Published online 5 December 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D21–D25***
doi:10.1093/nar/gkl986

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 15, 2006; Accepted October 26, 2006

ABSTRACT

GenBank (R) is a comprehensive database that contains publicly available nucleotide sequences for more than 240 000 named organisms, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. Most submissions are made using the web-based BankIt or standalone Sequin programs and accession numbers are assigned by GenBank staff upon receipt. Daily data exchange with the EMBL Data Library in Europe and the DNA Data Bank of Japan ensures 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, begin at the NCBI Homepage (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.

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 also contributes sequences from issued patents. GenBank, the EMBL Data Library (2) in Europe, and the DNA Databank of Japan (DDBJ) (3) comprise the International Nucleotide Sequence Databases, and are members of a long-standing collaboration in which information 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 and via a wide range of web-based retrieval and analysis services which operate on the GenBank data (4).

ORGANIZATION OF THE DATABASE

From its inception, GenBank has doubled in size about every 18 months. It currently contains over 65 billion nucleotide bases from more than 61 million individual sequences, with 15 million new sequences added in the past year. Contributions from whole genome shotgun (WGS) projects supplement the data in the traditional divisions to bring the total beyond 145 billion bases. Complete genomes (www.ncbi.nlm.nih.gov/Genomes/index.html) continue to represent a growing portion of the database, with over 120 of more than 370 complete microbial genomes in GenBank deposited over the past year. The number of eukaryote genomes for which coverage and assembly are significant continues to increase as well, with over 104 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 (www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html) developed by NCBI in collaboration with EMBL and DDBJ and with the valuable assistance of external advisers and curators. Over 240 000 named species are represented in GenBank and new species are being added at the rate of over 2900 per month. About 16% of the sequences in GenBank are of human origin and 13% 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, Bos taurus, Danio rerio, Zea mays, Oryza sativa, Strongylocentrotus purpuratus, Sus scrofa, Xenopus tropicalis,* and *Canis familiaris.*

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

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

(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 partitioned into 'divisions' that roughly correspond to 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. In recent years, divisions have been added to support specific sequencing strategies. These include divisions for expressed sequence tag (EST), genome survey (GSS), high throughput genomic (HTG), high throughput cDNA (HTC), and environmental sample (ENV) sequences, making a total of 18 divisions. For convenience in file transfer, the larger divisions, such as the EST and PRI, are partitioned into multiple files for the bimonthly GenBank releases on NCBI's FTP site.

Expressed sequence tags. ESTs continue to be a major source of new sequence records and gene sequences, comprising over 21 billion nucleotide bases in GenBank release 155. Over the past year, the number of ESTs has increased by over 40% to a total of 38.3 million sequences representing more than 1200 different organisms. The top organisms represented in the EST division are *H.sapiens* (7.8 million records), *M.musculus* (4.7 million records), *O.sativa* (1.2 million records), *Z.mays* (1.1 million records), *B.taurus* (1.1 million records), and *D.rerio* (1.1 million records). 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 database, dbEST (www.ncbi.nlm.nih.gov/dbEST/index.html) (5). The data in dbEST is processed further to produce the UniGene database (www.ncbi.nlm.nih.gov/UniGene/) of more than 1.2 million gene-oriented sequence clusters representing over 70 organisms, described more fully in (4).

Sequence-tagged sites (STSs), genome survey sequences (GSSs) and environmental sample sequences (ENV). The STS division of GenBank (www.ncbi.nlm.nih.gov/dbSTS/index.html) contains over 883 000 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 mapping information.

The GSS division of GenBank (www.ncbi.nlm.nih.gov/dbGSS/index.html) has grown over the past year by 22% to a total of 14.9 million records for over 600 organisms and comprises over 9.4 billion nucleotide bases. GSS records 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* (2.0 million records), *M.musculus* (1.5 million records), *H.sapiens* (970 000 records) and *C.familiaris* (854 000 records). Human GSS records have been used (www.ncbi.nlm.nih.gov/genome/clone) along with the STS records in tiling the BACs for the Human Genome Project (6).

The ENV division of GenBank accommodates non-WGS sequences obtained via environmental sampling methods in which the source organism is unknown. Records in the ENV division contain 'ENV' in the keyword field and use an '/environmental_sample' qualifier in the source feature. As of GenBank release 155, the ENV division of GenBank contained over 275 000 sequences, comprising 236 million base pairs, representing more than 4900 studies.

High-throughput genomic (HTC) and high-throughput cDNA (HTC) sequences. The HTG division of GenBank (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 155 of GenBank, the HTG division contained 15.9 billion base pairs of sequence, an increase of almost 3 billion bases over the past year.

The HTC division of GenBank accommodates HTC sequences. HTCs are of draft quality but may contain 5′-untranslated regions (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 155 contained more than 441 000 HTC sequences totaling over 539 million bases. One project generating HTC data is described in (8).

