GenBank

Dennis A. Benson*, Mark S. Boguski, David J. Lipman and James Ostell

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

Received October 4, 1996; Accepted October 8, 1996

ABSTRACT

The GenBank sequence database incorporates DNA sequences from all available public sources, primarily through the direct submission of sequence data from authors and from large-scale sequencing projects. Data exchange with the EMBL Data Library and the DNA Data Bank of Japan helps ensure comprehensive coverage. GenBank continues to focus on quality control and annotation while expanding data coverage and retrieval services. An integrated retrieval system, known as *Entrez*, incorporates data from the major DNA and protein sequence databases, along with genome maps and protein structure information. MEDLINE abstracts from published articles describing the sequences are also included as an additional source of biological annotation. Sequence similarity searching is offered through the BLAST family of programs. All of NCBI's services are offered through the World Wide Web. In addition, there are specialized server/client versions as well as FTP and e-mail server access.

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 Congress in 1988 to develop information systems, such as GenBank, 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 and secondarily from scanning the journal literature. A major source of data are bulk submissions of EST and other high-throughput data from sequencing centers. The data are supplemented by sequences from other public databases. Through an international collaboration with the EMBL Data Library in the UK and the DNA Databank of Japan (DDBJ), data are exchanged daily to ensure that all three sites maintain comprehensive sets of sequence information. The data are made available at no cost through the Internet, either by downloading database files or by text and sequence similarity search services.

ORGANIZATION OF THE DATABASE

GenBank has experienced another year of unprecedented growth. Over the past 12 months 420 000 new sequences have been added. As of Release 96 in August, 1996, GenBank contained 602 072 354 nucleotide bases from 920 588 different sequences. Notably, 1996 is the year in which the complete genome of a eukaryotic organism, the yeast *Saccharomyces cerevisiae* was completed and added to GenBank (2). The complete genomic sequences of an archeon, *Methanococcus jannaschii* (3) also entered the database this year (see Genomes division below). Historically, the database had been doubling in size about every 18 months, but that rate has rapidly accelerated due to the enormous growth in data from expressed sequence tags (ESTs). Over 65% of the sequences in the current release are ESTs and most of the growth in terms of sequence records over the past 2 years has come from the collaborative project between the Merck & Co. and Washington University (4,5). This growth is expected to continue because Washington University and the Howard Hughes Medical Institute are pursuing a mouse EST project of the same scale as human EST sequencing. Furthermore, the Human Genome Project has entered its pilot scale-up sequencing phase and 100 million nucleotides of human genomic DNA sequence data are expected during the next 2 years. Sequence records from several of the six US centers funded to do this work are already beginning to appear in GenBank.

Sequence-based taxonomy

Of the nearly one million sequences in GenBank, human entries predominate, constituting 59% of the total, but >16 000 species are represented and ~10 new organisms enter GenBank each day. Database sequences are processed, and can be queried, using a consistent and 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 via NCBI's home page.

After *Homo sapiens*, the top species in GenBank, in terms of the number of bases, include *Caenorhabditis elegans*, *Mus musculus*,

* To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: dab@ncbi.nlm.nih.gov

Saccharomyces cerevisiae, and *Arabidopsis thaliana*. For *Caenorhabditis* and *Saccharomyces* this is due to the impact of genomic sequencing projects; for *Mus* and *Arabidopsis*, sequence abundance currently reflects large-scale EST sequencing efforts.

Records and 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 included in the feature table. Bibliographic references are included along with links to the MEDLINE® unique identifiers for all published sequences.

The files in the GenBank distribution have traditionally been divided into 'divisions' which roughly correspond to taxonomic divisions, e.g. bacteria, viruses, primates, rodents, etc. In the past year a division has been added for Genome Survey Sequences, e.g. 'single pass' reads from cosmid/BAC/YAC ends, exon-trapped genomic sequences, and Alu PCR sequences. A new division, for High-Throughput Genomic (HTG) sequences, was added with Release 97 in October, 1996. The HTG division is designed for the efficient processing and distribution of long sequences expected to be primarily human and including computer-generated annotation.

