responsible for making decisions on content." And in turn, a properly linked array of community indexes could serve as a loose-knit index covering all subjects.

For information scientists, the challenge is developing software that allows each community to choose and retrieve documents for a do-it-yourself index. Schwartz's Harvest Program at the University of Colorado is considered the most ambitious attempt so far. The Harvest software, he says, makes it "pretty easy" for anyone "to list a set of uniform resource locaters and documents they want to have included in the index." The software then retrieves the data, extracts the content for indexing purposes, builds the index, and handles queries.

To make the program even more efficient, Schwartz and his colleagues have designed Harvest to be split into two distinct parts, a gatherer and a broker. The gatherer resides at a remote site, retrieving documents to be indexed, while the broker stays on the home machine and builds the index. "Imagine a scenario," says Schwartz, "where you're trying to index a lot of information at NASA, for instance. You can start distributing the process by putting gatherers on all the machines where the data are. Each gatherer can then extract data much more efficiently, because it doesn't have to go across the network to do it. ... Instead of a gatherer sitting on one machine reaching out, you can have gatherers on 100 different machines, each one boiling down the information locally and then sending it across [to a single broker]."

The project was started in late 1993, and Schwartz estimates that Harvest users scattered throughout the Web have put together almost 1000 independent indexes. To link them, Schwartz and his colleagues plan to create software that will allow each distinct index to have pointers leading to other Harvest indexes. And the Harvest researchers are also working on indexes that will support not just documents but actual scientific data, linked to programs that can interpret the data to produce meaningful information. "Right now the Internet is mostly used for data that humans look at," says Schwartz, "Eventually you want to build systems in which programs actually go along, collect data together, and do computations on it." Such indexes will do more than sort through what's on the Internet; they will make sense of it.

But in spite of the best efforts of Schwartz and other indexers, says Robert Kahn, one of the original architects of the Internet, who is now with the Corporation for National Research Initiatives in Reston, Virginia, there won't be any single best way to index the Internet. The solutions, he suggests, will be "technology dependent and sociology dependent," balancing such issues as whether the desired information is distributed throughout the Internet or is more concentrated and whether users care most about reliability, completeness, or speed.

"It's hard to generalize," he says. "It's like discussing transportation systems. You can ask me what I think will be the best transportation system in the future, and I'll say it depends on what you want to do and where you want to go."

–Gary Taubes

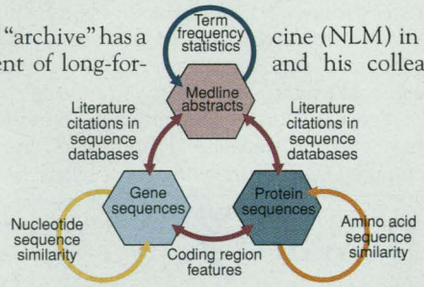
BIOINFORMATICS

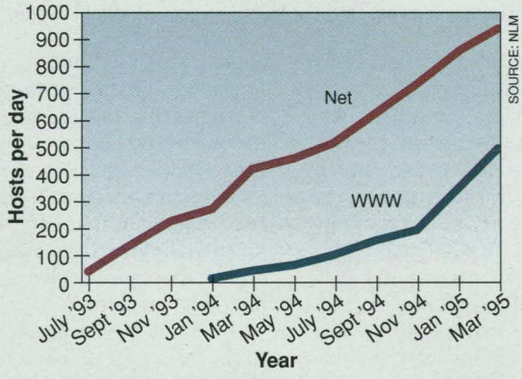
On-Line Archives Let Biologists Interrogate the Genome

To many of us, the word "archive" has a dry, dusty sound, redolent of long-forgotten library shelves. Not even computers necessarily change that perception; they just make electronic archives easier to search. To molecular biologists, however, on-line archives are anything but dry and dusty these days.

As researchers confront a fast-growing mountain of gene sequence data—it now stands at well over 200 million base pairs from humans and more than 8000 other species—electronic databases are fast becoming the lifeblood of the field. It's no coincidence that the major biomedical archives have found themselves on the cutting edge of database management. Witness the dual saga of GenBank and the Genome Sequence Data Base (GSDB), the former a nucleotide database that has been a fixture of molecular biology for more than a decade, and the latter its upstart rival.

At GenBank, "there's been a real turnaround in the past few years," says Mark Boguski, senior investigator at GenBank's current home, the National Library of Medicine (NLM) in Bethesda, Maryland. He and his colleagues have transformed GenBank from a simple archive into an intricately cross-linked array of databases where investigators can now search for similarities among gene and protein sequences, trace their evolution, and jump from sequence data to the relevant literature. By providing all these resources in a centralized archive, Boguski boasts, "our staff is playing a catalytic, proactive role in moving research ahead."