Whole genome shotgun sequence (WGS). Over 80 billion bases of WGS sequence appears in GenBank as sets of WGS contigs, many of them bearing annotations, originating from a single sequencing project. These sequences are issued accession numbers consisting of a four-letter project ID, followed by a two-digit version number, and a six-digit contig ID. Hence, the WGS accession number 'AAAA01072744' is assigned to contig number '072744' of the first version of project 'AAAA'. WGS sequencing projects have contributed over 18 million contigs to GenBank, a 64% increase over the past year. These primary sequences have been used to construct some 760 000 large-scale assemblies of scaffolds and chromosomes. WGS project contigs for *H.sapiens, C.familiaris, Pan trodlodytes, Macacca mulatta, Drosophila, Saccharomyces*, and more than 450 other organisms and environmental samples are available. For a complete list of WGS projects with links to the data, see www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi.

WGS projects may be annotated. However, many low-coverage genome projects do not contain annotation. Because these sequence projects are considered draft and not complete, these annotations may not be tracked from one assembly version to the next and should be considered preliminary.

Submitters of WGS sequences, and genomic sequences in general, are urged to use a new set of evidence tags of the form '/experimental=*text*' and '/inference=*TYPE:text*', where '*TYPE*' is one of a number of standard inference types and '*text*' is made up of structured text. These new qualifiers replace 'evidence=experimental' and 'evidence=non-experimental', respectively, which are no longer supported.

Special record types

Third Party Annotation. Third Party Annotation (TPA) records support the reporting of published sequence annotation by a scientist other than the original submitter of the primary sequence record in DDBJ/EMBL/GenBank. TPA records fall into one of two categories, 'experimental', in which case there is a direct experimental evidence for the

existence of the annotated molecule, and 'inferential', in which case the experimental evidence is indirect. TPA sequences may be created by assembling a number of primary sequences. 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 the 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.

Over 5000 TPA records are contained in GenBank release 155, including over 2170 for *Drosophila melanogaster*, 950 for *H.sapiens*, 330 for *O.sativa* and 290 for *M.musculus*. TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer-reviewed biological journal. TPA submissions to GenBank may be made using either BankIt, or Sequin. For more information on TPA, see www.ncbi.nlm.nih.gov/Genbank/TPA.html.

GenBank CON records for assemblies of smaller records. Although many genomes, such as bacterial genomes, are represented in GenBank as single sequences, it is desirable from the standpoints of data transfer and analysis to break some very long sequences, such as portions of eukaryotic genomes, into smaller segments. In these cases, CON division records for the entire sequence are produced that contain assembly instructions to allow the seamless display and download of the full sequence. Many CON records also include annotations.

BUILDING THE DATABASE

The sequences and biological annotations 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 EST, STS, GSS, HTC, WGS, or HTG sequences. Information is 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 electronic submission

Virtually all records enter GenBank as direct electronic submissions (www.ncbi.nlm.nih.gov/Genbank/index.html), 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 two working days of receipt, and do so at a rate of almost 1600 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 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 a 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.

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 'tbl2asn', described at www.ncbi.nlm.nih.gov/Sequin/table.html.

Submission using BankIt. About one-third of author submissions are received through NCBI's web-based data submission tool, BankIt (www.ncbi.nlm.nih.gov/BankIt). Using BankIt, authors enter sequence information directly into a form, and add biological annotations 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 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.

Submission using Sequin and tbl2asn. NCBI also offers a standalone multi-platform submission program called Sequin (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 sequences, such as that of the 5.6 Mb *Escherichia coli* genome, 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 e-mail the Sequin file to the address (gb-sub@ncbi.nlm.nih.gov).

Submitters of large, heavily annotated genomes may find it convenient to use 'tbl2asn', referenced above under 'Direct submission', to convert a table of annotations generated via an annotation pipeline into an ASN.1 record suitable for submission to GenBank.

Submission of barcode sequences. The Consortium for the Barcode of Life (CBOL) is an international initiative to develop DNA barcoding as a tool for characterizing species of organisms using a short DNA sequence derived from a portion of the cytochrome oxidase subunit I gene. NCBI, in collaboration with CBOL (barcoding.si.edu/index\s\do5(d)etail.htm), has created an online tool for the bulk submission of

barcode sequences to GenBank (www.ncbi.nlm.nih.gov/BankIt/barcode/) that allows users to upload files containing a batch of sequences with associated source information. It is anticipated that this tool will be used for other types of bulk submissions in the near future.

Sequence identifiers and accession numbers

Each GenBank record, consisting of both a sequence and its annotations, is assigned a unique identifier, the accession number, that is shared across the three collaborating databases (GenBank, DDBJ, EMBL) and remains constant over the lifetime of the record even when there is a change to the sequence or annotation. Each version of the DNA sequence within a GenBank record is also assigned a unique NCBI 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, contains the information present in both 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. These identifiers appear as qualifiers for CDS features in the FEATURES 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, e.g. /db_xref=' GI:1233445'.

