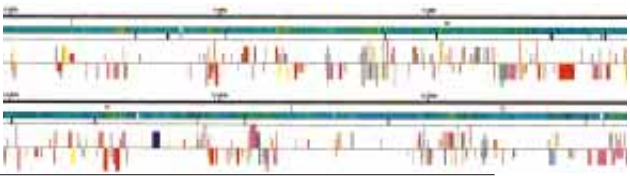
THE BIOINFORMATICS GOLD

A $300-million industry has emerged around turning raw genome data into knowledge for making new drugs

by Ken Howard



"Plastics." When a family friend whispered this word to Dustin Hoffman's character in the 1967 film *The Graduate,* he was advocating not just a novel career choice but an entirely different way of life. If that movie were made today, in the age of the deciphering of the human genome, the magic word might well be "bioinformatics."

Corporate and government-led scientists have already compiled the three gigabytes of paired A's, C's, T's and G's that spell out the human genetic code—a quantity of information that could fill more than 2,000 standard computer diskettes. But that is just the initial trickle of the flood of information to be tapped from the human genome. Researchers are generating gigantic databases containing the details of when and in which tissues of the body various genes are turned on, the shapes of the proteins the genes encode, how the proteins interact with one another and the role those interactions play in disease. Add to the mix the data pouring in about the genomes of so-called model organisms such as fruit flies and mice [see "The 'Other' Genomes," on page 53], and you have what Gene Myers, Jr., vice president of informatics research at Celera Genomics in Rockville, Md., calls "a tsunami of information." The new discipline of bioinformatics—a marriage between computer science and biology—seeks to make sense of it all. In so doing, it is destined to change the face of biomedicine.

"For the next two to three years, the amount of information will be phenomenal, and everyone will be overwhelmed by it," Myers predicts. "The race and competition will be who can mine it best. There will be such a wealth of riches."

A whole host of companies are vying for their share of the gold. Jason Reed of the investment banking firm Oscar Gruss & Son in New York City estimates that bioinformatics could be a $2-billion business within five years. He has compiled information on more than 50 private and publicly traded companies that offer bioinformatics products and services. These companies plug into the effort at various points: collecting and storing data, searching databases, and interpreting the data. Most sell access to their information to pharmaceutical and biotechnology companies for a hefty subscription price that can run into the millions of dollars.

The reason drug companies are so willing to line up and pay for such services—or to develop their own expensive resources in-house—is that bioinformatics offers the prospect of finding better drug targets earlier in the drug development process. This efficiency could trim the number of potential therapeutics moving through a company's clinical testing pipeline, significantly decreasing overall costs. It could also create extra profits for drug companies by whittling the time it takes to research and develop a drug, thus lengthening the time a drug is on the market before its patent expires.

"Assume I'm a pharmaceutical company and somebody can get [my] drug to the market one year sooner," explains Stelios Papadopoulos, managing director of health care at the New York investment banking firm SG Cowen. "It could mean you could grab maybe $500 million in sales you would not have recovered."

Before any financial windfalls can occur, however, bioinformatics companies must contend with the current plethora of genomic data while constantly refining their technology, research approaches and business models. They must also focus on the real challenge and opportunity—finding out how all the shards of information relate to one another and making sense of the big picture.

"Methods have evolved to the point that you can generate lots of information," comments Michael R. Fannon, vice president and chief information officer of Human Genome Sciences, also in Rockville. "But we don't know how important that information is."

