12–17 Nucleic Acids Research, 1999, Vol. 27, No. 1

GenBank

Dennis A. Benson*, Mark S. Boguski, David J. Lipman, James Ostell, B. F. Francis Ouellette+, Barbara A. Rapp 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 October 5, 1998; Accepted October 5, 1998

ABSTRACT

The GenBank (Registered Trademark symbol) sequence database incorporates DNA sequences from all available public sources, primarily through the direct submission of sequence data from individual laboratories and from large-scale sequencing projects. Most submitters use the BankIt (Web) or Sequin programs to format and send sequence data. Data exchange with the EMBL Data Library and the DNA Data Bank of Japan helps ensure comprehensive worldwide coverage. GenBank data is accessible through NCBI's integrated retrieval system, Entrez, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome and protein structure information. MEDLINE (Registered Trademark symbol) abstracts from published articles describing the sequences are included as an additional source of biological annotation through the PubMed search system. Sequence similarity searching is offered through the BLAST series of database search programs. In addition to FTP, Email, and server/client versions of Entrez and BLAST, NCBI offers a wide range of World Wide Web retrieval and analysis services based on GenBank data. The GenBank database and related resources are freely accessible via the URL: http://www.ncbi.nlm.nih.gov

INTRODUCTION

GenBank (1) is a public database of all known 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). NCBI was created by special legislation in 1988 to develop information systems in the field of molecular biology to support the biomedical research community. NCBI was also mandated to conduct basic and applied research and, as part of the NIH Intramural Program, NCBI scientists work in areas of gene and genome analysis, computational structural biology and mathematical methods for sequence analysis.

NCBI builds GenBank primarily from the direct submission of sequence data from authors. Another major source of data is bulk submission of EST and other high-throughput data from sequencing centers. The data are supplemented by sequences submitted to other public databases. Through a long-standing international collaboration with the EMBL Data Library (2) in the UK and the DNA Databank of Japan (DDBJ) (3), data are exchanged daily to ensure that all three sites maintain a comprehensive collection of sequence information. NCBI makes the data available at no cost over the Internet, by FTP access and by Web text and sequence similarity search services.

ORGANIZATION OF THE DATABASE

GenBank continues to grow at an exponential rate. Over the past 12 months 770 000 new sequences have been added. As of Release 108 in August 1998, GenBank contained almost 1.8 billion nucleotide bases from 2.5 million different sequences. Complete genomes represent a growing portion of the database, with 10 complete genomes added in 1998, compared with six in 1997, and two in 1996. A recent addition is the *Caenorhabditis elegans* genome of 100 Mb sequenced jointly by Washington University in St Louis and the Sanger Sequencing Center in Hinxton, UK. There are at least 20 additional microorganism genomes that are being sequenced, many of which are expected to be in the public databases over the coming year. Historically, GenBank had been doubling in size about every 18 months, but that rate has accelerated to doubling every 15 months due primarily to the enormous growth in data from expressed sequence tags (ESTs). Nearly 70% of the sequences in the current GenBank release are ESTs, and most of the growth in sequence records over the past three years has come from the collaborative project between Merck & Co. and Washington University (4,5). Human EST sequencing continues and is being supplemented by a mouse EST project supported by the Howard Hughes Medical Institute at Washington University.

Sequence-based taxonomy

Over 40,000 different species are represented in GenBank and new species are being added at the rate of 900 per month. Human sequences constitute 54% of the total sequences (42% of all sequences are human ESTs). After *Homo sapiens*, the top species

*To whom correspondence should be addressed. Tel: +1 301 435 5980; Fax: +1 301 480 9241; Email: dab@ncbi.nlm.nih.gov

+Present address: Centre for Molecular Medicine and Therapeutics, 950 West 28th Avenue, Vancouver, BC V5Z 4H4, Canada

in GenBank in terms of the number of bases include *Mus musculus*, *C.elegans*, *Arabidopsis thaliana* and *Drosophila melanogaster*. Database sequences are processed and can be queried using a comprehensive sequence-based taxonomy developed by NCBI in collaboration with EMBL and DDBJ and with the valuable assistance of external advisors and curators. Further details, along with a taxonomy browser and information on taxonomic resources, may be found on NCBI's home page.

