Nucleic Acids Research, 2003, Vol. 31, No. 1 **23–27**
DOI: 10.1093/nar/gkg057

GenBank

Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and David L. Wheeler*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 17, 2002; Accepted October 2, 2002

ABSTRACT

GenBank (R) is a comprehensive sequence database that contains publicly available DNA sequences for more than 119 000 different organisms, obtained primarily through the submission of sequence data from individual laboratories and batch submissions from large-scale sequencing projects. Most submissions are made using the BankIt (web) or Sequin programs and accession numbers are assigned by GenBank staff upon receipt. Daily data exchange with the EMBL Data Library in the UK and the DNA Data Bank of Japan helps ensure worldwide coverage. GenBank is accessible through NCBI's retrieval system, Entrez, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. BLAST provides sequence similarity searches of GenBank and other sequence databases. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. To access GenBank and its related retrieval and analysis services, go to the NCBI home page at: http://www.ncbi.nlm.nih.gov.

INTRODUCTION

GenBank (1) is a comprehensive public database of nucleotide and protein sequences with supporting bibliographic and biological annotation, built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD.

NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence (GSS), and other high-throughput data from sequencing centers. The US Office of Patents and Trademarks (USPTO) also contributes sequence data from issued patents. GenBank incorporates sequences submitted to the EMBL Data Library (2) in the UK and the DNA Databank of Japan (DDBJ) (3). Through a long-standing international collaboration between GenBank, EMBL and DDBJ, data are exchanged daily to ensure that all three sites maintain a comprehensive collection of sequence information. NCBI makes the GenBank data available at no cost over the Internet, via FTP access and a wide range of web-based retrieval and analysis services which operate on the GenBank data (4).

ORGANIZATION OF THE DATABASE

GenBank continues to grow at an exponential rate with 5.4 million new sequences added over the past 12 months. As of Release 131 in August 2002, GenBank contained over 22.6 billion nucleotide bases from 18.2 million different sequences. Complete genomes (http://www.ncbi.nlm.nih.gov/Genomes/index.html) represent a growing portion of the database, with 29 of the 86 complete microbial genomes in GenBank deposited over the past year. Many of these are the genomes of important human pathogens; an example of a recent addition is *Yersinia pestis*, the causative agent of the bubonic and pneumonic plagues. The number of eukaryote genomes for which both coverage and assembly are good is increasing rapidly and now includes *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Plasmodium falciparum*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Anopheles gambiae*, *Arabidopsis thaliana*, *Mus musculus* and *Homo sapiens*. There are now genomic sequencing projects underway for at least 134 additional microorganisms, and 33 additional eukaryotes, and many of these genomes are expected to appear in the public databases over the coming year. GenBank is currently doubling in size, both in terms of number of sequences and number of bases, about every 15 months. This is due in large part to the enormous growth in data from ESTs. About 69% of the sequences in GenBank Release 131 are ESTs, and current EST projects for human, mouse, rat and many other organisms will contribute still more data.

Sequence-based taxonomy

Over 119 000 species are represented in GenBank and new species are being added at the rate of over 1100 per month. About 33% of the sequences in GenBank are of human origin

*To whom correspondence should be addressed. Tel: +1 3014355950; Fax: +1 3014809241; Email: wheeler@ncbi.nlm.nih.gov

and 25% of all sequences are human ESTs. After *H. sapiens*, the top species in GenBank in terms of number of bases include *M. musculus*, *Rattus norvegicus*, *D. melanogaster*, *A. thaliana*, *Oryza sativa* and *C. elegans.* Database sequences are classified and can be queried using a comprehensive sequence-based taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html) developed by NCBI in collaboration with EMBL and DDBJ and with the valuable assistance of external advisors and curators.

GenBank records and divisions

Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references, and a table of features (http://www.ncbi.nlm.nih.gov/collab/FT/index.html) listing areas of biological significance, such as coding regions and their protein translations, transcription units, repeat regions, and sites of mutations or modifications.

