© 1993 Oxford University Press Nucleic Acids Research, 1993, Vol. 21, No. 13 2963–2965

GenBank

Dennis Benson, David J.Lipman and James Ostell
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

ABSTRACT

The GenBank sequence database has undergone an expansion in data coverage, annotation content and the development of new services for the scientific community. In addition to nucleotide sequences, data from the major protein sequence and structural databases, and from U.S. and European patents is now included in an integrated system. MEDLINE abstracts from published articles describing the sequences provide an important new source of biological annotation for sequence entries. In addition to the continued support of existing services, new CD-ROM and network-based systems have been implemented for literature retrieval and sequence similarity searching. Major releases of GenBank are now more frequent and the data are distributed in several new forms for both end users and software developers.

INTRODUCTION

GenBank™ is the NIH's database of all known nucleotide and protein sequences including supporting bibliographic and biological information. As of Release 76.0 in April, 1993, GenBank contained over 129,968,355 nucleotide bases from 111,911 different sequences. Entries include a concise description of the sequence, scientific name and taxonomy of the source organism, and a table of features specifying coding regions and other sites of biological significance. As part of the feature table, protein translations for coding regions are included.

GenBank has been the responsibility of the National Center for Biotechnology Information (NCBI) since October, 1992. The NCBI (1) is part of the National Library of Medicine (NLM), and, in turn, a part of the National Institutes of Health. Prior to October, 1992, GenBank was funded by the National Institute of General Medical Sciences as contracts to IntelliGenetics, Inc. (1987–1992), and Bolt, Beranek and Newman, Inc. (1982–1987). Los Alamos National Laboratory (LANL) has participated in GenBank since 1982 as a contractor with responsibilty for data entry and maintenance. An international collaboration with the EMBL Data Library in Heidelberg, Germany and the DNA Data Bank of Japan (DDBJ) in Mishima provides shared collection and exchange of sequence information.

The history of sequence databases, including GenBank, has been covered previously (2,3). This report describes current operations, recent developments and new services provided by NCBI.

CD-ROM DISTRIBUTION

The GenBank data is available on CD-ROM through a subscription service with the Government Printing Office (202-783-3238, 202-512-2233 FAX). Order forms are also included in each issue of NCBI News, a free subscription to which may be obtained by contacting NCBI. NCBI has increased the frequency of major releases from four to six per year. Each release contains a new, full copy of the database and is available in the following three versions.

NCBI-GenBank (flat file)

This version provides a continuation of the same flat file format in which GenBank has been distributed for many years. Each release is a full release incorporating all previous GenBank data supplemented by new data from direct submissions, NCBI journal scanning, patents and the EMBL and DDBJ databases. Conceptual translations of coding regions appear in feature tables. The release contains the standard index files and is organized into divisions. No retrieval software is provided.

Entrez

Entrez is a two CD-ROM set containing a Sequence disk and a References disk. *Entrez:* Sequences contains molecular sequence and a set of bibliographic references that are cited in the sequence databases. The DNA and protein sequence data are integrated from a variety of sources, including GenBank, EMBL, DDBJ, PIR, SWISS-PROT, Protein Research Foundation (PRF), the Brookhaven Protein Data Bank (PDB) and patents. The second disk contains a larger bibliographic subset of MEDLINE®, references to papers indexed under the NLM's Medical Subject Heading (MeSH), 'molecular sequence data'. The DNA sequence, protein sequence and bibliographic data are linked to provide easy traversal among the databases using a graphical user interface. The retrieval system allows for traditional keyword searching and uses pre-computed statistical measures of relatedness to allow queries that will find all articles or sequences similar to an article or sequence of interest.

Entrez contains retrieval software for the Apple Macintosh™ and for PC-compatible systems running Microsoft Windows™ (version 3.1 or later). A minimum of 2 Mbytes of memory is necessary. Documentation consists of a 30-page user's guide for installation and operating instructions. (Source code for an X11 version of the software for VMS and Unix platforms is available through anonymous FTP from 'ncbi.nlm.nih.gov' in the 'entrez' directory. Also, executables for these and other platforms are available on an unsupported basis.)

NCBI-sequences (ASN.1)

This title provides the integrated sequence dataset used on the *Entrez*: Sequences CD-ROM in the ISO ASN.1 standard data description format. DNA and protein sequence data are linked to journal citations appearing in MEDLINE. Files are provided that contain the inter-document/sequence linkage information and indices to the byte offsets of the beginning of sequence and bibliographic records. No retrieval software is provided.

NETWORK ACCESS

Anonymous FTP

Users on the Internet can use the file transfer protocol (FTP) program to download the entire GenBank release or the daily updates.

