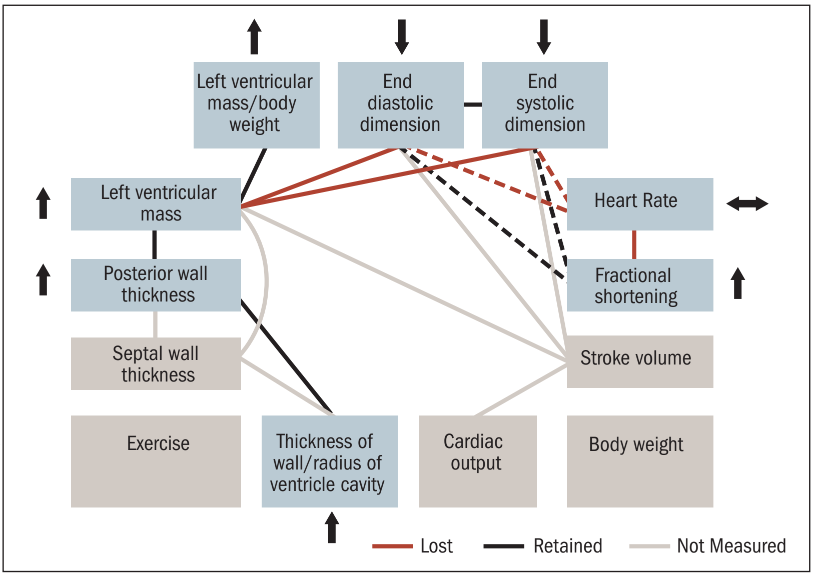
The National Institute of Health's International HapMap Project aims to generate a genomewide map of SNP blocks, or haplotypes, in the next few years. Genaissance already has its own haplotype database, based on the analysis of about 7,000 genes. According to Sequenom Chief Scientific Officer Charles Cantor, "Having the draft sequence has helped us do genetics far more efficiently. We want to study just a few, really important, disease genes ... Within the next six months, we will have more than we can study."

Using wafer arrays that hold 12 billion oligonucleotide probes, Perlegen has produced a haplotype map of 1.7 million SNPs, based on 25 individuals of diverse ethnic background. CEO David Cox's strategy is to identify "the top list of 100 to 200 regions in the genome involved with a trait and do our hardcore biology on those, instead of doing every possible technology that exists across the whole genome on everybody."

The Evolution Solution

One of the salutary lessons of the human genome sequence is the insight that evolution affords to the assembly of the genetic parts list. "Evolution tries to hold on to things that are functionally important," says Eric Green, the new intramural director of the National Human Genome Research Institute.

Green is designing algorithms that hunt for "multispecies conserved sequences" (MCSs) to unearth hard-to-recognize regulatory motifs. "Finding these noncoding functional elements will tell us a lot about how to make complex biological systems," Green says. His lab is currently hunting MCSs across 12 diverse genomes. "A lot of people thought that by sequencing the mouse, we would find all the important regions. Clearly that's not true," he says.



SOURCE: JOSEPH NADEAU, CASE WESTERN RESERVE UNIVERSITY SCHOOL OF MEDICINE

Heart and cell: *Data on the effects of elevated calsequestrin — a protein with cardiac implications — in inbred mice can be fed into a computational model of the heart, shown here, that predicts physiologic changes. The next step is to add gene expression and other genomic data to do the same type of predictions for cellular activity.*

This is not to malign the mouse genome, which is (to some) as important as the human sequence. "Not only do we know the sequence of the mouse genome, we know the variations in sequence among strains," says Joe Nadeau at Case Western Reserve University School of Medicine in Cleveland. Mouse strains can vary dramatically in their cognitive properties and susceptibility to genetic and infectious diseases, such as anthrax.

Nadeau's lab is using mouse genetics to study heart development. Crossbreeding mouse strains with characteristic traits, such as high heart rate, Nadeau has produced a computational model of the interrelationships between these cardiac properties (see "Heart and cell"). Says Nadeau: "The first time I showed this [model], a physician in the audience said, 'Yes, that's the heart, but we already know how it works.' Then someone pointed out that we had figured it out with computers and genetics in just months, rather than the 200 years it's taken physiologists!"

