NCBI News

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION
National Library of Medicine
National Institutes of Health

Volume 3, Issue 1 February 1994

Improving GenBank's Taxonomy

Taxonomic classification is an important factor in database organization, searching, and analysis. With assistance from taxonomy experts, NCBI has undertaken a comprehensive review of the GenBank taxonomy in order to correct errors, identify inconsistencies, and incorporate new scientific knowledge. The process of rebuilding the taxonomy has involved three steps: (1) developing a software tool for manipulating taxonomic trees, (2) merging existing taxonomies, and (3) enlisting expert assistance to review, revise, and maintain the taxonomy.

The first step was to prepare a merged view of existing taxonomies using TaxMan, a taxonomy database management tool developed by NCBI's Scott Federhen. TaxMan includes a set of functions for merging various taxonomies into a single structure, cross-mapping trees, and annotating taxonomy entries. With the taxonomy developed by Andrzej Elzanowski for PIR-International as a foundation, taxonomies from other comprehensive sequence databases (GenBank, EMBL, Swiss-Prot, and DDBJ) were added and merged. Next, specialized taxonomies such as the ICTV international standard taxonomy for viruses, the U.S. Department of Agriculture taxonomy for plants, and the FlyBase taxonomy for Drosophilidae were added at the appropriate branches of the emerging tree.

Following this integration phase, Mitchell Sogin of the Marine Biological Laboratory at Woods Hole organized a workshop to review and revise the taxonomy and to discuss mechanisms for continued maintenance as new species enter the database and taxonomic consensus develops. The workshop included representatives (see box on page 7) specializing in different branches of the taxonomic tree.

The taxonomy revision will proceed in two stages. First will be the task to formalize the use of organism names in the database by collecting all the variant spellings, synonyms, and misspellings and then selecting a preferred scientific name for each organism. Second will be a phase-in of new taxonomic classification lines from the revised tree as the review of subtrees by participating scientists is completed. Congruently, the GenBank database will be retrofitted with the new classification lines. Subsequent revisions will reflect new work in the field as well as the addition of synonyms and the correction of misspellings in organisms as they are identified.

Continued on page 7 ▶

Access NCBI Through World Wide Web

NCBI services are now accessible through World Wide Web (WWW). WWW is a rapidly growing network information system that permits easy access via hypertextlike links to factual information and database searching. The NCBI WWW server provides both information about and access to GenBank. Text searching and BLAST sequence similarity searching are provided as well as access to Network *Entrez*, which includes a subset of MEDLINE citations related to molecular sequence data.

The search services are front-ends to existing NCBI search systems, and their interfaces should be considered experimental. NCBI expects to be making changes to

Continued on page 2 ▶

In This Issue

NCBI News

NCBI News is distributed three times a year. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence and suggestions to *NCBI News* at the address below.

NCBI News
National Library of Medicine
Bldg. 38A, Room 8N-803
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 496-2475
Fax: (301) 480-9241
E-mail: info@ncbi.nlm.nih.gov

In 1988, Congress established the National Center for Biotechnology Information as part of the National Library of Medicine; its charge is to create automated systems for storing molecular biology, biochemistry, and genetics data, and to perform research into computational molecular biology.

NIH Publication No. 94-3272

ISSN 1060-8788

NCBI Creates New Database, New GenBank Division for STS Data

NCBI is creating a new database and a new GenBank division to facilitate access to the growing body of sequence tagged site (STS) data being generated by genome mapping laboratories. An STS is a short DNA sequence that has a single occurrence in the genome of an organism, thus serving as a physical mapping landmark.

GenBank currently contains about 3,200 STS sequences, most of which come from human sources. Beginning with Release 82.0 (April 1994), STSs will be consolidated into a new division. This reorganization will facilitate cross-comparison of STSs with sequences in other divisions for the purpose of correlating map positions of anonymous sequences with known genes.

In addition, a separate STS database, dbSTS, will be maintained to manage this special class of sequences and provide detailed information about the STS map locations and PCR conditions. Following the model of dbEST,[1] dbSTS will be a dynamic resource. Sequence similarity searches against other divisions of GenBank will be performed on a frequent periodic basis, and the annotations in the STS database will be updated automatically. Methods are under development to enable queries of dbSTS by chromosomal and subchromosomal location. Access will initially be by an e-mail server.

STS data may be submitted to GenBank via e-mail to the address bulk-sub@ncbi.nlm.nih.gov. Because STSs are usually submitted as a set of many sequences, we have designed a simple tagged "flat file" format (available on request) to simplify and streamline the direct submission process. Upon receipt of STS data, GenBank accession numbers will be issued and the data will then be released to the public unless the submitter requests that it be kept confidential until publication.