Divining that importance is the job of bioinformatics. The field got its start in the early 1980s with a database called GenBank, which was originated by the U.S. Department of Energy to hold the short stretches of DNA sequence that scientists were just beginning to obtain

SCIENCE, VOL. 287, NO. 5461; MARCH 24, 2000 (*above and opposite page*)

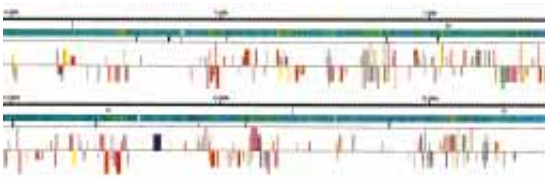
from a range of organisms. In the early days of GenBank a roomful of technicians sat at keyboards consisting of only the four letters A, C, T and G, tediously entering the DNA-sequence information published in academic journals. As the years went on, new protocols enabled researchers to dial up GenBank and dump in their sequence data directly, and the administration of GenBank was transferred to the National Institutes of Health's National Center for Biotechnology Information (NCBI). After the advent of the World Wide Web, researchers could access the data in GenBank for free from around the globe.

"The race and competition will be who can mine [the data] best. There will be such a wealth of riches."

Once the Human Genome Project (HGP) officially got off the ground in 1990, the volume of DNA-sequence data in GenBank began to grow exponentially. With the introduction in the 1990s of high-throughput sequencing—an approach using robotics, automated DNA-sequencing machines and computers—additions to GenBank skyrocketed. GenBank held the sequence data on more than seven billion units of DNA as this issue of *Scientific American* went to press.

Around the time the HGP was taking off, private companies started parallel sequencing projects and established huge proprietary databases of their own. Today companies such as Incyte Genomics in Palo Alto, Calif., can determine the sequence of approximately 20 million DNA base pairs in just one day. And Celera Genomics—the sequencing powerhouse that announced in April that it had completed a rough draft of the human genome [see "The Human Genome Business Today," on page 50]—says that it has 50 terabytes of data storage. That's equivalent to roughly 80,000 compact discs, which in their plastic cases would take up almost half a mile of shelf space.

But GenBank and its corporate cousins are only part of the bioinformatics picture. Other public and private databases contain information on gene expression (when and where genes are turned on), tiny genetic differences among individuals called single-nucleotide polymorphisms (SNPs), the structures of various proteins, and maps of how proteins interact [see "Beyond the Human Genome," on page 64].



Mixing and Matching

One of the most basic operations in bioinformatics involves searching for similarities, or homologies, between a newly sequenced piece of DNA and previously sequenced DNA segments from various organisms. Finding near-matches allows researchers to predict the type of protein the new sequence encodes. This not only yields leads for drug targets early in drug development but also weeds out many targets that would have turned out to be dead ends.

A popular set of software programs for comparing DNA sequences is BLAST (for Basic Local Alignment Search Tool), which first emerged in 1990. BLAST is part of a suite of DNA- and protein-sequence search tools accessible in various customized versions from many database providers or directly through NCBI. NCBI also offers Entrez, a so-called meta-search tool that covers most of NCBI's databases, including those housing three-

GENETIC DATA are the stuff of bioinformatics, which can be likened to looking for a needle in a haystack. In the fanciful example at the left, the needle is the word "DOG" buried amid a sequence of thousands of A's, C's, T's and G's, the four units that make up DNA. But bioinformatics also involves comparing genes from various organisms: the other illustrations on this page and on the preceding one are maps of fruit fly chromosomes alongside bar codes showing regions where the fly's genes are similar to those of others.

