Appendix I)

Databases, Repositories, and Informatics

Among the most useful products of genome projects will be information and materials—information about genes and their locations and sequences, and biological materials such as DNA fragments from chromosomes of known pedigree, ordered cosmids, and clones. Proper management of data and materials is essential to increase the efficiency and productivity of research and to reduce duplication of efforts so that genome projects can succeed in meeting the needs of medical scientists and molecular biologists in this century and the next.

Existing databases and repositories that gather, maintain, analyze, and distribute data and materials are already struggling to keep up with the exponential growth of molecular biology. Present capabilities will have to expand greatly to handle the increase of information resulting from a targeted set of genome projects. while it is logical to link computational needs to genome projects, however, funding devoted to storage of genetic data and materials and to sophisticated analysis of DNA will prove important in molecular biology even if a major mapping and se quencing initiative is not undertaken. Because the essential databases, repositories, and linking computer networks provide goods and services for the entire research community, the Federal Government has a long-standing tradition of supporting them and is in a unique position to further enhance the resources.

This appendix describes some existing databases and repositories and outlines present and future database needs relevant for human genome projects specifically and molecular biology in general.

Databases

Various databases exist that serve the needs of researchers in genome mapping and sequencing (see table D-l). One set of databases gathers, stores, and distributes information directly related to genetic maps and physical maps. Some databases specialize in map and sequence information from one specific genome—for example, there are databases exclusively devoted to the mouse, E. coli bacteria, drosophila, and nematode genomes—while others carry particular kinds of information from all the relevant genomes. Other databases gather data on the sequences and structures of proteins and amino acids that are not direct results of mapping and sequencing research but are necessary for addressing basic research problems underpinning genome research. The data from the differ ent types of maps and from different species have important interconnections, so it is essential that the information be linked for comparative studies.

Genetic Maps

Genetic maps can be generated in several ways (see ch. 2). Pedigree analysis of linked traits yields a map in which traits can be ordered sequentially and with a rough estimate of the distance between them. RFLPs and other DNA probes can help link the traits with specific genes or regions of DNA to produce more refined maps. Maps of the functional regions within individual genes aid in the search for underlying causes of genetic diseases and for the mechanisms by which genes control development and function. several different databases serve the different information needs for specific kinds of maps.

On-Line Mendelian Inheritance in Man (OMIM).–An atlas of human traits that are known to be inherited-expressed genes—has been compiled into a reference work known as Mendelian ***Inheritance in Man,*** **which has been published in seven editions. The listing has been edited by Victor McKusick of The Johns Hopkins University since 1966. As of March 1, 1988, 4,336 traits had been identified as genetically based, including over 2,000 diseases.**

Since 1986, the Howard Hughes Medical Institute (HHMI) has supported computerization of the list, and it is now accessible for on-line searches free of charge (4). It is cross-referenced in the Human Gene Mapping Library so that information on expressed genes can be linked to map data.

Human Gene Mapping Library (HGML).–A1so called the New Haven Database, HGML consists of five linked databases-one each for map information, relevant literature, RFLP maps, DNA probes, and contacts (researchers with information on data or materials). In addition, the map database is linked to OMIM. All of the databases are cross-referenced, so that data about a gene or probe of interest can be drawn from all five during the same search (10).

DNA Nucleotide Sequences

Databases containing raw DNA sequences, information about the origin of the DNA segment sequenced

Table D-- I.—Some Existing U.S. Databases and Repositories

	Location	Funding source	Annual budget[a]
Nucleotide sequence data:			
GenBank®	**Los** Alamos National Laboratory, Intelligenetics Corp., CA	**NIH,[b] DOE, NSF, USDA**	***$3,500,000***
Genetic map data:			
On-Line Mendelian Inheritance in Man (OMIM)	Johns Hopkins University Baltimore, MD	Johns Hopkins University, HHMI, NLM	$ 550,000[c]
Human Gene Mapping Library (HGML)	New Haven, CT	HHMI	$ 500,000
Protein and amino acid sequence and structure data:			
Protein Identification Resource (PIR)	**National Biomedical Research Foundation Washington, DC**	**NIH[d]**	**$ *500,000***
Protein Data Bank (PDB)	**Brookhaven National Laboratory Upton, NY**	NSF, NIH,[e] **DOE**	**$ 260,000**
Repositories:			
American Type Culture Collection (ATCC)/Human DNA Probe and Chromosome Library	**Rockville, MD**	NIH[f]	$ 300,000[c]
Human Genetic Mutant Cell ReDositorv	**Coriell Institute for Medical Research Camden, NJ**	**NIHg**	**$ 750,000**