To obtain a copy of the data input format specification or for further information, contact GenBank User Services at info@ncbi.nlm.nih.gov or (301) 496-2475.

[1]Boguski, MS, TMJ Lowe, and CM Tolstoshev. dbEST —database for "expressed sequence tags." *Nature Genet* 4:332–3, 1993. ■

***WWW**, continued from page 1*

the interfaces and to add more data following evaluation of user feedback and system performance.

WWW has client interfaces available at no cost for PC, Macintosh, Unix, and other systems. You can get additional information and client software (e.g., NCSA Mosaic) by Anonymous FTP to NCSA: ftp.ncsa.uiuc.edu in the /Web directory. For those familiar with WWW, the URL for establishing a link to NCBI is as follows: http://www.ncbi.nlm.nih.gov. ■

Entrez on CD-ROM Expands To Accommodate Growth

The sequence databases double in size about every 20 months. To accommodate this rapid growth, *Entrez* expanded from one disc to two in February 1993. With Release 8.0 in December 1993, another organizational change was required. Prior to December, the *Entrez:* Sequences disc contained the sequence database entries plus associated MEDLINE references and abstracts—approximately 50,000 records in October. A larger MEDLINE subset (about 150,000 records) was contained on the *Entrez:* References disc. In December all of the MEDLINE data were segregated on the *Entrez:* References disc to provide more room for growth of the sequence databases.

Because the *Entrez* software was modified to accommodate this change in data location, users were required to reinstall the *Entrez* software with Release 8.0. You can continue to use both discs with a single CD-ROM drive by specifying "Disc Swapping," which is now the default option, in the configuration program, *EntrezCf*. *Entrez* will prompt you to insert the appropriate disc as you search the sequence and reference databases. If your computer has two CD-ROM drives, *Entrez* makes it easy to use them both, with no need to exchange CD-ROM discs.

Entrez on Three Discs in October 1994

Given the rate of growth of the sequence databases, it is clear that by October 1994 it will no longer be possible to contain all the data on two discs. Consequently, *Entrez* will expand to three discs at that time. Our projections are that three discs will be sufficient to hold the data until October 1995.

What does this change mean for *Entrez* subscribers? First, there will be a price increase to account for production of the third disc. Pricing is determined each spring by the U.S. Government Printing Office. Based on past experience, however, it is likely that the increase in price will be quite modest, probably about $20 more per year. The second and more profound impact will be on hardware requirements.

Options for Accommodating Growth

Users will have at least four ways to accommodate the additional storage requirements of *Entrez* this October:

1. Install a total of three CD-ROM drives on your Mac or PC. A number of manufacturers now offer "tower" type CD-ROM configurations, which can accommodate three or more CD-ROM drives. Alternatively, for Macintosh users and for PC users who have SCSI host adapters, three internal and/or external SCSI CD-ROM drives may be "daisy-chained" together. With the cost of high-performance "double-speed" SCSI CD-ROM drives approaching $300, this approach is the most economical.

Continued on page 6 ▶

NCBI Data by FTP

NCBI maintains a repository of molecular biology databases and software development tools that are publicly available for network users through Internet FTP (file transfer protocol). The available directories include "repository", "toolbox", and "pub".

The repository directory holds more than 20 databases, such as

- Swiss-Prot (Amos Bairoch)
- ACeDB: A *C. elegans* Database (J. Thierry-Mieg, R. Durbin)
- FlyBase (Michael Ashburner)
- Eukaryotic Promoter Database (Philipp Bucher)
- REBASE (Restriction Enzyme Database - Richard Roberts)
- CarbBank/CCSD
- PROSITE (Dictionary of Protein Sites and Patterns - Amos Bairoch)

The toolbox directory contains a set of software and data exchange specifications that are used by NCBI to produce portable software, and includes ASN.1 tools and specifications for molecular sequence data.

The pub directory offers public-domain software, such as BLAST (a sequence similarity search program), MACAW (a multiple sequence alignment program), and Authorin submission software for Mac and PC systems.

All data in these directories can be transferred through Internet by using the Anonymous FTP program. To connect, type: **ftp ncbi.nlm.nih.gov** or **ftp 130.14.25.1**. Enter **anonymous** for the login name, and enter your e-mail address as the password. Change directories to "repository" to download databases (cd repository), "toolbox" to download ASN.1 tools (cd toolbox), or "pub" to download public-domain software (cd pub).