Files of the full release and daily updates of the NCBI-GenBank database are available for anonymous FTP from 'ncbi.nlm.nih.gov' (130.14.25.1). The full release in flat-file format is available as compressed files in the directory, 'ncbi-genbank'. A cumulative update file is contained in the sub-directory, 'daily', and a non-cumulative set of updates is in the sub-directory, 'daily-nc'. ASN.1 formatted data is in the directory, 'ncbi-asn1'. Software tools for handling the ASN.1 data and for developing ASN.1 applications can be found in the directory 'toolbox/ncbi_tools'.

E-Mail servers

Users with access to electronic mail can search GenBank and eight other databases using IRX-based text retrieval. To start, send a mail message containing the word 'help' to: 'retrieve@ncbi.nlm.nih.gov'. BLAST sequence similarity searching (4,5) is also available via e-mail through the address: 'blast@ncbi.nlm.nih.gov'. The two e-mail servers are averaging 2000 requests per day.

Network services

NCBI has begun to offer server-client services on the Internet beginning with BLAST clients for the PC, Macintosh, and Unix computers that make direct connections to a server at the NCBI. Over 1200 BLAST requests are processed daily through the server-client system. Client software for *Entrez* is undergoing testing and will be generally available in the fall, 1993. Information on registering hosts for these clients and obtaining software can be obtained by e-mail to the address: 'net-info@ncbi.nlm.nih.gov'. Preliminary reports on the use of the client-server approach have been extremely encouraging and support the position that this technology offers significant performance and operational advantages. NCBI will make a major commitment to insure that its servers maintain the most comprehensive, timely, and accurate version of sequence and sequence-related data. It will also supply stable programming interfaces to academic and commercial developers to facilitate the creation of additional clients for different platforms. Other directions planned for server-client services include: 1) access to a larger, molecular biology-related subset of MEDLINE, 2) greater flexibility in sequence record retrieval, and 3) more powerful similarity searching.

BUILDING THE DATABASE

The data in GenBank come from two primary sources: 1) annotators who extract the information from relevant journals and, 2) authors who submit data directly to the database. Approximately 36% of the records in GenBank are produced by the international collaborators, the EMBL Data Library (32%) and DDBJ (4%) with whom sequence information is exchanged on a daily basis. NCBI has developed a journal scanning operation in collaboration with Library Operations, the NLM division that creates MEDLINE. A team of annotators using NCBI software creates sequence records from journal articles. Sequence-containing articles are processed from all of MEDLINE's 3500 journals and from an additional set of journals that contain a significant number of sequences. Specially trained sequence indexers now have over two years of experience in reviewing the literature and creating new sequence entries. Approximately 9% of the current database consists of NCBI-created entries and 15% of the new entries are the result of journal scanning.

With NCBI scanning the literature and sending appropriate information to the database collaborators, the international 'journal split' is no longer necessary and, in fact, the collaborating databases now rely on NCBI to capture published sequences. Another major procedural change is that authors may now send their submissions to whatever database is most convenient instead of the database that was responsible for the journal in which their article was to be published.

Direct submission

The majority of entries continue to enter the database through direct author submission, a process pioneered by LANL over five years ago in response to editorial policies that limited the amount of sequence data that were publishable in an article reporting those data. The biological community has responded very positively to direct submission and this method has had the beneficial effect of involving authors closely in the data input phase of database production, thereby increasing the degree of data quality and integrity. Submissions are passed back to authors for their review before entering the database and are often returned with corrections or updates. Most importantly, authors have become active collaborators and have taken on responsibility for the accuracy and completeness of their data.

Therefore, NCBI strongly encourages authors to submit directly to the database prior to publication in order that the sequence can be available to the public no later than the time the paper appears in print. Authors, of course, have the right to request that their sequence be kept confidential until the time of publication. In these cases, authors are reminded to inform the database of the publication date in order to have a timely release of the data.

To help scientists submit sequences and to annotate their data, a program called Authorin was developed by IntelliGenetics, Inc. under the previous GenBank contract. It is still available free of charge through the NCBI by requesting a copy, preferably by e-mail (authorin@ncbi.nlm.nih.gov) or by telephone. Users should specify whether they prefer the PC or Macintosh versions. NCBI is also developing a platform-independent submission program which will enable users to submit sequences over the network and which will be available in source code form to enable developers to incorporate the code into their software.