The mouse is also invaluable for pharmacogenomics, as different strains exhibit different drug sensitivities, offering a strategy to identify genes related to drug response. "Right now, we can't do genomewide studies in humans," says Howard McLeod of Washington University in St Louis. "We just don't have enough patients to do the studies we want and still have statistically valid results."

The world of microbial genomics is undergoing a revolution. "The field was moving along slowly, then, suddenly everyone wanted to have their bacterium sequenced," says Steven Salzberg of The Institute for Genomic Research (TIGR). Since sequencing the first two bacterial genomes (*Haemophilus influenzae* and *Mycoplasma genitalium*) in 1995, TIGR has finished 20 more. "Studies of bacteria have revealed the power of genomics better than anything else," TIGR's Jonathan Eisen says. "Because those genomes *are* done."

Working in some cases with industry, Salzberg adds, "We've quickly turned this information into candidate drugs or vaccines." With the omnipresent bioterrorism threat, more opportunities will follow (see "Sequence Signatures and Homeland Security," page 34). But just as wondrous is the window this has opened onto evolution. "We are starting to get a better picture about very early events such as the origin of microbes," Eisen says.

Genome Glut

As the *Homo sapiens* sequence moves into the "complete" column, it joins 100 genomes already finished. "Enormous advances are being made filling in the data, understanding where all the genes

GEORGE POSTE

CEO, Health Technology Networks, and chairman, Orchid Biosciences

> *"In many instances, we are talking about not one, but a constellation or cassette of genes that are at the root of a disease."*

are in these organisms, and going beyond that," says Yale University's Michael Snyder. A generation ago, researchers could study only a gene at a time: These days, a graduate student can study an entire genome.

Julie Ahringer's team at the Wellcome Trust/Cancer Research Institute in Cambridge, England, has systematically turned off more than 85 percent of genes in the nematode *Caenorhabditis elegans* using RNAi (Kamath, R.S. *et al. Nature* 421, 231-237: 2003). Nematodes grazed on a lawn of bacteria containing small interfering RNA (siRNA) inhibitors designed against the nematode genome. "This would have been inconceivable five years ago," marvels Cold Spring Harbor's Lincoln Stein. The biggest surprise? "How many genes you can knock out without killing the worm," Stein says.

In complementary research, Mark Vidal's group at the Dana-Farber Cancer Institute is diligently engineering DNA constructs for all the nematode genes, or open reading frames (ORFs). These clones purposely lack the flanking regulatory DNA sequences that govern how and when the gene is switched on, making them clean experimental tools. Vidal's ambitious "ORFeome" project will confirm whether those ORFs predicted by gene-hunting software are genuine, and settle the worm's gene count.

At the Whitehead Institute, Richard Young has developed "genomewide location analysis." Young's group uses a combination of laboratory methods and informatics to understand how gene expression is regulated by DNA-binding transcription factors, mapping the binding sites of more than 100 of 141 such factors in baker's yeast (Lee, T.I. *et al. Science* 298, 799-804: 2002). Young's group is now turning its sights on the 1,700 or so transcription factors in humans. So, too, is Snyder: "This will tell you which genes these factors regulate. The ultimate goal is to understand the regulatory circuitry."

Protein interactions are popular targets for drug development. "We now recognize that a lot of intercellular signaling is between proteins themselves, rather than just between proteins and small molecules," says Walter Gilbert, Nobel laureate at Harvard University. Last year, two industry/academic consortia made impressive progress using mass spectrometry and other methods to map the yeast "interactome," which involves probably 30,000 protein-protein interactions (Gavin A-C. *et al. Nature* 415, 141-147: 2002; Ho, Y. *et al. Nature* 415, 180-183: 2002). Once again, this is just a prelude to similar studies in humans.

Using the yeast two-hybrid method, Hybrigenics is building libraries of validated protein-protein interactions, which can be viewed through maps called PIMRiders (PIM stands for Protein Interaction Map). For example, the company recently identified human proteins that interact with HIV. "All the drugs available today are directed against the viral protein," explains CEO Donny Strosberg. "But if you could target human proteins that interact with the virus, you could sidestep the problem of viral variability."