SLIM FILMS

THE MAJOR PLAYERS

Lion Bioscience

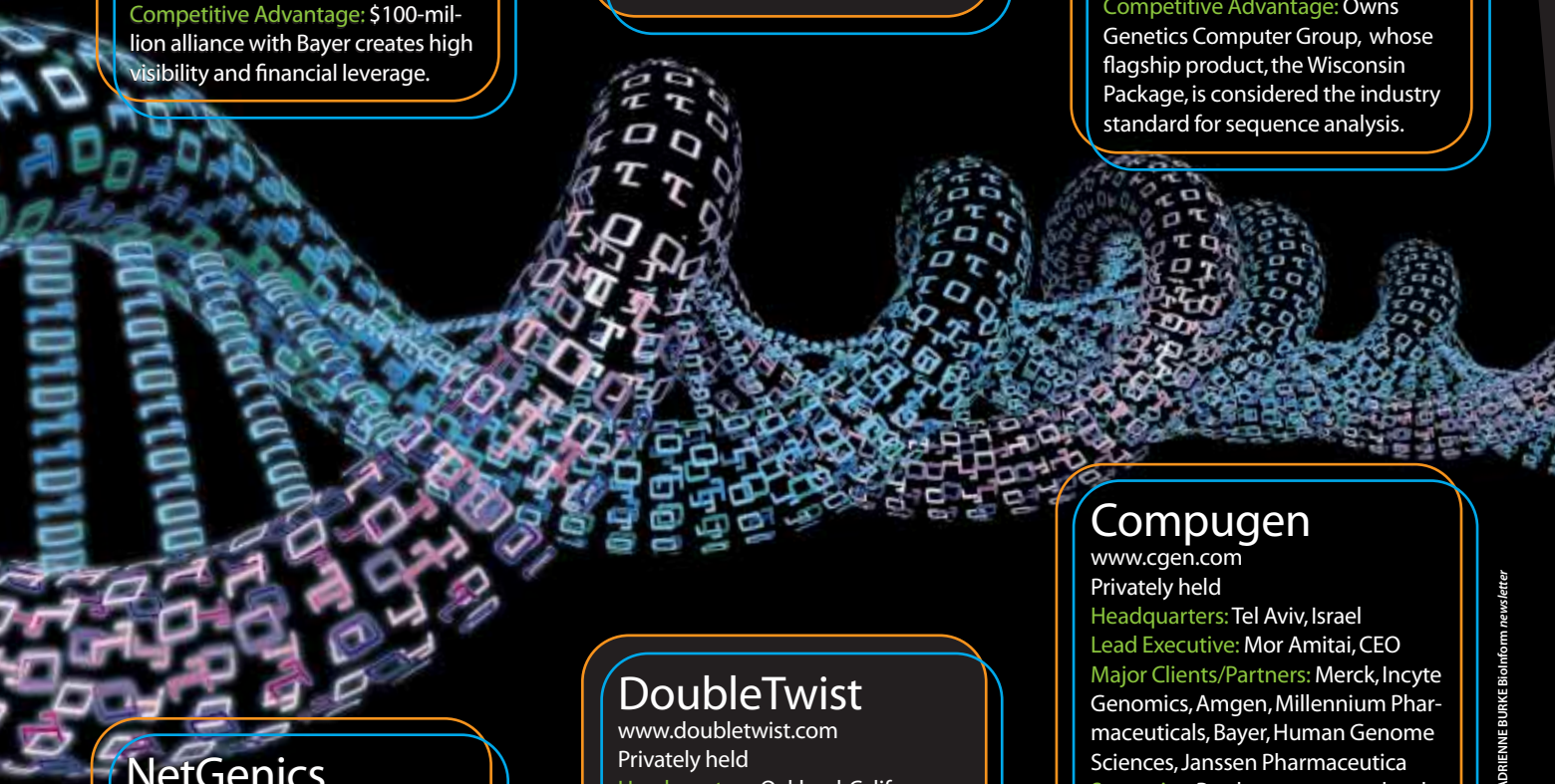
www.lionbioscience.com
Privately held
Headquarters: Heidelberg, Germany
Lead Executive: Friedrich von Bohlen, CEO
Major Clients/Partners: Bayer, Aventis, Pharmacia
Strategy: Provide enterprise-wide bioinformatics systems and services.
Financing This Year: None
Key Challenges: Continuing to penetrate large to midsize biotechnology and pharmaceutical client base; replicating its success with Bayer.
Competitive Advantage: $100-million alliance with Bayer creates high visibility and financial leverage.

