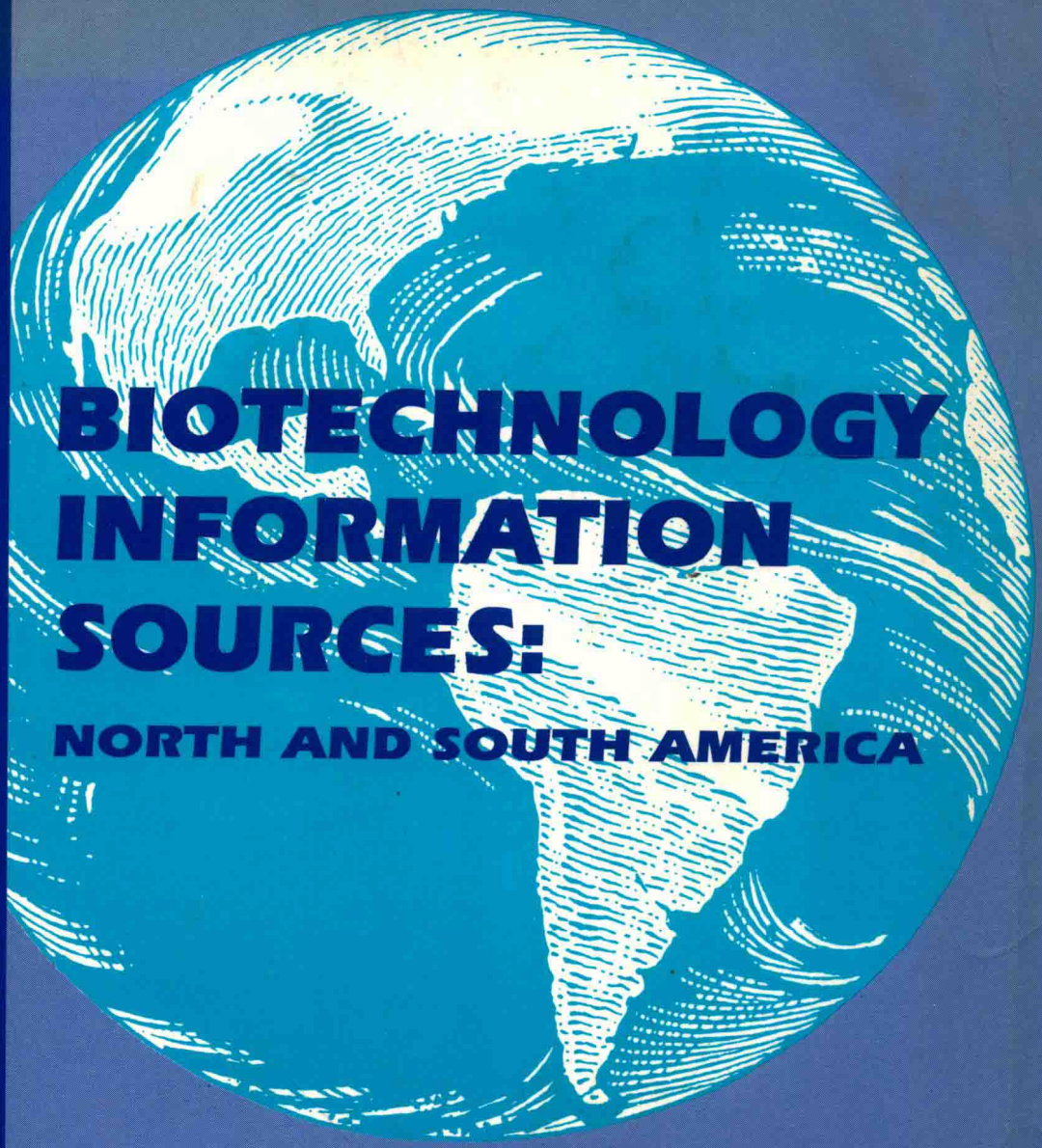

BIOTECHNOLOGY INFORMATION SOURCES:

NORTH AND SOUTH AMERICA

Compiled and Edited by

Barbara A. Rapp

icsti

BIOTECHNOLOGY INFORMATION SOURCES:

NORTH AND SOUTH AMERICA

Compiled and Edited by

Barbara A. Rapp
National Library of Medicine

Published for the International Council for Scientific and Technical Information

by Learned Information, Inc.
Medford, New Jersey
1994

ISBN: 0-938734-81-4
Price: $32.50
Cover Design: Jennifer Johansen

Printed in the United States of America

BIOTECHNOLOGY INFORMATION SOURCES:

