THE CHRONICLE OF HIGHER EDUCATION
Research & Publishing
From the issue dated June 29, 2001

Learning to Swim in the Rising Tide of Scientific Data

From astronomy to zoology, researchers face an unprecedented wealth of information

By LILA GUTERMAN

Philadelphia

In David S. Roos's lab at the University of Pennsylvania, biologists come and go, stereos play, electronic instruments hum, and scientists confer with one another as they work with microscopes or chemicals in vials. Downstairs, in another arm of Mr. Roos's research complex, the only motion and sound come from the clacking of keys in his computer lab.

ALSO SEE:

Diving Into the Data

Chemists See More Data, but Not the Deluge Experienced by Other Scientists

But everyone here is a biologist, and they're all working toward one goal: to understand a malaria-causing parasite and others like it.

The scientists have a lot to work with. Upstairs, researchers examine samples of human cells infected with malaria or create parasite proteins altered to glow green so they can be more easily tracked. Downstairs, biologists have at least 6,000 genes on the computer to manipulate and search.

What's happening in Mr. Roos's lab is mirrored around the world as scientists learn to navigate a swelling sea of data. Biology is not the only field that has become increas ingly tied to computers. As data flow in from automated processes, many fields of science are developing easily accessible central databases.

"Instead of spending six months doing an experiment which you can then understand in an afternoon when you're done, you can do an experiment in an afternoon and it takes you six months to figure out what you've got," says Mr. Roos.

Using information stored in the databases has saved scientists a vast

amount of time and effort, and what's more, it has allowed them to answer big-picture questions that seemed unapproachable just a few years ago.

But some scientists have avoided the shift and are letting their students or computer-science collaborators take on the waves of information. Other researchers have responded by retraining themselves and changing their methods -- slowly learning how to swim with the tide.

The revolution's precursors go back to the beginnings of modern science. For centuries, scientists have maintained a tradition of sharing their data and know-how -- whether fossil specimens or techniques for making oxygen -- through libraries, research universities, zoos, and natural-history museums.

In the 17th century, the first scientific journals appeared, introducing what is now the most common way for researchers to swap data and results. As the literature burgeoned, an early centralized database sprang out of a concern that scientists could not keep up. Chemical Abstracts Service began selling collected summaries of journal submissions in chemistry in 1907. *The Guinness Book of World Records* has cited it as the world's largest index.

Although some areas of science have used open databases to store information for decades, in recent years they have had to evolve, and other fields have created new ones -- all to keep up with the influx of new information.

Investigators foresee great rewards from pairing automated experimental techniques with the powerful computing capabilities that now link researchers around the world. In biology, scientists predict that discoveries making use of human-genome information will herald a new era of medicine tailored to an individual's genes. In astronomy, data streaming in from automated scans of the night sky are allowing discoveries of ever-farther objects, such as quasars. Atmospheric scientists hope that information collected by satellites will permit ever-better weather prediction and forecasts of climate change. Some neuroscientists envision a day when shared data will allow them to understand the function of every structure in the brain. And Mr. Roos hopes that studying the malaria parasite's genome could lead to new treatments, or even vaccines, for a disease that kills two million people per year.

Making such advances requires an intimate relationship between experiment and computation. When looking for a protein in the parasite that might make a good target for a drug, biologists like Mr. Roos use computer programs (which they often write themselves) to scan genomes for genes that have specific characteristics -- for instance, that produce proteins that appear on a cell's surface or take on a certain three-dimensional shape. By using several such criteria, the scientists can narrow a field of thousands of genes to several dozen.

Then they can use their knowledge of those genes or their relatives to select a few promising candidates and test experimentally whether the genes have attributes that could make them drug targets. Mr. Roos says that without the genome information, scientists would have to study many more genes individually. "Certainly it can save many years of effort" on any one project, he says.

Because they see the value of such work, Mr. Roos says, all biologists in his lab can now write at least rudimentary computer programs to search genome databases.

Researchers are also collaborating more with computer scientists, who are often eager to participate in many scientific areas. And the computer experts sometimes jump ship completely. "There's a definite cohort now of environmental scientists who are converted computer scientists," says Stephen D. Prince, a professor of geography at the University of Maryland at College Park.

Being able to produce data quickly through automated processes, or simply to grab it from established databases, has brought a significant change in culture for scientists. "The traditional way of doing science -- I bring all my data up on the workstation screen and know there's something interesting -- is not going to work," says Alexander Szalay, an astronomer at the Johns Hopkins University. Only computers can cope with and analyze the huge amount of information flowing out of automated sky surveys. "There will be a lot of data that no human eye can see. There's too much of it."

