 Nucleic Acids Research, 1998, Vol. 26, No. 1 1–7

GenBank

Dennis A. Benson*, Mark S. Boguski, David J. Lipman, James Ostell and B. F. Francis Ouellette

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

Received October 6, 1997; Accepted October 8, 1997

ABSTRACT

The GenBank® sequence database (http://www.ncbi.nlm.nih.gov/) 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 (WWW) or Sequin programs to send their 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® 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 series of database search programs. In addition to FTP, e-mail and server/client versions of *Entrez* and BLAST, NCBI offers a wide range of World Wide Web retrieval and analysis services of interest to biologists.

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. Another major source of data is bulk submission 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 continues to grow at an exponential rate. Over the past 12 months 690 000 new sequences have been added. As of Release 102 in August, 1997, GenBank contained over 1 billion nucleotide bases from 1.6 million different sequences. Complete genomes represent a growing portion of the database, with six complete genomes added in 1997 compared to two in 1996. Currently there are at least 32 complete 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 to the enormous growth in data from expressed sequence tags (ESTs). Over 72% of the sequences in the current GenBank 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 Merck & Co. and Washington University (2,3). Human EST sequencing continues and is being supplemented by a mouse EST project supported by the Howard Hughes Medical Institute at Washington University. Also, under the auspices of the Human Genome Project, several sequencing centers have been funded and over the next few years each center may be producing in the order of 100 000 bases of genomic sequence daily.

Sequence-based taxonomy

Over 30 000 different species are represented in GenBank and new species are added at the rate of 600 per month. Entries from humans constitute 57% of the total sequences (49% of all sequences are human ESTs). 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 on NCBI's home page. After *Homo sapiens*, the top

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

species in GenBank in terms of the number of bases include *Mus musculus*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*.

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; the most recent was added last year for high-throughput genomic sequences. For convenience in file transfer, the larger divisions, e.g., EST and primate, are divided into multiple files.

Sequence identifiers

To produce the GenBank database, NCBI tracks and indexes records from multiple sources of sequence data: DNA sequences from EMBL, DDBJ, Genome Sequence Database (GSDB) and the US Patent Office, plus amino acid sequences from PIR, SWISS-PROT, Protein Research Foundation (PRF) and the Protein Data Bank (PDB). In order to identify specific sequences from all the different sources, as well as changes in those sequences that may occur, NCBI assigns a stable identifier, termed a 'gi' number to each sequence. A new 'gi' number is assigned to every sequence version. These identifiers appear in the 'NID' field of a GenBank record, immediately following the ACCESSION field. The ACCESSION number, in contrast, is associated with each GenBank entry and does not change, even when the sequence or annotation changes.

There is now an agreement between the collaborative DNA sequence databases to introduce 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 referred to as '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:

```
ACCESSION  Z000001
      NID  g987654321
  VERSION  AZ000001.1      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 be introduced over the coming year in conjunction with EMBL and DDBJ. Protein sequences will have identification numbers (in the format of three letters followed by five digits, e.g., ABC78912) that does 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:

```
/protein_id='ABC78912.1'
```

Protein translations currently receive their own unique 'gi' number and that appears as a qualifier on the CDS feature.