NORTH AND SOUTH AMERICA

Acknowledgments

The author would like to acknowledge the contributions of the following people: Dr. Dennis Benson, who provided valuable suggestions and editorial review at all stages of the project; Dr. Trudi Bellardo Hahn, who did background research and writing for this report, provided important editorial guidance, and compiled the index; Lori Wagoner, who assisted with information gathering and organization, with special emphasis on journals, newsletters, and Internet information resources; Stephanie Lipow, who provided assistance with the section on Internet information resources and reviewed the full report; Mary Tackett and Stacey Arnesen, who reviewed resource descriptions; Gale Dutcher, who reviewed resource descriptions and provided assistance using the DBIR database; and the many other people who reviewed selected resource descriptions. The author also thanks the Life Sciences Working Group of the International Council for Scientific and Technical Information for providing the opportunity to compile this resource guide.

Contents

1. Introduction

The rapidly expanding field of biotechnology presents an enormous challenge in keeping pace with current developments. In response to the needs of the biotechnology community, hundreds of information products have emerged. A wide variety of databases and print products supplies information targeted to the research, technological, and commercial aspects of biotechnology.

For the research community, the most important development is the emergence of databases containing vast quantities of primary research data from genetic mapping and sequencing projects. Traditionally, the journal literature has been the major source of primary research information, and it continues to be indispensable. However, because of the vast amounts of research data generated by genetic sequencing and mapping projects, journals cannot reproduce the data in full, and the primary research databases have become extensions of experimental laboratories. The databases range in size and coverage from the comprehensive databases such as GenBank, which covers all known DNA sequences, to specialized databases focusing on a single organism, chromosome, or class of protein.

The development of the research databases has been a grass-roots phenomenon. Virtually all have grown from the work and initiative of individual scientists working alone or in small groups. Even now that many of the databases have become large-scale enterprises, their continued growth and success depends on the active support and participation of the contributing scientists. New databases continue to emerge as basic research progresses and new sets of questions are posed to databases.

The enormous organizational and computational challenge of providing the research data in a comprehensive and useable manner has created a growing infrastructure of information systems and resources. The major database producers are working together to create a system that will accommodate new databases and integrate them into the web of interrelated resources. Governmental initiatives are also contributing to these developments. The importance of information infrastructure for biotechnology is highlighted in the report *Biotechnology for the 21st Century*, prepared by the Federal Coordinating Council for Science, Engineering, and Technology in 1992, where information infrastructure is included as one of the important areas of activity.

For the business and technology development communities, many new journals and newsletters report current research trends, track patenting activity, and discuss issues related to technology transfer and business opportunities.

In the area of secondary information sources that facilitate retrieval or retrospective access to information, many new print and online products are devoted to biotechnology. In addition, many of the general bibliographic databases include specific features for defining biotechnology-related subsets.

The Internet is used heavily by scientists for searching databases, downloading data, and communicating among themselves individually or as members of discussion groups. In addition, CD-ROM is an important data distribution technology because of the large size of many of the databases. Online interactive access through the major database vendors continues to be important for accessing bibliographic databases, newsletters, and other text databases. The trend in the online industry toward providing network access to their systems may encourage increased use by scientists, who are becoming more accustomed to using the Internet as a regular part of their research activity.

1.1 Scope and Organization of This Report

This guide focuses on biotechnology information resources available in the countries of North and South America. Most of the resources in this guide are produced within this same geographic area, but some are produced in Europe or Japan. The latter group are included only if they are accessible through vendors and suppliers in North or South America. The global interconnectivity of modern electronic information resources has made geographic boundaries less confining, so for databases available on the Internet we include some additional computer addresses outside the geographic scope of this report.

Because of the broad reach of biotechnology, relevant information can be found in many types of information sources throughout science, engineering, and business. The scope of this report is limited to those resources explicitly designed to support biotechnology or others that are of obvious relevance to the field.

This report is organized by type of information resource, not by potential audience or use. Following this introduction, Chapter 2 covers the primary research databases that are used extensively every day as an integral part of molecular biology research. This chapter includes molecular sequence databases, genetic and physical mapping databases, molecular structure databases, all of which contain the actual research data in addition to bibliographic references to published papers in which the data are reported. The chapter also covers other databases with important primary information, including those with factual information about metabolic compounds and restriction enzymes and those with information on sources of research materials such as organisms, cell lines, and DNA probes.