Integrated database and sequence identifiers

In order to produce the GenBank database, NCBI maintains internally an Integrated Database, ID, to track and index records from the multiple sources of sequence data. These sources include submissions from EMBL, DDBJ, Genome Sequence Database (GSDB), and patents, plus amino acid sequences from PIR, SWISS-PROT, Protein Research Foundation (PRF) and the Protein Data Bank (PDB). ID represents the most current view that each data source has of its sequence data, and allows NCBI to assign stable identifiers. Such identifiers for nucleotide sequences are found in the 'NID' field of a GenBank record, directly following the ACCESSION field. Identifiers for encoded protein sequences are found in the FEATURES table under CDS and are labeled 'PID'. Through this approach, sequence information from a wide variety of sources can have a uniform identification system. These identifiers are stable and therefore help identify sequences which have changed.

Genomes division

A separate section of the database has been created for the placement of chromosome- or genome-length sequences. This section, the Genomes division, arose from the need to store sequences >350 000 bases, the upper limit for single database entries. The entries in this section are organized with respect to a 'reference sequence'. A reference sequence can be the complete sequence of a genome as contained in a single GenBank record, or a specific sequence that has been accepted by researchers in that particular field (e.g. HIV research) as a reference sequence. In some cases, the selection of a reference sequence is arbitrary. With respect to the reference, all other sequences in GenBank from that organism or organelle are aligned and the alignments are presented under the reference sequence in the Genomes division. This concept of a reference sequence is critical for human genome data, because the complete human sequence will actually be a composite of the sequences of several individuals (6).

In some cases, for example yeast and some bacterial genomes, the complete sequences of the chromosomes are known but exist as many separate records in GenBank. In these cases a virtual reference sequence is created that contains instructions on how to assemble the GenBank records to make the complete chromosome. The NCBI software tools can dynamically project the features annotated on the GenBank records to the coordinate system of the reference sequence, including entire chromosomes, for any region of interest. This provides the scientist with both a large-scale view of genomic sequence data but also a view of smaller regions around genes in the traditional GenBank record format. The breakdown of a genome sequence into separate records makes for a convenient unit of analysis and annotation updates because only a relatively small GenBank record need be changed and distributed. Nevertheless, any new information will appear automatically in the large-scale view as well.

The most complicated case is for chromosomes which are not completely sequenced such as eukaryotic chromosomes other than yeast. As one example, NCBI has obtained, in collaboration with a number of established mapping centers, genetic, physical and cytogenetic maps of human chromosomes and produced cross-referenced sets of aligned maps. Using STS markers on these maps, and the UniGene set of non-redundant human genes (7), aligned groups of sequences can be placed onto the framework of the whole chromosome. These sequences appear as 'islands' on the sequence map (Figs 1 and 2) and will eventually coalesce into contiguous sequence as the Human Genome Project is completed. This approach provides an important organizing principle for making optimal use of traditional annotated human sequence records in GenBank, as well as the specialized properties of EST, STS and HTG sequences.

EST data

ESTs or 'expressed sequence tags' are still the most rapidly expanding source of new sequence records and genes. Last year there were 328 905 sequences in the EST Division of GenBank (dbEST). Over the past year there has been a doubling in the number of ESTs for a current total of 658 698 sequences in dbEST representing 63 different organisms. The top five organisms include: 476 148 human ESTs (72%); 74 312 mouse ESTs (11%); 30 196 nematode ESTs (4.6%); 28 951 *Arabidopsis* ESTs (4.4%); and 11 316 rice ESTs (1.7%). ESTs continue to provide the major source of new gene discoveries and NCBI has serviced more than half a million queries (BLAST searches, e-mail retrievals, WWW accesses and anonymous ftp downloads) for dbEST data to date.