aBudget figures are approximate. Several of the databases have multiyear contracts; amount listed Is the average yearly allotment.
bNIH sponsors of GenBank®, pact and present, include the National Institute of General Medical Sciences (N IGMS), the Division of Research Resources (DRR), the National Institute for Allergy and Infectious Diseases, the National Cancer Institute, the National Library of Medicine, the National Eye Institute, and the National Institute of Diabetes and Diseases of the Kidney. The NIGMS administers the contract and coordinates the funding.
cThe Johns Hopkins University contribution to OMIM is difficult to measure, because it includes many indirect factors (Staff support, space, publication costs, etc.). HHMI contributes S318,000 and the NLM $100,000 annually.
dThe NIH sponsor is DRR.
eNIH sponsors are NIGMS and DRR.
fNIH sponsors are the National Institute of Child Health and Human Development (NICHD) and DRR; DOE has contributed some funds through DRR.
gThe NIH sponsor is NIGMS.

SOURCE: Office of Technology Assessment, 1988.

(which gene, which organism), and various annotations that summarize information about important features in the sequence (sites cut by DNA-cutting enzymes, regulatory sequences, protein-coding regions) will be directly affected by genome projects that emphasize sequencing. The major databases for nucleotide sequences are GenBank® and its European counterpart, EMBL (8). Each carries sequence data and related information for the human genome as well as bacterial, yeast, fruit fly, mouse, and other genomes. Since 1982, GenBank® and EMBL have split the task of data collection, with each database monitoring specific journals in molecular biology to locate and enter sequence data, and they cooperate closely in sharing and distributing it. They have recently been joined by the DNA Data Bank of Japan (DDBJ), which is in charge of monitoring Asian journals and contributing to the reciprocal exchanges. (DDBJ served primarily as an access node to GenBank" and EMBL starting in 1984, but did not start gathering its own data until 1987.)

GenBank".-GenBank® originated at the DOE's Los Alamos National Laboratory in 1979 and started to receive funding from the NIH in 1982. It is the major U.S. database for nucleic acid sequence information **from humans and other organisms (3). GenBank" is presently administered and receives a major portion of its funds from the National Institute of General Medical Sciences (NIGMS) of NIH. Data are entered and updated by curators at Los Alamos and are distributed by Intelligenetics Corp. (Mountain View, CA).**

The amount of data contained in GenBank" has grown exponentially since its inception. In addition, the number of users has increased from a small set of one hundred or so who accessed it when the first NIH contract started to tens of thousands of scientists who now access either directly or through commercial distributors. GenBank"'s new 5-year contract, which took effect in October 1987, significantly increases funding to meet the growing demand.

Protein and Amino Acid Sequences and Structures

Databases that gather information on protein and amino acid structure and function are crucial for the application of genomics research to clinical and pharmaceutical problems, as well as for advancing the understanding of basic problems in biology-how genes

function, how they code for proteins and enzymes, and how their protein products are structured and function (see ch. 2). The effects of map and sequence data on these databases will depend on the strategy followed for genome projects. For example, a concerted nucleotide sequencing effort would affect research on protein and amino acid structure more slowly than increased funding to researchers studying specific genes and their gene products—generally proteins (6).

Protein Identification Resource (PIR).-PIR is "a resource designed to aid the research community in the identification and interpretation of protein sequence information" (14). It contains sequence data for proteins and amino acids, with annotations that indicate known functional regions. PIR is run by the nonprofit National Biomedical Research Foundation and receives most of its funding from NIH's Division of Research Resources. Modest user fees cover the distribution costs; academic users pay a flat fee, while commercial users are charged by the amount of computer time they use. PIR has recently started cooperating with the Japan International Protein Database (JIPID) and the new European database, Martinsreid Institute for Protein Sequence Data (MIPS), to establish an international data network for protein sequences.

Protein Data Bank (PDB).–The Protein Data Bank was founded in 1971 as "an international computerized archive for structural data on biological macromolecules" (1), [t gathers information on the atomic coordinates of the structure of nucleic acids, messenger RNA, amino acids, proteins, and carbohydrates that have been derived from crystallographic studies. Structural information is a vital link in the understanding of how proteins function, which eventually leads to knowledge of the mechanisms of genetic disease and suggests possible directions for rational drug design.

PDB is based at DOE's Brookhaven National Laboratory and supported primarily by NSF, with additional funds from the National Institute of General Medical Sciences of NIH. Modest user fees help cover the costs of distribution. Use of the database has been growing rapidly and is predicted to continue growing in parallel with human genome projects. Linking PDB with databases that contain genetic map and sequence information will enhance the long-term goals of human genome research (12).