Chapter 3 covers indexes to the primary scientific literature, in both online and print form. Bibliographic databases that focus primarily on areas important to biotechnology, as well as more general scientific databases, are discussed. For the printed indexes, only those with a focus on biotechnology are covered. Because of the widespread importance of molecular sequence data, some bibliographic database producers now tag articles in which sequence data is published with specific subject headings to facilitate retrieval. This is being done currently for AGRICOLA, BIOSIS Previews, Chemical Abstracts (CA), and MEDLINE.

Chapter 4 covers applied biotechnology journals, online and printed newsletters, and other sources of current news information. These information sources are of particular importance for business-related information in biotechnology. Coverage of the journal literature is limited to journals that address the application of basic research and methodologies to new developments in the biotechnology industry. Journals were restricted to those that would be appropriate for a variety of audiences interested in keeping up with advances in biotechnology. It was not possible to cover, or even list, the thousands of journals in the various scientific fields that contribute to biotechnology.

Chapter 5 is devoted to patents as a special source of scientific information. The chapter covers indexes to the patent literature, in both online and print form, as well as sources for obtaining molecular sequence data reported in the patents. Online databases in this chapter are of two types: those that cover the patents themselves and those that contain information about patents, e.g., business or legal information related to patents, or information on the status of patents in a particular area. The print information resources in this chapter include newsletters and journals that describe patents and patent applications, and indexing services that include coverage of patents.

Chapter 6 covers directories of all types, including directories of organizations, sources of equipment and biological research materials, federally funded research projects, and biotechnology software and information resources.

Because of the importance of worldwide communications networks for accessing molecular biology research information, Chapter 7 discusses the use of international communication networks to access the primary research databases as well as other information sources.

Chapter 8 describes organizations that provide information services in biotechnology. A growing number of libraries, referral centers, clearinghouses, and other organizations include among their activities a variety of efforts to build and maintain collections, provide information services, or offer education programs related to biotechnology. Many of these organizations are biotechnology research institutes whose mission is to promote excellence in research; to build a stronger theoretical understanding of biological phenomena; to foster

economic and industrial development; to protect public health and safety; or to enhance agricultural productivity through genetic transformation and chemical treatments.

1.2 Descriptions of Information Resources

The descriptions of information resources in this guide are intended to be informative enough to permit readers with a variety of needs to make initial decisions about usefulness and suitability. For further information about content, availability, costs, or conditions of access, readers are advised to consult database producers, online services, CD-ROM suppliers, and publishers whose addresses, telephone numbers, and electronic mail (e-mail) addresses are listed in Appendix A, B, C, and D, respectively.

Titles of databases appear in various directories, catalogs, and other publications in a variety of forms. In this guide, each database is listed by the full name, followed by the most common acronym, if there is one, in parentheses. Cross-references to link variations in database titles have been made. Database names and acronyms are listed in the index in order to assist the reader with identification. Although databases are generally listed alphabetically, in some cases major comprehensive resources are listed first, followed by an alphabetic listing of more specialized resources.

The name of the database producer appears under each entry. When a single individual is responsible for creating or maintaining a database, that individual's name is included with the corporate name of the producer. In Appendix A, the entry appears under the corporate name. Each entry also indicates how the resource is available: through an online vendor, as a public resource available on the Internet, or on a physical medium such as CD-ROM or magnetic tape.

Printed publications are also listed alphabetically by title, followed by the publisher's name, frequency of publication, and a description of scope and content.

1.3 Sources Consulted

A large number of written and online resources were used to compile this directory. Complete citations for the written materials are provided in the Bibliography. A few sources must be accorded special mention for invaluable contributions to this work: "Genome Databases" by Jacqueline Courteau published in 1991 in *Science*; the 1992 *Journal of NIH Research* guide entitled "Directory: Databases," and the 1992 *Directory of Online Databases* published by Gale Research. Two online directories of databases and other information

resources in biotechnology were also relied on extensively: the Listing of Molecular Biology Databases (LiMB) and the Directory of Biotechnology Information Resources (DBIR), both of which are described in Chapter 6.

2. Primary Research Databases

Access to research information always has been an important component of scientific endeavor. The concept of sharing and publishing research results to create a foundation of knowledge on which new research can build is central to our model of scientific progress. In molecular biology, scientists have a critical need for rapid access to primary research information in a format that can be searched and compared to new experimental results. Molecular biology databases are literally extensions of modern laboratories.

