Nucleic Acids Research, 2004, Vol. 32, Database issue ***D23–D26***
DOI: 10.1093/nar/gkh045

GenBank: update

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 16, 2003; Revised and Accepted September 22, 2003

ABSTRACT

GenBank (R) is a comprehensive database that contains publicly available DNA sequences for more than 140 000 named organisms, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. Most submissions are made using the BankIt (web) or Sequin program 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 sequences and 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, USA.

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 sequences from issued patents. GenBank incorporates sequences submitted to the EMBL Data Library (2) in the UK and the DNA Data Bank of Japan (DDBJ) (3) as part of a long-standing international collaboration between the three databases in which data is exchanged daily to ensure a uniform and 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 9 million new sequences added over the past 12 months. As of Release 137 in August 2003, GenBank contained over 33.9 billion nucleotide bases from 27.2 million individual sequences. Complete genomes (http://www.ncbi.nlm.nih.gov/Genomes/index.html) represent a growing portion of the database, with over 40 of more than 130 complete microbial genomes in GenBank deposited over the past year. The number of eukaryote genomes for which coverage and assembly are good continues to increase with over a dozen such assemblies now available, including that of the reference human genome.

Sequence-based taxonomy

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. Over 140 000 species are represented in GenBank and new species are being added at the rate of over 1700 per month. About 26% of the sequences in GenBank are of human origin and 20% of all sequences are human ESTs. After *Homo sapiens*, the top species in GenBank in terms of number of bases are *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Oryza sativa*, *Drosophila melanogaster*, *Zea mays*, *Arabidopsis thaliana* and *Gallus gallus*.

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 mutation or modification.

The files in the GenBank distribution have traditionally been divided into 'divisions' that roughly correspond to

*To whom correspondence should be addressed. Tel: +1 301 435 5950; Fax: +1 301 480 9241; Email: wheeler@ncbi.nlm.nih.gov

taxonomic groups such as 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, high-throughput genomic (HTG) and high-throughput cDNA (HTC) sequences, making a total of 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 the NCBI's FTP site.

ESTs

ESTs continue to be the major source of new sequence records and gene sequences. Over the past year the number of ESTs has increased by over 45% to a total of 18.1 million sequences representing over 580 different organisms. The top five organisms represented in the EST division are *H.sapiens* (5.4 million records), *M.musculus* (3.8 million records), *R.norvegicus* (540 000 records), *Triticum aestivum* (500 000 records) and *Ciona intestinalis* (490 000 records). As part of its daily processing of GenBank EST data, the NCBI identifies through BLAST searches all homologies for new EST sequences and incorporates that information into the companion database, dbEST (http://www.ncbi.nlm.nih.gov/dbEST/index.html) (5). The data in dbEST is further processed to produce the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/) of gene-oriented sequence clusters described more fully in (4).

Sequence-tagged sites (STSs) and GSSs

The STS division of GenBank (http://www.ncbi.nlm.nih.gov/dbSTS/index.html) contains over 240 000 sequences including anonymous STSs based on genomic sequence as well as gene-based STSs derived from the 3′ ends of genes and ESTs. These STS records usually include primer sequences, annotations and PCR conditions.

The GSS division of GenBank (http://www.ncbi.nlm.nih.gov/dbGSS/index.html) has grown over the past year by 73% to a total of 6.4 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 *Z.mays* (1.3 million records), *M.musculus* (952 000 records), *H.sapiens* (893 000 records) and *Brassica oleracea* (595 000 records). Human data have been used (http://www.ncbi.nlm.nih.gov/genome/clone) along with the STS records in tiling the BACs for the Human Genome Project (6).

HTG and HTC sequences

The HTG division of GenBank (http://www.ncbi.nlm.nih.gov/HTGS/) contains unfinished large-scale genomic records that are in transition to a finished state (7). These records are designated as Phase 0–3 depending on the quality of the data. Upon reaching Phase 3, the finished state, HTG records are moved into the appropriate organism division of GenBank. As of release 137 of GenBank, the HTG division comprised some 12 billion bp of sequence.

The HTC division of GenBank accommodates high-throughput cDNA sequences. HTCs are of draft quality, but may contain 5′-UTRs and 3′-UTRs, partial coding regions and introns. HTC sequences that are finished and of high quality are moved to the appropriate organism GenBank division. GenBank release 137 contained more than 148 000 HTC sequences totaling over 200 million bases. A recent project generating HTC data has been described (8) and other projects are listed at: http://www.ncbi.nlm.nih.gov/genome/flcdna/.

