Managing the Genome Data Deluge

Molecular biologists are turning to computer technology to help them manage the growing flood of sequencing and mapping data their field is producing

In 1980, if you had mentioned the term "bioinformatics" to a typical molecular biologist, you almost certainly would have been met with little more than a blank stare. Plenty of labs had their resident computer nerd, who spent hours crouched over a terminal, agonizing over how the latest batch of data should be stored and analyzed, but few biologists viewed this eccentric activity as a legitimate scientific discipline. "It was O.K. if I worked on computers," recalls James Ostell, who was a biology graduate student at Harvard University in 1980, "as long as it didn't interfere with my benchwork."

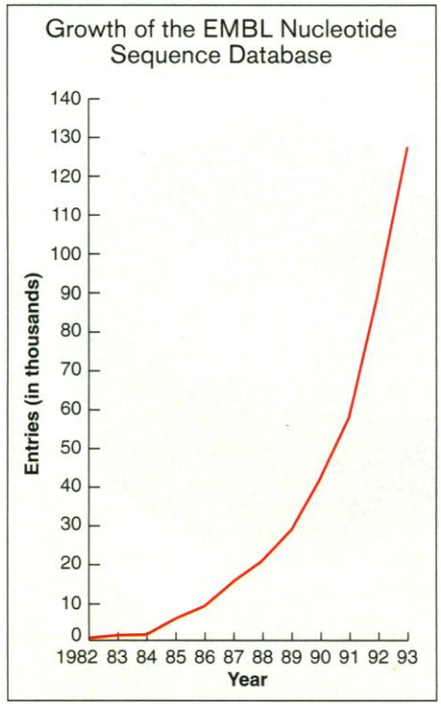
Today, however, that's all changed—and not just for Ostell, who's gone on to become chief of the information engineering branch at the National Center for Biotechnology Information (NCBI) at the National Institutes of Health campus in Bethesda, Maryland. Now, molecular biologists everywhere are increasingly turning to computer technology to help them deal with a major challenge: how to manage and interpret the flood of data being generated by the Human Genome Project and its companion efforts on model organisms from roundworms to mice. Entries in nucleotide sequence databases, such as the one run by the Heidelberg-based European Molecular Biology Laboratory (EMBL) data library, are growing exponentially (see figure). And it's a similar story for genetic and physical genome maps, protein structure information—and just about every other type of molecular biology data.

But as the data accumulate, a major problem is emerging. Researchers want instant access to all the information related to the genes they're studying. This would allow them, for instance, to gain clues to the function of a new gene that they've just sequenced by seeing whether other researchers have discovered similar genes and knew what their activities are. But the necessary data are usually spread over several molecular biology databases—there are now more than 50 in all—that don't communicate. It's the classic "Tower of Babel" situation, notes NCBI's Ostell. Moreover, it's an annoying bottleneck for research. When a molecular biologist sequences a new stretch of DNA, and discovers that it's similar to a gene from another organism, days of valuable research time can be wasted tracking down information on the function of this related gene. What's needed, says Cambridge University

The molecular biology data explosion has given rise to the new science of biological computing or "bioinformatics," explored by Peter Aldhous in a story beginning on this page. That the data can be a valuable commodity is also evident in the wrangle between DOE and NIH over GenBank, described by Leslie Roberts on p. 504.

geneticist Michael Ashburner, is an integrated system allowing a researcher to click on boxes on his or her computer screens and summon up all the relevant data instantly.

Producing such a system is a major goal for NCBI and its transatlantic counterpart, the European Bioinformatics Institute (EBI)—an expanded effort based on the EMBL data library, which will open in new quarters at Hinxton near Cambridge, U.K., in 1995 (*Science*, 18 June, p. 1741). In addition to distributing sequence databases to the biology community, both centers will boast major database research efforts that will place



them at the forefront of the field of database integration. About one-third of NCBI's $7.3 million-a-year budget is currently being spent on research to improve the databases and the software with which to search them, and EBI project leader Graham Cameron hopes to devote up to 20% of EBI's planned annual budget of some $7.5 million to similar applied database research.

Although NCBI and EBI are similar in overall conception, they are set to tackle the issue of database integration in different ways. NCBI has set about uniting the data from several databases in a central integrated databank. In contrast, EBI plans to weld a multitude of separate databases into a loose "federation," communicating over computer networks—an approach that's also favored by biocomputing experts involved in the U.S. Department of Energy genome project (see p. 504).

Building all of the important biology databases into a centrally integrated system will be a laborious task, but NCBI has already taken a first step down the road toward the goal. Since last fall, researchers using the databases distributed by NCBI on CD-ROM have been able to use a software package called Entrez to browse a central integrated database consisting of three types of data: nucleotide and protein sequences from the leading general sequence databases distributed by NCBI and EMBL, plus abstracts of papers from the Medline biomedical literature database.

To make the system work, the NCBI group first had to build into it cross references that record the connections between data that are biologically related—noting which protein is encoded by a particular genetic sequence, for example. "That's the really critical thing [for any integration project]," says NCBI director David Lipman. Entrez not only recognizes the links between nucleotide and protein sequences and between sequences and the papers that cite them, it also assesses the similarity between the sequences and includes word recognition routines that scan Medline abstracts to identify additional related papers. "We find [Entrez] an enormously useful program," says David Hillis, a regular user who heads a molecular evolution lab at the University of Texas at Austin.

NCBI staff are now working to incorporate three-dimensional structure information

from the Protein Data Bank run by the Brookhaven National Laboratory. But even when the protein structure data are incorporated, NCBI's system will still fall short of the extensive integration desired by researchers like Ashburner: Many of the most useful biological data are held not in the general databases, but in the myriad specialist databanks containing data on topics such as how gene activity is controlled, or catering to researchers who study specific model organisms, such as the roundworm *Caenorhabditis elegans*.

