

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

Received October 3, 1995; Accepted October 4, 1995

ABSTRACT

The GenBank sequence database continues to expand its data coverage, quality control, annotation content and retrieval services. GenBank is comprised of DNA sequences submitted directly by authors as well as sequences from the other major public databases. An integrated retrieval system, known as Entrez, contains data from GenBank and from the major protein sequence and structural databases, as well as related MEDLINE abstracts. Users may access GenBank over the Internet through the World Wide Web and through special client-server programs for text and sequence similarity searching. FTP, CD-ROM and e-mail servers are alternate means of access.

INTRODUCTION

GenBank® (1) is the NIH's database of all publicly available nucleotide and protein sequences including supporting bibliographic and biological information. Since 1992 it has been based at the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine, located on the NIH campus. NCBI was created by Congress in 1988 and specifically charged with developing automated information systems to support molecular biology and biotechnology. Its other mission is to conduct basic research and as part of the NIH Intramural Program, NCBI scientists pursue research in genome analysis, molecular structure modeling and prediction and mathematical methods for sequence analysis.

NCBI builds GenBank primarily from the direct submission of sequence data from authors and secondarily from scanning the journal literature. The data are supplemented by incorporating sequences from other public databases. Through an international collaboration with the EMBL Data Library in the UK and the DDBJ, data are exchanged daily ensuring that each of these resources maintains a comprehensive set of all known, public sequence information. The data are made available at no cost through the Internet, either by downloading database files or by text and sequence similarity search services.

ORGANIZATION OF THE DATABASE

GenBank has witnessed unprecedented growth over the past 12 months with more sequence records added in this single period

than in the entire previous 14 year history of GenBank. As of Release 90.0 in August, 1995, GenBank contained over 353 713 490 nucleotide bases from 492 483 different sequences. Although human entries predominate, constituting 54% of the total, >15 500 species are represented. After *Homo sapiens*, the top four species in terms of bases include *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Mus musculus* and *Arabidopsis thaliana*. Historically, the database has been doubling in size every 22 months, but that rate has rapidly accelerated due to the enormous growth in data from expressed sequence tags (ESTs). Over 56% of the sequences in the current release are ESTs and most of the growth over the past year has come from the collaborative project between Merck & Co. and Washington University (2).

Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism and a table of features that identifies coding regions and other sites of biological significance, such as transcription units, sites of mutations or modifications and repeats. Protein translations for coding regions are included in the feature table. Bibliographic references are included along with a link to the MEDLINE unique identifier for all published sequences.

The files in the GenBank distribution have traditionally been divided into 'divisions' which roughly correspond to taxonomic divisions, for example bacteria, viruses, primates, rodents, etc. In recent years, divisions have been added for patent sequences, ESTs and sequence tagged sites (STSs). The patent sequences are from the US 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 US, European and Japanese patent offices and the sequence databases.

A new genomes division will be created for the placement of chromosome- or genome-length sequences. The need for such a division arises from the production of sequences >350 000 bases, the upper limit for single database entries. Submitters of large entries are being encouraged to create multiple records of <350 kb, corresponding to the natural units in which the sequencing was done, often cosmid-size pieces or entries containing only a few genes. The genomes division will contain virtual records whose feature tables will specify how to assemble the individual entries into a single large, contiguous sequence. Thus, retrieval software will be able to provide customized views on demand of the complete genome, chromosome or subregion of interest.

* To whom correspondence should be addressed

EST data

ESTs are the most rapidly-expanding source of new genes. Last year there were 50 214 sequences in the EST Division of GenBank (dbEST) with 45% of these derived from human tissues (3). Currently dbEST contains 328 905 sequences with 80% coming from ~50 different human tissues or cell types. The remaining 20% are derived from 43 other organisms with *Arabidopsis thaliana* (21 056), *Caenorhabditis elegans* (12 102) and *Oryza sativa* (11 015) being abundantly represented by >10 000 ESTs each. Of the 262 108 human ESTs currently in dbEST, 185 480 (71%) have been contributed by the Washington University–Merck & Co. project. There have been >200 000 queries of dbEST data to date, including BLAST searches, e-mail retrievals, WWW accesses and anonymous FTP downloads. dbEST data assisted in the identification of a gene for familial Alzheimer's disease (4). In fact, 78% of the 45 human disease genes isolated by positional cloning to date, have at least one exact match to ESTs in dbEST (5). An international consortium is converting ESTs into gene-based STSs to develop transcript maps of the human genome (6). 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 dbEST and a query capability are available through the NCBI WWW home page listed above.