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, which remains constant over the lifetime of the record even when there is a change to the sequence or annotation. The DNA sequence within a GenBank record is also assigned a unique identifier, called a 'GI', that appears on the VERSION line of GenBank flatfile records following the accession number. A third identifier of the form 'Accession.version', also displayed on the VERSION line of flatfile records, consolidates the information present in the GI and accession numbers. An entry appearing in the database for the first time has an 'Accession.version' identifier equivalent to the ACCESSION number of the GenBank record followed by '.1' to indicate the first version of the sequence for the record, e.g. ACCESSION AF000001 VERSION AF000001.1 GI: 987654321. When a change is made to a sequence given in a GenBank record, a new GI number is issued to the sequence and the version extension of the 'Accession.version' identifier is incremented. The accession number for the record as a whole remains unchanged and the older sequence remains available under the old 'Accession.version' identifier and GI.

A similar system tracks changes in the corresponding protein translations using 'Accession.version' identifiers comprised of a protein accession number, e.g. AAA00001, followed by a version number. 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 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, are submitted primarily by individual authors to one of the three databases, or by sequencing centers as batches of ESTs, STSs, GSSs, HTCs or HTGs. 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 (http://www.ncbi.nlm.nih.gov/Genbank/index.html), with the majority of authors using the BankIt or Sequin program. 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 two 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 articles in which the sequence is cited to

retrieve the relevant data. Direct submissions receive a quality assurance review that includes checks for vector contamination, proper translation of coding regions, correct taxonomy and correct bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database. Authors may ask that their sequences be kept confidential until the time of publication. Since GenBank policy requires that deposited sequence data be made public when the sequence or accession number is published, authors are instructed to inform GenBank staff of the publication date of the article in which the sequence is cited in order to ensure timely release of the data. 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.

The NCBI works closely with sequencing centers to ensure timely incorporation of bulk data into GenBank for public release. GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission, including the program 'fa2htgs' and other tools (9).

Third party annotation

Third party annotation (TPA) refers to the annotation by third party authors of nucleotide sequences 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, 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, but data from secondary sources such as NCBI Reference sequences, or primary data from proprietary databases may not be used.

The format of a TPA record (e.g. BK000016) is similar to that of a conventional GenBank record 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; the Primary field provides the base ranges of the primary sequences that contribute to the TPA sequence.

TPA submissions to GenBank may be made using either BankIt or Sequin, but TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer-reviewed biological journal. For more information on TPA, see http://www.ncbi.nlm.nih.gov/Genbank/tpa.html.

BankIt

About a third of author submissions are received through the 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 such as coding regions or mRNA features. Free-form text boxes, list boxes and pull-down menus 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 to be submitted (7). BankIt can also be used by submitters to update their existing GenBank records.

Sequin

The NCBI has developed a stand-alone multi-platform submission program called Sequin (http://www.ncbi.nlm.nih.gov/Sequin/index.html) that can be used interactively with other NCBI sequence retrieval and analysis tools. Sequin handles simple sequences such as a cDNA, as well as segmented entries, phylogenetic studies, population studies, mutation studies, environmental samples and alignments 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 quality assurance. 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 at 'ftp.ncbi.nih.gov' in the 'sequin' directory. Once a submission is completed, submitters can email the Sequin file to the address: gb-sub@ncbi.nlm.nih.gov.

RETRIEVING GENBANK DATA

The ENTREZ system

The sequence records in GenBank are accessible via Entrez (http://www.ncbi.nlm.nih.gov/Entrez/), a robust and flexible 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 (10) and MEDLINE references via PubMed. The Entrez sequence databases are taken from a variety of sources and therefore include more sequence data than are available within the GenBank DNA sequence database alone.

BLAST sequence-similarity searching

Sequence-similarity searches are the most frequent and basic type of analysis performed on the GenBank data. 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 (11,12). BLAST searches may be performed on the NCBI's website, 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 distributes the GenBank releases in the traditional flat-file format as well as in the Abstract Syntax Notation (ASN.1) format used for internal maintenance. The full bimonthly GenBank release and the daily updates, which also incorporate sequence data from EMBL and DDBJ, are available by anonymous FTP from the NCBI at ftp.ncbi.nih.gov as well as from two mirror sites, at the San Diego SuperComputer Center

(ftp://genbank.sdsc.edu/pub/) and at the University of Indiana (ftp://bio-mirror.net/biomirror/genbank/). The full release in flat-file format is available as compressed files in the directory, 'genbank' with a non-cumulative set of updates contained in 'daily-nc'. A script is provided on the FTP site to convert a set of daily updates into a cumulative update.

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 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. and Wheeler,D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
2. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **31**, 17–22.
3. Miyazaki,S., Sugawara,H., Gojobori,T. and Tateno,Y. (2003) DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.*, **31**, 23–27.
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. *et al.* (2003) Database resources of the National Center for Biotechnology. *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. 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.
7. Kans,J.A. and Ouellette,B.F.F. (2001) Submitting DNA sequences to the databases. In Baxevanis,A. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Inc., New York, pp. 65–81.
8. Hayashizaki,Y. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
9. 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.
10. Chen,J., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **31**, 474–477.
11. 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.
12. 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.