GenBank divisions

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, intron/exon boundaries, sites of mutations or modifications and other sequence features. Protein translations for coding regions are in the feature table. Bibliographic references are included along with links to the MEDLINE abstracts for all published sequences.

The files in the GenBank distribution have traditionally been divided into 'divisions' that roughly correspond to taxonomic divisions, e.g., bacteria, viruses, primates, rodents, etc. There are currently 17 divisions. For convenience in file transfer, the larger divisions, e.g., EST and primate, are divided into multiple files.

Sequence identifiers and accession numbers

To produce the GenBank and Entrez databases, NCBI tracks and indexes records from multiple sources of sequence data: DNA sequences from EMBL, DDBJ and the US Office of Patents and Trademarks (USPTO), plus amino acid sequences from the nucleotide databases (CDS features with '/translations'), Protein Identification Resource (PIR), SWISS-PROT, Protein Research Foundation (PRF), the Protein Data Bank (PDB) and the USPTO. GenBank assigns each record an accession number, which is considered the unique identifier for each GenBank entry and does not change, even when there is a change to the sequence or annotation. In order to identify specific sequences from the different sources, as well as any changes in those sequences that may occur, NCBI additionally assigns a stable identifier, termed a 'gi' number, to each sequence. When a change in a sequence occurs, a new 'gi' number is assigned to the new sequence version. These identifiers appear in the 'NID' field of a GenBank record, immediately following the ACCESSION field.

There is an agreement among the collaborative DNA sequence databases to introduce in February 1999 a third identifier which will encompass the information present in both the 'gi' and ACCESSION number. GenBank will show this identifier on the VERSION line, which will appear below the NID line and will be in the form 'Accession.version'. For example, an entry appearing in the database for the first time would have a VERSION number equivalent to the ACCESSION number followed by '.1' to reflect that this is the first version of the sequence in this entry, e.g.:

ACCESSION AF000001
NID g987654321
VERSION AF000001.1 GI: 987654321

The VERSION line will also display the 'gi' number. If the nucleotide sequence changes, then so will the 'gi' number and the version, but the accession will remain the same. Although the NID line will carry redundant information, this line will remain in the file for an extended time to ensure compatibility with existing programs.

A similar system for tracking changes in the corresponding protein translations will also be introduced in February 1999. Protein sequences will have identification numbers (in the format of three letters followed by five digits, e.g., AAA00001) that do not change, followed by a version number which increases with each subsequent version of the sequence. This will appear as a qualifier for a CDS feature in the FEATURES table portion of a GenBank entry, e.g., /protein_id='AAA00001.1'

Protein translations currently receive their own unique 'gi' number, which appears as a qualifier on the CDS feature: /db_xref='PID:g1234567'. The letter prefix indicates the database of origin for these identifiers (d=DDBJ, e=EMBL, g=GenBank). Eventually, after a transition period of at least a year, this form will be phased out since the new 'protein_id' complete with the version number will represent both a stable identifier and a means to identify changes in the sequence.

Expressed sequence tag (EST) data

ESTs continue to be the major source of new sequence records and genes. Last year there were 1 247 603 sequences in the EST division of GenBank. Over the past year the number of ESTs has increased by nearly 50% to the current total of 1 765 860 sequences representing over 130 different organisms. The top five organisms include: 1 072 823 human sequences (61%); 351 852 mouse (20%); 72 569 nematode (4%); 57 227 rat (3%); and 37 848 fruit fly (2%).

As part of its daily processing of EST data, NCBI identifies through BLAST searches all homologies for new EST sequences and incorporates that information into a specialized database (dbEST). ESTs continue to provide the major source of new gene discoveries and NCBI has processed more than a million queries (BLAST searches, Email retrievals, Web accesses and anonymous FTP downloads) for dbEST in the past year.

