Model Organism Research

Most mapping and sequencing technologies were developed from studies of nonhuman genomes, notably those of the bacterium *Escherichia coli,* the yeast *Saccharomyces cerevisiae*, the fruit fly *Drosophila melanogaster*, the roundworm *Caenorhabditis elegans*, and the laboratory mouse *Mus musculus.* These simpler systems provide excellent models for developing and testing the procedures needed for studying the much more complex human genome.

A large amount of genetic information has already been derived from these organisms, providing valuable data for the analysis of normal gene regulation, genetic diseases, and evolutionary processes. Physical maps have been completed for *E. coli*, and extensive overlapping clone sets are available for *S. cerevisiae* and *C. elegans.* In addition, sequencing projects have been initiated by the NIH genome program for *E. coli, S. cerevisiae*, and *C. elegans.*

Mouse genome research will provide much significant comparative information because of the many biological and genetic similarities between mouse and man. Comparisons of human and mouse DNA sequences will reveal areas that have been conserved during evolution and are therefore important. An extensive database of mouse DNA sequences will allow counterparts of particular human genes to be identified in the mouse and extensively studied. Conversely, information on genes first found to be important in the mouse will lead to associated human studies. The mouse genetic map, based on morphological markers, has already led to many insights into human biology. Mouse models are being developed to explore the effects of mutations causing human diseases, including diabetes, muscular dystrophy, and several cancers. A genetic map based on DNA markers is presently being constructed, and a physical map is planned to allow direct comparison with the human physical map.

Informatics: Data Collection and Interpretation

Collecting and Storing Data

The reference map and sequence generated by genome research will be used as a primary information source for human biology and medicine far into the future. The vast amount of data produced will first need to be collected, stored, and distributed. If compiled in books, the data would fill an estimated 200 volumes the size of a Manhattan telephone book (at 1000 pages each), and reading it would require 26 years working around the clock (Fig. 14).

Because handling this amount of data will require extensive use of computers, database development will be a major focus of the Human Genome Project. The present challenge is to improve database design, software for

HUMAN GENETIC DIVERSITY:
The Ultimate Human Genetic Database

- Any two individuals differ in about 3×10^6 bases (0.1%).
- The population is now about 5×10^9.
- A catalog of all sequence differences would require 15×10^{15} entries.
- This catalog may be needed to find the rarest or most complex disease genes.