Making connections. The NLM's linked databases *(top)* draw increasing use, much of it over the Web.

Much the same feeling of excitement prevails in Santa Fe, New Mexico, where a new not-for-profit corporation called the National Center for Genome Resources is being funded by the Department of Energy (DOE) to develop the GSDB: an experimental, highly ambitious archive that originated as a spin-off of GenBank but takes a very different approach. Its goal, says the company's chief scientist, Christopher Fields, is to function not as a centralized archive but as a decentralized database, built up and maintained by the efforts of many people in many laboratories—in much the same way that science itself is. "Our task is to explore methods for building federations of distributed and autonomous databases," he says, the aim being to create an arena where scientists can pursue more complex and unexpected questions than they can at GenBank.

GenBank and its spin-off are hardly alone in the biological database world. Two others also try to maintain full archives of nucleotide sequences: the European Molecular Biology Laboratory's nucleotide sequence database (*Science*, 4 August, p. 630) and the DNA Data Bank of Japan. (The four large databases exchange newly submitted sequences every day so that each is complete.) And then there are many smaller scale, special-purpose databases, such as the archive of gene fragments at The Institute for Genomic Research (TIGR), a private research foundation in Gaithersburg, Maryland (*Science*, 16 December 1994, p. 1800). But GenBank and GSDB typify two leading approaches to coping with molecular biology's data flood.

The seeds of GenBank's transformation were planted back in 1992. That was when NIH decided to move GenBank to the NLM from Los Alamos National Laboratory, which had operated it with National Institutes of Health (NIH) funding through the 1980s. The database's home within the NLM was to be the National Center for Biotechnology Information (NCBI), where many of the staff members are themselves active researchers in various NIH laboratories. And once NCBI staffers took over, they quickly began to make changes.

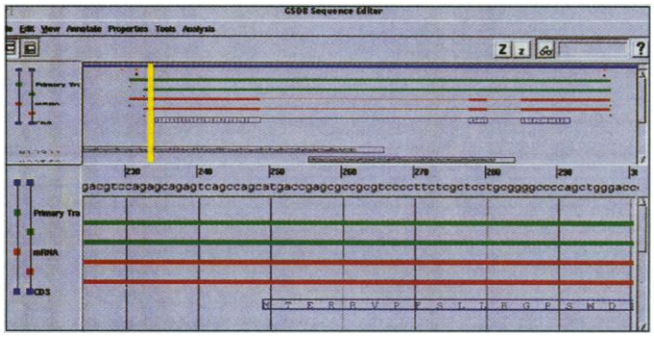
"The first decision we made was to broaden the scope," says NCBI's information engineering chief, James Ostell. "For example, DNA codes for protein. But the protein information was only an annotation on the DNA entry. So we made the proteins into their own database." The NCBI group provided each protein sequence with an electronic link back to the corresponding DNA sequence and vice versa. To the entries in both databases, they also added links into Medline, the National Library of Medicine's database of published research literature. And they devised a browsing system known as Entrez, which allowed users to jump between related items in the various databases.

As of last year users could submit queries over the World Wide Web, and now they can also add new sequences over the Web. Recently, the GenBank team has added a database on phylogeny, which will help in tracing evolutionary relations among the various organisms represented in GenBank. And this autumn they expect to introduce a database on three-dimensional protein structure, linked to the existing databases on protein and gene sequences.

Mining the sequences. Adding new databases to GenBank was only half the story, however. At the same time, the NCBI group was also moving to exploit the deep evolutionary relations that existed within the sequence databases themselves. "If you sequence a gene or a protein, the chances are high—about a third—that there will be a known homologous sequence in another species," explains NCBI Director David Lipman. These homologs, in turn, can provide invaluable information about the function and evolution of the original gene.

To aid users in finding these homologs, Lipman and his colleagues provided GenBank with software that takes each new protein or nucleotide sequence as it comes in and computes its "distance" from each of the existing entries. The number is based on sequence differences in the genes and proteins; it also takes into account such subtleties as noncoding introns in DNA and the hydrophobic-versus-hydrophilic nature of amino acids in proteins.

"This is data you can compute on," says NCBI's information resources chief, Dennis Benson, in a phrase often invoked for the NLM's other data archives (see box on p. 1358). A classic example came after an international collaboration identified *ATM*, the gene that, when mutated, causes ataxia telangiectasia, a rare but devastating disorder of the brain and immune system that generally kills its victims by their teenage years (*Science*, 23 June, p. 1749). To track down and clone the *ATM* gene, which has also been implicated in a heightened risk for breast cancer and an extreme sensitivity to radiation, the researchers worked for 13 years. But to find out if homologous genes existed in other species, they simply submitted a query to GenBank.



