Volume 14 Number 1 1986 Nucleic Acids Research

The GenBank® genetic sequence databank

Howard S.Bilofsky, Christian Burks[1], James W.Fickett[1], Walter B.Goad[1], Frances I.Lewitter, Wayne P.Rindone*, C.David Swindell and Chang-Shung Tung[1]

Computer and Information Sciences Division, BBN Laboratories Inc., 10 Moulton St., Cambridge, MA 02238, and [1]Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, University of California, Los Alamos, NM 87545, USA

Received 17 July 1985

ABSTRACT

The GenBank® Genetic Sequence Data Bank contains over 5700 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 database is distributed, and the services we offer to scientists who use the GenBank database.

INTRODUCTION

The GenBank® Genetic Sequence Data Bank is a U.S. Government-sponsored, computerized repository of all reported nucleic acid sequences, catalogued and annotated for sites of biological interest. The October 1985 GenBank release contains 5,731 entries, comprising a total of 5,248,932 nucleotide bases. These entries have been compiled from some 4400 articles and about 130 unpublished direct submissions.

The Data Bank 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. Cosponsors include the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, the National Institute of Arthritis, Diabetes and Digestive and Kidney Diseases, and the Division of Research Resources of the NIH, as well as the National Science Foundation, the Department of Energy, and the Department of Defense.

The Theoretical Biology and Biophysics Group at Los Alamos National Laboratory (LANL) gathers, annotates and organizes the data in the GenBank database. The Computer and Information Sciences Division of BBN Laboratories Inc.(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

• GenBank is a registered trademark of the U.S. Department of Health and Human Services

Table I. Summary of sequences contained in each of the twelve divisions of Release 37.0 of the GenBank database.

Database Division	Sequences	Bases
Primate Sequences	698	787,639
Rodent Sequences	956	734,546
Other Mammalian Sequences	198	169,397
Other Vertebrate Sequences	364	255,394
Invertebrate Sequences	479	306,504
Plant Sequences	406	397,912
Organelle Sequences	315	387,867
Bacterial Sequences	570	696,681
Structural RNA Sequences	528	58,443
Viral Sequences	895	1,190,657
Bacteriophage Sequences	138	207,668
Synthetic Sequences	184	56,224
Summary:	5,731	5,248,932

```
LOCUS       HUMANFH       345 bp ds-DNA              updated    11/11/85
DEFINITION  Human atrial natriuretic factor, partial cds.
ACCESSION   K02399
KEYWORDS    atrial natriuretic factor.
SOURCE      Human fetal liver genomic DNA library of Maniatis, clone
            lambda-HA22.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Vertebrata; Tetrapoda; Mammalia;
            Eutheria; Primates.
REFERENCE   1  (bases 1 to 345)
  AUTHORS   Maki,M., Parmentier,M. and Inagami,T.
  TITLE     Cloning of genomic DNA for human atrial natriuretic factor
  JOURNAL   Biochem Biophys Res Commun 125, 797-802 (1984)
FEATURES          from  to/span     description
    pept             1       123    atrial natriuretic factor, exon 1
                   244  >    345    atrial natriuretic factor, exon 2
    sigp             1        75    atrial natriuretic factor signal
                                     peptide
  SITES
    refnumbr         1         1    numbered 1 in [1]
    ->pept           1         1    ANF cds signal pept start
    pept/pept       76         0    ANF cds signal pept end/propept start
    pept/IVS       124         0    ANF cds exon A end/intron 1 start
    IVS/pept       244         0    ANF cds intron 1 end/exon B start
    pept/pept      346         0    ANF cds propept sequenced/unsequenced
BASE COUNT       75 a     93 c     88 g     89 t
ORIGIN      7 bp upstream of SacI site.
        1 atgagctcct tctccaccac caccgtgagc ttcctccttt tactggcatt ccagctccta
       61 ggtcagacca gagctaatcc catgtacaat gccgtgtcca acgcagacct gatggatttc
      121 aaggtagggc caggaaagcg ggtgcagtct ggggccaggg ggctttctga tgctgtgctc
      181 actcctcttg atttcctcca agtcagtgag gatccctttc cctggtattt tcctttttta
      241 aagaatttgc tggaccattt ggaagaaaag atgcctttag aagatgaggt cgtgcccccc
      301 caagtgctca gtgagccgaa tgaagaagcg ggggctgctc tcagc
//
```

Figure 1. A typical sequence entry, identified by primary accession number K02399, in GenBank distribution tape format.

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 and presented in the twelve major divisions listed in Table I. The total amount of information in the magnetic tape distribution, including index and directory files, totals more than 22 million characters, representing a 21-fold increase in size in the 37 months that the database has been available.

Most of the sequences included in the database were originally published in refereed scientific journals; a small but growing number of sequence reports are submitted directly to us or to our collaborators at the European Molecular Biology Laboratory(EMBL). 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 clean copy forms of the sequence data and associated annotations directly to the GenBank or EMBL database teams.

The information included with each sequence entry is illustrated in Figure 1. The format shown is that used on a standard GenBank distribution tape. Note that any reference to a GenBank entry should include the first number in the accession number list for the entry. A more detailed description of the organization, structure, and format of the database was published recently (1).

DATABASE DISTRIBUTION AND SERVICES

The database is distributed to investigators worldwide in several forms. Standard nine track magnetic tapes can be read on a wide variety of minicomputers and mainframe systems. A compressed format on floppy diskettes can be read on IBM-compatible personal computers. In 1984 and 1985 the GenBank and EMBL databases were published in combined form in special supplements to Nucleic Acids Research. Online access to the central GenBank computer is also available. All of the 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. The online tools are available through the

```
MAIN MENU:
1.LEARN        about the GenBank System
2.ORDER        a Users' Guide
3.COMMUNICATE  with other users
4.GROUPS       of sequences
5.SEQUENCE     operations
6.SOFTWARE     Clearinghouse
7.HANDY        operations
8.DONE         with GenBank menus, for now
Choice:
```

Figure 2. A typical menu in the GenBank Menu System.

GenBank Menu System, which provides versatile utilities for searching and downloading entries. A typical menu is shown in Figure 2.

The menus provide easy access to sophisticated methods for locating entries, including selection of entries containing a short subsequence with a specified level of base mismatches. They do not, however, provide database-wide homology searching or other sequence analytical utilities for sequence alignment or structure prediction. Many public, private, and commercial concerns do provide tools to analyze GenBank data.

Since the GenBank database is distributed with no restriction on how the data in it may be used or redistributed, some commercial sequence analysis packages deal directly with the GenBank floppy diskettes, others deal with sequence entries in the GenBank tape format, while still others provide their own diskettes containing some GenBank data. Information from the authors and distributors of such packages and services is available through the GenBank Software Clearinghouse menu, which can be chosen from the Main Menu illustrated in Figure 2.

ADDITIONAL INFORMATION

Inquiries about GenBank database services, orders for magnetic tape or floppy diskette copies of the database, and requests for online accounts 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. See also Burks *et al.* (1).

ACKNOWLEDGEMENTS

The GenBank project is funded by NIH contract NO1-GM-2-2127. Work at LANL is also supported under the auspices of the U.S. Department of Energy.

*To whom correspondence should be addressed

REFERENCES

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) CABIOS 1,in press.