Because of the nature of EST survey sequencing, there is much redundancy in dbEST. For example, there are >1300 sequences representing serum albumin mRNA. In order to organize the data in a more useful fashion, NCBI has created the UniGene collection of unique human genes (7). Briefly, UniGene starts with human entries in the primate (PRI) division of GenBank, combines these with human ESTs and creates clusters of sequences that share virtually identical 3′ untranslated regions (3′UTRs). In this manner, the nearly 500 000 human ESTs in dbEST have been reduced 10-fold to ~50 000 sequence clusters each of which may be considered as representing a single human

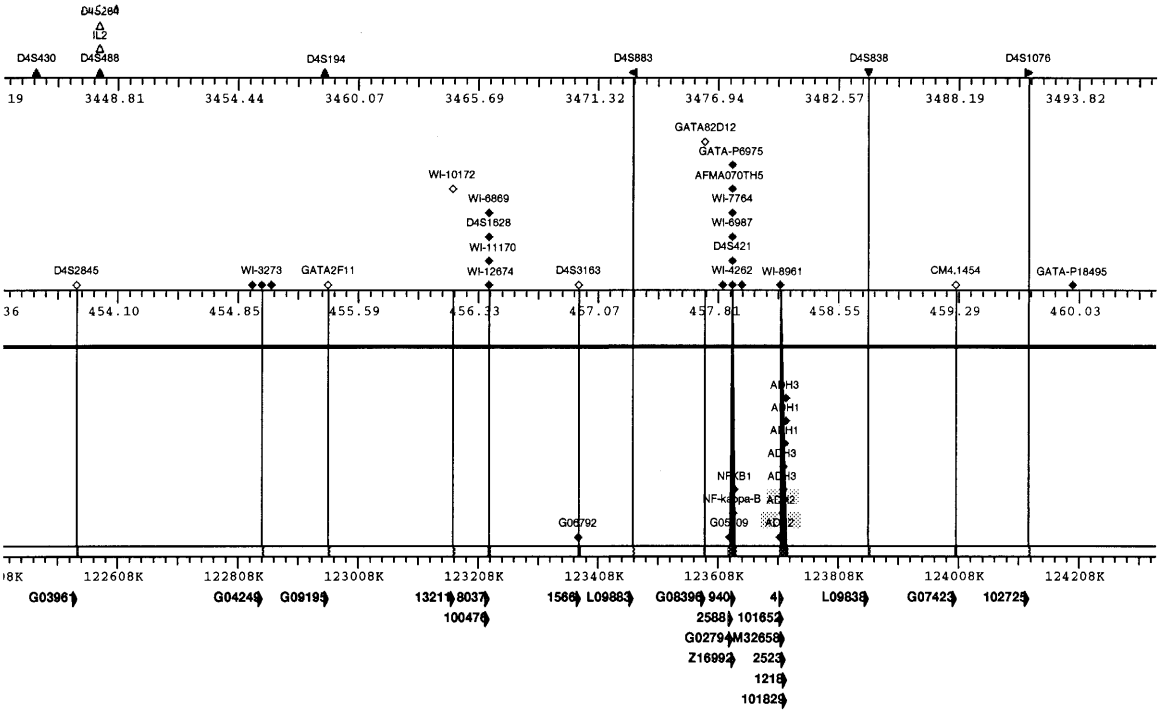

Figure 1. Detail of a region of Human Chromosome 4 around the ADH2 gene. Top is the Stanford radiation hybrid map, next the Whitehead Institute radiation hybrid/YAC content map, and third the NCBI sequence map. Généthon, CHLC and GDB maps are also present for this chromosome in *Entrez* but could not be shown in this figure. Vertical lines connect markers on different maps. Small arrows at bottom of figure show islands of aligned human sequence placed on the chromosome.