In order to organize the EST data in a useful fashion, NCBI has created the UniGene collection of unique human genes (6) and mouse genes. UniGene starts with entries in the Primate (or Rodent) division of GenBank, combines these with ESTs of that organism, and creates clusters of sequences that share virtually identical 3′ untranslated regions (3′ UTRs). In this manner, over one million human ESTs in GenBank have been reduced 20-fold to ~52 000 sequence clusters, each of which may be considered as representing a single human gene. In a similar fashion, the mouse ESTs have been organized as 10 000 clusters. The UniGene collection has been effectively used as a source of mapping candidates for the construction of a human gene map (7). In this case, the 3′ UTRs of genes and ESTs are converted to sequence-tagged sites (STSs) which are then placed on physical maps and integrated with preexisting genetic maps of the genome. The UniGene collection has also been used as a source of unique sequences for the fabrication of 'chips' for the large-scale study of gene expression (8). Access to the UniGene collection is provided through NCBI's home page on the Web.

Sequence-tagged site (STS) data

The ultimate purpose for creating high resolution physical maps of the human genome is to create a scaffold for organizing large scale sequencing (9). 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' to compare all human sequences with the contents of the STS division of GenBank; this identifies primer-binding sites on the human sequences that may be amplified in a PCR reaction. This 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 electronic PCR tool is also being made publicly available on the Web to enable any researcher with a new human sequence to relate that sequence to existing maps and HTG sequence data.

The STS division of GenBank currently contains 60 564 sequences and includes 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.

Genome survey sequence (GSS) data

The Genome Survey Sequences (GSS) division of GenBank has been the fastest growing division in the last year, having increased 5-fold to a total of 222 573 records with 103 970 022 nucleotides. GSS records represent 'random' genomic sequences, but are predominantly represented by 'BAC ends'—single reads from Bacterial Artificial Chromosomes used in a variety of genome sequencing projects, notably that of human (150 098 records), *Fugu rubripes* (32 075 records) and *A.thaliana* (26 605 records). It is expected that the human data will be used along with the STS records in tiling the BACs used for the Human Genome Project (10).

High throughput genomic (HTG) data

The high throughput genomic (HTG) sequences 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 (11). It is now clear that a great number of human sequences will continue to appear in the unfinished (HTG) division of GenBank, as they will in the corresponding finished (PRI) division. Together these two divisions should add some 2000 Mb of new genomic sequences from US-sponsored laboratories within the next three years.

Single nucleotide polymorphism (SNP) data

Single-base variations in the human genetic code called single-nucleotide polymorphisms (SNPs) promise to be helpful in large-scale association genetics studies. In collaboration with the National Human Genome Research Institute (NHGRI), NCBI has established a database (dbSNP) to serve as a central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms. One of the goals of the US Human Genome Project is to use these data to generate publicly available maps of at least 100 000 SNP markers within the next five years (12). The dbSNP database contains the experimental conditions used to detect each mutation and each mutation's observed variation for populations and individuals. A flexible file-based submission format has been developed to accommodate the typical large set of SNP submissions. Although dbSNP is not strictly a part of GenBank, appropriate links between dbSNP data and other genomic data, including GenBank, are being developed for NCBI's retrieval systems.

BUILDING THE DATABASE

The data in GenBank come from two sources: (i) authors who submit data directly to the collaborating databases, and (ii) bulk submissions from sequencing centers in the form of ESTs, STSs, GSSs or large genomic records (usually sequences from cosmids, BACs or YACs). Data are exchanged daily with DDBJ and EMBL, our collaborating databases, 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 submissions, with the majority of authors using the BankIt or Sequin programs. The nucleotide databases no longer routinely manually scan the published literature for articles containing sequences. Many journals have a policy of requiring authors with sequence data to submit data directly to the database as a condition of publication.

GenBank staff can usually assign an author an accession number within two working days of receipt, and do so at a rate of several hundred 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 checking for vector contamination, verifying proper translation of coding regions, and correct taxonomy and bibliographic citations. 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 sequences be kept confidential until the time of publication. In these cases, authors are reminded to inform the database of the publication date of the article in which the sequence is cited 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 lags in releasing data, errors or omissions using the Email address update@ncbi.nlm.nih.gov

Several large-scale sequencing projects are producing megabases of human genomic DNA sequence. NCBI works closely with sequencing centers 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., through options in BLAST to mask Alu and other types of repetitive elements). GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission using the program 'fa2htgs' and other tools (13).

