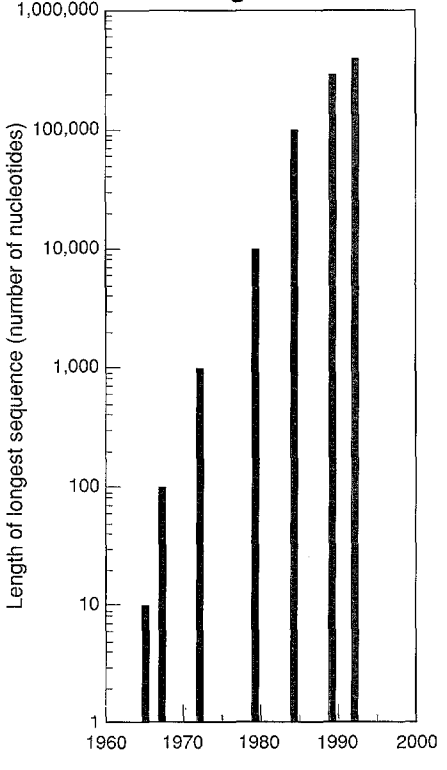
Decades of Nonlinearity: *the growth of DNA sequence data*

Christian Burks, Michael J. Cinkosky, and Paul Gilna

The first nucleotide sequence was published in 1965; it was the sequence of an RNA molecule less than 100 nucleotides long. The methods used were so arduous that until the mid-1970s a person could determine the sequence of only about a hundred bases in a year. Then Maxam and Gilbert in the U.S. and Sanger in England developed new sequencing techniques that were a hundred times faster (see "DNA Sequencing" in "Mapping the Genome"). Figure 1 shows that today biologists are determining the complete sequences of pieces of DNA over 100,000 nucleotides in length. Almost 100,000,000 nucleotides of sequence data have been published—a wealth of information that has formed the basis for many scientific discoveries. How has the enormous and rapidly growing quantity of data been maintained and managed?

As shown in Figure 2a, the rate of sequence-data accumulation was increasing rapidly in the late 1970s. (Data for Figure 2a were compiled from the GenBank database, which includes the publication date and length of each sequence entered.) In response to the growing interest in gathering and analyzing the data, the biology community held several discussions in 1978 on establishing a database facility to collect, organize, and distribute sequence data and annotation about each sequence. For design purposes, the operation of a database can be compared to industrial processes in which a set of input objects is transformed into a set of output objects. In a sequence database, the input is DNA sequences generated by individual laboratories and stored in individual formats with varying amounts of annotation; the output is a collection of DNA sequences stored at a central facility in a uniform format with a precisely defined degree of annotation. For any such process to be workable and efficient, the mechanism for the process must match the volume of the input stream.

Figure 1



During the planning stages for the public sequence databases, how fast did biologists expect the amount of data to grow? Up to 1981 the few recorded projections generally assumed linear growth. Figure 2b shows a linear projection—based on the average annual rate from 1975 to 1977, 25,000 nucleotides per year—for the period up to 1986. (Note that the scale of Figure 2b compresses the previously impressive growth up to 1978.) The linear model predicts that under 300,000 nucleotides of sequence data would have been accumulated by 1986, and that a database project would have had to handle no more than 30,000 in any year. Funding-agency planning and subsequent project proposals to the agencies were based on that linear model. In 1982 the GenBank project, the American sequence database, was established at Los Alamos through a five-year contract with the NIH. (Also in that year a database storing essentially the same information was established at the European Molecular Biology Laboratory; Japan developed a similar institution a few years later.) Because a steady rate of data accumulation was expected, GenBank was staffed with only a few people who were expected to search the literature and enter into a database all the DNA and RNA sequence data that would appear.