Browsing the gene sequence. A prototype screen display from the Genome Sequence Data Base pulls together sequence data from the genome, its RNA transcript, and the resulting protein (CDS), at coarse *(top)* and fine resolution.

SOURCE: CHRIS FIELDS/NATIONAL CENTER FOR GENOME RESOURCES

"I couldn't even imagine going through all those entries without the computer," says one of the paper's 29 co-authors, Danilo Tagle of the National Center for Human Genome Research. But with the computer, says Tagle, "it was very fast work, probably no more than 2 hours. The very first hit we got in the search, with the best similarity score, was a homologous sequence in yeast." This sequence was actually one of several yeast homologs, which turned out to encode a group of enzymes, known as PI-3 kinases, that help control cell growth and division. This information, in turn, gave the team a crucial hint that the enzyme encoded by an unmutated *ATM* gene does something similar—thus helping to explain how *ATM* creates such a wide variety of problems.

Other researchers tell similar stories: the growth factor that turned out to be related to certain fibroblast proteins, for example, or the oncogene that was homologous to another yeast cell-cycle control gene. GenBank currently fields some 10,000 search requests per day. "This is a real revolution in biology, and that's no exaggeration," says Stanford University geneticist David Cox. "You can go into the database and very quickly find out how what you have done fits in with everything else that has been done."

While GenBank and its Entrez query system are remarkably fast and efficient at answering the kind of straightforward homology questions posed by the vast majority of users, many "power users," who tend to be from the major sequencing laboratories, find it a tad inflexible. "It doesn't allow you to construct a very complicated query," points out Anthony Kerlavage, TIGR's database chief. "For example: 'Show me all the sequences and their map locations for cytochromes from the frog genus *Bufo* that were collected in the southwest United States.' This kind of question might be important for biological diversity studies or evolution studies. But it's impossible to do in GenBank," he says, in part because GenBank is set up as a simple "flat file" database and in part because it restricts a user to specific search paths.

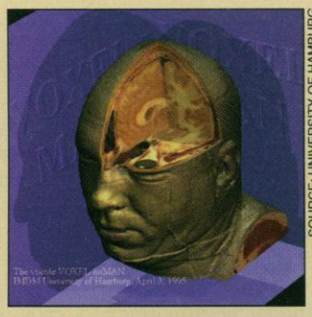
What's more, says Kerlavage, "you can imagine wanting sequence information on a gene, plus mapping information, plus information about the original sample—where the specimen was collected, where it was stored, and so forth. These databases might be archived at three different places. So you can envision a system that takes a query and brings back all the information without the user having to worry about where it came from. Yet you can't do that easily through GenBank," because it's a centralized archive, he says.

A confederacy of data. Both Kerlavage and Lipman point out that GenBank architects had good reasons for centralizing it and adopting the flat file architecture. Collecting all the data in a centralized repository allows for better quality control, and the flat file format is a simpler architecture that minimizes the computational demands of each query. But there's an alternative configuration, one that could open the way for users to submit more complex queries and summon information from many different sources: a so-called relational database, which allows users to follow links between chunks of information that might be distributed over several files or even several databases.

Indeed, the Los Alamos group did create an experimental, relational version of GenBank in the early 1990s, shortly before the database was transferred to NCBI. Lipman and his colleagues elected not to incorporate relational features into the official GenBank, but Los Alamos persuaded DOE to continue supporting its experimental version on the basis of its potential contribution to the Human Genome project. That version has now evolved into GSDB.

Chief scientist Fields is the first to admit

The Visible Man Steps Out



SOURCE: UNIVERSITY OF HAMBURG

Here's looking in you. A three-dimensional reconstruction of Visible Human data, created at the University of Hamburg.

As seen in cyberspace—which is the only place he *can* be seen anymore—the gentleman in question is a burly, bull-necked individual with a dragon tattooed across his chest. He seems disconcertingly lifelike as he stares from the computer screen. But he is even more disconcerting when a mouse click sends the virtual point of view deep inside his body. Suddenly, it's as if he were sliced head to toe by a digital knife. Heart, lungs, brain, bones, liver—everything stands revealed.

Meet the Visible Human. In the real world he was a 39-year-old prisoner who was executed by lethal injection in Texas. But now in the virtual world he has been resurrected—with his prior consent—to star in the National Library of Medicine's (NLM's) gruesomely fascinating effort to create a comprehensive digital atlas of the human body.

