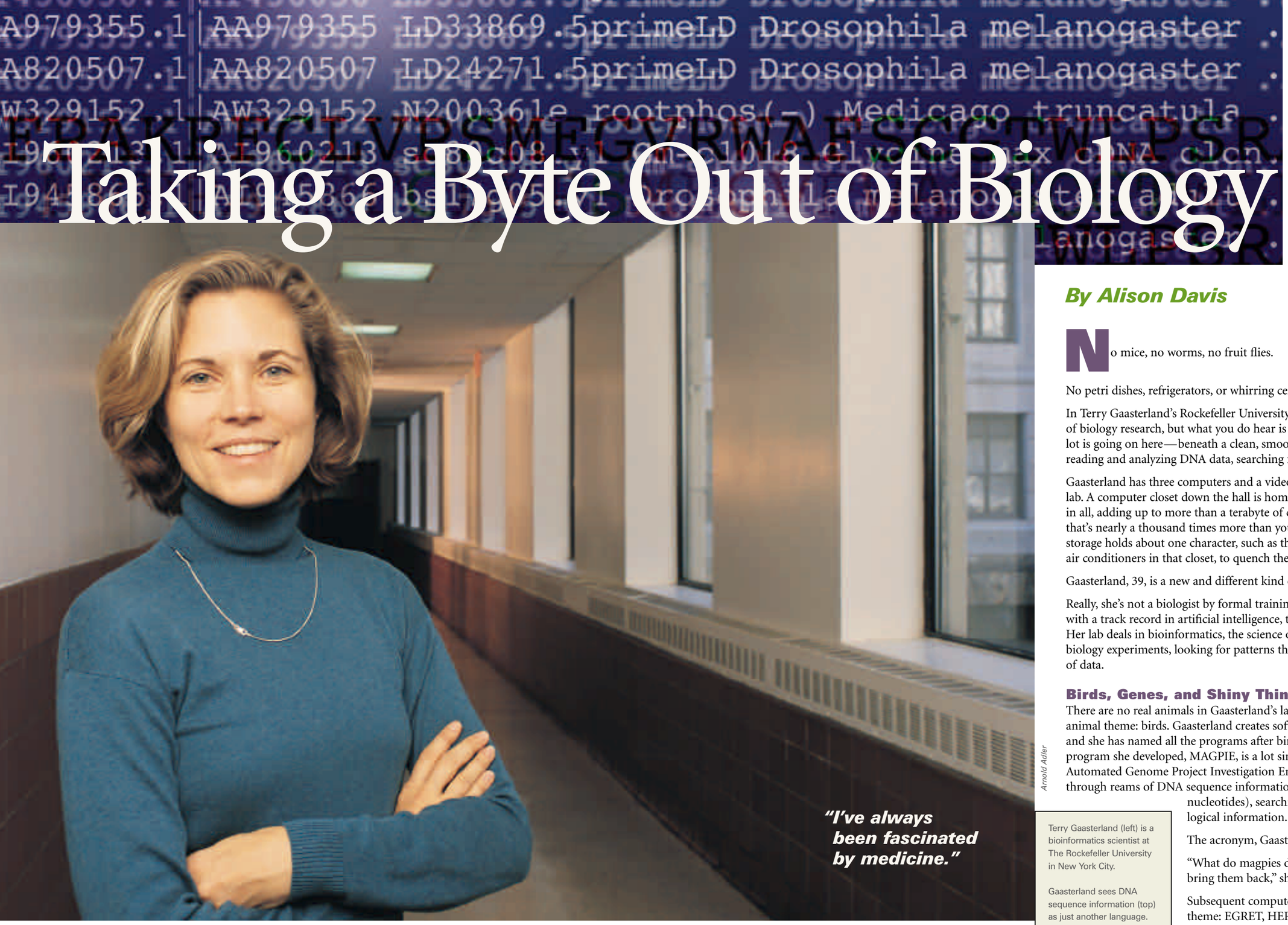
Taking a Byte Out of Biology



Terry Gaasterland

Arnold Adler

"I've always been fascinated by medicine."

Terry Gaasterland (left) is a bioinformatics scientist at The Rockefeller University in New York City.