Suppliers are keeping up with the genomewide trend. Invitrogen is generating a human ORFeome, allowing customers to select ORFs over the Web. Applied Biosystems (ABI) has stockpiled about 120,000 SNP assays, while it continues polishing Celera Genomics' genome sequence data. "We've had a big program to catalog the functional variation in the genome," says Mark Adams, vice president of informatics.

Software developers have also caught the genomewide bug. Gene-IT's Biofacet software can compare so many sequences, "We haven't found an upper limit yet," claims Richard Resnick, vice president of services. Using the Dutch National Supercomputer TERAS (a 1,024-CPU SGI Origin 3800), Biofacet took 520,000 CPU hours to perform 70 million protein sequence alignments across 82 organisms, searching for proteins that are common in bacteria but not in humans, and hence might make good targets for antibiotics.

Debugging Bioinformatics

Despite the oft-cited "data deluge," the ability of the best bioinformatics algorithms to predict sequences and structures leaves much to be desired, as ongoing efforts to determine the total number of genes attest (see "The Dark Side of Genomics").

For example, it is standard practice to use DNA or protein sequence to predict a protein's structure and, possibly, function. But this remains an imprecise science. "Once the similarity between two sequences drops below 30 percent, most of the proteins will change their function in sometimes

The Dark Side of Genomics

Identifying genes is one of the more glamorous aspects of genomics, but it's difficult picking out the 1 percent to 2 percent of coding sequences amid all the nonsense and junk DNA.

Most genes undergo alternative splicing, producing two or more different proteins. Then, there are pseudogenes — genes that look functional but aren't. Some of these are "dead genes," according to University of California at Santa Cruz bioinformatician David Haussler. "They once had a function," he says, "but they have accumulated enough mutations to slowly decay and become worthless." Distinguishing functional genes from pseudogenes is not trivial.

Thanks in part to the mouse genome, completed last year (see Paper View, Feb. 2003 *Bio • IT World*, page 46), the emerging consensus is that the total number of human genes is less than the initial 30,000 estimate. But uncertainty remains, and Haussler concedes that "computational prediction of genes is simply very hard." New programs such as TwinScan from Washington University in St. Louis, and Genomix's EXP6, are assisting efforts to not only identify novel genes but also confirm the existence of "dark genes" — those predicted by computer programs but unverified by other techniques.

AnVil and Applied Biosystems (ABI) collaborated to shed light on the "dark" genome using PCR (polymerase chain reaction) methods with TaqMan, microarrays, and AnVil's advanced analytics, to study about 10,000 putative genes. AnVil designed the analytical program that "determines whether the lab data are hard evidence that these are genes," AnVil's John McCarthy says. Once "found again," those genes will help finalize the actual number of genes. ABI reports that 2,400 to 2,500 genes were confirmed after screening 9,500 predicted ones. Aside from the inherent scientific interest, "The assays for those genes will make a nice addition to our offering," ABI's Raymond Samaha says.

Ultimately, identifying all genes will be a walk in the park compared to what lies ahead. Eric Green of the National Human Genome Research Center says: "We will find all the genes. But we aren't even in diapers yet in terms of finding all the other stuff — the functionally important sequences and the regulatory elements."

—M. B.

JANET THORNTON

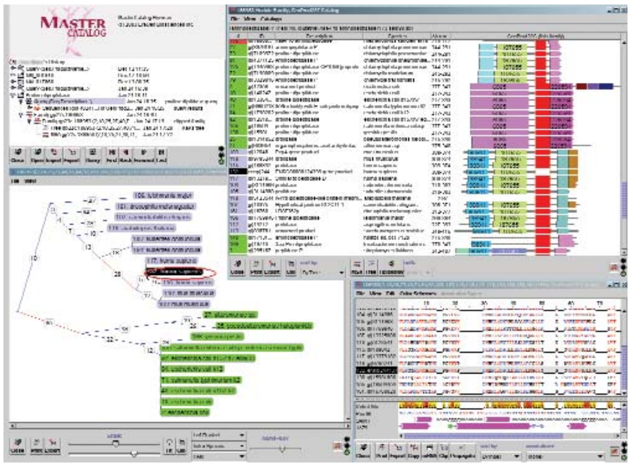
professor of biomolecular structure, University College, London