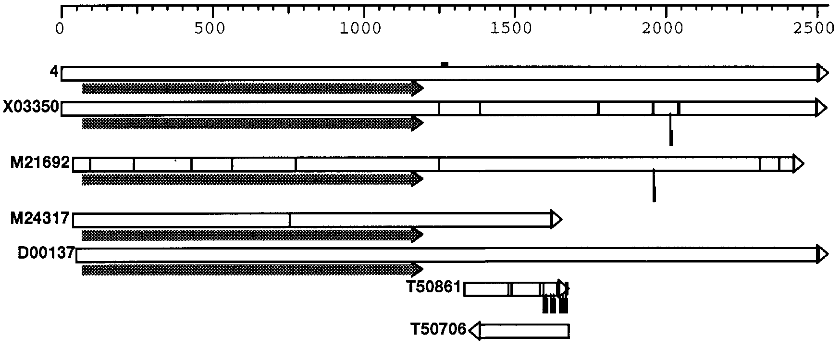

Figure 2. Detail of ADH2 mRNA cluster from previous detailed figure. Open arrows indicate sequence records in GenBank. Solid arrows show CDS features annotated on those records. Vertical lines within open arrow represent mismatches and deletions relative to reference sequence (top). Lines extending below open arrows are insertions. Two short records at bottom are ESTs. Alignment can also be seen as text alignment with single base resolution and translated CDS shown in register in alignment (not shown).

gene. Access to the UniGene collection is provided through NCBI's home page on the WWW. The UniGene collection has been effectively utilized 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 STSs which are then placed on physical maps and integrated with preexisting genetic maps of the genome.

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 (8). Physical maps based on 'sequence tagged site' (STS) landmarks are used to develop so-called 'sequence-ready' clones consisting of overlapping cosmids or

BACs. As the high-throughput sequence data derived from these clones are submitted to GenBank, STSs become crucial reference points for organizing, presenting and searching the data. NCBI uses a process of 'electronic PCR' in which all human sequences are compared with the contents of the STS division of GenBank (dbSTS) to see if the former sequences contain primer binding sites in the correct orientation and at an appropriate spacing that the expected product would be amplified in a PCR reaction. This tool permits us to assign an initial location on the map for sequence data and to relate previous GenBank entries to the new reference sequence. The electronic PCR tool developed for this task is also being made publicly available on the WWW 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 37 261 STS 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 and PCR reaction conditions.

BUILDING THE DATABASE

The data in GenBank come from three sources: (i) authors who submit data directly to the collaborating databases; (ii) bulk submissions from sequencing centers; and (iii) annotators at NCBI who extract the information from relevant journals. Data are exchanged daily with collaborating databases so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct submission

The majority of entries continue to enter the database through direct author submission. Many journals have the policy of requiring authors with sequence data to submit data directly to the database as a condition of publication. Even for those journals without a mandatory submission policy, author submission has the positive benefits of acquiring annotation information directly from the authors and reducing the time-lag between publication and the appearance of the sequence in the database.

Over the past year several large-scale sequencing projects have begun scaling up to the goal of producing hundreds of megabases of human genomic DNA sequence over the next 1–3 years. NCBI is working 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. 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.

Over 80% of individual submissions are now received through a Web-based data submission tool, BankIt. With BankIt, authors enter sequence information directly into a form, edit as necessary, and add biological annotation (e.g. coding regions, mRNA features). Free-form text boxes provide the option of using your own words to describe the sequence, without having to learn formatting rules or use restricted vocabularies. BankIt creates a draft record in GenBank flat file format for the user to review and revise.

GenBank has developed a platform-independent submission program called Sequin, which runs stand-alone or over the network. The advantages of Sequin over the previous submission program, Authorin, include the capacity to handle long sequences and segmented entries, easier editing and updating, and complex annotation capabilities. In addition, Sequin contains a number of built-in validation functions for enhanced quality assurance. Versions for Macintoshes, PCs and Unix are available at no charge. It can be obtained by anonymous FTP to 'ncbi.nlm.nih.gov' in the 'pub/sequin' directory. Once a submission is completed, users can e-mail it to the address: gb-sub@ncbi.nlm.nih.gov