Gaasterland sees DNA sequence information (top) as just another language.

By Alison Davis

No mice, no worms, no fruit flies.

No petri dishes, refrigerators, or whirring centrifuges.

In Terry Gaasterland's Rockefeller University lab, you do not hear these routine sounds of biology research, but what you do hear is a quiet, constant hum of electronics. Yet a lot is going on here—beneath a clean, smooth surface, her computers are relentlessly reading and analyzing DNA data, searching for genes and figuring out what they do.

Gaasterland has three computers and a video monitor in her uptown Manhattan office-lab. A computer closet down the hall is home to stacks upon stacks of hard drives, 300 in all, adding up to more than a terabyte of computer disk space. If you didn't know, that's nearly a thousand times more than your home PC. A single byte of computer storage holds about one character, such as the letter "a." Oh, and there are two 10-ton air conditioners in that closet, to quench the intense heat produced by the machines.

Gaasterland, 39, is a new and different kind of biologist.

Really, she's not a biologist by formal training—Gaasterland is a computer scientist, with a track record in artificial intelligence, the art of training computers to "think." Her lab deals in bioinformatics, the science of piecing together data from thousands of biology experiments, looking for patterns that themselves are a new and different kind of data.

Birds, Genes, and Shiny Things

There are no real animals in Gaasterland's lab, but sooner or later you do notice an animal theme: birds. Gaasterland creates software to analyze experimental genetic data, and she has named all the programs after birds. The acronym for the first computer program she developed, MAGPIE, is a lot simpler than its real name, Multipurpose Automated Genome Project Investigation Environment. MAGPIE's job is to comb through reams of DNA sequence information (millions of DNA "letters" called nucleotides), searching for patterns that signify hidden biological information.

The acronym, Gaasterland explains, fits perfectly.

"What do magpies do? They go and collect shiny things and bring them back," she says.

Subsequent computer programs have perpetuated this bird theme: EGRET, HERON, SANDPIPER. The names of each of

these programs spell out different computer-performed tasks, each aiming to discern multidimensional meaning from two-dimensional genomic information, which resembles words on a page.

To bioinformatics scientists like Gaasterland, bits of information and patterns among the data are "shiny things." Her far-reaching goal, and that of other computer scientists who concentrate on biological mysteries, is to get computers to apply logic to biology, which by nature is incredibly complicated. It will take many years, she says, but ultimately future versions of computer programs like MAGPIE will swallow huge amounts of genetic data and spit out predictions about how biology works.

Same or Different?

One project Gaasterland is currently working on involves analyzing the gene readout data from the lungs of smokers and non-smokers. She is trying to figure out why some people get lung cancer and some don't, so she wrote a computer program to analyze the two sets of data, asking an elementary question: What's the same and what's different?

Gaasterland's experiments require a terabyte of computer disk space—nearly a thousand times more than a home computer.

What Gaasterland found is that some of the differences can be very, very tiny, showing up only when you look at gene activity in individual cells and ask which genes are turning on or off. Genes that are turning on are getting ready to make proteins. These dynamic changes in gene readout, or "expression," can be so subtle that they may not themselves lead to any noticeable change in appearance and/or behavior—something scientists call a "phenotype." Lots of little changes can add up, though, to make a new phenotype.

Here's how it works. The gene readout data that feeds into Gaasterland's computer comes from biologists all over—to date, a few hundred researchers. The data is usually posted on the researchers' Web sites, from where she can download it. In setting up a collaboration, Gaasterland first talks to the scientists, asking them questions about what they are trying to discover. The researchers hand over their experimental results, she says, in return for Gaasterland's promise to work with them to figure out what new information they can pull out of the data. She then uses computer logic to come up with new ways to "query" the data.

But it's not just a matter of "smoothing out" the edges with standard statistical tools, Gaasterland explains. Rather, she applies principles of artificial intelligence to get the data to reveal its hidden secrets, and importantly, to pose new questions to researchers.

Gaasterland's experiments have the potential to make new discoveries by linking information from different fields of study. For example, she explains, she has huge bodies of data from scientists studying heart disease, obesity, and diabetes. In combining and analyzing these different data sets, Gaasterland and her students and postdoctoral researchers essentially "connect the dots." It's exciting, she says, because she has a bird's-eye view that individual researchers—each with their own data alone—do not have.

