The GenBank® genetic sequence data bank

Howard S. Bilofsky^{1*} and Christian Burks²

¹Life Sciences/Research Systems Department, BBN Laboratories Incorporated, 10 Moulton St., Cambridge, MA 02238 and ²Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, University of California, Los Alamos, NM 87545, USA
Submitted August 17, 1987

ABSTRACT

The GenBank Genetic Sequence Data Bank contains nearly 15000 entries for DNA and RNA sequences that have been reported since 1967. This paper briefly describes the contents of the database, the forms in which the data are distributed, and the services available to scientists using the GenBank database.

INTRODUCTION

The GenBank Genetic Sequence Data Bank is a U. S. Government-sponsored, computerized repository of nucleotide sequences, annotated to provide information about the bibliographic, functional, and physical contexts of the sequences. As of August 1987 the database held approximately 15,000 entries containing nearly 15,000,000 nucleotides. These entries have been compiled from some 18,000 articles and unpublished direct submissions.

GenBank was created in 1982 by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) in response to a critical scientific need for a timely, centralized, accessible repository for genetic sequences. Current co-sponsors include the National Institute of General Medical Sciences, the National Cancer Institute, the National Library of Medicine, the National Institute of Allergy and Infectious Diseases, the National Heart, Lung, and Blood Institute, the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases, the National Institute of Dental Research, the National Eye Institute, the Division of Research Resources of the NIH, The National Science Foundation, the Department of Energy, and the U.S. Department of Agriculture.

The Theoretical Biology and Biophysics Group at Los Alamos National Laboratory (LANL) gathers, annotates and organizes the data in the GenBank database. The Life Sciences/Research Systems Dept. of BBN Laboratories Incorporated (BBN), a research, development, and consulting firm, maintains the computerized data center and distributes the database.

The GenBank project's sponsors have appointed database curators from the scientific community. Each curator assists the team at LANL in collecting and organizing the information in one portion of the database. The government also appoints scientists to an advisory panel that provides suggestions on the project's future directions.

DATABASE CONTENTS

The sequences in the GenBank database are organized roughly along taxonomic lines with consideration given to reflect areas of scientific interest. Currently there are twelve major taxonomic divisions as indicated in Table I. Many sequences first appear in the database as unannotated entries, the thirteenth division. These entries, which include initial sequence reports and consequently have not been annotated for biological features, represent an effort to make new sequences accessible with a minimum of delay. The rapid growth of the database is shown in Figure 1. During the first two years of its existence, it took roughly nine months to increase by 1 million bases. The last such increase took only nine weeks. In addition to sequence data, entries include biological annotations of structural and functional features of sequences, bibliographic citations,

Table I. Summary showing the taxonomic distribution of sequences in Release 52.0 of GenBank.

Taxonomic Division	Sequences	Bases
PRIMATE	1803	2258373
RODENT	2113	1998618
Other MAMMAL	419	458484
Other VERTEBRATE	612	579596
INVERTEBRATE	937	865954
PLANT	938	1252036
ORGANELLE	498	957487
BACTERIAL	1147	1687519
STRUCTURAL RNA	738	84066
VIRAL	1284	1998510
BACTERIOPHAGE	207	326586
SYNTHETIC	268	81876
UNANNOTATED	3056	2306042
Summary:	14020	14855147

taxonomic information, comments about specific observations from different researchers, and cross-referencing information for related databases, such as the Howard Hughes Medical Institute Human Gene Mapping Library at New Haven. The format of a typical entry has been shown previously (1). The information included in the magnetic tape distribution format consists of indices, summaries, and directories in addition to the nucleotide sequences and has grown to more than 50 million characters.

Most of the sequences included in the database continue to appear first in refereed journals. A growing number, approaching a third of total entries, is being submitted directly to us and to our collaborators at the European Molecular Biology Laboratory (EMBL). With the ever-increasing rate of generation of new nucleotide sequences, such direct submissions have become increasingly important. Even when a sequence has been published, the timeliness and correctness of its appearance in the database are enhanced if the authors provide computer readable or clear copy forms of the sequence data and associated annotations directly to the GenBank or EMBL database teams.

An important aspect of each entry is an accession number by which each sequence can be uniquely identified. Scientists are encouraged to refer to GenBank and EMBL database sequences in print by these accession numbers.

Figure 1.

DATABASE DISTRIBUTION AND SERVICES

The database is distributed to researchers worldwide in several forms. Currently it is available on standard nine track tape for use on microcomputers and mainframe systems. A compressed binary format on floppy diskettes can be used with IBM-compatible personal computers. An eight volume compendium of sequences found in the GenBank and EMBL databases was published in 1987. The database is also accessible online to subscribers. All GenBank services are available for the cost of providing the service.

Online GenBank access provides users with flexible tools for locating and retrieving data of interest on the same NIH-owned computer used to prepare each release. This computer, also used to provide access to the NIH PROPHET system, can be accessed by direct long-distance dialup or over the Telenet telecommunications network. Online users utilize the GenBank Menu System, which provides versatile utilities for searching and downloading entries. Although these menus do not provide database-wide homology searching, or structure prediction algorithms, they do allow sequence selection based on comparison with a short subsequence, even permitting some mismatching. Many public, private and commercial concerns do provide tools to analyze GenBank data. Since there are no restrictions on the use or redistribution of GenBank data, many commercial providers of software for sequence analysis offer tools for handling the database in any of its widely distributed formats, including the tape and floppy diskette formats provided by GenBank. A listing of providers of such tools and services is available online through the GenBank Software Clearinghouse menu option.

ADDITIONAL INFORMATION

Effective October 1, 1987, GenBank operations are being administered by IntelliGenetics, Inc. Inquiries concerning database services and placing orders should be directed to them at (415)-962-7360. Questions concerning tapes or floppy diskette copies of the database distributed prior to October 1 or about placing an order for floppy diskettes based on Release 52.0 of GenBank should be directed to BBN at (617)-497-2742. Inquiries regarding the contribution of sequence data to the GenBank database should be directed to LANL at (505)-667-7510 or by electronic mail to genbank@lanl.gov.

ACKNOWLEDGEMENTS

GenBank® is a registered trademark of the U.S. Department of Health and Human Services. The GenBank project is funded by NIH contract N01-GM-2-2127. Work at LANL is also supported under the auspices of the U.S. Department of Energy.

¹ Additional GenBank staff at BBN: June Bossinger, Carl Foeller, Frances I.Lewitter, Wayne P.Rindone and C.David Swindell

² Additional GenBank staff at LANL: Edwin J.Atencio, Michael J.Cinkosky, Victor I.Esekogwu, James W.Fickett, Brian T.Foley, Walter B.Goad, Jamie E.-D.Hayden, Natalie Lopez, Kersti A.MacInnes, Thomas G.Marr, Annette V.Martinez, Frances A.Martinez, Mia J.McLeod, Santosh K.Mishra, Debra Nelson, Carol R.Schermer, Maura T.Smith, Brian L.Trujillo, and Chang-Shung Tung

*To whom correspondence should be addressed

REFERENCE

1. Burks,C., Fickett,J.W., Goad,W.B., Kanehisa,M., Lewitter,F.I., Rindone,W.P., Swindell,C.D., Tung,C.-S., and Bilofsky,H.S. (1985) *Comp. Appl. Biosci.* 1,225-233.