Present and Future Needs

The many types of information that are produced in molecular biology necessitate the maintenance of a ~'arietuv of specialized databases. At the same time, however, the information in different databases must often be combined in order to understand the full dimensions of an-y specific research problem. It is crucial for the scientific community to be able to access information on a topic of interest from a variety of databases that may handle different aspects of the problem. Thus databases must use standardized or easily translatable formats and they must be interconnected. The problem of format has been recognized and is being addressed in scientific meetings, by database advisers, and by funding agencies. Several programs are underway to improve the linkages between databases. An experimental project at the National Library of Medicine, discussed below, will develop a system to link a variety of databases relevant to molecular biology.

The speed with which data are entered into the databases has been a major concern. The exponential increase in data has not always been matched by increases in the support for databases and personnel to operate them, causing a lag time of several months or even years between the publication of data and their entry, in fully annotated form, into databases. If the lag time is excessive, the efficiencies of centralized data management and retrieval are lost. One solution that is being explored is the direct submission of data to the databases by the researchers as a requirement for publication in journals. At least one journal has already agreed to cooperate with GenBank® and EMBL in an attempt to speed acquisitions in this way (19). Another possibility is to encourage funding agencies to make the submission of data or materials to the appropriate databases a condition of receiving research grants. The automation of data entry will be necessary as the amount of data increases. Automated methods are already under development; the capacity to enter data may be built into some automated sequencing machines.

The timely exchange of data is also affected by issues of intellectual property rights and technology transfer. Open and rapid exchange of information and materials speeds research and is particularly important when the data have medical or clinical implications. If the data and materials become commercially valuable, however—and many researchers predict that they will—the values of open access and free exchange could clash with the desire to protect proprietary rights on potentially patentable data or materials. Because access to databases and repositories is international, there are concerns that U.S.-funded research could be commercialized by other countries. The problems are not intractable, however: There are set'eral successful precedents of advance contracts that specifv how data will be contributed to databases while protecting property rights (4,21). (See also ch. 8.)

A major problem faced by databases for the past decade has been insufficient funding for handling the exponential increase of data. Costs will continue to rise

83-360 0 - 88 - 8 Ql. 3

as more map and sequence data are generated. The government agencies and other organizations that support genome projects appear to recognize the importance of continued funding for relevant databases. For example, the increased budget in the new GenBank® contract (for 1987 through 1991) indicates that funding agencies are aware of the need to enhance database maintenance. An initiative within the National Library of Medicine to strengthen information resources for molecular biology and biotechnology (discussed below) should lend further support to databases needed for genome projects. The Howard Hughes Medical Institute has been particularly active in supporting database resources and networks to link them. It is essential that financial support continue to keep pace with the growing body of data.

Repositories

Genome projects will generate biological materials as well as sequence and map data. Access to these materials is a key element in making the map information useful. A scientist searching for a gene of unknown location would want to have access to a panel of DNA markers that could give an approximate location, then a more closely spaced set of markers to locate it more precisely. Once the gene's location was established on the genetic map, the investigator would select DNA clones covering that region of the human chromosomes from a repository, thus obtaining the DNA encoding the gene. Each of these steps would require access to a set of cloned DNA fragments. Existing repositories are hardly sufficient, but how much must be invested in them will depend on conclusions on the value of centralized sources rather than housing materials in individual labs.

Companies developing a new product derived from or related to a human gene would also wish to have access to such materials in many instances. Storage and handling of such DNA resources is thus a crucial function. The materials will be most widely useful if they are stored at national collection and storage facilities. DNA probes, vectors, and some other materials are best maintained at a facility such as the American Type Culture Collection (ATCC). Others, such as cell lines derived from individuals and families with genetic diseases, are stored in the Human Genetic Mutant Cell Repository in Camden, New Jersey. Other materials that are unlikely to have substantial demand from a wide variety of investigators might be stored at the laboratories that generated them and distributed on a more informal basis to those requesting them. Present methods and technologies for the amplification, characterization, storage, and distribution of materials are expensive and time-consuming; the costs of storage could become a major component of mapping and sequencing projects. Newer and cheaper storage methods will have to be developed as production of DNA fragments increases. The development of automated techniques for organizing, managing, and accessing materials will be necessary; research on automated repository management is already underway at ATCC and at DOE's Los Alamos National Laboratory (11, 21).