InforMax

www.informaxinc.com
Privately Held
Headquarters: Bethesda, Md.
Lead Executive: Alex Titomirov, CEO
Major Clients/Partners: Products used by 19 drug companies
Strategy: Provide desktop and enterprise-wide bioinformatics tools.
Financing This Year: None
Key Challenge: Evolving business into enterprise-wide systems.
Competitive Advantage: High market penetration with desktop line of bioinformatics tools.

Oxford Molecular Group

www.oxmol.co.uk
Stock Symbol: OMG (London)
Headquarters: Oxford, England
Lead Executive: N. Douglas Brown, chairman
Major Clients/Partners: Novartis, Glaxo Wellcome, Merck, Pfizer, Smith-Kline Beecham, Abbott Laboratories
Strategy: Provide broad range of drug-discovery research software and services.
Financing This Year: None
Key Challenge: Expanding business into more enterprise-wide products and services.
Competitive Advantage: Owns Genetics Computer Group, whose flagship product, the Wisconsin Package, is considered the industry standard for sequence analysis.



NetGenics

www.netgenics.com
Privately held
Headquarters: Cleveland, Ohio
Lead Executive: Manuel J. Glynias, president and CEO
Major Clients/Partners: Abbott Laboratories, Aventis, IBM
Strategy: Provide enterprise-wide bioinformatics systems and services.
Financing This Year: $21.3 million
Key Challenge: Continuing to penetrate large and midsize biotechnology and pharmaceutical client base.
Competitive Advantages: Well funded and has relationships with large pharmaceutical companies.

DoubleTwist

www.doubletwist.com
Privately held
Headquarters: Oakland, Calif.
Lead Executive: John Couch, president and CEO
Major Clients/Partners: Derwent Information, Clontech Laboratories, Myriad Genetics, AlphaGene, University of Pennsylvania
Strategy: Provide on-line access to a variety of bioinformatics tools and databases.
Financing This Year: $37 million
Key Challenges: Providing unique proprietary tools and attracting enough customers to support an Internet-portal business model.
Competitive Advantage: High visibility and potentially large market.