Basic research in molecular biology involves gene mapping, gene sequencing, protein identification and sequencing, determining the biological function of proteins, determining the structure of proteins, and ultimately, understanding the relationship between protein structure and function. This research cuts across a broad spectrum of disciplines, yielding advances in the basic life sciences, medicine, agriculture, and the environmental sciences. Three types of databases are now being created as central archives of the accumulated research data: molecular sequence databases; physical and genetic mapping databases; and molecular structure databases. The increasing quantity and complexity of sequences and structural data for proteins and nucleic acids create both challenges and opportunities for biomedical researchers. A new generation of practical computer tools for data analysis and integrated information retrieval is now emerging, and recent developments in rapid database searching, multiple sequence alignment, and molecular modeling provide the tools that scientists need to make use of the vast quantities of information in the databases (Boguski, 1992).

This chapter describes major databases that are important for molecular and structural biology. Section 2.1 covers comprehensive and specialized molecular sequence databases, and Section 2.2 covers physical and genetic mapping databases. Databases of both types have grown substantially because of the Human Genome Project, the Plant Genome Research Program, and related efforts to map and sequence model organisms. Section 2.3 covers molecular structure databases, which contain crystallographic and NMR (nuclear magnetic resonance) data for proteins and carbohydrate molecules whose structures have been experimentally determined. These databases, along with sequence databases, are used extensively in protein modelling and structure prediction. Section 2.4 covers databases that serve as catalogs for research materials such as tissue cultures, cell lines, and DNA probes. Section 2.5 covers other databases that contain information important for conducting molecular biology experiments, including databases with factual information about enzymes and other metabolic compounds, metabolic pathways, and restriction enzymes.

Each database is identified by its full name, followed by a common acronym or shorter name in parentheses, if applicable. Under each entry appears the name of the database producer if the resource represents the efforts of a particular organization, or the name of the curator if it represents primarily the work of one or two individual scientists. Addresses, telephone numbers, and e-mail addresses of the producers and curators appear in Appendix A. Each entry also indicates at least one way to access the database, under the following headings: *Online, E-mail Server, FTP, CD-ROM, Diskette,* and *Magnetic tape. Online* refers to online interactive access through modem or over the Internet. (Internet services, including e-mail, file transfer, and retrieval facilities, are described in detail in Chapter 7.) Online access usually is offered through subscriptions to commercial online vendors, but there are also free services, and these are noted. *E-mail Server* and *FTP* refer to network access services, and the network address and name of the organization providing the service are provided. These two services are provided free of charge. For CD-ROM, diskette, and magnetic tape, the name of the distributor is provided. Addresses, telephone numbers, and e-mail addresses for the online services appear in Appendix B, and in Appendix C for CD-ROM and magnetic tape suppliers who are not also the database producers.

2.1 Molecular Sequence Databases

In a recent review article on molecular biology databases (Boguski, 1992), molecular sequence data was referred to as "the common currency of modern biomedical research [that] often provides exciting and unexpected links between diverse systems that accelerate research progress." A molecular biologist today cannot afford to be unfamiliar with the DNA and protein sequence collections, and significant effort is currently being expended to make these data more accessible to the general scientific community. The number of sequences has been doubling about every 20 months and software tools for information retrieval and analysis are achieving a new level of sophistication, integration, and accessibility.

Molecular sequence databases contain the textual representation of nucleic acid and protein sequences, as well as descriptive and bibliographic information. For nucleic acids, either DNA or RNA, the sequence is the exact linear order of the nucleotides along the backbone chain of a particular segment of nucleic acid. In DNA, the backbone chains are composed of a combination of four nucleotides: adenine, thymine, guanine, and cytosine, represented as A, T, G, and C. In RNA, the constituent nucleotides are adenine, uracil, guanine, and cytosine, represented as A, U, G, and C. For proteins, the sequence represents the linear order of amino acid residues that make up the protein. The approximately 20 amino acids, which appear in hundreds of thousands of combinations to form the universe of proteins, are generally represented as single letters in the databases, although in some cases they are represented as three-character abbreviations.

In addition to the sequence data itself, a database record typically includes such information as source organism, gene locus, literature reference(s) to papers in which the sequence was published or cited, and annotations describing specific biological features of the sequence. One of the important features is the specification of the coding regions for a gene and the corresponding conceptual translation, resulting in a theoretical protein sequence.