Earth scientists face similar issues, with satellites constantly beaming down data about the land, oceans, and atmosphere. "All these gigabytes are just falling out on the floor with nobody to ever look at them," says Mr. Prince. But he says that problem is vastly better than the alternative -- insufficient data to do research.

In biology, the influx of data has spawned a new approach, which its practitioners call systems biology. Having information about all of the genes or all of the proteins in a cell means that scientists can ask questions of such complexity that they were unfathomable just a few years ago.

Leroy Hood, a founder of the Institute for Systems Biology, in Seattle, describes systems biology by analogy in explaining the private research institute's work. "Suppose biologists wanted to figure out how a car works," he says. "In the past, one would study the ignition, another the transmission, a third the brakes." Instead, in a systems-biology approach, scientists would use automated experimental techniques and computation to list all the car parts and analyze how each is related to one another.

A goal of the work is to be able to write mathematical equations that describe biological processes, such as how a cell receives and

generates signals. "At the end of the day, if you can put together a mathematical model ... you're getting close to understanding how the whole system works," says Thomas D. Pollard, a professor of structural biology at the Salk Institute for Biological Studies, in La Jolla, Calif. In the car analogy, says Mr. Hood, the objective would be to predict the structure of the vehicle from its parts and to guess how it would react to certain situations, such as a foot slammed on the brake.

A similar transition is taking place in astronomy, according to S. George Djorgovski, a professor of astronomy at the California Institute of Technology. "You can start asking different questions," he says. "How can you ask questions about the sky as a whole?" By using large data sets from sky surveys.

In genetics, the surge in data has even given rise to a new type of scientist: a bioinformatics specialist. Those researchers know enough biology and computer science to make sense of the enormous amounts of information.

Several universities have begun or are planning graduate programs or undergraduate majors in the field. This fall, the University of California at San Diego is starting a Ph.D. program headed by Shankar Subramaniam, a professor of bioengineering, chemistry, and biochemistry. He says a dire need exists for such programs and for integrating quantitative biological concepts and techniques into standard courses as well.

Sylvia Spengler, the National Science Foundation's program director for biological databases, agrees. "We need to start requiring statistics [for students in biology], requiring a basic understanding of what databases can and can't do for you, and what software can and can't do for you."

It's not enough, say many scientists, for only informatics specialists to understand the techniques. Because of the importance of the new databases for research, all scientists need exposure to computer programming and statistical methods. James N. Gray, a computer scientist at the Microsoft Corporation who is working to build public databases in astronomy, says that should be no problem for astronomers. "The astronomers think nothing of learning new programming languages and new tools," he says.

Many earth scientists are similarly computer literate, having used or created their own digital data archives for many years, says Gene R. Major, a senior scientist with the National Aeronautics and Space Administration's Global Change Master Directory, a clearinghouse for online data sets. "We were online before the Web came around."

But in biology, which traditionally has placed less emphasis on mathematics, many midcareer scientists are having to retrain to become computer savvy. Jessica Kissinger, a lecturer in biology who works on

the computational side of Mr. Roos's lab, taught herself the statistical methods required to work with databases and the programming needed to set up and run a genomic database. "I've been on the steepest learning curve of my life," she says.

Those seeking to retrain have several options outside of registering for semester-long courses alongside undergraduates. Several universities and organizations run popular hands-on immersion courses that run one to several weeks. Material is available through the Web and in books for scientists to cobble together. And, says Ms. Kissinger, "by doing, you learn a lot."

Others have relied on their collaborators or students for programming or other computer work. That's true even in astronomy. "People of my generation grew up programming computers," says Mr. Djorgovski, who has taught at Caltech since 1987. "I imagine the next generation of graduate students will have to become data-mining experts because that's the technology they'll be using."

Biologists' levels of computer literacy are too low, says Ms. Spengler of the N.S.F. She worries that biologists limit the questions they can ask by learning only one or a few informatics techniques. "People learn how to do something and then they use that tool all the time because that's the only tool they ever learned how to use. I know people who treat screwdrivers that way," she says. "Screwdrivers are wonderful, but they work only for some sorts of things. So if you insist on using them for something they're not made for, like driving nails, they're nowhere near as effective."

Others seem less concerned. "Biologists are smart enough to figure out how to run some user-friendly software," says Salk's Dr. Pollard.

"In all aspects of society, whenever something radical happens, there's always going to be a certain amount of upheaval," says Charles DeLisi, a professor of science and engineering at Boston University. "There's less of that in the academic community, where people are constantly retraining. People are adapting rather rapidly."

Some biologists play down the changes and instead view genomics and its offshoots as little more than the latest hot trend.