An Eye on Science

How does a computer scientist learn biology in the first place? Gaasterland says that half the fun is picking it up along the way. She concedes to having on hand a few "consultants" (computationally minded biologist friends) who she can freely ask to fill in gaps in her knowledge about cells, organisms, and how they function.

Despite gaps in her detailed knowledge of biology, however, Gaasterland has never had a lack of enthusiasm for biomedical science.

"My schooling has been totally and completely computers," she says, "but I've always been fascinated by medicine."

"My father introduced me at an early age to the idea of studying animals to figure out how to treat people," Gaasterland says. Her father, Douglas Gaasterland, is a physician-scientist at Georgetown University in Washington, DC. When Terry was growing up, he had a research lab at the National Institutes of Health in Bethesda, Maryland, where he used lasers to study and treat glaucoma in monkeys.

"At 5 years old, I was used to seeing eyeballs in the lab fridge," she laughs.

She grew up with science, but math has always been front and center in Gaasterland's life. She took algebra in 7th grade and completed calculus by 10th grade. During her junior and senior years in high school, she was done with classwork by lunchtime, leaving afternoons for ballet classes and evenings for differential equations at the local community college. On Saturdays, she traveled an hour north to Baltimore, where she took a neurology course at The Johns Hopkins University.

James C. Leupold, U.S. Fish and Wildlife Service

Gaasterland's gene-analyzing computer programs, such as MAGPIE and EGRET, are all named after birds.

Today, Gaasterland breaks the stereotypic mold of a computer scientist. If you saw her roller-blading with friends in Central Park after dark ("It's safe!" she insists) or hanging out in New York's Soho jazz and blues clubs, her zest for living would be apparent. A perfect Saturday afternoon is spent strolling around New York's museums and art galleries, she says, "where you can find truly cutting-edge art."

Strolling is the operative word, since she doesn't own a car anymore. Gaasterland enjoys watching Manhattan life by walking its vibrant streets or hopping around in cabs. Liking city apartment living so much surprises even her. "I thought living in a crowded apartment would be awful," Gaasterland says, remembering her childhood days in a quiet, tree-lined Washington, DC suburb.

Gaasterland is passionate about Manhattan, but also about the marriage of computers and biology. She is on a mission to train computers to help scientists understand how genes mastermind the precision functioning of organisms ranging from bacteria to people. So much information is hidden in our genes, Gaasterland says, and researchers simply need to learn how to interpret it.

"DNA is just another language."

Shaping Up

DNA is indeed the language of our lives, spelling gene "words." Genes instruct the body how to make worker molecules called proteins, which combine in wondrous ways to allow us to think and to sense the world around us.

But while DNA represents two-dimensional information (akin to words on a page), proteins are three-dimensional things. Each protein has a characteristic shape that suits it to its unique biological task. A protein in the wrong shape can be a problem, sometimes causing illness and disease. In order to understand how misshapen proteins affect our health and to figure out ways to mimic or block protein shapes to fight disease, scientists need to see up close what proteins actually look like. To do this, researchers called structural biologists rely on high-energy physics techniques. Such researchers blast X-rays at protein samples and, based on how the X-rays are scattered, the scientists can piece together the shape of a protein.

Gaasterland is getting computers to help with that problem, too. She is part of an organized effort, called structural genomics, that aims to predict protein shapes from their DNA (genomic) sequence. Gaasterland is a member of the New York Structural Genomics Consortium, which gets research funding from the National Institute of General Medical Sciences.

"Three-dimensional properties of proteins are lurking in two-dimensional sequences," Gaasterland says, describing the prevalence of sequence "signatures" that point to telltale genetic directions for making recurring protein shapes.

By comparing gene properties and examining the DNA information from creatures throughout the vast biological kingdom, Gaasterland can figure out which characteristics have proven indispensable for the proper functioning of organisms spanning millions of years of evolutionary time. For example, she says, certain pairs of amino acids, the building blocks of proteins, always change together across species. Since amino acids fit together much like LEGO® pieces, recognizing recurring pairs of them hints that such molecular duos translate three-dimensionally into signature folds or bends in protein shapes.

