Electronic Data Publishing in GenBank

Michael J. Cinkosky, James W. Fickett, Paul Gilna, and Christian Burks

Improvements in DNA-sequencing technology in the mid-1970's enabled researchers around the world to determine the exact sequence of nucleotides in samples of DNA much more easily than before (see "Decades of Nonlinearity: The Growth of DNA Sequence Data" above). Computers were the most convenient way to handle the large quantities of sequence data discovered using the new methods. Furthermore, since many people became interested in applying computer technology to interpreting those data, the data needed to be readable by computers. To meet those needs, Walter Goad created the Los Alamos Sequence Library in 1979, which in 1982 became GenBank.

Like many scientific databases at that time, GenBank was designed as a curated data repository. For its first several years of operation, the data were collected from published articles containing DNA sequence data in figures. The sequence data and related annotation (for example, information about the function and structure of the sequence) were typed into a computer and formatted into complete database entries, which were then distributed to users in both electronic and printed form.

The limitations of this style of operation became obvious fairly early. The volume of data being generated continued to grow dramatically. It became increasingly difficult for the database staff to keep up with the flow of data, and the delay between publication of an article and appearance of the data in the database grew accordingly. At the same time, the data were becoming increasingly important to biologists, which aggravated the problem of slow turn-around time for data processing.

Another problem was that a growing body of data would never, as the situation stood, appear in the database because it would never appear in print. Journals were already beginning to limit the amount of space that they would devote to printing nucleotide sequences; therefore, authors began omitting "uninteresting" sequence data (such as introns and other non-coding regions) from their papers. For computational biologists, however, those data are potentially of great interest and not having them in the public database would severely hinder some types of studies. Furthermore, in 1986 both the DOE and NIH began to talk about the Human Genome Project. If undertaken, that project would result in the generation of at least a thousand times the quantity of data that was already in the database, and probably far more. It was becoming critical to develop a different approach to building and maintaining the database.

Electronic Data Publishing

Reconsidering the problem made it clear that sequence data and results based on those data should be handled by completely separate communication methods. Whereas

scientific results needed peer review and an essentially free-form medium like the printed page, sequence data needed a largely automatic form of quality control and a highly structured, electronic format to be useful. To meet this need, we created what we call Electronic Data Publishing.

In Electronic Data Publishing, the originators of the data retain responsibility for the data in much the same way that they retain responsibility for the contents of published articles. Rather than being communicated primarily through journal articles, the data are deposited directly into an electronic database, and a separate article referring the reader to the appropriate database entries is published in a traditional journal. The database staff provides tools to help the originators get their data into the database, as well as software to provide automatic checks on the quality and integrity of the data.

To speed the transition to this new model, we enlisted the aid of many of the editors of the journals in which most of the sequence data were appearing. Because they were as acutely aware of the problems as we were (they were particularly interested in reducing the number of pages devoted to the printing of sequence data), many agreed to require submission of the data to the database before a paper discussing the data could appear in their journals. Within a year we were receiving a significant percentage of our data in electronic form before the related article appeared in print.

Table 1. Divisions of GenBank

Division	Number of entries (June 1992)	Change in number of entries since March 1992	Number of bases (June 1992)	Change in number of bases since March 1992
Bacteriophage	779	18	1,102,766	-13,880[1]
Other viruses	7,750	1,238	11,883,566	1,007,715
Bacteria	7,965	760	13,732,370	1,290,821
Organelles	2,241	130	3,721,811	409,921
Plants and fungi	6,196	682	10,713,664	1,436,907
Invertebrates	6,079	868	8,422,573	977,127
Rodents	12,737	909	13,942,988	964,730
Primates	15,996	1,257	17,258,180	1,620,375
Other mammals	2,660	215	3,537,274	355,010
Other vertebrates	3,250	276	3,915,314	342,341
RNA	2,698	162	1,517,776	134,686
Unannotated	1,649	-360[2]	1,532,138	-297,009[2]
Synthetic[3]	1,282	27	857,738	42,302
Total	**71,282**	**6,220**	**92,165,158**	**8,270,506**