GenBank staff can usually assign an author an accession number within 1 working day of receipt. The accession number serves as confirmation that the sequence has been submitted and allows readers of the article to retrieve the relevant data. All direct submissions receive a systematic quality assurance review including screening against GenBank to identify full or partial matches, checking for vector sequence and verifying proper translation of coding regions. A draft of the GenBank record is passed back to the author for review before entering the database. Authors have the right to request that their sequences be kept confidential until the time of publication. In these cases, authors are reminded to inform the database of the publication date in order to have a timely release of the data. Although only the submitting scientist is permitted to modify sequence data or annotations, all users are encouraged to inform the database of possible errors or omissions using the e-mail address: update@ncbi.nlm.nih.gov

Molecular Modeling Database (MMDB) and Cn3D

NCBI has begun distributing a database of macromolecular three-dimensional structure information, specifically aimed at molecular modeling research. MMDB is based upon the data in the Brookhaven PDB. By reorganizing and validating the PDB data, MMDB provides explicit descriptions of a biopolymer's spatial structure, its chemical organization, and the linkage between the two. With MMDB, there is a clear cross-referencing between the three-dimensional structure and the chemistry of a macromolecule. Explicit linkages have been made to the sequence entries within the ID database so that, with the appropriate graphics software, users are able to view the three-dimensional structure of proteins identified via text or similarity searching of the sequence database. During the past year, NCBI added a new three-dimensional structure viewer called Cn3D ('See in 3D') to the *Entrez* retrieval system.

RETRIEVING GENBANK DATA

The *ENTREZ* system

Entrez is an integrated database retrieval system which accesses DNA and protein sequence data, related MEDLINE references, genome data from the GenBank genomes division, the NCBI taxonomy and three-dimensional structures from MMDB (9). The DNA and protein sequence data are integrated from a variety of sources via the ID database previously described. The MEDLINE references are approximately one million citations indexed under the NLM's Medical Subject Heading (MeSH), 'genetics'.

The linkages among data sources are shown in Figure 3. The DNA sequence, protein sequence, MEDLINE, genome and three-dimensional structure data are linked to provide easy traversal among the data sets. *Entrez* provides an entry point into

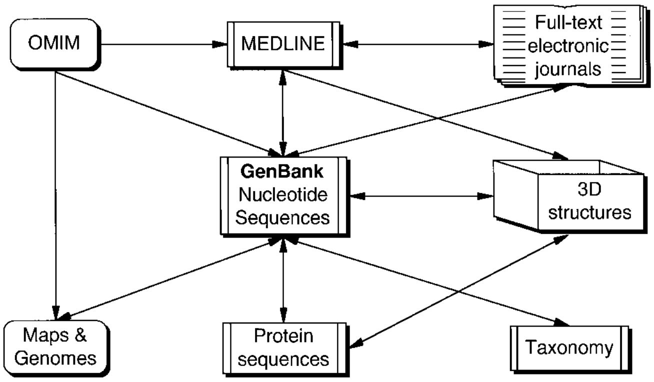

Figure 3. Data sources and interconnecting links among the various information components which comprise the *Entrez* integrated retrieval system.

sequence or bibliographic records by simple Boolean queries. From a record, hypertext links may be used to navigate through the information space using a point-and-click interface. Some of the links are simple cross-references, for example, between a sequence and the abstract of the paper in which the sequence was reported, or between a protein sequence and its corresponding DNA sequence. Among these cross-references are external links to the full-text versions of articles when these are available from publishers' WWW sites. Other links are based on computed similarities among the sequences or among the textual documents. The pre-computed 'neighbors' allow very rapid access for browsing groups of related records.

Entrez is available over the Internet through the Web and in a server/client version. A CD-ROM version has been discontinued because of the size of the database and the inconvenience of multiple CD-ROM discs. The server/client version of *Entrez* operates with a client program on a user's machine over the Internet connected 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, plus the added capability of linking to full-text versions of journal articles. Viewers are available in both versions to visualize genome and related map information as well as three-dimensional structures. Since the three-dimensional structure information has been linked to the set of protein sequences, users can easily determine a set of sequence neighbors for a given sequence and then locate and visualize structures for members of the neighbor set.