The files in the GenBank distribution have traditionally been divided into 'divisions' that roughly correspond to taxonomic groups, e.g., bacteria (BCT), viruses (VRL), primates (PRI), and rodents (ROD). In recent years, divisions have been added to support specific sequencing strategies. These include divisions for EST, GSS, and high throughput genomic (HTG) sequences. A new division, the high throughput cDNA (HTC) division, has recently been added, bringing the total to 17 divisions. For convenience in file transfer, the larger divisions, such as the EST and PRI, are partitioned into multiple files when posting the bimonthly GenBank releases on NCBI's FTP site.

ESTs

ESTs continue to be the major source of new sequence records and gene sequences. Last year there were 8.6 million ESTs in GenBank. Over the past year, the number of ESTs has increased by over 45% to the current total of 12.5 million sequences representing over 430 different organisms. The top five organisms include: *H. sapiens*, with 4.6 million sequences (37% of the total); *M. musculus*, with 2.6 million sequences (21%); *R. norvegicus*, with about 350 000 sequences (2.8%); *Glycine max*, with about 268 000 sequences (2.1%); and *D. melanogaster*, with about 256 000 sequences (2.0%).

ESTs also continue to be the major source of new gene discoveries. As part of its daily processing of GenBank EST data, NCBI identifies through BLAST searches all homologies for new EST sequences and incorporates that information into the companion dbEST database (5). In order to better organize redundant EST data, NCBI maintains the UniGene (http://www.ncbi.nlm.nih.gov/UniGene/) collection of gene-based sequence clusters for 10 animals including human (6), mouse, rat, cow, frog, zebrafish and 7 plants, including rice, wheat, barley and corn. Additional information about UniGene is included in a separate article in this issue (4).

Sequence-tagged sites (STSs)

The STS division of GenBank contains over 124 000 sequences including anonymous STSs based on genomic sequence as well gene-based STSs derived from the 3′ ends of genes and ESTs. These STS records usually include primer sequences, annotations and PCR reaction conditions.

The purpose for creating high resolution physical maps of the human genome is to create a scaffold for organizing large scale sequencing (7). Physical maps based on STS landmarks are used to develop so-called 'sequence-ready' clones consisting of overlapping cosmids or BACs. As the HTG sequence data derived from these clones are submitted to GenBank, STSs become crucial reference points for organizing, presenting and searching the data. NCBI uses 'electronic PCR' (e-PCR) to compare all human sequences with the contents of the STS division of GenBank in order to identify primer-binding sites on the human sequences that may be amplified in a PCR reaction. The e-PCR tool permits the assignment of an initial location on the map for sequence data and the association of existing GenBank entries to the new reference sequence. The version of the e-PCR tool available on the web enables any researcher with a new human sequence to relate that sequence to existing maps and HTG sequence data.

GSSs

The GSS division of GenBank continues to grow rapidly, having grown over the past year by 40% to a total of 3.7 million records with over 2.0 billion nucleotides. GSS records represent 'random' genomic sequences, and are predominantly single reads from Bacterial Artificial Chromosomes ('BAC-ends') used in a variety of genome sequencing projects. The most highly represented species in the GSS division are *M. musculus* (942 000 records), *H. sapiens* (872 000 records), *Brassica oleracea* (299 000 records), and *Tetradon nigroviridis* (about 189 000) records). The human data is being used (www.ncbi.nlm.nih.gov/genome/clone) along with the STS records in tiling the BACs for the Human Genome Project (8).

HTG sequences

The HTG sequences in the HTG division of GenBank are unfinished large-scale genomic records that are in transition to a finished state, after which they will be placed in the appropriate organism division (9). These records are designated as Phase 0–3 depending on the quality of the data. Phase 0 records consist of survey sequences generated to characterize clones and may or may not progress to Phase 1. Phase 1 records contain unfinished sequence, and may consist of unordered, unoriented contigs with gaps. Phase 2 records contain unfinished sequence as ordered, oriented contigs, with or without gaps. Phase 3 records consist of finished sequence, with no gaps and may have annotations; upon reaching Phase 3, HTG records are moved into the appropriate organism division of GenBank. As of release 131 of GenBank, the HTG division comprised some 8 billion base pairs of sequence.