"It's clear now that we have a relatively small set of protein folds. The complexity of life comes from how they go together and how they have evolved to perform different functions."

subtle or radical ways," explains Janet Thornton of the European Bioinformatics Institute. Indeed, sequences sharing 90 percent identity may have different functions. The leap from structure to function is equally precarious. Thornton and colleagues have documented a haphazard relationship between the sequences, structures, and functions of proteins containing one very common type of fold known as TIM barrels.

Since predictions about structure and function are often filed alongside sequences in databases as "annotation," misinformation gets propagated. "We found a whole number of proteins that were annotated by virtue of their similarity to proteases, but weren't proteases," says Brandeis University structural biologist Gregory Petsko. "I have a terrible feeling that the number of wrong annotations is huge."

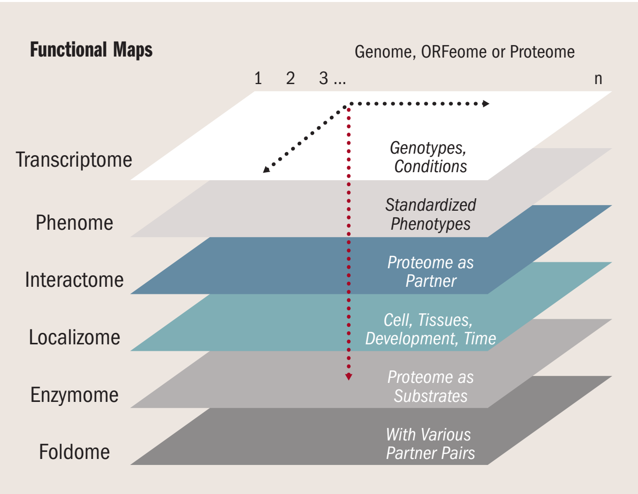
The difficulty of filtering critical data from background noise is affecting new fields such as metabonomics, the study of an organism's metabolites. "The metabonome for every organism is different, and it changes with age," says Jeremy Nicholson of Imperial College London, whose group coined the term. (The metabolome, by contrast, is the full complement of metabolites in a given cell.) There are 600 to 700 known major metabolites, leading to vast numbers of potential combinations and different sets in each cell. "The metabonome is so big, we may never know how big it really is," Nicholson says. His group has developed a metabonomics blood test that can non-invasively diagnose coronary heart disease (Brindle, J.T. *et al. Nature Medicine* 8, 1439-1444: 2002). Patients are being recruited for a larger trial (see www.magicad.org.uk).



Family breakup: *EraGen's MasterCatalog clusters sequence databases into more evolutionary families than traditional methods, allowing better discrimination between related and unrelated proteins.*

The explosion of genomewide data from DNA microarray studies is striking (see "The Maturation of Microarrays," page 46). But proteomics is catching up. "We've already seen exciting signs that proteomics can find biomarkers, and this is a rich area of research," says Ruedi Aebersold of the Institute for Systems Biology.

Roland Eils, co-founder of bioinformatics firm Phase-it (recently acquired by Europroteome), concurs. At the German Cancer Research Center, Eils has been mining data generated from mouse breast tumors using Ciphergen's Surface Enhanced Laser/Desorption Ionization (SELDI) system. "Mouse tumors are similar to human in terms of how they progress," Eils says, "but lab mice are more homogeneous," such that differentially expressed proteins are more clearly defined.



ADAPTED WITH PERMISSION FROM VIDAL, M. *CELL*, 104, 333-339: 2001

Oodles of omics: *Genomic data is interrelated in many ways. Piecing together a single cellular effect requires multiple types of data.*

But these are baby steps. "The huge challenge ahead for proteomics and genomics has to do with how we interpret, analyze, and store the data," Aebersold says. "How do you interrelate the information from two types of experiments, such as microarrays and proteomics?"