According to Gaasterland, her role in the structural genomics effort is in finding so-called protein targets—"families" of proteins whose three-dimensional structures are likely to be similar and can be used as benchmarks in predicting the structures and functions of other proteins. The goal of structural genomics is to find the three-dimensional shapes of all the parts (proteins and other large molecules) in a cell.

Terry Gaasterland

Up close, a protein molecule (right) has all sorts of twists and turns.

A Full House

To begin to understand the concept of selecting cellular targets, consider a metaphor of a cell as a house full of contents and activity. By taking an inventory of what's inside a cell, or a house, and observing when and where things happen, you can make certain assumptions about where to look for new information. These can serve as new "targets."

In a fictitious house-cell, structural biologists might piece together the molecular parts that make up the furniture and appliances, revealing the identity of the couches, chairs, and refrigerator, for example. In a real cell, the "furniture" might be the cell's protein scaffolding, and the appliances mini-molecular machines that generate energy for the cell. Such an effort generates a "parts list" for the interior of the house, or the inside of a cell.

Context is key—you can often infer the function of an object by observing other nearby objects and checking out the conditions under which they are used. For example, in a house, a room with a flat surface and two chairs could be either a dining room or an office—or both, at different times of day. If the room is used in the late evening or very early in the morning, the surface is more likely to be used as a desk than as a formal dining table. A search for other contextual clues, like a bookshelf, would strengthen the assumption that this room is used as an office.

Central Park is one of Gaasterland's favorite roller-blading spots.

Along with the placement of things, activity can also point to possible function.

"If you see the lights go on in the garage at 7:00 p.m., you can infer that the people are doing something in there," Gaasterland says. "That's a new target for study."

Further looking may uncover details about what exactly is going on in the garage in the early evening, as would analyzing more contents of the dining room/office.

Proteins, Proteins, Proteins

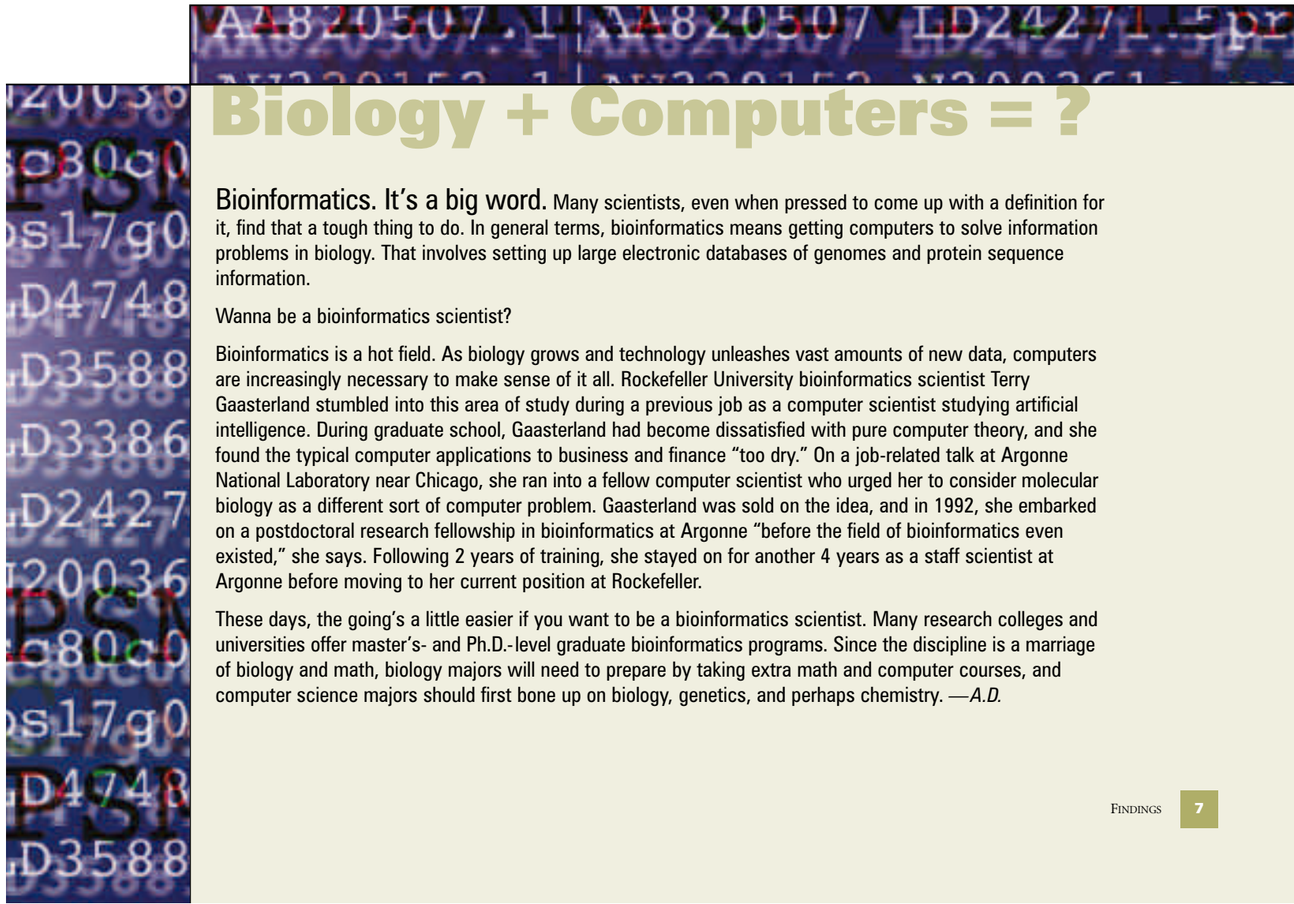
Hidden deep within two-dimensional genomic information are many clues about cell function. Using knowledge in hand to make assumptions and predictions about what is not known can speed the pace of biological discovery, leading to better ways to diagnose and treat disease.