```
/db_xref='PID:g1234567'
```

Eventually, after a transition period, this form will be phased out since the new 'protein_id' complete with the version number represents both a stable identifier and a means to identify changes in the sequence.

Genomes in GenBank

The retrieval, representation and querying of data from completely sequenced genomes is becoming a key function for sequence databases. Several hundred complete genomes already exist in GenBank ranging from viruses and organelles to free-living organisms with representatives from archaea, eubacteria and eukaryotes. Prominent examples are the recent additions of the 1.7 megabases of *Helicobacter pylori* and the 4.6 megabases of *Escherichia coli*. It is a challenge to organize and present the wealth of information offered by even the simpler genomes. On submission, completely sequenced genomes are split into overlapping 10-kilobase records for inclusion into the distributed release of GenBank. However, it is clear that a higher-level view is also necessary and such views are now offered in *Entrez*. The genomic-level view (Fig. 1) shows the spatial organization of genes and provides immediate access to the underlying individual sequence records (Fig. 2).

A separate section of the database (Genomes Division) contains full-length genome sequences (4). This special division was created because the GenBank database has followed the convention that single entries can be no larger than 350 kilobases. In contrast, the Genomes Division contains the full-length sequences in several formats—FASTA, GenBank flat file and ASN.1. Entries are available for *E.coli*, *Haemophilus influenzae*, *H.pylori*, *Mycoplasma genitalium*, *Methanococcus jannaschii*, *Mycoplasma pneumoniae*, *Rhizobium sp. NGR234* and *Saccharomyces cerevisiae*. For these genomes, where the complete sequence is known but exists as many separate records in GenBank, 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 individual GenBank records to the coordinate system of the whole chromosome for any region of interest. This provides the scientist with both a view of the chromosome as a whole and of smaller regions around genes in the traditional GenBank record format.

Representing incomplete genome data, such as eukaryotic chromosomes other than yeast, is more complicated. 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 (5), aligned groups of sequences can be placed onto the framework of the whole

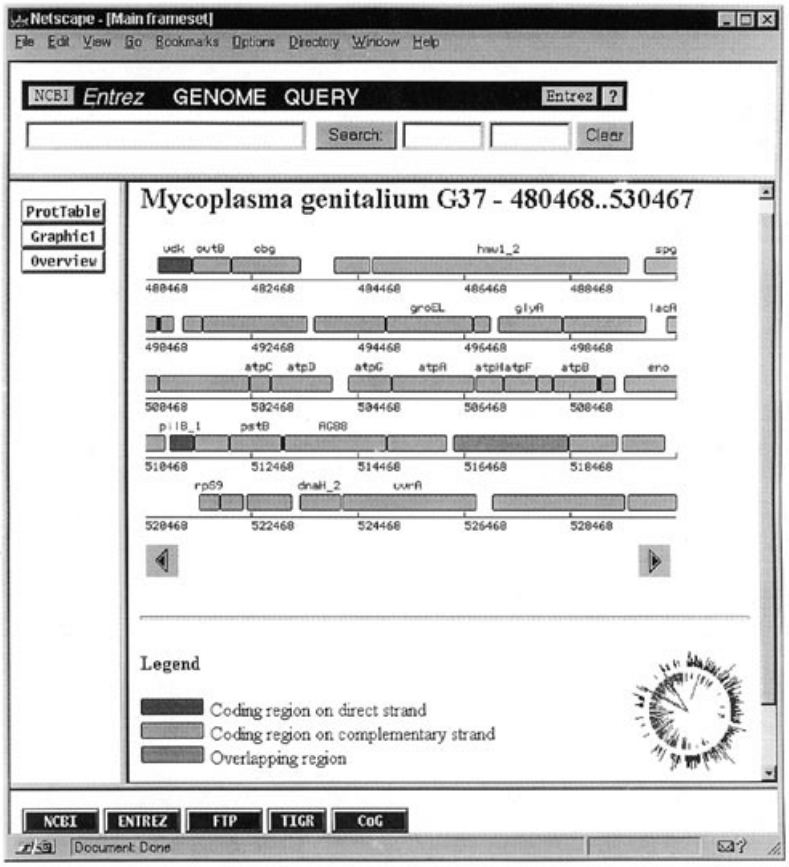

Figure 1. Graphical view of a segment of 50 000 bp from the complete genome of *M.genitalium*. This WWW presentation shows a compact view of all the CDSs with the gene name labels. Clicking on one of these will generate an additional browser page with the corresponding GenPept formatted record. The ProtTable button presents the same genomic region in a tabular format, presenting access to GenBank records as well as FASTA sequences allowing transparent migration to the blast server.