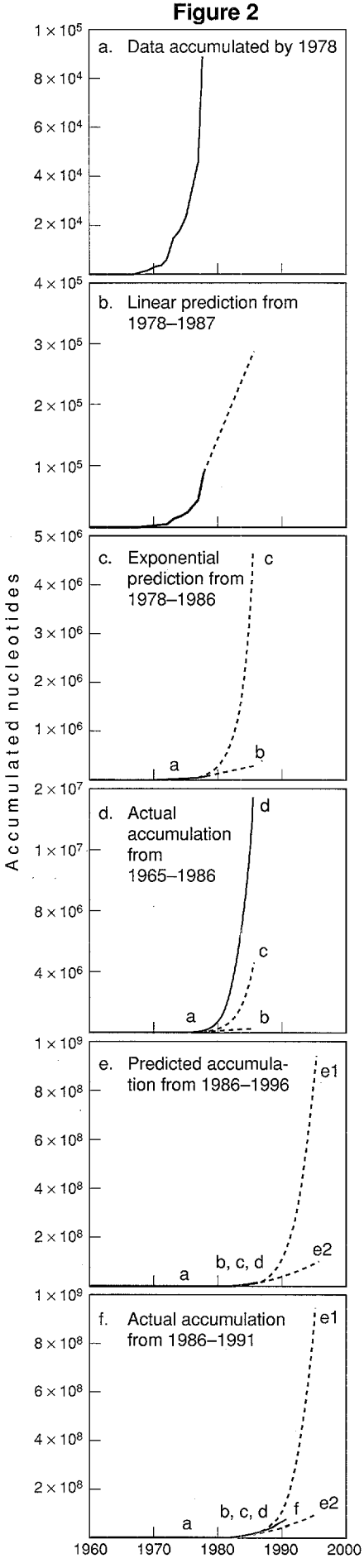
Suppose the community had instead projected exponential growth for the sequence data. Figure 2c shows that if we use the annual rate increase for the years 1975-77 (64 percent per year) to project the accumulation over the period 1978-86, an exponential model predicts an accumulation 15 times that of the model in Figure 2b, and a rate of accumulation orders of magnitude higher. Clearly, in that scenario a database project could not rely on a constant number of staff members each processing data at a constant speed.

What really happened? As can be seen in Figure 2d, the increase of sequence data far outstripped even the exponential model, and completely dwarfed the linear model that was actually used to design GenBank. This created a crisis for the scientific community wanting access to all these data and in particular for the GenBank project, which was responsible for providing access.

In 1986-87, as we planned and developed proposals for the second five-year GenBank contract, we revisited the issue of modeling the growth of sequence data. Figure 2e presents the envelope in which we expected the growth to lie. The lower limit is an extrapolation from the previous three years assuming a constant rate of acceleration. The upper limit is based on the assumption that seven billion bases of sequence, twice the total of the human genome, will be determined by 2005 (consistent with the goals of the Human Genome Project). The rate of acceleration is assumed to increase linearly to bring the curve to that endpoint. With the genome project in mind, we developed a new strategy—and corresponding mechanisms—for the flow of data in and out of the database (see "Electronic Data Publishing in GenBank" below) that we believed would accommodate growth within the projected envelope shown in Figure 2e.

Five years later, Figure 2f shows that actual growth of sequence data has indeed remained within this envelope, and that the accumulation of nucleotide sequence data continues to accelerate. It is worth noting that if the Human Genome Project goals for sequencing are to be met, the rate of sequencing will have to accelerate considerably over the next decade. ■

Figure 2



Further Reading

Walter B. Goad. 1983. GenBank—and its promise for molecular genetics. *Los Alamos Science* 9 (Fall): 52–61.

C. Burks, J. W. Fickett, W. B. Goad, M. Kanehisa, F. Lewitter, W. P. Rindone, C. D. Swindell, and C.-S. Tung. 1985. The GenBank nucleic acid sequence database. *Computer Applications in the Biosciences* 1:225–233.

Christian Burks. 1989. How much sequence data will the data banks be processing in the near future? In *Biomolecular Data: A Resource in Transition*, edited by R. R. Colwell, pp. 17–26. Oxford University Press, England.

Christian Burks. 1989. The flow of nucleotide sequence data into data banks: role and impact of large-scale sequencing projects. In *Computers and DNA*, edited by G. Bell and T. Marr, pp. 35–45. Addison-Wesley, Reading, MA

Michael S. Waterman. 1990. Genomic sequence databases. *Genomics* 6:700–701.