Gene Myers, of the University of California at Berkeley, sees a more fundamental problem: "We are building bigger and bigger computers, but, frankly, we haven't made the computer easier for biologists to use." The software is often too difficult for "busy, senior molecular biologists" to learn, creating "a barrier to discovery."

Another impediment is data hoarding. Many researchers argue that genomic data should be made public instantly, even if this causes problems for scientists who, naturally, seek first dibs on their own data. "People are scared of bioinformatics," asserts Peter Weisner, a consultant and formerly of LION Bioscience and Phase-it. "People would rather keep their data in their computers, because they were afraid that good bioinformatics would find more information than they had."

But Steve Lincoln, of Invitrogen subsidiary Informax, says, "Bioinformatics isn't magic; it's just a tool like everything else." Among the more promising tools is a new algorithm developed by R. Mark Adams of Variagenics (now part of Nuvelo) that predicts the effects of a given SNP on protein structure and function, which will help winnow the number of SNPs used in clinical trials. A program called Multiprospector, written by Jeffrey Skolnick (University of Buffalo), models interactions between proteins of undetermined structure and "predicts the 3-D structure of the complex." And EraGen Biosciences' new "evolutionary proteomics tool" contains data from GenBank and other private databases clustered into 150,000 protein families. The platform enables several types of analysis, including multiple sequence alignment and evolutionary trees.

While the genome may never be 100 percent complete, progress over the past two years is such that, "The sequence is now in much better shape," says Richard Durbin of the Wellcome Trust Sanger Institute, "and we have better tools for dealing with it." But large tracts of repetitive DNA remain "impossible to sequence with current technologies," says David Haussler of the University of California at Santa Cruz, host of the "Golden Path" genome portal.

With the sequence now secure, there is a shift away from raw data generation to a more holistic approach, called systems biology. Says Perlegen's David Cox: "Everyone was told, 'If you have the information, [drugs] will fall out like a gold nugget.' They are still waiting for the clunk! Finding the 'needles' in the data haystacks has proved hard. The new trend is to coalesce data into computer models of known biological pathways and networks."

(CONTINUED ON PAGE 46)

ALLEN ROSES

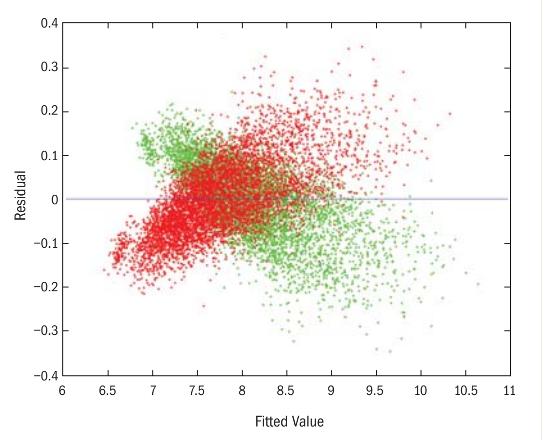
senior vice president, GlaxoSmithKline

"I think a decade from now pharmacogenomic testing will be a requirement. Before you try something on the public, you'll have to do genotyping."

The Maturation of Microarrays

The advent of DNA chips in the late 1990s came, conveniently, just as genome-sequencing projects were gaining momentum. At last, scientists could study the expression or sequence of thousands of genes simultaneously.

"Three years ago, people had the idea that studying gene expression ... would somehow open our eyes to disease pathways never seen before," says Anthony Brookes, of the Karolinska Institute. But there were problems — in chip manufacture, in training, and particularly in data analysis. "This turned part of biology into a data-rich area," says Terry Speed, at the University of California at Berkeley. "Where before you had a notebook to write your results in, now you have megabytes in the computer."



COURTESY KATHLEEN KERR, UNIVERSITY OF WASHINGTON, SEATTLE

Something fishy: *The fishtail pattern, elicited by advanced statistical analysis, alerted researchers to how the two dyes behave differently — just one of many key error promulgators lurking in microarray data.*

Biologists weren't used to that, and it showed. In the early days, says Nat Goodman of the Institute for Systems Biology, researchers declared gene expression as "significantly different if you saw a twofold or threefold change, without any attention to replication or statistical criteria."

