

## Primer on Molecular Genetics

database access and manipulation, and data-entry procedures to compensate for the varied computer procedures and systems used in different laboratories. Databases need to be designed that will accurately represent map information (linkage, STSs, physical location, disease loci) and sequences (genomic, cDNAs, proteins) and link them to each other and to bibliographic text databases of the scientific and medical literature.

## Interpreting Data

New tools will also be needed for analyzing the data from genome maps and sequences. Recognizing where genes begin and end and identifying their exons, introns, and regulatory sequences may require extensive comparisons with sequences from related species such as the mouse to search for conserved similarities (homologies). Searching a database for a particular DNA sequence may uncover these homologous sequences in a known gene from a model organism, revealing insights into the function of the corresponding human gene.

Correlating sequence information with genetic linkage data and disease gene research will reveal the molecular basis for human variation. After a disease gene is identified, however, the altered protein specified by the flawed gene must still be compared with the normal version to identify the abnormality that causes disease. Once the error is pinpointed, researchers must try to determine how to correct it in the human body, a task that will require knowledge about how the protein functions and in which cells it is active.

**Fig. 14. Magnitude of Genome Data.** If the DNA sequence of the human genome were compiled in books, the equivalent of 200 volumes the size of a Manhattan telephone book (at 1000 pages each) would be needed to hold it all. New data-analysis tools will be needed for understanding the information from genome maps and sequences.