Even with the advent of automated repository management techniques, however, the high cost of storing and maintaining materials makes the selection of materials to collect particularly crucial. While it might be desirable to keep large collections of clones generated in an attempt to develop libraries of overlapping clones or contigs (see ch. 2), the curators of repositories and the scientists who use them will have to choose which materials are of utmost importance, and these decisions should be periodically reviewed (22,23).

American Type Culture Collection

The ATCC maintains a variety of different collections of animal, plant, and bacterial cell lines, hybridomas, phage, and recombinant DNA vectors, as well as an NIH-sponsored repository of human DNA probes and chromosome libraries (20). The collection of chromosome libraries includes materials from DOE's National Gene Mapping Library (see ch. 5). The ATCC amplifies and stores samples and distributes them, along with pertinent information, to investigators for a nominal fee. Investigators must agree not to use the materials for commercial purposes nor to sell them,

The repository maintains a database of information on the source and characteristics of the material in its collection. Its advisory committee has recommended that the database be included in a mapping database such as HGML.

Human Genetic Mutant Cell Repository

Sponsored by the National Institute of Genera] Medical Sciences of NIH, the Human Genetic Mutant Cell Repository was founded in 1972 to maintain a collection of well-characterized human cell cultures (2,17). The cultures are available to investigators worldwide at a nominal fee. The repository contains over 4,000 individual cultures, which represent more than 400 genetic diseases and 700 to 800 chromosomal aberrations (7). The curators of the collection have increasingly sought to include material from multigenerational family groups for linkage analysis; the repository now maintains cell lines from the Venezuelan Huntington's pedigree (see box 7-A) and others such as cystic fibrosis families, families with fragile X-linked mental retardation, and so on.

Data Analysis, In formatics, and Computer Resources

Development of analysis methods to search for and compare sequence information, to predict sequences that code for proteins and the structures of those proteins, and to aid in other aspects of the analysis of data from genome projects will eventually need to utilize parallel processing techniques and the capacity of supercomputers. Most researchers agree that the hardware to tackle the complex problems of sequence analysis and comparison already exists but that satisfactory software must be developed. The DOE, the NIH, the NLM, and the NSF support various programs and grants for the development of software to represent and analyze data and for the development of computer resources such as supercomputing centers and computer networks. Several of these resources are described below. Numerous private firms are developing or marketing computer programs that search databases or analyze data on nucleic acid or protein sequences.

BIONET™

BIONET™ is a nonprofit computer network run by Intelligenetics, Inc. [Mountain View, CA) and funded by the Division of Research Resources of NIH and by modest user fees (13). Its goals are to "provide computation assistance in data analysis and problem solving to molecular biologists and researchers in related field, to serve as a focus for the development and sharing of new software, and to promote rapid sharing of information and collaboration among a national community of scientists" (9). BIONET™ provides access to several major databases (GenBank", EMBL, PIR, PDB, and databases of restriction enzymes and plasmid vectors) as well as to software for analyzing nucleic acid and protein sequences. The network also aids communication between its members through a series of bulletin boards on topics of user interest and through an electronic mail system. BIONET™ serves users in the United States, Canada, and Europe.

National Biotechnology Information Center

The National Biotechnology Information Center is an initiative to develop and enhance a range of tools for molecular biology information that is being sponsored by the National Library of Medicine (NLM) (18). The project is presently the subject of several authorizations bills but has already received some appropriations for a range of projects, including the building and maintenance of databases, developing a comprehensive listing of existing databases, and improving information retrieval systems. NLM has already developed a prototype of a retrieval system, called the Information Retrieval Experiment (IRX) that connects data from several different databases and graphic and visual sources. For example, a database search for a specific disease gene will yield information on whether the gene has been mapped, the map of the gene in graphic form, bibliographic information on publications about the map, as well as information on clinical symptoms, diagnosis, and visual representations of affected patients (X-rays, diagrams, photos, and so on). The NLM initiative will enhance the management of data from genome projects and will forge links between information from many areas of molecular biology to aid in basic and biomedical research (15). The NLM is in an advantageous position to coordinate database activities through its expertise in handling information through existing literature databases such as MEDLINE.

The Matrix of Biological Knowledge Workshop

The Matrix of Biological Knowledge Workshop, a month-long conference held during the summer of 1987, was an attempt to formulate models and make recommendations for the organization of knowledge and data from all disciplines in biology (16). R was sponsored by the NIH, the DOE, the Sloan Foundation, and the Santa Fe Institute.