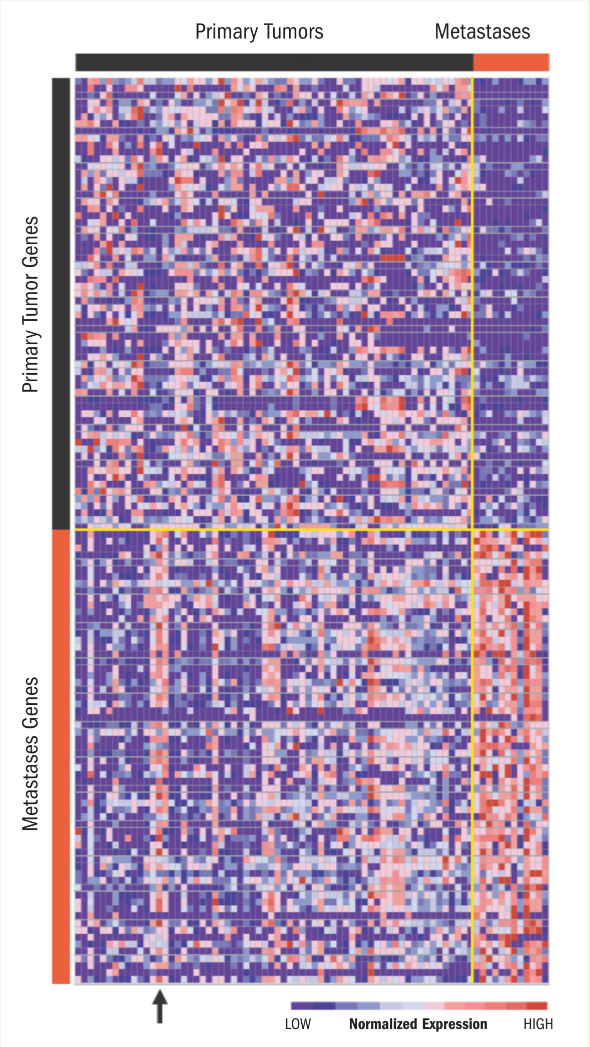
Gary Churchill, a staff scientist at the Jackson Laboratory, chanced upon one of Stanford University microarray pioneer Pat Brown's groundbreaking *Science* papers while surfing the Web: "A lot of statisticians like me noticed and downloaded the data. I carried those [results] around on my laptop for a long time, trying to figure out what was going on."

The breakthrough came when postdoc Katie Kerr showed him a striking graph (see "Something fishy"). "With a data set like that, you need replication so you can sort out the signal from the noise," Churchill says. Kerr's graph showed that the two fluorescent dyes typically used in microarray experiments had different intensities, and that was skewing the data plot. "A lot of excellent statisticians have migrated to the field," Goodman says, "but Churchill and Speed were among the first, and pushed the hardest." Both scientists maintain useful Web sites (see www.jax.org/staff/churchill/labsite/index.html and stat-www.berkeley.edu/users/terry/zarray/html).

Results from commercial and homemade DNA chips have improved dramatically with experience. "The goal was to get the noise from the chip to become irrelevant," says Affymetrix President Steve Fodor. The payoff is coming, although in a slightly different way than anticipated. "Rather than a research tool that gets you to the primary cause [of a disease], it's another type of phenotype," Brookes says. In other words, chips reveal important differences, but not necessarily the reasons for those differences.

The past two years have witnessed a stream of reports dissecting gene expression "signatures" in cancer. At the Whitehead Institute Center for Genome Research, Sridhar Ramaswamy, Todd Golub, and colleagues described a signature that predicts whether tumors are likely to metastasize.

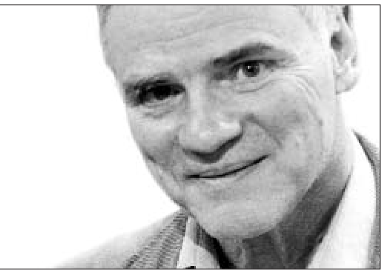
This study used data from a variety of DNA chip platforms, and several tumor types, showing how much more robust the data and the analytical tools have become (Ramaswamy, S. *et al. Nature Genet.* 33, 49-54: 2003).