BankIt

Over 65% of individual submissions are 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 free text 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.

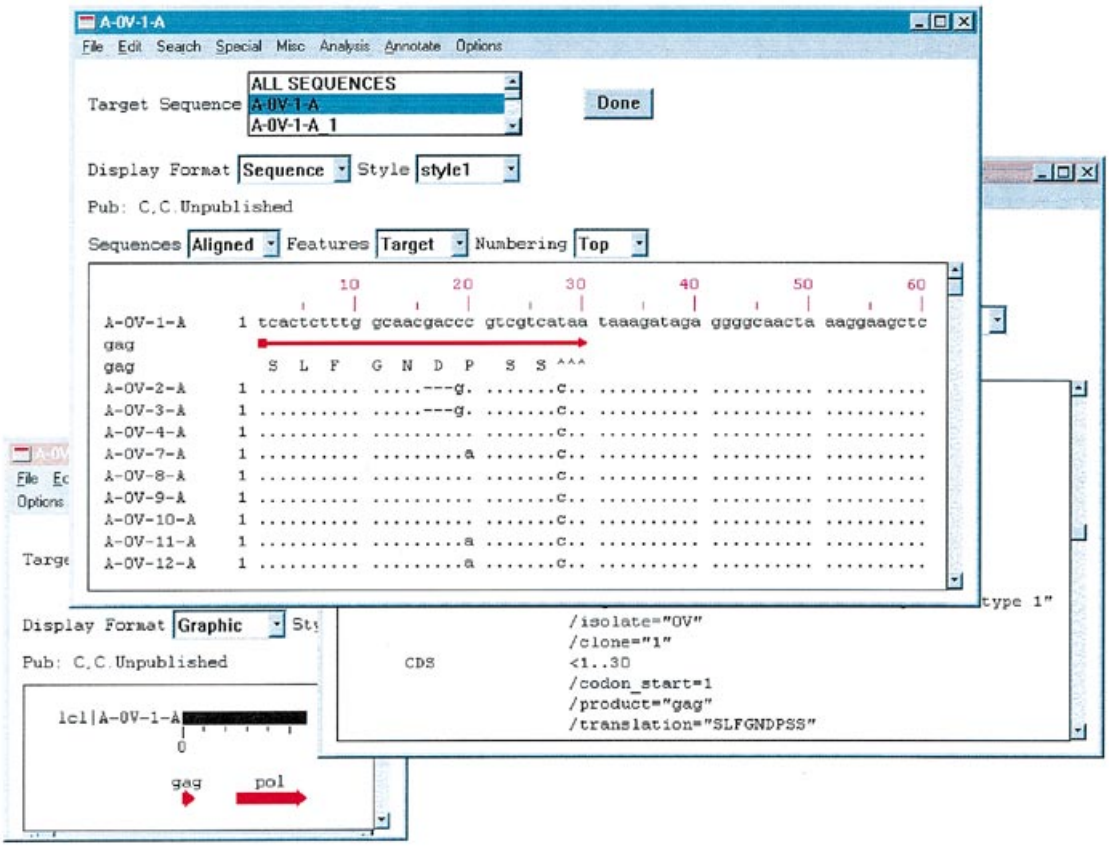

Figure 1. Aligned sequence view from Sequin for the submission of HIV1 sequences representing a population study. Sequin allows authors to prepare and submit nucleotide sequences for which they may have already generated an alignment, and also allows editing of these alignments. In addition, various views of the data are available, including a graphical view (lower left), standard GenBank text view (lower right), or a view showing features and alignments (upper center). Sequin has a number of functions to deal with population, phylogenetic and mutational studies, as well as standard records (e.g., cDNA with one CDS feature) or very large ones (e.g., *E.coli* genome with 4000 CDS features).

BankIt is the tool of choice for simple submissions, especially when only one or a small number of records is submitted (11).