ORGANIZATION OF THE DATABASE

GenBank contains over 129 million bases as of April, 1993, an increase of over 46 million bases over the previous 12-month period. Historically, the database has doubled in size about every

22 months. The traditional flat file is distributed in 14 different divisions, such as bacterial, viral, mammalian, primate, etc. Two new divisions were added this year: PAT for for patent sequences and EST for 'Expressed Sequence Tags' (see below). The patent sequences are from the U.S. Patent and Trademark Office and from the European Patent Office and are being entered into the database as part of an ongoing collaboration between the U.S. and European patent offices and the GenBank and EMBL databases, respectively.

EST data

EST sequences are 'single pass,' partial DNA sequences that are derived from clones that are randomly selected from cDNA libraries and EST data are the most rapidly-expanding source of new genes (6). Because these data differ from traditional GenBank entires and thus require special processing and annotation, NCBI also makes them available in a separate database, dbEST, in addition to the EST Division of GenBank. dbEST now includes nearly 20,000 sequences from humans, model organisms for genome research, and other species. NCBI accepts bulk submissions of cDNA data and assigns GenBank accession numbers. Data are also accepted from direct submissions to Los Alamos, EMBL and DDBJ. ESTs are automatically screened upon entry and then periodically searched against the nucleotide and protein sequence databases in order to identify matches with known genes. The data are stored in a relational database and reformatted as a separate (EST) division of the GenBank database. dbEST sequences can be searched by the BLAST e-mail server and full reports of EST records can be obtained by querying an e-mail server (est_report@ncbi.nlm.nih.gov). The full reports contain information on the availability of physical cDNA clones and mapping data in collaboration with the Genome Data Base at Johns Hopkins.

The Integrated Database (ID)

In order to produce the GenBank database, NCBI maintains internally an Integrated Database, ID, to track and index ASN.1 records from the multiple sources of sequence data. These sources include submissions from LANL, EMBL DDBJ, dbEST, and patents plus amino acid sequences from PIR, SWISS-PROT, PRF and PDB.

ID represents the latest view that each data source has of its data, and allows NCBI to assign stable identifiers to sequences. Through this approach, sequence information from a wide variety of sources will have a uniform identification system. These identifiers will be stable and therefore it will be easier to know when sequences have changed. This approach will make the ID database useful as an archive and will also allow scientists to retrace the history of revisions for every entry.

ID will also allow NCBI to reduce unnecessary redundancy in the database and thus produce a more useful view of the data for biologists. Consider a DNA sequence entry with a coding region and conceptual translation, and the corresponding SWISS-PROT entry. If the conceptual translation is identical to the SWISS-PROT entry, it will be replaced with the SWISS-PROT entry. A biologist retrieving the DNA entry will automatically get the full annotation of the SWISS-PROT protein. Likewise, retrieving the SWISS-PROT protein would immediately allow one to map to the corresponding DNA sequence and coding region. ID will allow more comprehensive approaches for reducing redundancy, and will provide the most biologically relevant entries for release. By maintaining all previous records, it will always be possible for scientists to re-evaluate the evolution of the current view of every sequence record.

MAILING ADDRESS

GenBank
National Center for Biotechnology Information
Bldg. 38A, Rm. 8S-803
8600 Rockville Pike
Bethesda, MD 20894, USA
301-496-2475

E-MAIL ADDRESSES

info@ncbi.nlm.nih.gov	(general information about NCBI and services)
gb-sub@ncbi.nlm.nih.gov	(submission of sequence data to GenBank)
update@ncbi.nlm.nih.gov	(revisions to GenBank entries and notification of release of 'hold until published' entries)

ACKNOWLEDGEMENTS

We would like to acknowledge all of the GenBank staff presently at the NLM and Los Alamos National Laboratory and all those who have contributed to the database since its beginnings over 12 years ago. The support of Dr Ruth Kirchstein and her staff at NIGMS in the first ten years was critical for its existence. Also, the GenBank advisors, both national and international, have contributed their interest and energy towards making the database a valued resource.

REFERENCES

1. Benson, D., Boguski, M., Lipman, D.J. and Ostell, J. (1990) *Genomics*, **6**, 389–391.
2. Smith, T.F. (1990) *Genomics*, **6**, 702–707.
3. Burks, C., Cinkosky, M.J., Fischer, W.M., Gilna, P., Hayden, J.E.D., Keen, G.M., Kelly, M., Kristofferson, D. and Lawrence, J. (1992) *Nucleic Acids Research*, **20** (Suppl.), 2065–2069.
4. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
5. Gish, W. and States, D.J. (1993) *Nature Genetics*, **3**, 266–272.
6. Sikela, J.M. and Auffray, C. (1993) *Nature Genetics*, **3**, 189–191.