The major DNA sequence databases are the GenBank Genetic Sequence Database (GenBank), the EMBL Nucleotide Sequence Database (EMBL), and the DNA Data Bank of Japan (DDBJ). Since 1987, the producers of these databases in the United States, Europe, and Japan have worked together to collect and distribute a comprehensive international database of all reported genetic sequences. The databases include sequence data that is published in the scientific literature as well as data submitted directly to the databases by scientists. In fact, many journal editors require that authors submit their sequence data to the databases as a condition of publication. In mid-1993, the databases started including sequence data from United States, European, and Japanese patents.

In the United States, the National Center for Biotechnology Information (NCBI) assumed responsibility for the production and distribution of GenBank in October 1992. NCBI is located in the National Library of Medicine at the National Institutes of Health. For the previous ten years, GenBank had been administered elsewhere at NIH, in the National Institute for General Medical Sciences, and had been produced and distributed through contracts with IntelliGenetics, Inc. (1987-1992) and Bolt Beranek and Newman (1982-1987). During the entire ten-year period, Los Alamos National Laboratory (LANL), as subcontractor, collected and maintained the sequence data.

Since assuming responsibility for GenBank, the NCBI has developed an integrated approach to building and managing molecular sequence databases. Using the Abstract Syntax Notation (ASN.1) standard data description language issued by the International Standards Organization, NCBI has defined a data representation structure for sequence databases. Using this structure, NCBI has converted data from multiple sources to a common format to create integrated resources. This integrated approach is exemplified in NCBI's *Entrez: Sequences* CD-ROM, which provides a common view of sequences from several DNA and protein sequence databases, as well as MEDLINE abstracts for papers in which the sequences were published. NCBI also continues to distribute GenBank as a discrete database in its traditional flat file format, since many software packages depend upon that format. The GenBank database, as well as other sequence databases, is also redistributed by third parties, generally as part of a commercial software package that includes sequence analysis programs.

There are two major protein sequence databases. The Protein Information Resource (PIR) is maintained in the United States by the National Biomedical Research Foundation (NBRF) at Georgetown University, in collaboration with the Martinsried Institute for Protein Sequences

in Germany and the International Protein Information Database in Japan. As an international resource, it is also known as the Protein Sequence Database. The SWISS-PROT Protein Sequence Data Bank (SWISS-PROT) is maintained collaboratively by the EMBL Data Library and Dr. Amos Bairoch at the University of Geneva. Although the protein databases originated independently from the DNA databases, there are cooperative activities among the producers of both types of databases because of the natural relationship between DNA and protein.

The comprehensive sequence databases are listed and described in Section 2.1.3. In addition, there are a growing number of specialized sequence databases, covering, for example, specific organisms or particular classes of proteins. A representative selection of specialized databases is described in Section 2.1.4. These databases are typically maintained by one or two curators as an integral part of their research activities. Virtually all are distributed free of charge, often directly by the curators. Many are available from Internet FTP sites and through WAIS or Gopher servers. The network access servers are managed individually by interested scientists or by the large database providers such as NCBI, EMBL, DDBJ, and NBRF.

2.1.1 Using the Sequence Databases

Sequence databases are used in two different ways. First, in order to retrieve a particular sequence or set of sequences, the database can be searched by text terms such as accession number, author name, title words, gene name, or organism name. This is the same type of text retrieval search as used for bibliographic or other text-based retrieval. However, molecular biologists also have a need to search the sequence data directly, comparing a given sequence to all other sequences in the database in order to identify sequences that are similar, and therefore potentially homologous, to the query sequence. This can be a computationally intense operation, requiring multiprocessor computers and sophisticated pattern-matching algorithms. Two public-domain programs are widely used for sequence similarity searching: FASTA and BLAST. The BLAST algorithm developed in 1989 at NCBI, permitting a 1,000-base sequence to be compared against more than 100,000 sequences in less than 15 seconds, provides significant increases in search speed over FASTA. These algorithms or variants are used by several free e-mail servers. In the United States, the NCBI provides BLAST e-mail searches at the address blast@ncbi.nlm.nih.gov, and NBRF provides FASTA e-mail searches at the address: fileserv@nbrf.georgetown.edu.

Once particular sequences are identified, they can be used and analyzed in a number of ways. Software tools for many applications ranging from sequence assembly to multiple alignment to protein modelling are available in the public domain as well as from commercial molecular biology software producers.