Nucleic Acids Research, 1994, Vol. 22, No. 17 3441–3444

GenBank

Dennis A.Benson*, Mark Boguski, 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 continues to expand its data coverage, quality control, annotation content and retrieval services for the scientific community. Besides handling direct submissions of sequence data from authors, GenBank also incorporates DNA sequences from all available public sources; an integrated retrieval system, known as *Entrez*, also makes available data from the major protein sequence and structural databases, and from U.S. and European patents. MEDLINE abstracts from published articles describing the sequences are also included as an additional source of biological annotation for sequence entries. GenBank supports distribution of the data via FTP, CD-ROM, and E-mail servers. Network server-client programs provide access to an integrated database for literature retrieval and sequence similarity searching.

INTRODUCTION

GenBank® is the NIH's database of all known nucleotide and protein sequences including supporting bibliographic and biological information. As of Release 83.0 in June, 1994, GenBank contained over 191,393,939 nucleotide bases from 182,753 different sequences. Although human entries predominate, constituting 27% of the total, more than 8,000 species are represented. 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.

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. A new release of the database appears every two months. Each release contains a new, full copy of the database and is available in the following two versions.

NCBI-GenBank (Flat File)

This version provides 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 other sequence 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. Beginning with Release 84.0 in August, the distribution will require two CD-ROMs.

Entrez

Entrez is an integrated database and retrieval CD-ROM which accesses DNA and protein sequence data, plus a set of related MEDLINE references. The DNA and protein sequence data are integrated from a variety of sources, including GenBank, EMBL, DDBJ, LANL, PIR, SWISS-PROT, Protein Research Foundation (PRF), the Brookhaven Protein Data Bank (PDB) and patents. The data are also organized by taxonomic classification. A database of taxonomic information is being assembled at NCBI with assistance from a panel of taxonomy experts. The linkage among data sources is shown in Figure 1. The MEDLINE references are 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 4 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/software' directory. Also, executables for these and other platforms are available on an unsupported basis.) Since the data files on the *Entrez* CD-ROM are in the ISO ASN.1 standard data description format, NCBI no longer distributes a separate CD-ROM ASN.1 version of the database for software developers.

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

*To whom correspondence should be addressed

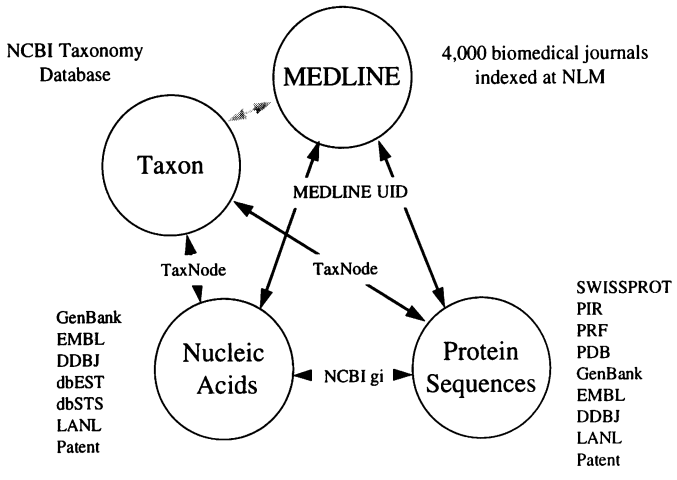

Figure 1. Data sources and interconnecting links among the four information components which generate the integrated *Entrez* retrieval system.

updates (which also incorporate sequence data from other public databases). Files of the full release and daily updates of the 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, '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 ten 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 (2) is also available via e-mail through the address: 'blast@ncbi.nlm.nih.gov'. The two e-mail servers average over 2800 requests per day.

Network services

NCBI offers client programs for executing BLAST and *Entrez* queries directly over the Internet. Client software for the PC, Macintosh, and Unix computers makes direct connections to a server at the NCBI. *Network Entrez* offers the same interface as the CD-ROM version with the enhancement of searching a larger subset of MEDLINE (approximately one million articles in molecular biology). Information on registering hosts for BLAST or *Entrez* clients and obtaining software can be obtained by e-mail to the address: 'net-info@ncbi.nlm.nih.gov'. World Wide Web/Mosaic access is also available for *Entrez* or BLAST searching. The URL for the GenBank home page is: http://www.ncbi.nlm.nih.gov/. Network searching is growing at a faster pace than the e-mail services and averages approximately 3000 requests per day.

BUILDING THE DATABASE

The data in GenBank come from two primary sources: 1) authors who submit data directly to the collaborating databases, and 2) annotators at NCBI who extract the information from relevant journals. Data are exchanged daily with the collaborating databases so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct submission

The majority of entries continue to enter the database through direct author submission. Many journals have the policy of requiring authors with sequence data to submit data directly to the database as a condition of publication. Even for those journals without a mandatory submission policy, author submission has the positive benefits of acquiring annotation information directly from the authors and reducing the time-lag between publication and the appearance of the sequence in the database.

GenBank staff can usually assign an author an accession number within one working day of receipt. The accession number serves as confirmation that the sequence has been submitted and allows readers of the article to retrieve the relevant data. All direct submissions receive a systematic quality assurance review including screening against GenBank to identify full or partial matches, checking for vector sequence and verifying proper translation of coding regions. A draft of the GenBank record is passed back to the author for review before entering the database. Authors 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. Although only the submitting scientist is permitted to modify sequence data or annotations, all users are encouraged to inform the database of possible errors or omissions using the e-mail address, 'update@ncbi.nlm.nih.gov'.