Sequin

NCBI has developed a platform-independent submission program called Sequin which runs stand-alone or over the network. Sequin handles simple sequences (e.g. a cDNA), as well as long sequences and segmented entries (with which BankIt and other Web-based forms have difficulty). Sequin has convenient editing and complex annotation features and contains a number of built-in validation functions for enhanced quality assurance. It is also designed to facilitate submission of sequences from phylogenetic, population (Fig. 1) and mutation studies. Sequin's strength is in editing and updating sequence records, as well as its ability to function as a sequence analysis tool. For example, Sequin can now incorporate any analysis tool available on the Web that accepts FASTA or ASN.1 as its input format. In addition, Sequin is able to work on large records (e.g. the *Escherichia coli* genome at 5.6 Mb) and read in all of its annotations via simple tables. Versions for Macintosh, PC and Unix computers are available via anonymous FTP to 'ncbi.nlm.nih.gov' in the 'sequin' directory. Once a submission is completed, users can Email it 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 is an integrated database retrieval system that accesses DNA and protein sequence data, MEDLINE references (PubMed), genome data, the NCBI taxonomy, and protein structures from the Molecular Modeling Database, MMDB (14). The DNA and protein sequence data are integrated from a variety of sources and therefore include more sequence data than are available within GenBank alone. PubMed was developed at NCBI and allows text searching of the 9 million references in MEDLINE as well as links to the full-text of over 200 journals that are available on the Web.

Entrez provides an entry point into sequence or bibliographic records by simple Boolean queries. From a record, a user can 'point-and-click' via hypertext links to reach different information sources. 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 are based on computed similarities among the sequences or among MEDLINE

abstracts. The resulting pre-computed 'neighbors' allow very rapid access for browsing groups of related records.

Entrez is available over the Internet both through the Web and in a server/client version. The Web version, including PubMed MEDLINE searching, is used by over 60 000 different users per day. The server/client version of Entrez operates with a client program on a user's machine connected 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. The Web version has essentially the same functionality as the server/client, with the added capability of linking to full-text versions of journal articles. Both versions allow viewing of genome and related map information as well as 3D structures (14).

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 family of search programs to locate good alignments between a query sequence and database sequences. Each BLAST alignment is accompanied by an alignment score and a measure of statistical significance, called the Expectation value, for judging the quality of the alignment. The Web version of the standard BLAST 2.0 program accepts a query sequence or accession number which is pasted into an input window. The search for similarity is performed using a PAM or BLOSUM scoring matrix and results in a set of gapped alignments containing hyperlinks to database hits. The Web versions of BLAST now provide a graphical overview consisting of an array of bars, representing database hits, which are aligned to a master bar, which represents the query sequence. These bars are color-coded by expectation value and clearly show the extent and quality of the sequence similarities detected by BLAST as well as the disposition of gaps in the alignments.

Specialized versions of Web BLAST facilitate other approaches to similarity searching. Position Specific Iterated PSI-BLAST (15) initially performs a conventional BLAST search to produce alignments from which it constructs a position specific profile. Subsequent BLAST iterations use this profile in place of the query and scoring matrix to find similarities in a database. The newly implemented Pattern Hit Initiated PHI-BLAST (16) takes both a peptide query sequence and a peptide pattern, or motif, found within the query, as input. The motif specifies an obligatory match between query and database sequences about which optimal local alignments are constructed.

The default databases searched by BLAST are the non-redundant (nr) nucleotide and protein subsets of GenBank. However, specialized databases or subsets may also be searched. Standard Web BLAST allows the restriction of searches to the 'month' database, comprised of sequences submitted over the last 30 days, dbEST (all of dbEST, human only or mouse only), *E.coli* or *Saccharomyces cerevisiae* sequences, as well as several other GenBank subsets. A new specialized BLAST page allows a nucleotide query against any combination of 16 complete and 20 incomplete microbial genomes. Another recent addition, 'Blast 2 Sequences', can display the similarity between two DNA or peptide sequences by producing a dot-plot representation of the alignments it reports.