Selected Recent Publications by NCBI Staff

Boguski, MS, TMJ Lowe, and **CM Tolstoshev.** dbEST—database for "expressed sequence tags." *Nature Genet* 4:332–3, 1993.

Claverie, JM. Detecting frame shifts by amino acid sequence comparison. *J Mol Biol* 234:1140–57, 1993.

Rudd, KE. Maps, genes, sequences, and computers: an *Escherichia coli* case study. *ASM News* 59(7):335–41, 1993.

Koonin, EV, and TV Ilyina. Computer-assisted analysis of the relationships between protein and DNA sequences involved in rolling circle DNA replication. *Biosystems* 30:241–68, 1993.

Wootton, JC, and **S Federhen.** Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 17:149–63, 1993.

Koonin, EV, and VV Dolja. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. *Crit Rev Biochem Mol Biol* 28:375–430, 1993.

Lawrence, CE, SF Altschul, MS Boguski, JS Liu, **AF Neuwald,** and **JC Wootton.** Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262:208–14, 1993.

Gregory, PE, DH Gutmann, A Mitchell, S Park, **M Boguski,** T Jacks, DL Wood, R Jove, and FS Collins. Neurofibromatosis type 1 gene product (neurofibromin) associates with microtubules. *Somat Cell Mol Genet* 19:265–74, 1993.

Major, F, D Gautheret, and R Cedergren. Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proc Natl Acad Sci U S A* 90:9408–12, 1993.

GenBank: Focus on Quality

NCBI continues to concentrate on improving GenBank quality. Several steps have been taken to correct errors, reduce redundancy, promote consistency, and ensure the quality of data at the point of entry into the system.

Controlling Quality of New Direct Submissions

Quality control begins with NCBI's applied database staff. GenBank annotators who build and maintain the database entries are trained in molecular biology and skilled in database production operations. Senior scientists provide daily scientific guidance and review.

When a new submission arrives at NCBI on disk or by e-mail (gb-sub@ncbi.nlm.nih.gov), an annotator assigns a GenBank accession number after a quick review of format and general content. However, before the submitter receives the full GenBank record for review and comment, the sequence is screened against GenBank using BLAST to identify full or partial matches, followed by a search to detect vector contamination. Algorithms that check for internal consistency are used to confirm coding regions, detect open reading frames, and verify amino acid translations. Using GenBank content and data representation guidelines, annotators then review the descriptive parts of the entry; assign the locus name, definition line, keywords, and taxonomy classification; and, in consultation with the submitter, add and modify features if needed.

Reducing Redundancy and Regularizing Features

Although staff at the international DNA sequences databases have been working together to remove duplicate entries and merge records where appropriate, analyses by NCBI researchers reveal that there is still considerable internal redundancy in the database: more than 5 percent of the entries have duplicate sequences and another 5 percent have close matches. There are also many records that contain coding sequences with no features, translations, or protein product names. A project to create an enhanced view of GenBank, called GenBank Select, aims to reduce redundancy and regularize feature annotation. For example, where exact or nearly exact matches exist between Swiss-Prot and translated GenBank entries, the Swiss-Prot translations, names, and descriptions will be substituted in the GenBank records. In addition, subset sequences that exist as separate GenBank entries, such as identical cDNAs or a cDNA and its corresponding genomic sequence, will be merged into a single record for the view presented in GenBank Select. Thus, GenBank Select will be an optional view of GenBank data, which will continue to be distributed and accessed as before.

Keeping Current

With support from the National Library of Medicine's MEDLINE indexers, who review more than 3,500 journals, GenBank annotators scan the current

Continued on page 7 ▶

Frequently Asked Questions

I am in a hurry to submit a sequence, but I don't have a copy of Authorin. Is Authorin available electronically for downloading?

Authorin is available by Anonymous FTP through the Internet. You will find the Mac version (3.0) in the /pub/authorin/mac directory and the DOS version (1.2) in the /pub/authorin/dos directory. If you would like a printed copy of the documentation or have any questions, contact us at authorin@ncbi.nlm.nih.gov.

Authorin Macintosh release 3.0 crashes at the end of the "Prepare Submission" process and may or may not flash an error message and produce an .sbt submission file. What is the problem?

This problem occurs when the path to the ASNLOAD folder is not properly specified during the Authorin installation procedure. The ncbi.cnf file must be edited to reflect the user's hard disk name and any possible intervening folders. If this step is not carried out when following the installation instructions, the system may crash when preparing data for submission.

I need an accession number for an article that was accepted for publication and I am in a hurry. How long does it take?