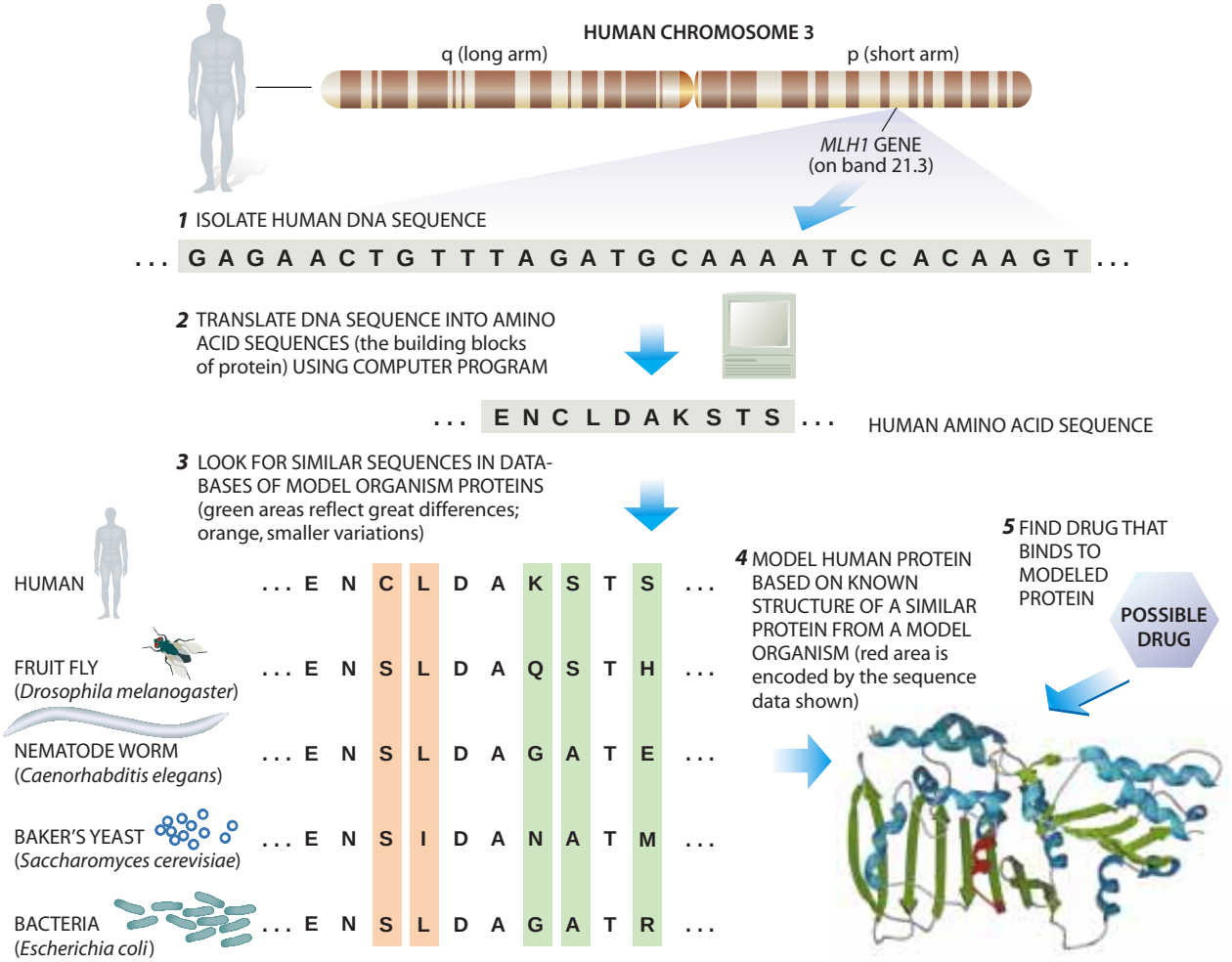
Compugen

www.cgen.com
Privately held
Headquarters: Tel Aviv, Israel
Lead Executive: Mor Amitai, CEO
Major Clients/Partners: Merck, Incyte Genomics, Amgen, Millennium Pharmaceuticals, Bayer, Human Genome Sciences, Janssen Pharmaceutica
Strategies: Produce computer hardware and software to accelerate bioinformatics algorithms; engage in gene discovery and drug development; offer bioinformatics tools via Internet portal.
Financing This Year: None
Key Challenges: Evolving business model to drug discovery; expanding product lines; developing Internet-portal business model.
Competitive Advantages: One of the first companies to develop specialized bioinformatics tools, giving it expertise in data mining. Has a stable of proprietary biological data for use in developing drug targets.

SLIM FILMS; SOURCES: THE COMPANIES LISTED HERE; JASON REED *Oscar Gruss & Son*; ADRIENNE BURKE BioInform *newsletter*

Using Bioinformatics to Find Drug Targets

By looking for genes in model organisms that are similar to a given human gene, researchers can learn about the protein the human gene encodes and search for drugs to block it. The *MLH1* gene, which is associated with colon cancer in humans, is used in this example.



LAURIE GRACE, WITH ASSISTANCE FROM MARK GERSTEIN AND PAT FLEMING *Yale University* AND DAVID WHEELER AND JENNIFER VYSKOCIL *NCBI*

dimensional protein structures, the complete genomes of organisms such as yeast, and references to scientific journals that back up the database entries.

An early example of the utility of bioinformatics is cathepsin K, an enzyme that might turn out to be an important target for treating osteoporosis, a crippling disease caused by the breakdown of bone. In 1993 researchers at SmithKline Beecham, based in Philadelphia, asked scientists at Human Genome Sciences to help them analyze some genetic material they had isolated from the osteoclast cells of people with bone tumors. (Osteoclasts are cells that break down bone in the normal course of bone replenishment; they are thought to be overactive in individuals with osteoporosis.)