"We've experienced a number of waves of technological innovation in biology over the past 30 years, and each of these new technologies was incorporated into the armamentarium of the practicing biologist," says Robert A. Weinberg, a professor of biology at the Massachusetts Institute of Technology. "I think this is going to be another one of those waves."

The new approaches, however, will not replace the reductionist strain in biology of studying one or a few molecules at a time, he predicts.

"There are still a lot of details that can only be worked out using more-traditional strategies," he says. Such details include deciphering how molecules perform their functions in cells or working out causes and effects. In problems such as those, genomic data could even be distracting.

"Obtaining such data is quite seductive," he says. But "often it is clear that one gets enormous amounts of data which, simply put, are not interpretable."

Dr. Pollard agrees. "It's a lot of fun to go out there and look for all these new things," he says. "That's fine. That should be done. Then people will have to get back to work on the mechanisms." In fact, says Dr. Pollard, the completion of many organisms' genome sequences means that more scientists, not fewer, should concentrate on old-fashioned single-molecule work, rather than on continuing to take an "inventory."

Even Mr. Roos, one of the converts to the new style of biology, agrees that reductionist biology will remain important -- and many of the experimentalists in his lab do exactly that type of work. "My personal belief is that biology has not fundamentally changed," he says.

But members of his lab note a great shift. Just three years ago, according to Ms. Kissinger, the lab group used only two computers. Now it owns 18 and taps into greater computing power housed in the computer-science department. Michael Crawford, a postdoctoral researcher, says that although his graduate training was purely experimental, at some points in Mr. Roos's lab he has spent as much as three-fourths of his time at a computer.

Scientists in other fields note the change, too. "Why are we doing this? Why are we taking so much data and spending so much time cleaning it up and making it available?" asks Mr. Szalay of Hopkins. "It's subconsciously like creating a legacy. Normally a scientific paper gets referenced for two years or so, and then it starts to die off. But to be there and be able to say, 'I was there when astronomy changed,' or 'What we did changed astronomy,' is something incredibly exciting."

DIVING INTO THE DATA

Researchers in many disciplines are increasingly looking online for data, rather than using scientific instruments to collect the information themselves. Here are some of those databases:

Astronomy

HUBBLE SPACE TELESCOPE: one of the first major projects in the field to share observations in a database (http://hubble.stsci.edu).

SLOAN DIGITAL SKY SURVEY: released its first data sets this month, including measurements of 14 million objects (http://www.sdss.org).

NATIONAL VIRTUAL OBSERVATORY: being planned by some astronomers, to combine data from many sky surveys.

Biology

GLOBAL BIODIVERSITY INFORMATION FACILITY: announced in March to connect smaller databases and create a directory of three billion specimens in museums and seed banks (http://www.gbif.org).

SPECIES ANALYST: provides access to natural-history databases (http://habanero.nhm.ukans.edu).

SPECIES 2000: aims to index all the world's known species (http://www.sp2000.org).

DEEP GREEN: presents data on the genetics and evolution of plants (http://ucjeps.berkeley.edu/bryolab/greenplantpage.html).

Earth Science

GLOBAL CHANGE MASTER DIRECTORY: descriptions of and links to data sets on many aspects of earth science (http://gcmd.gsfc.nasa.gov).

TERRA: data from a satellite the National Aeronautics and Space Administration uses to collect information about the earth (http://terra.nasa.gov).

NCEP/NCAR REANALYSIS: weather data at six-hour intervals over a 50-year period, from 1948 to 1998, from the National Center for Atmospheric Research and the National Centers for Environmental Prediction (http://dss.ucar.edu/pub/reanalysis).

Genetics

GENBANK: sequence data on the human and other genomes (http://www.ncbi.nlm.nih.gov.libezproxy2.syr.edu/Genbank).

HUMAN GENOME PROJECT WORKING DRAFT: human-sequence data assembled into full-length genes and chromosomes (http://genome.ucsc.edu).

PLASMODB: genome information for a parasite that causes malaria, *Plasmodium falciparum* (http://plasmodb.org).

Neuroscience

NEURONDB: information about nerve-cell properties (http://senselab.

med.yale.edu/senselab/NeuronDb).

FMRI DATA CENTER: recently started collecting data sets from functional brain imaging (http://www.fmridc.org).

Paleontology

PALEOBIOLOGY DATABASE: in May, started providing access to information about fossils collected worldwide (http://www.nceas.ucsb.edu/public/pmpd).

SOURCE: *Chronicle* reporting

http://chronicle.com.libezproxy2.syr.edu
Section: Research & Publishing
Page: A14


Easy-to-print version

E-mail this article