STS data

The dbSTS continues to expand rapidly. The most notable development of the past year is the availability of ~2000 gene-based STSs derived from ESTs and mapped on human genomic DNA cloned in YAC libraries by the Whitehead/MIT genome center by T. Hudson and colleagues. This laboratory is part of an international consortium that is developing transcript maps of the human genome (6). In addition to YAC maps, a major goal of the project is to create high-resolution radiation hybrid maps of human chromosomes. dbSTS contains PCR primers, reaction conditions and map locations for gene-based markers derived from this project. Such data will be used to support the visualization of transcript maps in the new genomes division of GenBank.

BUILDING THE DATABASE

The data in GenBank come from two primary sources: (i) authors who submit data directly to the collaborating databases; and (ii) annotators at NCBI who extract the information from relevant journals. Data are exchanged daily with 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.

In early 1996, funding agencies in the US, Europe and Japan will begin to support pilot sequencing projects with the goal of producing tens to hundreds of megabases of human genomic DNA sequence over the next 1–3 years. NCBI will be working closely with a number of centers performing this work to ensure timely incorporation of these data for public release. In parallel, NCBI has developed methods to display these data integrated with genetic and physical map data and to search the sequences more effectively (e.g. options in BLAST to mask Alu and other types of repetitive elements). Already, GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission.

Most individual submissions are now received through a Web-based data submission tool, BankIt. With BankIt, authors enter sequence information directly into a form, edit as necessary and add biological annotation (e.g. coding regions, mRNA features). Free-form text boxes provide the option of using your own words to describe the sequence, without having to learn formatting rules or use restricted vocabularies. BankIt creates a draft record in GenBank flat file format for the user to review and revise. A standalone program for Macintoshes and PCs is also available at no charge. 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. Once a submission is completed, users can e-mail it to the address: 'gb-sub@ncbi.nlm.nih.gov'.

GenBank staff can usually assign an author an accession number within 1 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'.

GenBank is developing a platform-independent submission program called SEQUIN which will run stand-alone and over the network and 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 early 1996.

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

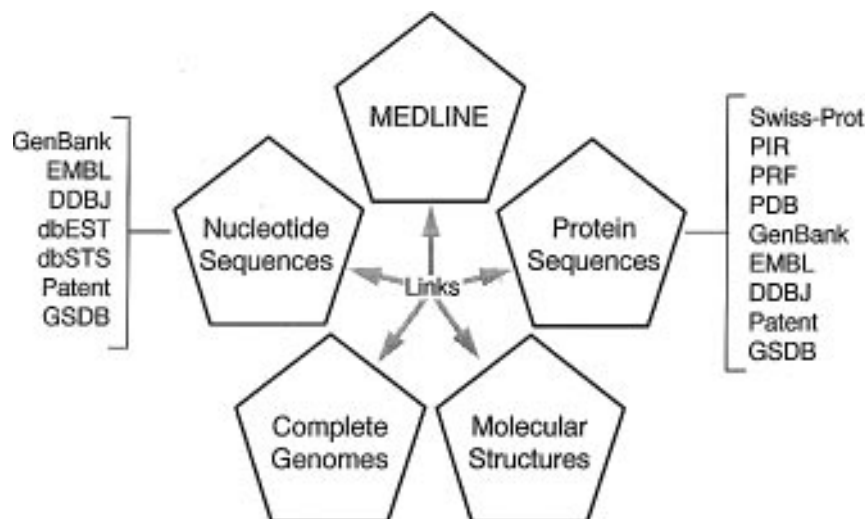


Figure 1. Data sources and interconnecting links among the five information components which generate the integrated Entrez retrieval system.

information and in identifying sequences that had been submitted confidentially and should be released.

The Integrated Database (ID)

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

Molecular Modeling Database (MMDB)

NCBI has begun producing a database of macromolecular 3-D structure information, specifically aimed at molecular modeling research. MMDB is based upon the data in the Brookhaven Protein Data Bank (PDB). By reorganizing and validating the PDB data, MMDB provides explicit descriptions of a biopolymer's spatial structure, its chemical organization and the linkage between the two. With MMDB, there is a clear cross-referencing between the 3-D structure and the chemistry of a macromolecule. Explicit linkages have been made to the sequence entries within the ID database so that, with the appropriate graphics software, users are able to view the 3-D structure of proteins identified via text or similarity searching of the sequence database.