The workshop grew out of the efforts of a committee sponsored by the NIH that attempted to set forth and evaluate models used in biomedical research. Several scientific meetings prior to the workshop had addressed the particular complexities of biological data; at the workshop, biologists, computer scientists, and database experts actually tried to work out some of the problems raised at earlier meetings. Participants at the workshop issued the following general recommendations:

> . . . that support for a centrally coordinated effort to establish a knowledge base of databases in the biological sciences be aggressively pursued; that the current independent efforts to establish interdatabase structures and analysis tools be coordinated with a long-term view towards maximum integration; . . . that these coordinated efforts incorporate the most up-todate computer science and analytical methods; and finally, that these activities directly involve the experimental and biotechnology communities in order to ensure the uti]-ity of the ensuing developments (16).

These recommendations appear to reinforce the direction of ongoing efforts in agencies that sponsor databases. The specific recommendations issued by work-

ing groups in each of seven broad categories may prove useful for the future management of databases in all of biology.

Appendix D References

1. Abola, E. E., Bernstein, F.C., and Koetzle, T.F., "The Protein Data Bank," in P.S. Glaeser (cd.), *The Role of Data in Scientific Progress (New* York, NY: Elsevier Science Publishers, 1985).
2. Aronson, M. M., Miller, R.C., Nichols, *W .W.*, et al., "Breakpoint Map of Human Translocation Cell Cultures Available From the NIGMS Human Genetic Mutant Cell Repository," Journal *of Cytogenetics* and Cell *Genetics* 30:179-189, 1981.
3. Burks, C., Fickett, J.W., Goad, W .B., et al., "The GenBank" Nucleic Acid Sequence Database," *Computers Applications in Bioscience 1:225-233, 1985.*
4. Cahill, G.C., Vice President for Scientific Training and Development, Howard Hughes Medical Institute, Bethesda, MD, personal communication, December 1987.
5. Cassatt, J., National Institutes of Health, personal communication, April 1987.
6. George, D., Protein Identification Resource, National Biomedical Research Foundation, Washington, DC, personal communication, December 1987.
7. Greene, A. E., Human Genetic Mutant Cell Repository, Coriell Institute for Medical Research, Camden, NJ, personal communication, March 1988.
8. Harem, G., and Cameron, G., "The EMBL Data Library," Nucleic *Acids Research 14:5-9, 1986.*
9. *Introduction to BZONET* Mountain View, CA: Intelligenetics, Inc., 1987).
10. Kidd, K., Lecture at "Genomics Meeting: Assessing the Repository, Informatics, and Quality Control Needs," Lawrence Livermore National Laboratory, Livermore, CA, Aug. 26-27, 1987.
11. Knobeloch, D., and Beugelsdijk, T.J., '(Automated Sample Management: Organizing, Managing, and Accessing the DNA Fragments Generated in the Process of Mapping and Sequencing the Human Genome," in *Repository, Data Management, and Quality Assurance Needs for the National Gene Library and Genome Ordering Projects, see* ref. 23.
12. Koetzle, T., Protein Data Bank, Brookhaven National Laboratory, Upton, New York, NY, personal communication, September 1987.
13. Kristofferson, D., "The BIONET™ Electronic Network," *Nature 325:555-556, 1987.*
14. Ledley & Barker, Protein Identification Resource project description, November 1987.
15. Masys, D., National Library of Medicine, Bethesda, MD, personal communication, December 1987.
16. Morowitz, H.J., and Smith, T. (eds.), *Report of the Matrix of Biological Knowledge Workshop* (Santa Fe, NM: Santa Fe Institute, 1987).
17. "The National Institute of General Medical Sciences Human Genetic Mutant Cell Repository, " *Somatic Cell and Molecular Genetics 12:421, 1986.*
18. National Library of Medicine, "Biotechnology Information: A Plan for the National Library of Medicine," unpublished report, 1987.
19. "A New System for Direct Submission of Data to the Nucleotide Sequence Data Banks," *Nucleic Acids Reseamh 15, No. 18, 1987.*
20. Nierman, W.C., Benade, L. E., and Maglott, D. R., *American Type Culture Collection: NIH Repository of Human DNA Probes and Libraries* (Rockville, MD: American Type Culture Collection, 1987).
21. Stevenson, R., American Type Culture Collection, Rockville, MD, personal communication, October 1987.
22. U.S. Congress, Office of Technology Assessment, "Costs of Human Genome Projects," workshop held Aug. 7, 1987 (see app. A).
23. U.S. Department of Energy, Office of Health and Environmental Research, and National Institutes of Health, *Repository, Data Management, and Quality Assurance Needs for the National Gene Library and Genome Ordering Projects,* unpublished report from a workshop, August 1987.