RAMASWAMY, S. ET AL. *NATURE GENETICS* 33. © NATURE PUBLISHING GROUP. USED WITH PERMISSION.

Spot the difference: *Expression of 128 genes analyzed in 64 primary and 12 metastatic cancers reveals a subset of 64 genes overexpressed in metastases (red, bottom right) and some primary tumors (where the arrow points).*

Further enhancements are on the horizon. "If protein chips could really work, they could be very valuable," says Scott Patterson, who just joined Farmal Biomedicine from Celera Genomics. "The real future lies in the integration of data from many different sources — genotypes, proteins, metabolites, and arrays," Churchill predicts. "It is really essential to find new ways to tie them all together." —**M. B.**



LEROY HOOD

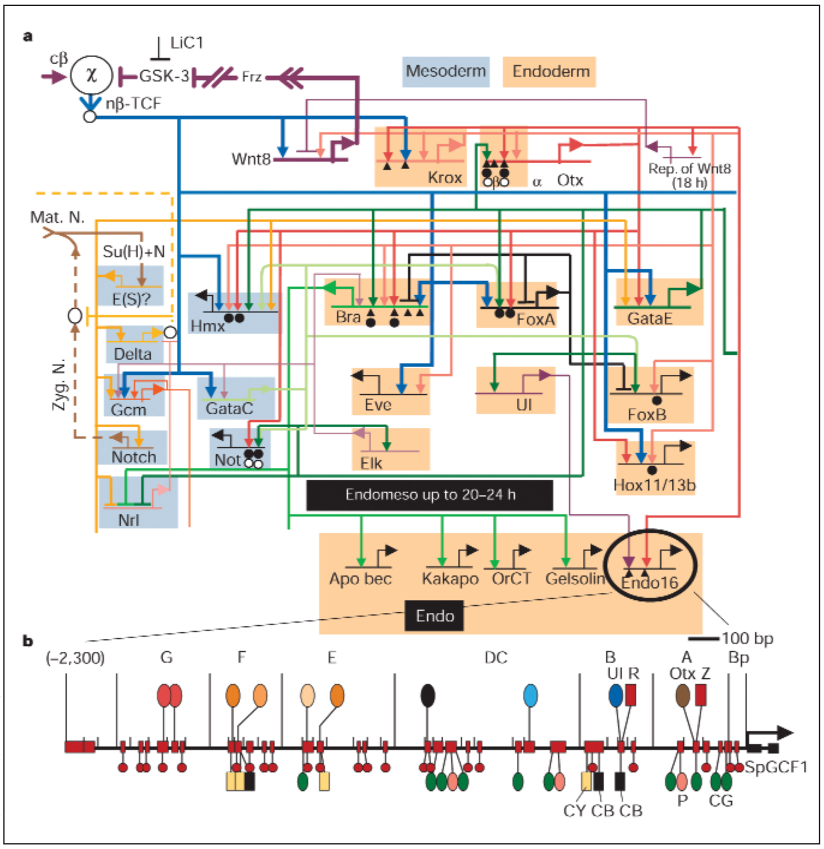
founder, the Institute for Systems Biology

> *"The genome is the beginning of this wonderful new adventure into systems biology and towards revolutionizing medicine."*

(CONTINUED FROM PAGE 44)

"We need computational techniques that will not only help us decipher genomes, but that can integrate the many different levels of information coming out of the genome," says Leroy Hood, founder of the Institute for Systems Biology. He cites Eric Davidson's studies of sea urchin development, which "have transformed our understanding of gene regulatory networks" (see "Major networks" and Davidson, E.H. *et al. Science* 295, 1669-1678: 2002).