RETRIEVING GENBANK DATA

The Entrez system

Entrez is an integrated database retrieval system which accesses DNA and protein sequence data, related MEDLINE references, a collection of genome data and 3-D structures from MMDB (7).

The DNA and protein sequence data are integrated from a variety of sources, including GenBank, EMBL, DDBJ, dbEST, dbSTS, GSDB, PIR, SWISS-PROT, Protein Research Foundation (PRF), PDB and patents. The MEDLINE references are papers indexed under the NLM's Medical Subject Heading (MeSH), 'genetic'. The genome dataset includes information on the large-scale organization of completely sequenced chromosomes or genomes, such as physical and genetic maps and aligned sequences.

The linkages among data sources are shown in Figure 1. The DNA sequence, protein sequence, MEDLINE, genome and 3-D structure data are linked to provide easy traversal among the datasets. Entrez provides an entry point into sequence or bibliographic records by simple Boolean queries. From the record, hypertext links may be used to navigate through the information space using a point-and-click interface. Some of the links are simple cross-references, for example, between a sequence and the abstract of the paper in which the sequence was reported or between a protein sequence and its corresponding DNA sequence. Other links, however, are based on computed similarities among the sequences or among the textual documents. The precomputed neighbors allow very rapid access for browsing groups of related records.

Entrez is currently available in three versions: CD-ROM, an Internet client-server applications and through the World Wide Web. Entrez currently occupies five CD-ROMs. Since the usefulness of the CD-ROM version decreases with each additional disc, CD-ROM users are encouraged to adopt one of the Internet-based versions.

The client-server version of Entrez operates with a client program on a user's machine over the Internet to a server located at the NCBI. Client programs for Macintosh, PC and Unix computers can be obtained by downloading from 'ncbi.nlm.nih.gov' in the 'entrez/network' directory. This version combines the full functionality of the Entrez interface with access to an extended set of 1.2 million MEDLINE citations and to the MMDB structure database. The sequences are updated daily and the MEDLINE weekly. The Web version of Entrez also provides access to the same sequence and MEDLINE datasets through standard Web browsers such as Netscape and Mosaic. Additional