HTC sequences

The recently created HTC division of GenBank is designed to accommodate HTC sequences. HTCs are of draft quality, but may contain 5′-UTRs and 3′-UTRs, partial coding regions, and introns. HTC sequences which are finished and of high-quality are moved to the appropriate organism GenBank division. GenBank release 131 contained more than 38 000 HTC

sequences totaling over 46 million bases. Three organisms now provide the bulk of the HTC sequences; *M. musculus*, 55%, *Zea mays*, 24%, and *H. sapiens*, 14%. A recent project generating HTC data has been described (10).

Sequence identifiers and accession numbers

Each GenBank record, consisting of both a sequence and its annotations, is assigned a stable and unique identifier, the accession number. The accession number remains constant over the lifetime of the record even when there is a change to the sequence or annotation. Each DNA sequence in GenBank is assigned another unique identifier, called a 'gi'. The gi numbers appear on the VERSION line of GenBank records following the accession number. When a change is made to a sequence given in a GenBank record, a new gi number is assigned to the new sequence version associated with the record while the accession number for the record remains unchanged. The older sequence version retains the old gi.

By agreement among the collaborative DNA sequence databases, a third identifier was introduced in February of 1999 which consolidates the information present in both the gi and accession numbers. GenBank displays this identifier on the VERSION line, which appears below the ACCESSION line in the GenBank flat file format and is of the form 'Accession.version'. For example, an entry appearing in the database for the first time has a VERSION number equivalent to the ACCESSION number followed by '.1' to reflect that this is the first version of the sequence for this entry, e.g.:

ACCESSION AF000001
VERSION AF000001.1 GI: 987654321

If the nucleotide sequence changes, then so will the gi number and the version, but the accession will remain the same.

A similar system was adopted for tracking changes in the corresponding protein translations using identification numbers (in the format of three letters followed by five digits, e.g. AAA00001) that do not change, followed by a version number that increases with each subsequent version of the sequence. These identifiers appear as qualifiers for CDS features in the FEATURES table portion of a GenBank entry, e.g. /protein_id = 'AAA00001.1' Protein sequence translations also currently receive their own unique gi number, which appears as a second qualifier on the CDS feature: /db_xref = 'GI:1233445'.

BUILDING THE DATABASE

The data in GenBank, and the collaborating databases EMBL and DDBJ, is submitted primarily by individual authors to one of the three databases, or by sequencing centers as batches of ESTs, STSs, GSSs, HTCs or HTGs (usually sequences from cosmids, BACs or YACs). Data are exchanged daily with DDBJ and EMBL so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct submission

Virtually all records enter GenBank as direct electronic submissions, with the majority of authors using the BankIt or Sequin programs. Many journals require authors with sequence data to submit the data to a public database as a condition of publication.

GenBank staff can usually assign an accession number to a sequence submission within 2 working days of receipt, and do so at a rate of almost 700 per day. 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 checks for vector contamination, verification of the proper translation of coding regions, and checks for correct taxonomy and bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database. Authors have the right to request that their sequences be kept confidential until the time of publication. In these cases, authors are instructed to inform GenBank staff of the publication date of the article in which the sequence is cited in order to ensure a timely release of the data. GenBank policy requires that deposited sequence data be made public when the sequence or accession number is published. Although only the submitting scientist is permitted to modify sequence data or annotations, all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank at update@ncbi.nlm.nih.gov.

Several large-scale sequencing projects are producing megabases of genomic DNA sequence from human, mouse and other organisms. NCBI works closely with sequencing centers to ensure timely incorporation of these data into GenBank for public release. In parallel, NCBI has developed methods to integrate these sequences with genetic and physical map data and to search the sequences more effectively (e.g. through specialized BLAST variants such as MegaBLAST and options to mask Alu and other types of repetitive elements). GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission, including the program 'fa2htgs' and other tools (11).