To build data from a range of databases into a centralized integrated system, it's first necessary to convert all the data records into a standard format—a computing equivalent of the invented universal language Esperanto, as NCBI's Ostell puts it. NCBI so far has been using as its Esperanto a language called ASN.1 that was developed in the computer industry to exchange information. If the curators of the specialized databases were to routinely convert their data into ASN.1, it would be a relatively simple task for NCBI staff to extend their system and produce a more comprehensive integrated database. So far, however, they have proved reluctant to do this. Many don't like ASN.1, which is more complex than the formats used by most specialist databases. "It's not human readable," complains Richard Durbin, chief informaticist at the Sanger Center, the genome institute that will be EBI's neighbor at Hinxton. Durbin's opinion carries weight in this field, since he is co-originator of one of the most widely used database systems in genome research: The ACeDB system developed to manage data from the *C. elegans* genome project, which has since been adopted by several other model organism communities.

Ostell takes a pragmatic view of these difficulties. It's early in the game, he says, and many people are taking a wait and see attitude. And even without data from the specialist databases, says Ostell, "we have the most complete integrated resource that's available right now."

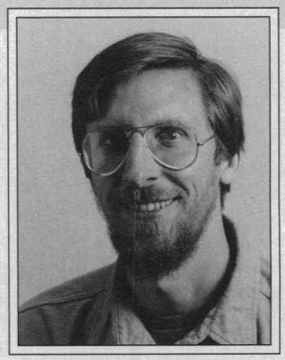
Other teams are also pursuing integration projects, using alternative formats. A European consortium, for instance, led by biomathematician Otto Ritter of the German Cancer Research Center in Heidelberg, is working on the Integrated Genome Database project, an attempt to unite sequence, mapping, and disease gene phenotype data. And at the Cold Spring Harbor Laboratory, computational biologist Thomas Marr is developing a similar system called Genome Topographer. So far, no single integration project has emerged as the indisputable front runner among these efforts. "Technically, all of these solutions could work," says Nat Goodman, chief informaticist at Eric Lander's genome center at the Whitehead Institute.

EMBL's informaticists, however, believe that any attempt to produce a central integrated database faces a major difficulty. "Trying to get all database producers to agree on one data model...is a hopeless exercise," argues Rainer Fuchs of the EMBL data library. Database curators, he says, worry that the need to convert all of their data routinely into a standard format will restrict their freedom to alter their own internal formats as they see fit. That's why Fuchs and EBI project leader Cameron favor a loose databank federation—one that doesn't require the data to be converted en masse into a common format, and so should allow the participating databases greater autonomy.

Making participation more attractive to the individual curators, however, means paying a price elsewhere in terms of technical obstacles. Creating such a federation will require sophisticated software engineering to produce "mediator" programs for converting questions asked of the integrated system into the separate query language understood by each participating database. Individual data records, temporarily converted into a standard format, would be relayed back to the user, so that he or she can browse through related data just as in a centrally integrated system.

"We have the most complete integrated resource that's available right now."

–James Ostell



The problem is that the mediator programs Fuchs envisions are still the subject of cutting-edge research in computer science. And many of the smaller specialist biology databases don't yet include the sophisticated query routines that they would require to become part of a federated system. "We need at least 5 years" before such a federation becomes feasible, Fuchs estimates.

Nevertheless, biologists using the CD-ROM release of EMBL's databases are now provided with a software package called EMBL-Search that provides a taste of what Fuchs and Cameron have in mind. Staff at the EMBL data library have built cross references into the databases they distribute so that entries in the Swiss-Prot protein sequence database, for instance, also give the accession numbers of corresponding DNA sequences in the EMBL nucleotide database. EMBL-Search now recognizes these cross references so that researchers browsing one database can summon up related entries from the other and view the corresponding protein and genetic sequences side by side.

Of course, this is a long way from an extensive federated integrated system. And with both the centralized and federated approaches to database integration facing formidable hurdles, it's as yet unclear which will emerge as the favored model. Indeed, some database producers have a foot in each camp. An example is molecular geneticist Philip Bucher of the Swiss Institute for Experimental Cancer Research in Lausanne, who produces the Eukaryotic Promoter Database (EPD). This databank contains detailed information on the promoter sequences (which regulate gene expression) that are in the EMBL nucleotide database. With the new release of EMBL's CD-ROM, it's possible for the first time to view the EPD annotations alongside the genetic sequences to which they refer. Yet Bucher was also the first specialist database curator to produce an ASN.1 version of his databank.

While working on their long-range goal of comprehensive database integration, NCBI and EBI will also provide some more immediate help for the molecular biology community. For instance, both intend to produce software that can serve as a flexible filter to screen out unwanted sequences when viewing the general databases. Currently, researchers searching the databases for nucleotides or proteins that share features with a particular sequence often get bombarded with hits against scores of sequences in which they have no interest. For example, a researcher who already knows that a sequence looks like an immunoglobulin but wonders what else it resembles would not be interested in getting back all immunoglobulin sequences in response to his or her query.

Such filters may be available in a year or two, but the consensus is that integrated systems linking most of the important molecular biology databases probably won't be in general use before the end of the decade. Informaticists agree that it's for the market to decide which system this will be. "None of us are so foolish so that if it looks as if someone else's approach is taking off, we wouldn't switch to it," says NCBI's Ostell. But it's possible, says the Whitehead Institute's Goodman, that the favored solution will be a hybrid of the centralized and federated models—with a large central database containing the most frequently used data operating at the hub of a wider database federation. For the time being, he says, "We need to try both [approaches] in parallel."

–Peter Aldhous