Implementation of the Electronic Data Publishing model also required the development of a large software system with several major components. First, we designed and built a relational database to store the data in a far more structured manner than was practical with our original ASCII-text database format. Then we built an interactive, window-based interface to this database, called the Annotator's WorkBench, which enables people to work directly on the contents of the database. We also worked with the European Molecular Biology Laboratory and the DNA

[1]As part of our curation of GenBank, we often combine duplicated sequence data into a single representation. In the Bacteriophage division between January and March 1992, the amount of data submitted was less than the amount of duplicate data merged, so the net change during that period was a decrease.

[2]The Unannotated division of the database was formerly used to distribute data quickly by releasing them to the public in raw form prior to the more detailed work of annotation. No data have been added to this division for some time. We continue to relocate sequences from this division to their appropriate taxonomic division through annotation, resulting in a decrease of the amount of data classed as unannotated.

[3]Synthetic DNA includes such laboratory-constructed DNA as short oligonucleotide probes, cloning vectors, expression vectors, synthetic genes, etc., which cannot readily be considered as originating from single taxonomic species.

Table 2. Amount of Sequence Data from Well Studied Organisms

Organism	Bases sequenced	Number of genome equivalents sequenced	Percent of total data in database
C. elegans (nematode)	0.54×10^6	0.007	0.7
E. coli (bacterium)	2.81×10^6	0.597	3.6
S. cerevisiae (yeast)	2.95×10^6	0.203	3.8
D. melanogaster (fruit fly)	3.02×10^6	0.018	3.9
M. musculus (mouse)	6.89×10^6	0.002	8.9
H. sapiens	13.44×10^6	0.005	17.4

Databank of Japan to develop systems for sharing data, so that researchers need enter data into only one of the three databases. Finally, we created a format for automatically processable database submissions and wrote software to aid in the preparation of these submissions, which is distributed freely to anyone requesting it. Data submitted in that format are run directly into the database, where the database staff can easily use validation software that we have written to check the data for biological consistency. (As a simple example, the software checks that exons do not contain stop codons).

The impact of these changes on our operation has been dramatic. We now receive about 95 percent of our data directly from researchers, mostly in automatically processable form. In 1984, we processed sequences containing approximately 1.38 million nucleotides. At that time, it was taking, on average, more than one year from publication for data to appear in the database at a cost of approximately $10 per base pair. In 1990, we processed 10 times as much data (about 14.1 million nucleotides) with an average turn-around time of two weeks at a cost of roughly $0.10 per base pair. Further, we have been able to maintain this performance since 1990, despite the fact that the rate of submissions has more than doubled to 30 million base pairs per year in the first half of 1992.

A brief survey of the contents of GenBank indicates the extent of sequence data and the areas in which biologists have been particularly interested. Table 1 shows the contents as of release 72 (June 1992) broken down by taxonomic and other categories of origin. Approximately half the data are from expressed regions, the rest being primarily introns and sequences immediately upstream and downstream of genes. A new development is the submission of thousands of rough sequences, each a few hundred base pairs long, from human cDNAs (see pages 136–139 in "Mapping the Genome").

About 2850 organisms (including viruses) are represented in GenBank. The only completely sequenced genomes are from viruses and cell organelles (mitochondria and chloroplasts), ranging in size from a few hundred base pairs for certain plant viruses to more than 200 kilobase pairs for the cytomegalovirus. Table 2 gives information (as of December 1991) on the organisms to which the most sequencing effort has been devoted. (The heading, "number of genome equivalents," means the ratio of the number of bases sequenced from that organism to the number in its genome, without the subtraction of any duplications in the database.) In one notable recent change, the amount of sequence in the database from the nematode *Caenorhabditis elegans* increased by a factor of about 7.7 between December 1988 and December 1991, 2.5 times larger than the increase of the database as a whole. ■