chromosome. These sequences appear as 'islands' on the sequence map and will eventually coalesce into contiguous sequence on completion of the Human Genome Project. 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 continue to be the major source of new sequence records and genes. Last year there were 658 698 sequences in the EST division of GenBank (dbEST). Over the past year the number of ESTs has nearly doubled to the current total of 1 247 603 sequences in dbEST representing over 100 different organisms. The top five organisms include: 819 082 human (65%); 226 999 mouse (18%); 70 628 nematode (5%); 32 671 *Arabidopsis* ESTs (3%); and 19 681 rice ESTs (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 within dbEST. ESTs continue to provide the major source of new gene discoveries and NCBI has serviced more than a million queries (BLAST searches, e-mail retrievals, WWW 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 (5). UniGene starts with human entries in the primate 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 800 000 human ESTs in dbEST have been reduced 20-fold to ~41 000 sequence clusters, each of which may be considered as representing a single human gene. Access to the UniGene collection is provided through NCBI's home page on the WWW. 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 pre-existing 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 (6). 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

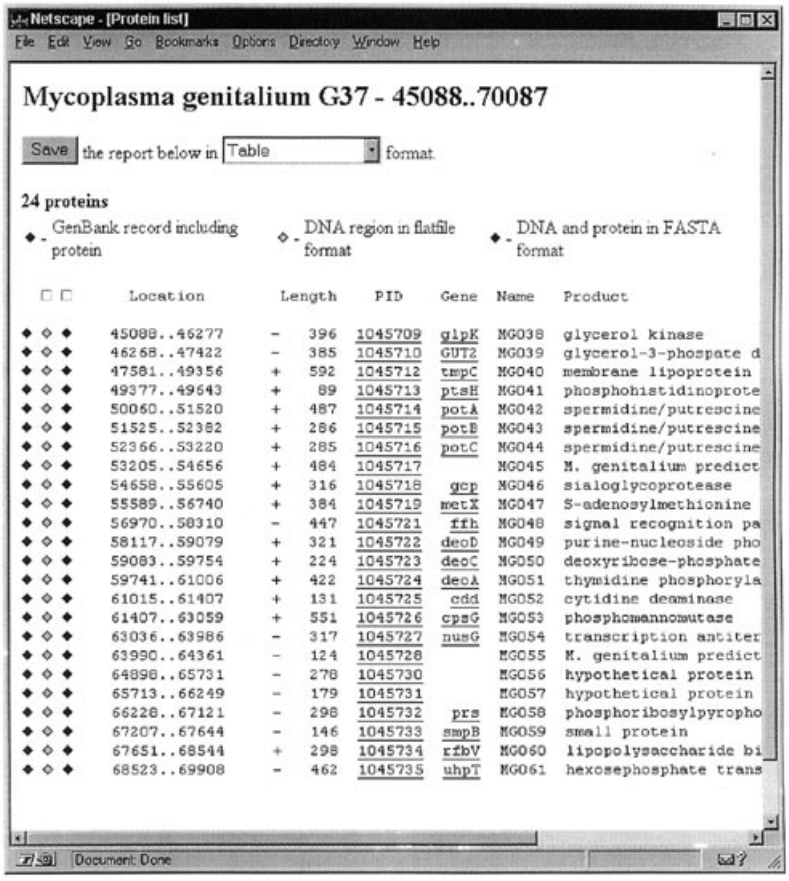

Figure 2. Detailed listing of genes corresponding to the segment shown in Figure 1. Each record is available in GenBank or FASTA format.

searching the data. NCBI uses 'electronic PCR' to compare all human sequences with the contents of the STS division of GenBank (dbSTS); this identifies primer-binding sites on the human sequences that may 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 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 51 292 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 two sources: (i) authors who submit data directly to the collaborating databases, and (ii) bulk submissions from sequencing centers in the form of ESTs or large genomic records (usually sequences from cosmids, BACs or YACs). 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 records enter the database through direct author submission using the BankIt or Sequin programs. Many journals have a 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.

Several large-scale sequencing projects have begun to produce hundreds of 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., 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 (4).

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,

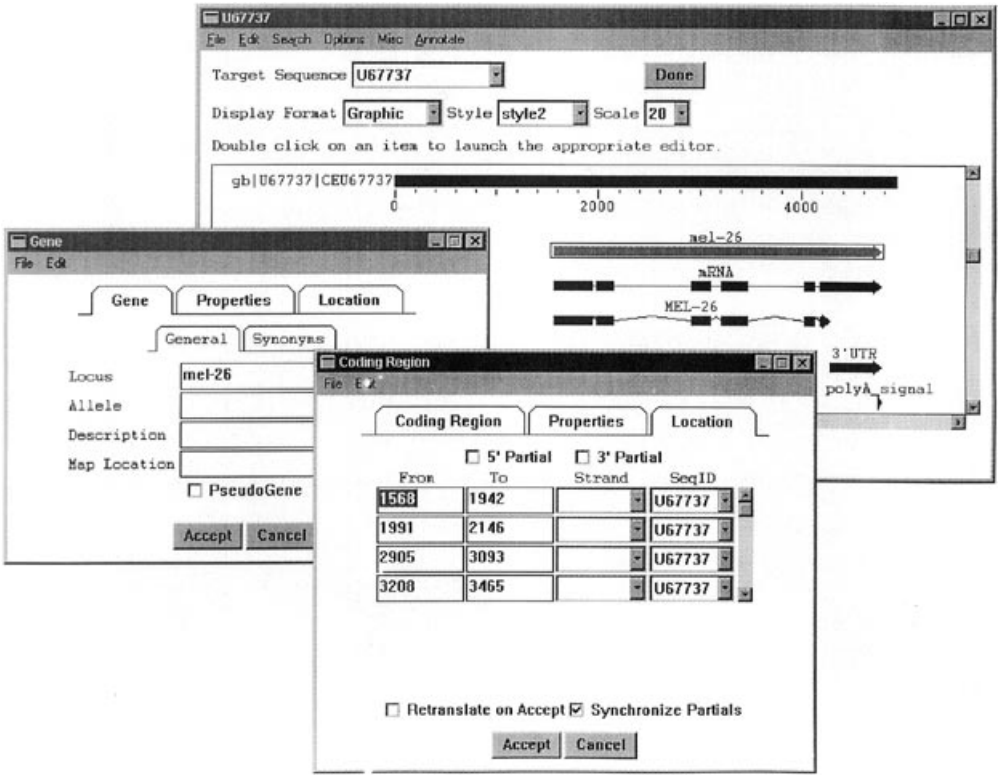

Figure 3. Graphical view from a GenBank record update using Sequin. Sequin presents alternative views, graphical in this case, all of which allow an easy way to start up the editor necessary for the displayed items by double clicking. In this case the gene is highlighted, and the gene editor is active. Also present is the CDS editor, showing a spreadsheet-like table allowing the editing of the positions of the various exons.

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.

Sequin