The fastest way to get an accession number is to send the Authorin submission file (the file ending in the extension .sbt) by e-mail to gb-sub@ncbi.nlm.nih.gov. The accession number will be issued within 24 hours and e-mailed back to the address from which the submission was sent.

After loading my Authorin 2.1 program on the Mac, it crashes when I begin, gives the following error message TXCL_UNLOCK, and will not let me continue. Why does this happen?

You are trying to use Authorin 2.1 with System 7.0x on the Macintosh. With System 7.0x you must use the latest version of Authorin, Release 3.0. This release will work on both System 6.0x and 7.0x.

I have numerous submissions to make and much of the data is the same. Do I have to retype author and citation information for each submission?

You can create a template file with the information that will be the same in all the submissions. By opening the template file and saving under a new name, you can retain the master file and add additional information to the new file. Directions are provided in both the Mac and PC manuals for creating and using a template file.

The term e-mail server *is used to describe the RETRIEVE and BLAST e-mail addresses, but I don't understand exactly what this term means.*

The e-mail (electronic mail) servers at the NCBI are host computers that receive specifically formatted e-mail queries, process these queries, and return the search results to the address from which the message was sent. No specific password or account is needed only the ability to send e-mail to an Internet site. ■

NCBI's Board Invites Community Comment

The NCBI—and all research centers at the National Institutes of Health—has a Board of Scientific Counselors (BOSC). The Board meets regularly to review NCBI activities and to provide guidance and advice on how best the Center can meet the needs of the molecular biology community.

The eight-member board consists of molecular biologists and computer scientists from academia and industry. At their most recent meeting, BOSC members, listed below, invited input from the user community on NCBI databases, software, and services.

Robert T. Sauer, Ph.D. (Chairman)
Department of Biology
Massachusetts Institute of Technology
Phone: (617) 253-3163

Helen M. Berman, Ph.D.
Center for Computational Chemistry
Rutgers University
Phone: (908) 932-4667
E-mail: berman@dnarna.rutgers.edu

Charles R. Cantor, Ph.D.
Center for Advanced Biotechnology
Boston University
Phone: (617) 353-8504
E-mail: crc@buenga.bu.edu

John R. Devereux, Ph.D.
Genetics Computer Group, Inc.
Phone: (608) 231-5200
E-mail: devereux@gcg.com

Paula Fitzgerald, M.D., Ph.D.
Merck Research Laboratories
Phone: (908) 594-5510
E-mail: paula_fitzgerald@merck.com

Michael W. Hunkapiller, Ph.D.
Applied Biosystems Division
Perkin-Elmer Corporation
Phone: (415) 570-6667

Sung-Hou Kim, Ph.D.
Melvin Calvin Laboratory
Lawrence Berkeley Laboratory
Phone: (510) 486-4333

Myra N. Williams, Ph.D.
Glaxo, Inc.
Phone: (919) 990-5686
E-mail: mnw30117@glaxo.com ■

Entrez *Expands*, *continued from page 3*

2. Install *Entrez* index files on the hard disk and use two CD-ROM drives for the datafiles. The design and organization of the three-disc *Entrez* product is not yet complete (more details will be available in the next *NCBI News*). But we expect that it will be possible to put all the index files and other files necessary to locate sequence and reference records on one CD-ROM disc, leaving the data records on the remaining two discs. Therefore, users who can dedicate approximately 500 MB of magnetic disc space to *Entrez* will be able to copy the index and link files to a hard disk and use two CD-ROM drives to hold the databases. With hard disk costs less than $1 per megabyte, this alternative is only slightly more expensive than having three CD-ROM drives and will provide a considerable improvement in performance.

3. For the best performance, use the CD-ROMs only as distribution media and put all the data on magnetic disks. Two gigabytes of hard disk storage should accommodate *Entrez* through October 1995. Two gigabyte SCSI drives currently cost less than $2,000, and 16-bit SCSI-2 host adapters for PCs cost about $250. Users with newer Macintoshes or Unix workstations should already have the necessary disk controller. Thus, for a cost of about $2,000, PCs, Macs, or Unix systems can hold all of *Entrez* on magnetic disk. The performance of these systems should be more than adequate for use as departmental servers on local area networks.

4. For users with direct Internet connections, Network *Entrez* is available at no cost. As discussed previously in the *NCBI News* (August 1993), Network *Entrez* requires that a local network administrator take responsibility for (1) establishing and maintaining the Internet connection, (2) installing the necessary network software (MacTCP for Macintosh or one of several TCP/IP software packages for Windows PCs), and (3) installing the retrieval software (available via Anonymous FTP from ncbi.nlm.nih.gov). For more information about Network *Entrez*, send electronic mail to net-info@ncbi.nlm.nih.gov. Users with Internet access who do not have the local support necessary for Network *Entrez* can use a version of *Entrez* that has been adapted for use with the Mosaic/World Wide Web hypertext-based information service (see related article on page 1).