Third Party Annotation (TPA)

As a result of an agreement among the collaborative nucleotide sequence databases, a new type of database submission, called a TPA sequence, is being accepted to allow the annotation of existing nucleotide sequences by third party authors. A TPA submission is derived or assembled from public primary sequence data found in the DDBJ/EMBL/GenBank International Nucleotide Sequence Collaboration Databases. Examples of TPA submissions include: an mRNA sequence assembled from overlapping ESTs; an mRNA sequence derived from an unannotated section of genomic sequence by comparison with another known mRNA from a different organism; or the annotation of exons, introns and coding regions on an unannotated genomic sequence. Trace data sequences or Whole Genome Shotgun (WGS) sequences in DDBJ/EMBL/GenBank may also be used as the basis of a TPA submission. Data from secondary sources such as NCBI Reference sequences or model organism databases, or primary data from proprietary databases may not be used as the basis of a TPA submission.

The format of TPA records is similar to that of a conventional GenBank records but includes the label 'TPA': at the beginning of each Definition Line and the keywords 'Third Party

Annotation; TPA' in the keywords field. The comment field of TPA records lists all primary sequences used to assemble the TPA sequence, while the primary field provides the base ranges of the primary sequences that contribute to the TPA sequence. For an example of a TPA record, see accession number 'BK000016'.

TPA submissions to GenBank may be made using the web-based submission tool, BankIt, or using Sequin. TPA sequences are not released to the public until their accession numbers and/or sequence data and annotation appear in a peer-reviewed publication in a biological journal.

BankIt

About a third of individual submissions are received through NCBI's web-based data submission tool, BankIt (http://www.ncbi.nlm.nih.gov/BankIt). Using BankIt, authors enter sequence information directly into a form, edit as necessary, and add biological annotation (e.g. coding regions, mRNA features). Recent revisions of BankIt allow for the entry of more fielded information through the use of listboxes and pull-down menus. Free-form text boxes allow the submitter to further describe the sequence, without having to learn formatting rules or use restricted vocabularies. BankIt validates submissions, flagging many common errors, and checks for vector contamination using a variant of BLAST called Vecscreen, before creating a draft record in GenBank flat file format for the submitter to review. BankIt is the tool of choice for simple submissions, especially when only one or a small number of records is submitted (9). BankIt can also be used by submitters to update their existing GenBank records.

Sequin

NCBI has developed a stand-alone multi-platform submission program called Sequin (http://www.ncbi.nlm.nih.gov/Sequin/index.html) which can also be used interactively with other NCBI sequence retrieval and analysis tools. Sequin handles simple sequences (e.g. a cDNA), as well as long sequences and segmented entries, for which BankIt and other web-based submission tools are not well-suited. Sequin has convenient editing and complex annotation capabilities and contains a number of built-in validation functions for enhanced quality assurance. It is also designed to facilitate the submission of sequences from phylogenetic, population studies, mutation studies and environmental samples, and can incorporate alignment data. Sequin can be used to edit and update sequence records, as well as to perform sequence analysis. For example, Sequin can now incorporate any analysis tool available on the web that accepts FASTA or ASN.1 (Abstract Syntax Notation 1) formatted data as its input. In addition, Sequin is able to accommodate large sequence records, such as the *Escherichia coli* genome of 5.6 Mb and read in a full complement of annotations via simple tables. Versions for Macintosh, PC and Unix computers are available via anonymous FTP to 'ftp.ncbi.nih.gov' in the 'sequin' directory. Once a submission is completed, users can email the Sequin file to the address gb-sub@ncbi.nlm.nih.gov. Additional information about Sequin can be found through the NCBI home page.

