Recent changes in the GenBank® On-line Service

David Benton
GenBank/IntelliGenetics, 700 E. El Camino Real, Mountain View, CA 94040, USA

Received November 28, 1989; Revised and Accepted February 1, 1990

ABSTRACT

The GenBank On-line Service provides access to the GenBank and EMBL nucleic acid sequence databases and to the Swiss-Prot and GenPept protein sequence databases. Users can query the databases by sequence similarity and annotation keywords and retrieve entries of interest. This access is available through e-mail servers, anonymous FTP, anonymous interactive login, and login to established, password-protected, individual accounts.

INTRODUCTION

In November, 1989, GenBank added substantial new functionality to its On-line Service. This article describes the current services provided by the GenBank On-line Service (GOS) and methods of obtaining access to these services. Currently, the on-line databases include the most recent quarterly releases of GenBank, EMBL, GenPept, and Swiss-Prot, and the data added to each of these (except Swiss-Prot) since their most recent releases (in the New Data databases). The GenPept database is a protein and peptide sequence database derived by automatically translating the protein coding regions of annotated GenBank entries. The three New Data databases are updated daily. Access to all GOS services is available to both commercial and non-commercial users at the same cost. On-line help is available for all aspects of this Service and user manuals are available from the GenBank On-line Service at the author's address.

INTERACTIVE ACCESS

Interactive access to the GOS databases is provided through the Telenet public data network, via remote login over the Internet, and by direct-dial telephone lines. At present, the IRX (Information Retrieval Experimental Workbench) program (1) is the primary interactive database retrieval program. Fifteen-minute IRX sessions are provided to users via anonymous logins over the Telenet public data network and (using telnet) over the Internet. Extended use is available by login to either of two classes of established user accounts. The GenBank contract requires that the cost of providing these accounts be recovered from the users. Complete information on costs and application forms are available from the GenBank On-line Service at the author's address or (415) 962-7364.

Class 0 Accounts

Anonymous users of the interactive system are provided with fifteen-minute complimentary sessions using the IRX retrieval program. With this program, entries in any of the on-line databases can be located by searching for a keyword or combination of keywords appearing in any of the fields of the entries' annotations. Located entries can be displayed on the terminal or downloaded to the user's computer with the Kermit file-transfer program. (The Kermit program is available for a wide variety of computers from numerous software bulletin boards, user groups, and from Columbia University. MS-DOS and Macintosh versions are available from GenBank on request.) New users of the IRX program should read the on-line introduction and users guide which can be displayed by answering 'Y' to the first question the program asks ('Do you want help?').

To log in to the GOS Class 0 account, one must have a supported terminal or a computer with software for emulating one of those terminals (see the list in Figure 1) and a modem capable of communicating at 300, 1200, or (preferably) 2400 baud. Instructions for and an example of dialing a local Telenet number to access the GenBank computer are shown in Figure 1. The Telenet customer service number is (800) 336-0437 and can be used to obtain the number of the nearest Telenet node. After completing the login procedure shown in Figure 1, the IRX database query program is immediately started.

Class 1 Accounts

To avoid the fifteen-minute limitation on Class 0 interactive sessions, users of the GOS may wish to establish accounts on the GOS computer. These accounts provide access to the GOS computer, 1 Mbyte of disk space for user files, access to the Unix utilities, IRX, and interactive and batch mode use of FASTA and TFASTA (a version of FASTA that compares peptide sequences with nucleic acid sequence databases by translating the database sequences in up to six reading frames 'on the fly'). Class 1 accounts also provide electronic mail access for contacting other users of the GOS and users of computers connected to the Internet and other computer networks. (Restrictions on the use of the Internet for commercial purposes apply to GOS users. Please read the application materials carefully.) Access to a wide variety of electronic bulletin boards is also provided. A listing of some of the bulletin boards (also known as newsgroups) is provided in Figure 2. Of special interest are the bionet.journals.contents newsgroup which provides tables of contents of several important journals on-line before publication, bionet.sci-resources which provides on-line copies of the NIH Guides to Grants and Contracts, and bionet.molbio.genome-program which provides access to the staff of and announcements from the National Center for Human Genome Research of the NIH and the U.S. Department of Energy's Genome Program.

Class 2 Accounts

For an additional fee, Class 2 users are provided with access to the IntelliGenetics Suite of sequence analysis programs and databases formatted for those programs. Additional databases (e.g., the PIR Protein Sequence Database, KeyBank™, and VectorBank™) are also available to Class 2 users. Class 2 users also have access to all the facilities available to Class 1 users.