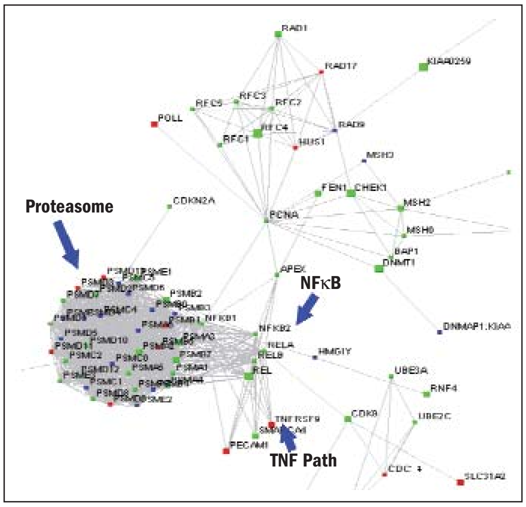
Whitehead Institute Fellow Trey Ideker is building a systems model called Cytoscape. "We are starting to get a glimpse of how genes and proteins interact with drugs and hormones to dictate function," Ideker says. Simpler organisms — notably, yeast — are the proving ground where Ideker and his ilk test their programs, plugging experimental data into the model to see if it can account for the observations. The current success rate is about 40 percent. The big payoff, of course, will be if they can do this for multiple data types and in humans. "Interaction data is the hot area of research now," he says.



HOOD, L. & GALAS, D. NATURE, 421. © NATURE PUBLISHING GROUP; USED WITH PERMISSION

Major networks: *a) Transcription factors control a cascade of gene regulation (arrows, activation; ⊥ , inhibition. b) The regulatory region of gene* endo16 *enlarged, showing 34 DNA-binding sites in six clusters.*

Human data are already being analyzed in some places. Gene Networks Sciences has built a detailed model of a cancer cell, representing the actions of 500 genes and proteins. Bernhard Palsson and colleagues at the University of California at San Diego, meanwhile, have produced a model of red blood cell metabolism. Analyzing sequence data from patients with hemolytic anemia, the model accurately predicted whether particular SNPs were linked to severe or mild forms of this disorder. Palsson, a co-founder of Genomatica, cautions that "other cells are 20 to 50 times more complicated" than red blood cells.



COURTESY MILLENNIUM PHARMACEUTICALS

Paris match: *Millennium scientists compared the signature of a nonresponder patient with that of a group of good responders, revealing possible molecular mechanisms affecting response, such as the proteasome complex, NFκB, and the TNF pathway.*

The turning point for these models will be "when they advance from rediscovering pathways — finding what we already know — to making novel predictions we can then prove," says Andrea Califano of First Genetic Trust.

Millennium Pharmaceuticals has obtained hints of such results from "Paris," its pathways analysis platform that includes literature, data analysis, and pathway visualization tools. Millennium scientist George Mulligan recently used Paris to study patient responses to the cancer drug Velcade. Mulligan sifted through "an avalanche of data" — microarray results on tens of thousands of genes from about 50 patients — to understand why some patients on the drug fared better than others. In a comparison of patients with different responses, Paris revealed physiologically important sets of genes that are turned off or on. Most importantly, the proteasome complex — Velcade's cellular site of action — appeared to be working at different levels in the patients.

In another approach to answering this question, Millennium is evaluating a subset of 30 genes that may indicateof response. "The idea that there might be a genetic component to response [to Velcade] evolved with the understanding of the genome and microarray technology," says the drug's developer, Millennium's Julian Adams. "That was not in my consciousness when this was begun."

Beyond the Genome

Whether systems biology lives up to its promise, everyone wants to know when the fruits of the genome will be translated into new and better medicines. Signs of progress are subtle, but growing. "Everything in our pipeline is genomics-based," says Human Genome Sciences CEO Bill Haseltine. He means genomics in the new sense — woven in with the rest of drug discovery and development. Other companies hope to follow suit, from Celera to big pharmaceuticals. GlaxoSmithKline still has a heart-disease drug in Phase II trials that sprang from Smith-KlineBeecham's groundbreaking 1993 deal with Human Genome Sciences — one of the few genomics-derived drugs that hasn't been dropped in early trials.

Even if progress is too incremental for Wall Street, the future of genome-based science and medicine is wondrous. "Clearly, we have not peaked in appreciating the true value of the genome projects," says Tom Cech, president of the Howard Hughes Medical Institute.

Fifty years from now, we may be looking back nostalgically at the genome revolution, just as we have celebrated the 50th anniversary of the double helix. Whatever breakthroughs lie ahead, they will owe a profound debt to this pair of historic feats. ●