Ensuring stable access to sequence data

It is becoming increasingly popular for research groups to share new biological sequences and update existing sequences by directly posting the data on the Web. While this is a convenient and effective way to share the data among a set of collaborators, if original data and updates are not also submitted to a central repository, three significant problems arise; the access lifetime of the data may be reduced, the full biological context of the data may not be realized, and existing data in heavily used centralized databases will become outdated.

The ephemeral nature of much of the content on the web is part of the common experience of web users. In one attempt to quantify content lifetime, 360 randomly selected web pages were tracked for a period of 4 years, and a half-life of only 2 years was measured for the set (9). Although a well-maintained web page can certainly persist for longer than 2 years, the relatively short half-life reported for this set of pages reflects the many factors that can intervene to affect access to web-posted data.

Even during the accessible lifetime of web-posted sequence data, however, the full biological context of a sequence may not be realized if the sequence cannot be conveniently compared with others—perhaps derived from distantly related organisms that are beyond the scope of the host web page.

In addition, if updates to sequences contained within centralized databases are made to a web page, but not also made to corresponding records in the central database, the newer data will not reach the wider research community and much of the impact of the data will be lost.

Submission of sequence data to a centralized repository such as GenBank solves these three problems. Researchers are ensured stable access to the data via versioned bimonthly releases available by FTP, NCBI-maintained as well as numerous third party interfaces to a uniform dataset, and the archival redundancy offered by the tripartite International Nucleotide Sequence Databases collaboration. Combining new data with that of other researchers worldwide within a central database provides a broad biological context that stimulates discovery—keeping each sequence current magnifies the utility of all the sequences in the database.

RETRIEVING GenBank DATA

The Entrez system

The sequence records in GenBank are accessible via Entrez (www.ncbi.nlm.nih.gov/Entrez/), a flexible database retrieval system that covers over 30 biological databases. These include DNA and protein sequences derived from GenBank and other sources, genome maps, population, phylogenetic and environmental sequence sets, gene expression data, the NCBI taxonomy, protein domain information, protein structures from the Molecular Modeling Database, MMDB (10); each database linked to the scientific literature via PubMed and PubMed Central.

BLAST sequence-similarity searching

Sequence-similarity searches are the most fundamental and frequent type of analysis performed on the GenBank data. NCBI offers the BLAST (www.ncbi.nlm.nih.gov/BLAST/) family of programs to detect similarities between a query sequence and database sequences (11,12). BLAST searches may be performed on the NCBI's website, or via a set of standalone programs distributed by FTP. BLAST is discussed in a separate article in this issue (4).

Obtaining GenBank by FTP

NCBI distributes GenBank releases in the traditional flat-file format as well as in the Abstract Syntax Notation (ASN.1) format used for internal maintenance. The complete 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 a mirror site at the University of Indiana (ftp://bio-mirror.net/biomirror/genbank/). The complete release in the 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 in the 'tools' directory of the GenBank FTP site to convert a set of daily updates into a cumulative update.

MAILING ADDRESS

GenBank, National Center for Biotechnology Information, Building 38A, Room 3N-301-B, 8600 Rockville Pike, Bethesda, MD 20894, USA. Tel: +1 301 496 2475; Fax: +1 301 480 9241.

ELECTRONIC ADDRESSES

NCBI Home Page: info@ncbi.nlm.nih.gov

Submission of sequence data to GenBank: gb-sub@ncbi.nlm.nih.gov

Revisions to or notification of release of 'confidential' GenBank entries: update@ncbi.nlm.nih.gov

General information about NCBI and services: info@ncbi.nlm.nih.gov

CITING GenBank

If you use the GenBank database in your published research, we ask that this paper be cited.

ACKNOWLEDGEMENT

Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids Res*., **34**, 16–20.
2. Cochrane,G., Aldebert,P., Althorpe,N., Andersson,M., Baker,W., Baldwin,A., Bates,K., Bhattacharyya,S., Browne,P., van denBroek,A. *et al*. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res*., **34**, 10–15.
3. Okubo,K., Sugawara,H., Gojobori,T. and Tateno,Y. (2006) DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res*., **34**, 6–9.
4. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **34**, 173–180.
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. and Ouellette,B. (2001) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins Chapter Submitting DNA Sequences to the Databases*. John Wiley and Sons, Inc., NY, pp. 65–81.
8. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al*. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
9. Koehler,W. (2002) Web page change and persistence—a four-year longitudinal study. *J. Am. Soc. Inform. Sci. Technol*., **53**, 162–171.
10. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D., Ke,Z. *et al*. (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res*., **33**, 192–196.
11. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., 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., Schäffer,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**, 3986–3990.