E-MAIL SERVERS

In addition to providing interactive access, GenBank currently offers two electronic mail servers, one for sequence similarity searching and one for database entry retrieval. These are freely available to anyone who can send mail to an Internet address. The following networks have gateways to the Internet: BITNET, EARN, NETNORTH and JANET. Users of computers on these networks may need to change the format of the addresses given below to send the message through a forwarding gateway. Users should consult their computer system managers or administrators to determine the proper forwarding gateway and address form. Questions regarding the use of the e-mail servers (or other aspects of the GOS) may be addressed to: CONSULTANT@GENBANK.BIO.NET. (For users without nameservers, GENBANK.BIO.NET has the IP address: 134.172.1.160.)

FASTA Server

The GenBank FASTA Server receives mail messages containing a nucleic acid or protein query sequence and instructions for the search, performs a FASTA (2) sequence similarity search against the specified database, and returns the results by electronic mail.

To access the program, send an electronic mail message containing the formatted query sequence (as described below) to the following Internet address: SEARCH@GENBANK.BIO.NET. To receive instructions on using the FASTA Server, send a mail message to this address containing the word **HELP** as its only text.

Queries consist of a mail message with search parameters identifying the database to be searched, values related to the search and the query sequence to be used in the search. Figure 3 is an example of a mail message sent for a FASTA search. The mail message has two mandatory lines, three optional lines, a line identifying the query sequence as described below, and lines containing the query sequence. These lines start with the keywords shown and appear in the body of the mail message in the following order:

```
ATDT8569995(CR)          Use ATDP for a pulse dialing phone. Dial local Telenet number.
CONNECT 1200             This message tells you that you have reached a data line number,
                         and it gives you the baud rate of the connection. Depending on your
                         modem, a CONNECT message may not appear.
(CR)(CR)                 (if connecting at 300 or 1200 baud)
                         or
@(CR)                    (if connecting at 2400 baud)
TELENET 415 118A
TERMINAL=(CR)
@ c genbank,genbank(CR)
PASSWORD = 4nigms(CR)    This is the current Telenet password; it may be entered in either
                         upper or lowercase.
GENBANK CONNECTED
SunOS UNIX (GENBANK.BIO.NET)

login:genbank(CR)
Password:4nigms(CR)      This is the password for the GenBank computer; it must be entered
                         in lowercase characters.
Last login...            This message includes a date showing the last anonymous login, as
                         well as other system information.
OS/MP 4.0B (GENBANK/root) #1: Mon Jan 15 14:26:43 PST 1990

The following is a list of commonly used terminals
Designation     Terminal Type
adm3a           Lear Siegler (ADM)
aaa-48          Ann-Arbor Ambassador in 48 line mode
aaa-60          Ann-Arbor Ambassador in 60 line mode
dm3025          Datamedia 3025a
h19             Heath H19 or Zenith
hp2621          Hewlett Packard HP2621
hp2648-iv       Hewlett Packard HP2648A
sun             Sun Microsystems Workstation console
tvi912          Televideo 912, 920
tvi950          Televideo 950
vi200           Visual 200
vt100           Digital Equipment VT100 (default)
vt102           Digital Equipment VT102
vt200           Digital Equipment VT200
Press Return to select vt100, or enter the appropriate terminal
TERM = (vt100)           (type the designation of the appropriate terminal type followed by (CR))
```

Figure 1. Typical Class 0 GOS Telenet login procedure. In the example above, the user types **bold** characters exactly as shown, types user-specific information similar to that shown in ***bold-italic***, and sees the messages shown in typewriter font. The symbol (CR) indicates that the key labelled 'Return', 'RET', or 'Enter' should be pressed. Comments and instructions are shown in *italic*.

The remainder of the message contains the query sequence identification line and the sequence data. In Figure 3 the query sequence and identifier are shown in 'Pearson format'. The query sequence identification line is a mandatory comment line which begins with a greater-than sign ('>') followed by the name of the sequence, a space, and an optional note about the sequence. The sequence data begin on the next line. Sequences in IntelliGenetics/BIONET format (3) are also acceptable.

The query message format must be followed precisely, but note that either upper or lower case letters may be used. Each line of the message must contain less than 80 characters. Longer lines will be truncated. The message text begins with the keyword DATALIB and should not contain blank lines. The message should contain only one query sequence.