"We think this is one of the first of the true digital libraries," says Visible Human project director Michael Ackerman. Not only were the data collected digitally, but the final product, released last December, is among the first to be designed from the ground up for distribution over the Internet and manipulation on ordinary users' computers. As a result, it's a vivid illustration of a slogan applied to the NLM's other digital projects (see main text): "data you can compute on."

The original impetus, says Ackerman, came in about 1987 or 1988, when he was lecturing on the uses of computers in medical education. "As I went around the country, in a few anatomy departments I was taken aside and shown a computer where someone had scanned in the image of a brain, or an elbow, or whatever," he recalls.

This is how you should teach anatomy, his hosts would say. Once you dissect a cadaver you can't put it back together and dissect it again in a different way to understand how the organs fit together in space. The obvious answer was to display a full, anatomically correct human on a computer, says Ackerman.

The result was a plan to digitize two cadavers, a "midlife normal" male and female between the ages of 20 and 60. Each would first be imaged top to bottom via high-resolution MRI and CT, the most common scanning techniques used on live patients. And finally, with the cadavers solidly frozen in blocks of gel, they would each be ground away millimeter by millimeter with high-precision machine tools. At each step the slice would be photographed with digital cameras.

In practice, says Ackerman, it took several years to gear up for the project and select a contractor—the University of Colorado Medical School in Denver—and then several years more to find satisfactory cadavers. Thus, it was only at the end of last year that the NLM was finally able to release the data set from the Texas prisoner; the virtual woman, a 59-year-old female who died of a heart attack, won't be released until later this year. But the response so far has been far greater than Ackerman and his colleagues ever imagined, he says: "We've already licensed the data to almost 300 sites in 23 countries."

Applications have ranged from the straightforward (creating a multimedia textbook of anatomy), to the unexpected (using the data on tissue thickness to calibrate radiation treatment programs), to the unclassifiable: "There are three different artists using the data for—I don't know what," says Ackerman. At the same time, the Visible Human data set is also proving to be a laboratory for research.

At the General Electric Research and Development Laboratory in Schenectady, New York, for example, graphics engineer William E. Lorensen has been working on algorithms that could aid planning for surgery by going from two-dimensional CT and MRI scans to a three-dimensional reconstruction of organs, muscle, and bone. The Visible Human data set makes an ideal testing ground for those algorithms, Lorensen has found, because it is complete, easy to study, and anatomically ideal. Says Lorensen, "This guy was pumping a lot of iron in jail."

Ultimately, some researchers would like to do not just a static three-dimensional reconstruction of the Visible Human, but a full-bore, supercomputer-level simulation in which the bones articulate, the muscles contract, the blood flows, and the organs shift. In short, says Victor Spitzer, principal investigator for the Visible Human project at the University of Colorado Medical School, researchers want to make this guy *walk*.

Why? "Because if you had such a model," says Spitzer, "and if you had all the muscles balanced in a normal walking motion, then you could inflict a limp, atrophy a muscle—add things, subtract things, and see the consequences." He says the payoff, in terms of ergonomic design, better prostheses, and better physical therapy, could be enormous. Computer models of biomechanics have been built before, says Spitzer. "But what this data set brings to the problem is realism. ... When we map this image onto the image of a real patient's body, it's all there."

–M.M.W.

that GSDB is still in the embryonic stages. "We are in the midst of a major development process," he says. "So currently the usage by outside researchers is very low." Later this month, however, the company plans to demonstrate a preliminary version of its new database architecture at the annual Genome Sequencing and Analysis Conference on Hilton Head Island in South Carolina. "My expectation is that usage will increase rapidly after that," he says.

That, of course, remains to be seen. Nonetheless, many power users in the database community are watching the GSDB group closely and give it generally high marks for ambition. "I've been very impressed with them so far," says David T. Kingsbury, principal investigator for the Human Genome Database at Johns Hopkins University. "They seem to be in touch with what the genome community wants."

At SRI International, bioinformatics expert Peter Karp agrees—and, like Kingsbury, points out that it's not a matter of one approach being better than the other. "My overall impression is that the GenBank group views their job as mirroring the published biological literature as accurately as they can," he says, "whereas GSDB is taking more of an electronic-publishing paradigm." Like the most innovative electronic publications, he explains, it offers users more freedom to choose their own paths through the information: "Both groups have the goal of providing high-quality data. But they have different ideas of how best to achieve that goal."

And neither database is in any danger of gathering dust.

–M. Mitchell Waldrop

M. Mitchell Waldrop is working on a book on computers and networking.