Human Genome Sciences scientists sequenced the sample and conducted database homology searches to look for matches would give them a clue to the proteins that the sample's gene sequences encoded. Once they found near-matches for the sequences, they carried out further analyses and discovered that one sequence in particular was overexpressed by the osteoclast cells and that it matched those of a previously identified class of molecules: cathepsins.

For SmithKline Beecham, that exercise in bioinformatics yielded in just weeks a promising drug target that standard laboratory experiments could not have found without years and a pinch of luck. Company researchers are now trying to find a potential drug that blocks the cathepsin K target. Searches for compounds that bind to and have the desired effect on drug targets still take place mainly in a biochemist's traditional "wet" lab, where evaluations for activity, toxicity and absorption can take years. But with new bioinformatics tools and growing amounts of data on protein structures and biomolecular pathways, some researchers say, this aspect of drug development will also shift to computers, in what they term "in silico" biology [see "Forget In Vitro—Now It's 'In Silico,'" on next page].

It all adds up to good days ahead for bioinformatics, which many assert holds the real promise of genomics. "Genomics without bioinformatics will not have much of a payoff," states Roland Somogyi, former director of neurobiology at Incyte Genomics who is now at Molecular Mining in Kingston, Ontario.

Forget In Vitro—Now It's "In Silico"

With the human genome essentially complete, futurists are suggesting that scientists will soon be able to use bioinformatics to model the astronomical number of biochemical reactions that add up to human life. Ken Howard discusses the possibility of such "in silico" biology with complexity expert Stuart A. Kauffman, an external professor at the Santa Fe Institute in New Mexico who is also founder and chief scientific officer of Bios Group in Santa Fe.

Q: What is the promise of bioinformatics and "in silico" biology?
Kauffman: We're entitled to think of the 100,000 genes in a human cell as some kind of parallel-processing chemical computer in which genes are continuously turning one another on and off in some vastly complex network of interaction. Cell-signaling pathways are linked to genetic regulatory pathways in ways we're just beginning to unscramble. The most enormous bioinformatics project in front of us is unscrambling this regulatory network, which controls cell development from the fertilized egg to the adult.

Q: What is the payoff?
Kauffman: We will know which gene to perturb—or which sequences of genes to perturb, and in what order—to guide a cancer cell to nonmalignant behavior or to apoptosis [programmed cell death]. Or to guide the regeneration of some tissue, so that if you happen to have lost half of your pancreas we'll be able to regenerate your pancreas. Or we'll be able to regenerate the beta cells in people who have diabetes.

Q: What needs to happen to achieve that goal?
Kauffman: It's not going to be merely bioinformatics—there has to be a marriage between new kinds of mathematical tools. Those tools will in general suggest plausible alternative circuits for bits and pieces of the [cell's] regulatory network. And then we're going to have to marry that with new kinds of experiments to work out what the circuitry in cells actually is. And bioinformatics has to be expanded to include experimental design. What we're going to get out of each of these pieces of bioinformatics is hypotheses that need to be tested.

Q: What challenges lie ahead?
Kauffman: Suppose I pick out 10 genes that I know regulate one another, and I try to build a circuit about their behavior. It's a perfectly fine thing, and we should do it. But the downside is the following: those 10 genes have inputs from other genes outside that circuit. So you're taking a little chunk of the circuitry that's embedded in a much larger circuit with thousands of genes in it. You're trying to figure out the behavior of that circuit when you do not know the outside genes it impacted. And that makes that direct approach hard, because you never know what the other inputs are. We've known for years what every neuron is in the lobster gastric ganglia [a nerve bundle going to the animal's digestive

Michael N. Liebman, head of computational biology at Roche Bioscience in Palo Alto, agrees. "Genomics is not the paradigm shift; it's understanding how to use it that is the paradigm shift," he asserts. "In bioinformatics, we're at the beginning of the revolution."