DATALIB — This line specifies the database or database division to be searched. (mandatory). To search the entire EMBL, GenPept, or Swiss-Prot database enter **EMBL, GenPept,** or **Swiss-Prot,** respectively. The New Data databases can be searched by specifying **GenBank/new, EMBL/new,** or **GenPept/new,** respectively. GenBank can be searched in full or in part by specifying the division you wish to search. The choices are listed below and should be entered as shown:

GenBank/primate	GenBank/plant	GenBank/phage
GenBank/rodent	GenBank/organelle	GenBank/synthetic
GenBank/other_mammalian	GenBank/bacterial	GenBank/unannotated
GenBank/other_vertebrate	GenBank/structural_rna	GenBank/all
GenBank/invertebrate	GenBank/viral	GenBank/new

KTUP — This line identifies the k-tuple value which specifies the sensitivity of the search. Values range from 3 to 6 for nucleic acid searches (defaults to 4 if not entered) and from 1 to 2 for protein searches (defaults to 1 if not entered). Lower values result in more sensitive searches but require more time to complete. (optional)

SCORES — This line specifies the number of best-ranked sequences to be listed in the results. The default value is 100. (optional)

ALIGNMENTS — This line identifies the maximum number of best-ranked sequences to be aligned in the results. The default value is 20. (optional)

BEGIN — Contains only the keyword BEGIN. (mandatory)

INTERNET BBOARD Name	USENET Newsgroup Name
AGEING	bionet.molbio.ageing
AGROFORESTRY	bionet.agroforestry
BIONEWS	bionet.general
BIOTECH	bionet.technology.general
BIO-CONVERSION	bionet.technology.conversion
BIO-JOURNALS	bionet.journals.contents
BIO-MATRIX	bionet.molbio.bio-matrix
BIO-SOFTWARE	bionet.software
EMBL-DATABANK	bionet.molbio.embldatabank
EMPLOYMENT	bionet.jobs
GENBANK-BB	bionet.molbio.genbank
GENOMIC-ORGANIZATION	bionet.molbio.gene-org
HUMAN-GENOME-PROGRAM	bionet.molbio.genome-program
METHODS-AND-REAGENTS	bionet.molbio.methds-reagnts
MOLECULAR-EVOLUTION	bionet.molbio.evolution
PIR	bionet.molbio.pir
POPULATION-BIOLOGY	bionet.population-bio
PROTEIN-ANALYSIS	bionet.molbio.proteins
RESEARCH-NEWS	bionet.molbio.news
SCIENCE-RESOURCES	bionet.sci-resources
SWISS-PROT	bionet.molbio.swiss-prot

Figure 2. International BIOSCI network bulletin boards available on the GenBank On-line Service. Internet and Usenet names of each bulletin board are shown.

```
From:  drbob@someaddress.somewhere.edu Tue Jun 14 21:36:38 1988
Date:  14 Jun 1988 2129:02-PDT
To:    SEARCH@GENBANK.BIO.NET
Subject:

DATALIB GenBank/other_mammalian
KTUP  4
SCORES 100
ALIGNMENTS 20
BEGIN
>BOVPRL GenBank entry BOVPRL from gbmam file.907 nucleotides.
tgcttggctgaggagccataggacgagagcttcctggtgaagtgtgtttcttgaaatcat
caccaccatggacagcaaa
```

Figure 3. Example FASTA Server e-mail query message. The first four lines are a mail header that is automatically created by the mailer program. The Subject may be left blank (it is ignored by the server).

The weekly update files are available as standard ASCII files or as compressed ASCII files. The compressed files are about one-third the size of the standard files. They can be distinguished by the .Z suffix and can be uncompressed after transfer with the standard Unix uncompress utility. In addition to the weekly incremental updates, a cumulative update file, updated daily and containing all entries which have entered the database or been revised since the previous release, is maintained in each of the new data directories. These files, which are provided in compressed form only, are named gbseq.all.Z, gpseq.all.Z, and emseq.all.Z. The current GenBank and GenPept release files (in gb-rel*NN* and gp-rel*NN*, respectively, where *NN* is the release number) are provided in compressed form only. The software in the dos, mac, and vms directories is not supported by the GenBank On-line Service.

To access any of the directories available for anonymous FTP, one should use the FTP protocol to connect to GENBANK.BIO.NET, using **anonymous** as the Username and one's surname as the Password. The files in these directories are also available for downloading to users who access the GOS computer via Telenet or direct-dial. High-speed modems are provided for the direct-dial lines to facilitate file downloading.

ENVIRONMENT

All components of the GOS operate on a Solbourne Series 4/802 multiprocessor superminicomputer capable of 80 million instructions per second (MIPS). This computer is currently supplied with 64 Mbytes main memory, 3.5 Gbytes of magnetic disk storage, 2 Gbytes of on-line archival tape storage, a 9-track tape drive, and optical disk drives. This computer is dedicated to the GOS; other GenBank work, including GOS development work, is performed on other computers. The GOS computer uses the Unix operating system (Sun OS 4, a dialect of bsd 4.3). The GOS computer is connected to the Bay Area Regional Research Network (BARRNet) by a T1 (1.54 Mbit/sec) communication channel. BARRNet is the local branch of the NSFNet.