GenBank has developed a platform-independent submission program called Sequin which runs stand-alone or over the network (Fig. 3). Sequin handles long sequences and segmented entries, has convenient editing and complex annotation features and contains a number of built-in validation functions for enhanced quality assurance. Sequin's strength is the way in which it can edit and update sequence records. Newer versions will enhance its ability to function as a sequence analysis tool. For example, PowerBLAST (7) is now incorporated in Sequin. 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'. Additional information about Sequin can be found from the NCBI home page.

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 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 errors or omissions using the e-mail address: update@ncbi.nlm.nih.gov.

3D structure data in *Entrez*: MMDB, VAST and Cn3D

The *Entrez* retrieval system also contains all publically available information on the three-dimensional structures of proteins and nucleic acids. Structural data are taken from the Brookhaven Protein Data Bank, following a once-per-month electronic update cycle. These are subjected to certain automatic validation checks and uniform, machine-derived annotation on secondary structure and domain organization is added. Cross references to MEDLINE and to protein and nucleic acid sequences are also added. This information is cast in a 'computer friendly' form, to speed network transmission and facilitate visualization with local 3D browsing software. The collection is referred to as MMDB, a Molecular Modeling Database (8). NCBI also distributes with

MMDB a dedicated 3-dimensional viewer, CN3D (for 'see in three dimensions'), that functions as a helper application to WWW *Entrez* (9).

A new initiative undertaken this year has been to add to MMDB a complete taxonomic assignment for the proteins and nucleic acid structures it contains. The assignments are based on scanning the free-text descriptions of the molecular 'source' by a program, 'PDBeast', comparing to the NCBI taxonomy and displaying matches to a human-expert operator for validation. These assignments allow taxonomy-based queries of the structural database. An *Entrez* user may now easily select among structural 'neighbors', for example, those coming from a particular lineage, such as prokaryotes.

Another addition to *Entrez* is a complete set of structural neighbors. All structures in MMDB (and individual domains from these structures) have been compared to one another using the VAST algorithm (10), a method that searches for similar substructures shared between those proteins. The significance threshold employed by VAST is conservative, in the sense that the substructure similarities it records are very surprising in a statistical sense. However, since 3-dimensional structure is well conserved in protein evolution, VAST often detects homologous proteins that are not detected by sequence neighboring. It is to provide access to these possible remote homologs that the structural neighbor data have been added to *Entrez*. The system also now supports, via helper applications launched from the 'structure summary' page, display of structural alignments and 3-dimensional superpositions of all structural neighbors. These displays allow a user to examine in detail the structural similarity detected by VAST, and in particular to check for conserved features in functionally important sites, which may allow some inference with respect to conservation of protein function.