BLAST sequence similarity searching

One of the most frequent uses of GenBank is sequence similarity searching. NCBI offers the BLAST family of search programs to perform fast searching with rigorous statistical methods for judging the significance of matches. WWW access to BLAST currently offers two interfaces, a 'Basic' version with default search parameters and an 'Advanced' option which allows customization of the parameters. A new graphical version called PowerBLAST, designed for rapid analysis and annotation of large contigs of genomic sequence data (J. Zhang, personal communication), is available as a server/client application on NCBI's ftp site and it is being modified for use via the BLAST WWW pages. PowerBLAST is particularly useful for assessing multiple EST matches with a query sequence.

The primary databases for sequence similarity searching are non-redundant ('nr') versions of nucleotide and protein sequences. ESTs are available for separate searching, as is a set of those sequences added during the previous 30 days (designated 'month' on the BLAST web pages). Frequent users may find the server/client version of BLAST more convenient; clients are available for several platforms. BLAST client software also incorporates advanced features such as one-to-many alignments of the query sequence with all the matching sequences (as opposed to the standard results that show the query sequence aligned individually against each matching sequence). Another feature of the client software is the ability to generate organism-specific output, for example, searches restricted to human sequences. Information on BLAST client software can be obtained by e-mail to the address: blast-help@ncbi.nlm.nih.gov

Anonymous FTP

Users on the Internet can use the file transfer protocol (FTP) program to download the entire GenBank release or the daily updates (which also incorporate sequence data from other public databases). Files of the full release and daily updates of the GenBank database are available for anonymous FTP from: ncbi.nlm.nih.gov. The full release in flat-file format is available as compressed files in the directory, 'genbank'. A cumulative update file is contained in the subdirectory, 'daily', and a non-cumulative set of updates is in the subdirectory, 'daily-nc'. ASN.1 formatted data is in the directory, 'ncbi-asn1'. Software developers creating their own interfaces or analysis tools for GenBank data are offered the NCBI toolkit to assist in developing specialized applications. Software can be found in the directory: toolbox/ncbi_tools

E-mail servers

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. The RETRIEVE e-mail server is still operating (retrieve@ncbi.nlm.nih.gov), but will eventually be superseded by QUERY, which also supports the RETRIEVE query syntax. BLAST sequence similarity searches can be performed by e-mail through the address: blast@ncbi.nlm.nih.gov. Documentation can be obtained for each of the servers by sending the word 'help' in the body of an e-mail message to the addresses above.

CD-ROM

The flat file version of GenBank is available on CD-ROM through a subscription service with the Government Printing Office (Tel: +1 202 512 1800; Fax: +1 202 512 2233). Order forms are also included in each issue of *NCBI News*, a free subscription to which may be obtained by contacting NCBI. A new release of the database appears every 2 months. Each release is a full release incorporating all previous GenBank data supplemented by new data from direct submissions, NCBI journal scanning, patents and the other sequence databases. Conceptual translations of coding regions appear in feature tables. The release contains the standard index files and is organized into divisions. No retrieval software is provided. The distribution currently requires six CD-ROMs. (The CDROM version of *Entrez* has been discontinued because of the size of the distribution files.)

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. Currently five GenBank fellows are in the program and applications are reviewed on a continuing cycle.

MAILING ADDRESS

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

ELECTRONIC ADDRESSES

http://www.ncbi.nlm.nih.gov/ (NCBI 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., Lipman, D. J. and Ostell, J. (1996) *Nucleic Acids Res.*, **24**, 1–5.

2 Williams, N. (1996) *Science*, **272**, 481.

3 Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D. *et al.* (1996) *Science*, **273**, 1058–1073.

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 Marshall, E. (1996) *Science*, **273**, 1788–1789.

7 Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tomé, P., Aggarwal, A., Bajorek, E. *et al.* (1996) *Science*, **274**, 540–546.

8 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.

9 Schuler, G. D., Epstein, J. A., Ohkawa, H. and Kans, J. A. (1996) *Methods Enzymol.*, **266**, 141–162.