The revolution involves many different players, each with a different strategy. Some bioinformatics companies cater to large users, aiming their products and services at genomics, biotechnology and pharmaceutical companies by creating custom software and offering consulting services. Lion Bioscience, based in Heidelberg, Germany, has been particularly successful at selling "enterprise-wide" bioinformatics tools and services. Its $100-million agreement with Bayer to build and manage a bioinformatics capability across all of Bayer's divisions was at press time the industry's largest such deal.

Other firms target small or academic users. Web businesses such as Oakland, Calif.–based DoubleTwist and eBioinformatics, which is headquartered in Pleasanton, Calif., offer one-stop Internet shopping. These on-line portals allow users to access various types of databases and use software to manipulate the data.

In May, DoubleTwist scientists announced they had used their technology to determine that the number of genes in the human genome is roughly 105,000, although they said the final count would probably come in at 100,000. For those who would rather have the software behind their own security firewalls, Informax in Rockville, Oxford Molecular Group in England, and others sell shrink-wrapped products.

Making Connections

Large pharmaceutical companies—"big pharma"—have also sought to leverage their genomics efforts with in-house bioinformatics investments. Many have established entire departments to integrate and service computer software and facilitate database access across multiple departments, including new product development, formulation, toxicology and clinical testing. The old model of drug development often compartmentalized these functions, ghettoizing data that might have been useful to other researchers. Bioinformatics allows researchers across a company to see the same thing while still manipulating the data individually.

In addition to making drug discovery more efficient, in-house bioinformatics can also save drug companies money in software support. Glaxo Wellcome in Research Triangle Park, N.C., is replacing individual packages used by various investigators and departments to access and manipulate databases with a single software platform. Robin M. DeMent, U.S. director of bioinformatics at Glaxo Wellcome, estimates that this will save approximately $800,000 in staffing support over a three- to five-year period.

To integrate bioinformatics throughout their companies, pharmaceutical giants also forge strategic alliances, enter into licensing agreements and acquire smaller biotechnology companies. Using partners and vendors not only allows big pharma to fill in the gaps in its bioinformatics capabilities but also gives

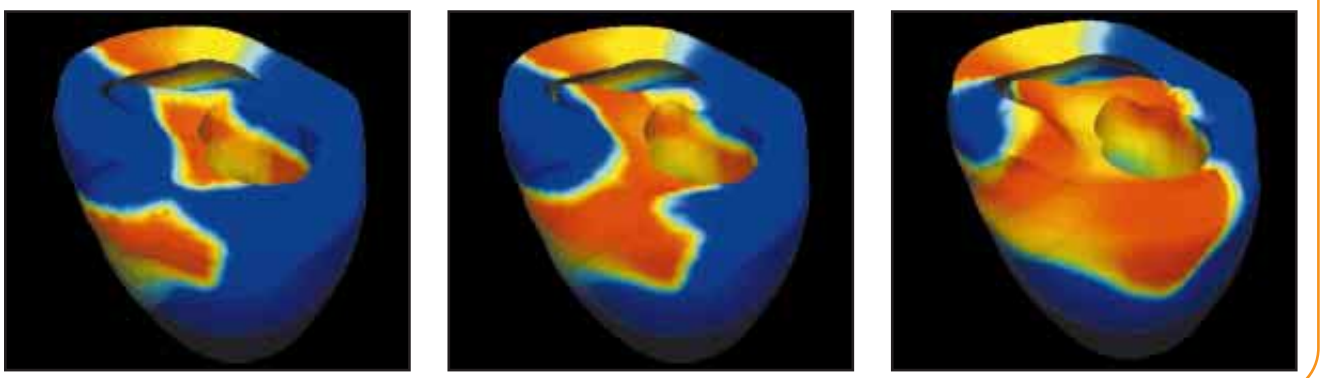
system], what all the synaptic connections are and what the neurotransmitters are. You have maybe 13 or 20 neurons in the ganglion, and you still can't figure out the behavior of the ganglion. So no mathematician would ever think that understanding a system with 13 variables is going to be an easy thing to do. And [now with the human genome] we want to do it with 100,000 variables. Let me define the state of the network as the current on-and-off values of all 100,000 genes. So how many states are there? Well, there's two possibilities for gene one and two possibilities for gene two and so on, so there's $2^{100,000}$ states, which is roughly $10^{30,000}$. So even if we treat genes as on or off—which is false because they show graded levels of activity—that's $10^{30,000}$ possible states. It is mind-boggling because the number of particles in the known universe is 10^{80}.