To help scientists submit sequences and to annotate their data, a program called Authorin is available free of charge from GenBank. It can be obtained by anonymous FTP to 'ncbi.nlm.nih.gov' in the 'pub/authorin' directory or by a phone or e-mail (authorin@ncbi.nlm.nih.gov) request. Users should specify whether they prefer the PC or Macintosh version. Once a submission is completed, users can e-mail it to the address: 'gb-sub@ncbi.nlm.nih.gov'.

GenBank is developing a platform-independent submission program called SEQUIN, which will run stand-alone and over the network. It will provide internal consistency checks as well as access to feature validation tools so that the process of sequence submission will not only be easier, but will have the potential of offering the author additional biological information about the sequence. SEQUIN will be introduced in late 1994.

GenBank also works closely with genome centers to facilitate the submission of high-volume data; likewise special procedures are arranged with groups for bulk submission of data such as ESTs and STSs.

Journal scanning

GenBank has a journal scanning operation to scan the current literature from over 3500 journals and identify sequences which have been published but were not submitted directly by authors. This operation has also proven successful in updating publication information and in identifying sequences that had been submitted confidentially and should be released.

ORGANIZATION OF THE DATABASE

GenBank contains over 191 million bases as of June, 1994, an increase of over 53 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 which generally correspond to taxonomic divisions, e.g., bacterial, viral, mammalian, primate. Over the past couple years separate divisions have been added for patent sequences and for expressed sequence tags (ESTs). A separate STS division for Sequence Tagged Sites will be added later in 1994. 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 cooperative project among the U.S., European, and Japanese patent offices and the sequence databases.

EST data

ESTs, or 'expressed sequence tags' (3), are the most rapidly-expanding source of new genes. 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 (4). dbEST has been operating for over two years and now includes nearly 40,000 sequences mostly from humans but with 15 other species represented. 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 GenBank. dbEST sequences can be searched by the BLAST e-mail server and full reports of EST records can be obtained by querying the NCBI e-mail server (retrieve@ncbi.nlm.nih.gov). Summary information on the database and a query capability are available through the NCBI WWW/Mosaic home page listed above.

STS data

Based upon the dbEST model, a new database (dbSTS) and new GenBank division for STS (5) and microsatellite marker data has been created. The first release of this new division in GenBank 85.0 will consolidate STS data from other GenBank divisions. For new data, streamlined direct submission procedures have been developed. This reorganization of GenBank facilitates cross-comparison of STSs with sequences in other divisions for the purpose of correlating map positions of STS sequences with known genes or ESTs. It also enhances the interoperability of GenBank with other databases containing more extensive mapping information. WWW/Mosaic access, e-mail retrieval by map location and updated homology search results are provided as for dbEST.

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 EMBL, DDBJ, LANL, dbEST, and patents, plus amino acid sequences from PIR, SWISS-PROT, PRF and PDB.

ID represents the most current view that each data source has of its sequence data, and allows NCBI to assign stable identifiers (known as 'gi identifiers' and which appear in the Comments and Feature Table portion of GenBank entries). Through this approach, sequence information from a wide variety of sources can have a uniform identification system. These identifiers are stable and therefore help identify sequences which 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 produce an enhanced view of GenBank, called GenBank Select, which will reduce unnecessary redundancy in the database and regularize feature annotation. For example, where exact or nearly exact matches exist between Swiss-Prot and translated GenBank entries, the Swiss-Prot translations, names, and descriptions will be substituted in the GenBank Select records. In addition, subset sequences that exist as separate GenBank entries, such as identical cDNAs or a cDNA and its corresponding genomic sequence, will be merged into a single record for the view presented in GenBank Select.

GenBank Fellows

The GenBank Fellowship Program is a new NCBI initiative to improve the quality of the database and also to serve as a bioinformatics training program. GenBank fellows are selected for strong backgrounds in biology and for a motivation to apply computational tools to the organization of electronic data in molecular and structural biology, genetics and phylogeny. Training is provided in the Unix operating system, software tools for manipulating data, files and processes, sequence analysis methods and statistics, and database management systems. GenBank Fellows, under the supervision of a mentor from NCBI's Computational Biology Branch, will pursue various applied research projects to improve the the quality and annotation of GenBank entries, to reduce sequence redundancy, and to establish and maintain links to other databases such as those containing genetic and physical mapping data and three-dimensional macromolecular structures. Approximately 200 applications were received for the initial five positions and the first GenBank Fellows began the program in July, 1994.

Mailing address

GenBank
National Center for Biotechnology Information
Bldg. 38A, Rm. 8S-803
8600 Rockville Pike
Bethesda, MD 20894. USA
Tel: 301-496-2475
Fax: 301-480-9241

E-mail addresses

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)
info@ncbi.nlm.nih.gov	(general information about NCBI and services)

Citing GenBank

If you use GenBank as a tool in your published research, we ask that this paper be cited.

REFERENCES

1. Benson, D., Lipman, D.J., and Ostell, J. (1993) *Nucleic Acids Research*, 21,2963–2965.

2. Altschul, S.F., Boguski, M.S., Gish, W., and Wootton, J.C. (1994) *Nature Genetics*, 6, 119–129.
3. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., and Venter, J.C. (1991) *Science*, 252, 1651–1656.
4. Boguski, M.S., Lowe, T.M.J., and Tolstoshev, C.M. (1993) *Nature Genetics*, 4, 332–333.
5. Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989) *Science*, 245, 1434–1435.