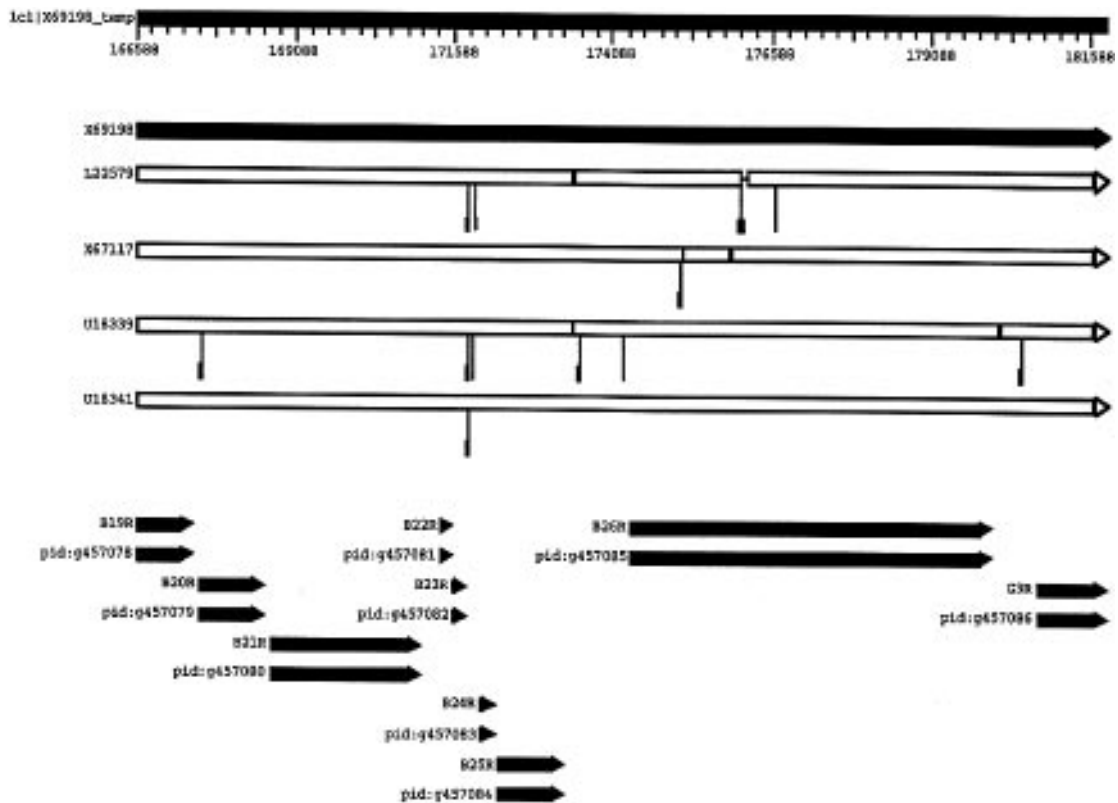


Figure 2. A graphical view from Network Entrez of a region of Variola virus. The top of the figure displays a reference sequence with base pair positions; in the middle are four aligned sequences with insertions and deletions indicated; below are the associated feature table representations of the genes and coding regions.

links are provided to external data sources such as map information from FlyBase for *Drosophila* sequence records and to the full text of publicly available online journal articles.

Network Entrez can display sequences in a graphical view in order to provide an easy way to visualize complex annotations such as segmented sequences or alternative splicing in coding regions (Fig. 2). The graphical viewer is also convenient for visualizing regions of interest in genome-length sequences and for presenting related physical and genetic map information.

BLAST sequence similarity searching

One of the most common uses of GenBank is sequence similarity searching. NCBI offers the BLAST family of search programs to perform fast searching with rigorous statistical methods for judging the significance of matches. NCBI offers client programs for executing BLAST (8,9) queries directly over the Internet. Information on registering hosts for BLAST clients and obtaining software can be obtained by e-mail to the address: 'blast-help@ncbi.nlm.nih.gov'. World Wide Web access is also available for BLAST searching.

Anonymous ftp

Users on the Internet can use the ftp program to download the entire GenBank release or the daily 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 are 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 10 other databases by accession number or Boolean combinations of text words. To start, send a mail message containing the word 'help' to: 'retrieve@ncbi.nlm.nih.gov'. BLAST sequence similarity searching is also available via e-mail through the address: 'blast@ncbi.nlm.nih.gov'.

CD-ROM

GenBank data are available on CD-ROM through a subscription service with the Government Printing Office (Tel: +1 202 783 3238, FAX: +1 202 512 2233). 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 2 months. Each release contains a new, full copy of the database and is available in the following two versions.

The Flat File 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 supplement-

ented 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. The distribution currently requires two CD-ROMs.

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, pursue various applied research projects to improve 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 3-D macromolecular structures. Currently five GenBank fellows are in the program and applications are reviewed on a continuing cycle.

Mailing address

GenBank, National Center for Biotechnology Information, Building 38A, Room 8S-803, 8600 Rockville Pike, Bethesda, MD 20894, USA. Tel: +1 301 496 2475; Fax: +1 301 480 9241.

Electronic addresses

<http://www.ncbi.nlm.nih.gov/>, NCBI World Wide Web Home Page.

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 GenBank 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., Boguski, M., Lipman, D.J. and Ostell, J. (1994) *Nucleic Acids Res.*, **22**, 3441–3444.
- 2 Boguski, M.S. (1995) *Trends Biochem. Sci.*, **20**, 295–296.
- 3 Boguski, M.S., Tolstoshev, C. M. and Bassett, D. E. (1994) *Science*, **265**, 1993–1994.
- 4 Levy-Lahad, E., Wasco, W., Poorkaj, P., *et al.* (1995) *Science*, **269**, 973–977.
- 5 Bassett, D.E., Boguski, M.S., Spencer, F., Reeves, R. Goebel, M. and Hieter, P. (1995) *Trends Genet.*, **11**, 372–373.
- 6 Boguski, M.S. and Schuler, G.D. (1995) *Nature Genet.*, **10**, 369–371.
- 7 Schuler, G.D., Epstein, J.A., Ohkawa, H. and Kans, J.A. (1995) *Methods Enzymol.*, in press.
- 8 Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) *Nature Genet.*, **6**, 119–129.
- 9 Madden, T.L., Tatusov, R.L. and Zhang, J. (1995) *Methods Enzymol.*, in press.