Q: Where are we in terms of that problem?
Kauffman: We're at the very beginning, but there's going to be a day when somebody comes in with cancer, and we diagnose it with accuracy not just on the morphology of the cancer cell but by looking at the detailed patterns of gene expression and protein-binding activities in that cell.

Q: How far away is this? One year or 200 years?
Kauffman: The tools will mature within the next 10 to 12 years, and then we'll really start making progress, getting the circuitry for big chunks of the genome and actually understanding how it works. I think 30 to 40 years from now we will have solved major chunks.

For the full transcript of Ken Howard's interview with Stuart A. Kauffman, visit the Scientific American Web site at www.sciam.com/interview/2000/060500/kauffman

COMPUTER MODEL OF A HEART in fibrillation shows waves of uncoordinated electrical activity sweeping the organ. The model is based on changes in the expression of four genes whose function is altered during chronic heart failure.



it the mobility to adapt new technologies as they come onto the market rather than constantly overhauling its own systems. "If a pharmaceutical company had a large enough research budget, they could do it all themselves," Somogyi says. "But it's also a question of culture. The field benefits as a whole by providing different businesses with different roles with room to overlap."

Occupying some of that overlap—in resources, products and market capitalization—are companies such as Human Genome Sciences, Celera and Incyte. They straddle the terrain between big pharma and the data integration and mining offered by specialist companies. They have also quickly seized on the degree of automation that bioinformatics has brought to biology.

But with all this variety comes the potential for miscommunication. Getting various databases to talk to one another—what is called interoperability—is becoming more and more key as users flit among them to fulfill their needs. An obvious solution would be annotation—tagging data with names that are cross-referenced across databases and naming systems. This has worked to a degree. "We've been successful in bringing databases together by annotation: database A to database B, B to C, C to D," explains Liebman of Roche Bioscience. "But annotation in A may change, and by the time you get down to D the references may not have changed, especially with a constant stream of new data." He points out that this problem becomes more acute as the understanding of the biology and the ability to conduct computational analysis becomes more sophisticated. "We're just starting to identify complexities in these queries, and how we store data becomes critical in the types of questions we can ask," he states.

Systematic improvements will help, but progress—and ultimately profit—still relies on the ingenuity of the end user, according to David J. Lipman, director of NCBI. "It's about brainware," he says, "not hardware or software." SA

KEN HOWARD is a freelance science writer based in New York City.

Further Information

TRENDS IN COMMERCIAL BIOINFORMATICS. A report issued March 13, 2000, by Jason Reed of Oscar Gruss & Son. To obtain a free copy, log onto www.oscargruss.com/reports.htm

USING BIOINFORMATICS IN GENE AND DRUG DISCOVERY. D. B. Searls in *Drug Discovery Today,* Vol. 5, No. 4, pages 135–143; April 2000.

BioInform, a biweekly newsletter on the subject of bioinformatics, can be accessed at www.bioinform.com

To access the bioinformatics databases maintained by the National Center for Biotechnology Information (NCBI), go to www.ncbi.nlm.nih.gov

RAIMOND L. WINSLOW AND DAVID SCOLLAN *Johns Hopkins University* AND PHYSIOME SCIENCES, PRINCETON, N.J.