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 protein structures from MMDB. The DNA and protein sequence data are integrated from a variety of sources and therefore includes more sequence data than is available within GenBank alone. This past year, the entire set of 9 million MEDLINE references were made accessible through the PubMed system. PubMed was developed at NCBI and not only allows text searching of MEDLINE but also links to the full-text of over 25 journal titles 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 40 000 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, plus 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 3-dimensional structures.

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. The recently released Version 2.0 of BLAST allows the introduction of gaps (deletions and insertions) into alignments. With a gapped alignment tool, homologous domains do not have to be broken into several segments and tend to be more biologically meaningful than ungapped results. Also included in Version 2.0 is Position-Specific-Iterated BLAST (or PSI-BLAST). Given a query, PSI-BLAST performs an initial gapped BLAST search of the database. In subsequent iterations, it uses statistically significant alignments from the previous search to construct a position-specific score matrix for use, in place of the query and standard amino acid substitution matrix, in the next round of searching (11).

A new graphical version called PowerBLAST (7), designed for rapid analysis and annotation of large contigs of genomic sequence data 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 are sets of human-only, mouse-only and 'other' EST's. Those sequences added during the previous 30 days (designated 'month' on the BLAST web pages) are also available. 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.

Other ways to access GenBank

The full GenBank release (issued every 2 months) or 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 in the sub-directory, '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. 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 e-mail 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 will no longer be available on CD-ROM due to declining demand and the number of disks needed to contain a single release. Existing GenBank subscribers are being referred to the CD-ROM service available through the EMBL Databank.

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 a 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., Lipman,D.J. and Ostell,J. (1997) *Nucleic Acids Res.*, **25**, 1–6.
2 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.
3 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.
4 Ouellette,B.F.F. and Boguski,M.S. (1997) *Genome Res.*, **7**, 952–955.
5 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.
6 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.
7 Zhang,J. and Madden,T.L. (1997) *Genome Res.*, **7**, 649–656.
8 Hogue,C.W.V., Ohkawa,H. and Bryant,S.H. (1996) *Trends Biochem. Sci.*, **21**, 226–229.
9 Hogue,C.W.V. (1997) *Trends Biochem. Sci.*, **22**, 314–316.
10 Gibrat,J.-F., Madej,T. and Bryant,S.H. (1996) *Curr. Opin. Struct. Biol.*, **6**, 377–385.
11 Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.