we search for single base changes that may cause a genetic defect, part of the problem is distinguishing which change(s) is responsible for the disease. The second reason is that, as argued below, the data quality from large sequencing projects also requires a change in our current concept of sequence. In fact, the concept of "the genome" as a unique entity is not quite firm, which further complicates matters. Humans differ from one anther in about one nucleotide in one thousand. In addition, recombination makes it difficult to maintain genomic material in a static condition. For these reasons, genomic sequence databases must necessarily be more fluid than our current database "world view." New models of sequence are required, and some people, including database staffs, have already begun to think about these problems.

While most discussions of genomic sequencing center on volume or number of nucleotides, the real situation is much more complex. For example, a clone will be shotgun sequenced

An effort to reduce the backlog of all sequences from 1960 to 1987 that are not included is well along, and this effort will be complete by the end of 1990. GenBank contains 95% of the sequences published in the last 2 years in journals for which it is responsible. Today, about 80% of the published sequences are entered and annotated within 3 months, and efforts are underway to improve this percentage. An effort is made to have journals require or encourage submission of sequences to GenBank in computer-readable form. While 65% of the GenBank entries come directly from the authors, about 45% of the submissions are in computer-readable form. The program Authorin has been designed to help scientists enter and annotate their sequences. Relational database management systems are being tried as a replacement for the older, flat file system. Others are exploring object-oriented databases.

None of this is easy. Collecting and managing data that matical analyses and the testing and refinement of theoretical hypotheses. This is not to suggest that mathematical analyses or deep theoretical concepts have not played an important role in formulating our modern view of biology. Rather, we are witnessing a natural metamorphosis in which the new and, until recently, unanticipated mountain of highly syntactically structured data is opening vast new analysis and

evolutionarily related homologous functional families. This demonstrated the utility of organizing sequence data under a major theoretical construct.

It is important to note that although Dayhoff's work was supported by the National Institutes of Health, it was not as database activity but as basic research. In fact, the NIH, with two relevant exceptions, did not support database activities and Newman Inc. in Boston and the Molgen project under the Stanford University Medical Experimental Computer Resource, SUMEX, itself a DRR-supported project. The latter was included in a local attempt to make sequence analysis

was never published, preventing any broadly based discussion within the research community prior to NIH's "sources sought" announcement almost 2 years later.

In August 1979 Bell and Goad organized a small meeting erkandl and Pauling (1965), Dayhoff and Eck (1966), Fitch and Margoliash (1967), and others foresaw the importance of computer support of databases and sequence analysis, these were not to become commonplace in molecular biology until the mid 1980s!

With prodding from a number of researchers, including

establish a centralized database in collaboration with the Europeans and potentially the Japanese. The database was to be accessible electronically and distributed via magnetic media (as the protein sequence and structure databases by then were). Phase II was to establish an analysis and software library coupled to the database. For reasons that are not clear database it would surely have been a better choice. This was particularly felt by Margaret Dayhoff. Others were still concerned that the database was not at an academic research center. The community showed some surprise and concern that only three proposals had been submitted. This was in part because three of the four players—Bolt, Beranek and Newman with PROPHET, IntelliGenetics as an outgrowth of the Stanford SUMEX/Molgen project, and the NBRF with the Protein Information Resource—were organizations with past links to the NIH infrastructure. No university or non-NIH-associated commercial centers applied. The question still remains whether this was only because no one else was in a position to attempt such a project or that somehow NIH

ware, and the computing costs at Los Alamos National Laboratory became excessive. These funding problems, along with limited computer science- and database-experienced staff, led to both the maintenance of the database in a flat file format (dropping the relational table form) on a very limited minicomputer system and the eventual introduction of incomplete or unannotated data entries. Network access through the PROPHET system, and later through BioNet, proved to be of secondary importance, as most large research laboratories and academic departments accessed the database through local installation. With more and more commercial and academic search and analysis packages becoming available on the new powerful computer workstations, this trend can only included. One of the major database integration efforts recently initiated is that by the Howard Hughes Medical In-

lay behind the original phase II, the BioNet, the Dana-Farber Cancer Institute's MBCRR (Smith *et al.*, 1986), and other

[redacted]

11. MAXAM, A. M., AND GILBERT, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* **74**: 560-564.

12. MAXAM, A. M., AND GILBERT, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *In* "Methods in Enzymology" (L. Grossman and K. Moldave, Eds.), Vol. 65, pp. 499–560, Academic Press, New York.

13. MOROWITZ, H. J., AND SMITH, T. F. (1987). "Report of the Matrix of Biological Knowledge Workshop," Santa Fe Institute, Santa Fe, NM.

14. NATHANS, D., AND SMITH, H. O. (1975). Restriction endonucleases in the analysis and restructuring of DNA molecules. *Annu. Rev. Biochem.* **44**: 273–293.

15. ROBERTS, R. J., *et al.* (1977). Restriction and modification enzymes and their recognition sequences. *In* "DNA Insertion Elements, Plasmids and Episomes" (A. I. Bukari *et al.*, Eds.),

18. SANGER, F., NICKLEN, S., AND COULSON, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**: 5463–5467.

19. SMITH, T. F., GRUSKIN, K., TOLMAN, S., AND FAULKNER, D. (1986). The molecular biology computer research resource. *Nucleic Acids Res.* **14**: 25–29.

20. SMITH, R. F., AND SMITH, T. F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. USA*, in press.

21. SUTCLIFFE, J. G. (1979). Complete nucleotide sequence of the *Escherichia coli* plasmid pBR322. *Cold Spring Harbor Symp. Quant. Biol.* **43**: 77.

22. WATERMAN, M. S., SMITH, T. F., AND BEYER, W. A. (1976). Some biological sequence metrics. *Adv. Math.* **20**: 367–387.