PowerBLAST (17) is a network BLAST client designed for the analysis of large contigs of genomic sequence. PowerBLAST is capable of processing 100 kb of sequence per hour to produce a set of gapped alignments which can be filtered by organism.

Other ways to access GenBank

The full GenBank release (issued every 2 months) and the daily updates (which also incorporate sequence data from other public databases) are available by anonymous FTP from 'ncbi.nlm.nih.gov'. 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 contained in 'daily-nc'. Software developers creating their own interfaces or analysis tools for GenBank data are offered the NCBI ToolKit to assist in developing specialized applications. The ToolKit software can be found in the directory 'toolbox/ncbi_tools'.

Users with access to electronic mail can search GenBank and several other databases by accession number or Boolean combinations of text words. The QUERY server (query@ncbi.nlm.nih.gov) performs text-based searches of the integrated Entrez databases. It allows access not only to sets of sequence or MEDLINE records, but also to the neighbored data. Various output formats, such as FASTA for sequence data, are available. BLAST sequence similarity searches can be performed by Email through the address: blast@ncbi.nlm.nih.gov. Documentation can be obtained by sending the word 'help' in the body of an Email message to the addresses above. The flat file version of GenBank is no longer available on CD-ROM due to declining demand and the large number of disks needed to contain a single release.

GenBank Fellows

The GenBank Fellowship Program is an 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 motivation to apply computational tools to the organization of electronic data in molecular and structural biology, genetics and phylogeny. 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. 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 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 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., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F.F. (1998) *Nucleic Acids Res*., **26**, 1–7.
2 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res*., **26**, 8–15.
3 Yoshio,T., Fukami-Kobayashi,K., Miyazaki,S., Sugawara,H. and Gojobori,T. (1998) *Nucleic Acids Res*., **26**, 16–20.
4 Aaronson,J.S., Eckman,B., Blevins,R.A., Borkowski,J.A., Myerson,J., Imran,S. and Elliston,K.O. (1996) *Genome Res*., **6**, 829–845.
5 Hillier,L., Lennon,G., Becker,M., Bonaldo,M., Chiapelli,B., Chissoe,S., Dietrich,N., DuBuque,T., Favello,A., Gish,W. *et al.* (1996) *Genome Res*., **6**, 807–828.
6 Schuler,G.D. (1997) *J. Mol. Med*., **75**, 694–698.
7 Deloukas,P., Schuler,G.D., Gyapay,G., Beasley,E.M., Soderlund,C., Rodriguez-Tome,P., Hui,L., Matise,T.C., McKusick,K.B., Beckmann,S. *et al.* (1998) *Science*, **282**, 744–746.
8 Ermolaeva,O., Rastogi,M., Pruitt,K.D., Schuler,G.D., Bittner,M.L., Chen,Y., Simon,R., Meltzer,P., Trent,J.M. and Boguski,M.S. (1998) *Nature Genet*., **20**, 19–23.
9 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) *Science*, **270**, 1945–1954.
10 Smith,M.W., Holmsen,A.L., Wei,Y.H., Peterson,M. and Evans,G.A. (1994) *Nature Genet*., **7**, 40–47.
11 Kans,J.A. and Ouellette,B.F.F. (1998) 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, NY, pp. 319–353.
12 Collins,F.S., Patrinos,A., Jordan,E., Chakravarti,A., Gesteland,R., Walters,L. and the members of the DOE and NIH planning groups (1998) *Science*, **282**, 682–689.
13 Ouellette,B.F.F. and Boguski,M.S. (1997) *Genome Res*., **7**, 952–957.
14 Marchler-Bauer,A., Addess,K.J., Chappey,C., Geer,L., Madej,T., Matsuo,Y., Wang,Y. and Bryant,S.H. (1999) *Nucleic Acids Res*., **27**, 240–243.
15 Altschul,S.F, Madden,T.L., Schaffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res*., **25**, 3389–3402.
16 Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V., Altschul,S.F. (1998) *Nucleic Acids Res*., **26**, 3986–3991.
17 Zhang,J. and Madden,T.L. (1997) *Genome Res*., **7**, 649–656.