There are lots and lots of proteins we know little about, Gaasterland explains. She estimates that researchers have a hunch about what roughly 50 percent of our protein-making genes do, based upon previous experiments. Another 20 percent can be guessed because they look so much like the genes of other organisms, like fruit flies or mice.

But for the remaining 30 percent, Gaasterland says, "we don't have a clue."

Much like Manhattan, Gaasterland sees a cell as a lively neighborhood, bustling with constant activity. Communications and negotiations between proteins are constantly going on. She is confident that, with their "absolute precision," computers can make sense of the mayhem, using logic to find rules and order hiding in the letters of our genome.

"DNA is just another language," she says. "We are only just beginning to learn how to hear the individual words—let alone listen, understand, and speak." ■



Biology + Computers = ?

Bioinformatics. It's a big word. Many scientists, even when pressed to come up with a definition for it, find that a tough thing to do. In general terms, bioinformatics means getting computers to solve information problems in biology. That involves setting up large electronic databases of genomes and protein sequence information.

Wanna be a bioinformatics scientist?

Bioinformatics is a hot field. As biology grows and technology unleashes vast amounts of new data, computers are increasingly necessary to make sense of it all. Rockefeller University bioinformatics scientist Terry Gaasterland stumbled into this area of study during a previous job as a computer scientist studying artificial intelligence. During graduate school, Gaasterland had become dissatisfied with pure computer theory, and she found the typical computer applications to business and finance "too dry." On a job-related talk at Argonne National Laboratory near Chicago, she ran into a fellow computer scientist who urged her to consider molecular biology as a different sort of computer problem. Gaasterland was sold on the idea, and in 1992, she embarked on a postdoctoral research fellowship in bioinformatics at Argonne "before the field of bioinformatics even existed," she says. Following 2 years of training, she stayed on for another 4 years as a staff scientist at Argonne before moving to her current position at Rockefeller.

These days, the going's a little easier if you want to be a bioinformatics scientist. Many research colleges and universities offer master's- and Ph.D.-level graduate bioinformatics programs. Since the discipline is a marriage of biology and math, biology majors will need to prepare by taking extra math and computer courses, and computer science majors should first bone up on biology, genetics, and perhaps chemistry. —*A.D.*