RETRIEVING GENBANK DATA

The Entrez system

Entrez (http://www.ncbi.nlm.nih.gov/Entrez/) is an integrated database retrieval system that accesses DNA and protein sequence data, genome mapping data, population sets, phylogenetic sets, environmental sample sets, gene expression data, the NCBI taxonomy, protein domain information, protein structures from the Molecular Modeling Database, MMDB (12), and MEDLINE references via PubMed. The DNA and protein sequence data are integrated from a variety of sources and therefore include more sequence data than are available within GenBank DNA sequence database alone. Entrez searching is provided on NCBI's web site, via the Query email server (query@ncbi.nlm.nih.gov), and as a network client that can be downloaded by FTP. Entrez is also discussed elsewhere in this issue (4).

BLAST sequence-similarity searching

The most frequent type of analysis performed using GenBank is the search for sequences similar to a query sequence. NCBI offers the BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) family of programs to locate regions of similarity between a query sequence and database sequences (13,14). BLAST searches may be performed on NCBI's web site, or using a set of stand-alone programs distributed by FTP. BLAST is discussed in more detail in a separate article in this issue (4).

Obtaining GenBank by FTP

NCBI uses the ASN.1 data format for internal maintenance of GenBank, but distributes the GenBank releases in the traditional flat-file format as well as in ASN.1. The full bimonthly GenBank release and the daily updates, which also incorporate sequence data from EMBL and DDBJ, are available by anonymous FTP from NCBI at 'ftp.ncbi.nih.gov' as well as from two mirror sites, at the San Diego SuperComputer Center (genbank.sdsc.edu/pub) and at the University of Indiana (bio-mirror.net/biomirror/genbank). The full release in flat-file format is available as compressed files in the directory, 'genbank' with a cumulative update file contained in the sub-directory, 'daily', and a non-cumulative set of updates contained in 'daily-nc'. A set of sequence-only files in FASTA format, corresponding to the GenBank database subsets searched by BLAST and including the non-redundant nucleotide and protein databases, is available in the 'blast/db' directory.

MAILING ADDRESS

GenBank, National Center for Biotechnology Information, Building 38A, Room 8S-803, 8600 Rockville Pike, Bethesda, MD 20894, USA. Tel: +1 3014962475; Fax: +1 3014809241.

ELECTRONIC ADDRESSES

http://www.ncbi.nlm.nih.gov/ (NCBI 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 'confidential' 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.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
2. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R., Redaschi,N., Stoehr,P., Tuli,M.A. and Vaughan, R. (2002) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **30**, 21–26.
3. Tateno,Y., Imanishi,T., Miyazaki,S., Fukami-Kobayashi,K., Saitou,N., Sugawara,H. and Gojobori,T. (2002) DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.*, **30**, 27–30.
4. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **31**, 28–33.
5. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.
6. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
7. Hudson,T.J., Stein,L.D., Gerety,S., Ma,J., Castle,A.B., Silva,J., Slonim,D.K., Baptista,R., Kruglyak,L., Xu,S.-H. *et al.* (1995) An STS-based map of the human genome. *Science*, **270**, 1945–1954.
8. Smith,M.W., Holmsen,A.L., Wei,Y.H., Peterson,M. and Evans,G.A. (1994) Genomic sequence sampling: a strategy for high resolution sequence-based physical mapping of complex genomes. *Nature Genet.*, **7**, 40–47.
9. Kans,J.A. and Ouellette,B.F.F. (2001) In Baxevanis,A. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, New York, NY, pp. 65–81.
10. Hayashizaki,Y. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
11. Ouellette,B.F.F. and Boguski,M.S. (1997) Database divisions and homology search files: a guide for the perplexed. *Genome Res.*, **7**, 952–957.
12. Marchler-Bauer,A., Anderson,J.B., Fedorova,N., DeWeese-Scott,C., Geer,L.Y., Hurwitz,D., Jackson,J.J., Jacobs,A., Lanczycki,C., Liebert,C.A., Madej,T., Marchler,G.H., Mazumder,R., Nikolskaya,A., Panchenko,A.R., Shoemaker,B.A., Song,J., Sridhar Rao,B., Thiessen,P.A., Vasidevan,S., Wang,Y., Yamashita,R.A., Yin,J. and Bryant,S.H. (2002) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **30**, 249–252.
13. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
14. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3991.