CONCLUSION AND FUTURE DIRECTIONS

Recent improvements in the GenBank On-line Service now make it possible for users to access daily updates of the databases via electronic mail or through easy-to-use and powerful interactive programs. Those who maintain local copies of the database can now update these copies by anonymous FTP on a weekly basis. In addition, similarity searching and analytical software are now available to users of the GOS. Within the next year, we plan to add access to the relational database version of GenBank. This addition will make possible much more complex queries than can be posed using existing database query programs.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the major contributions of the following to the work described in this article: Stephen Barnhouse, Michael Cinkosky, David Horner, Rand Huntzinger, David Kristofferson, Will Nelson, Eliot Lear, Julie Ryals, Peter Stoehr, Spencer Yeh, and Katherine Yudin. The author thanks Doug Brutlag and John Moore for helpful comments on the manuscript. Provision of the IRX program by Dr. Dennis Benson and the National Library of Medicine and the FASTA program by Dr. William Pearson are also gratefully acknowledged. This work is funded by NIH Contract N01-GM-7-2110. GenBank is the registered trademark of the National Institutes of Health. KeyBank and VectorBank are trademarks of IntelliGenetics Incorporated. PIR is a registered trademark of the National Biomedical Research Foundation.

When a query message is received, it is placed in a batch queue, processed in the order received, and the results delivered by return mail. The status of a job being processed can be determined by sending a mail message to the SEARCH address above containing only the word **QUEUE**. No individual is permitted to have more than one search waiting in the queue at a time. If a user sends a second query message before his first request has been processed, the initial search will be cancelled and replaced by the search submitted second.

Entry Server

E-mail access to sequence database entries is provided for three primary reasons: 1) to enable users of the FASTA Server to retrieve entries identified by sequence similarity searches; 2) to enable users of the Class 0 interactive system described above, who access it by network remote login (e.g., telnet) to retrieve copies of entries of interest; and 3) to enable readers of journals that identify published sequences by accession number to retrieve

computer-readable versions of those sequences. To retrieve a database entry, send a mail message containing only the entry name or the accession number (not both) to the address: RETRIEVE@GENBANK.BIO.NET. The on-line databases are searched and the entries (if any) that correspond to the supplied entry name or accession number are returned by electronic mail. To receive instructions on using the Entry Server, send a mail message to the RETRIEVE address (above) containing only the word **HELP**. Note that, due to the order in which the databases are searched, if both GenBank and EMBL data banks contain entries with the same primary accession number (the usual case), a query on the accession number will return the GenBank version of the entry. If the EMBL-format version of the entry is required, it can be retrieved from the EMBL file server (4).

ANONYMOUS FTP

In addition to interactive access and electronic mail servers, GenBank also currently provides files for anonymous FTP (File Transfer Protocol), including GenBank and EMBL new data and contributed software. Each week, the new entries created in the GenBank database are collected into an update file. The file has a name of the form gb*MMDD*.seq, where *MM* is the number of the month and *DD* is the date of file creation. Likewise, new EMBL entries are collected into files with names of the form em*MMDD*.seq. The weekly update files are kept in the new data directories until they are superseded by a new quarterly release of the database.
Currently, the following directories are available for anonymous FTP:

pub/db/alu	Jurek Jurka's Alu sequence and alignment database
pub/db/embl-newdata	EMBL weekly update files
pub/db/gb-newdata	GenBank weekly update files
pub/db/gb-rel62	GenBank release 62.0
pub/db/gp-newdata	GenPept weekly update files
pub/db/gp-rel62	GenPept data derived from GenBank release 62.0
pub/db/seqanalref	Amos Bairoch's bibliography of sequence analysis literature
pub/dos	IBM PC compatible programs
pub/mac	Macintosh programs
pub/vms	VAX/VMS programs

REFERENCES

1. Harman, D., Benson, D., Fitzpatrick, L., Huntzinger, R., and Goldstein, C. (1988) In Proceedings of RIAO 88 Conference *User-Oriented Content-Based Text and Image Handling*, Cambridge, MA., pp. 840–848.
2. Pearson, W.R. and Lipman, D.J. (1988) Proc. Natl. Acad. Sci. U.S.A. **85,** 2444–2448.
3. Brutlag, D.L., Clayton, J., Friedland, P., and Kedes, L.H. (1982) Nucleic Acids Res. **10,** 279–294.
4. Stoehr, P.J. and Omond, R.A. (1989) Nucleic Acids Res. **17,** 6763–6764.