The expansion of the sequence databases is an inevitable reflection of the rapid progress in gene and protein sequencing. An additional CD-ROM drive or hard disk is the cost of ready local access to more than a gigabyte of sequence and related bibliographic data. Contact us via e-mail at info@ncbi.nlm.nih.gov, or by phone at (301) 496-2475, if you have any questions about upcoming releases of *Entrez* on CD-ROM. ■

***Taxonomy**, continued from page 1*

In addition, taxonomy revisions are now being incorporated into *Entrez*. With Release 10.0, you will be able to use the organism field to restrict searches to any level in the taxonomic hierarchy, not just the genus and species level. For example," eukaryotes," "fungi," and "primates" will all be valid search terms.

From its inception, the taxonomy review has been a collaborative project. Representatives of the major sequence databases and taxonomy experts will continue to work together to maintain a current and accurate taxonomy resource. The international DNA sequence databases, EMBL and DDBJ, have agreed in principle to adopt the revised taxonomy as a database standard. We invite your suggestions and participation. Anyone interested in more information or in curating a segment of the taxonomy should contact Scott Federhen at NCBI (federhen@ncbi.nlm.nih.gov).

***Focus on Quality**, continued from page 4*

journal literature to locate new sequences, update publication information, and identify GenBank sequences that should be released to the public database.

In addition, retrospective projects to release published sequences and systematically update the numerous "in press" entries with full citation data are under way.

Although only the submitting scientist is permitted to modify sequence data or annotations, NCBI encourages all GenBank users to point out possible errors or omissions, provide updated publication information, or request the release of data that have been published. Send update notices to update@ncbi.nlm.nih.gov. ■

Taxonomy Project Participants

Robert Baker, Texas Tech. University
Lois Blaine, American Type Culture Collection
Russell Chapman, Louisiana State University
Michael Donaghue, Smithsonian Institution
Andrzej Elzanowski, PIR-International
David Freshwater, University of Miami
John Gunderson, Marine Biological Laboratory at Woods Hole
David Hillis, University of Texas
Jack Holt, Bergey's Trust
Rodney Honeycutt, National Science Foundation
Shung-Chang Jong, American Type Culture Collection
Eugene Koonin, NCBI
Gary Olsen, University of Illinois at Urbana-Champaign
Kate Rice, EMBL
Hugh Robertson, University of Illinois at Urbana-Champaign
Mitchell Sogin, Marine Biological Laboratory at Woods Hole
John Taylor, University of California at Berkeley
Ward Wheeler, American Museum of Natural History

TO RECEIVE INFORMATION FROM NCBI, PLEASE SEND THIS FORM TO:

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION
National Library of Medicine
National Institutes of Health
Bldg. 38A, Room 8N-803
8600 Rockville Pike
Bethesda, MD 20894

Queries about services and software may also be sent via electronic mail to info@ncbi.nlm.nih.gov or by fax to (301) 480-9241.

Name (please print) _____
Department _____
University or Institution _____
Mailing Address _____

Phone _____
Fax _____
E-mail _____

Comments _____

NCBI News Mailing List (Check one)

❑ Add new name to *NCBI News* mailing list
❑ Remove name from *NCBI News* mailing list
❑ Make corrections or changes as shown (please enclose label)

NCBI Services

CD-ROM Products

Entrez: Sequences

Integrated sequence data from GenBank, EMBL, DDBJ, Swiss-Prot, PIR, PRF, and PDB linked to MEDLINE abstracts. Text retrieval software for Macintosh and PC-compatible systems running Windows 3.1 is included, but there is no sequence similarity search software.

NCBI-GenBank (Flat File)

Data distribution disc containing the GenBank DNA database in flat file format. *No software included.*

NCBI-Sequences (ASN.1)

Data distribution disc containing the *Entrez:* Sequences integrated sequence dataset in the ASN.1 standard data description format. Intended primarily for software developers. *No software included.*

CD-ROM Orders

To place orders with the Superintendent of Documents (U.S. Government Printing Office), phone (202) 783-3238 or fax (202) 512-2233. **To check on orders,** phone (202) 783-3238 or fax (202) 512-2168

Order forms available from NCBI.

Network Access

GenBank Submissions

New Submissions	gb-sub@ncbi.nlm.nih.gov
Corrections/Updates	update@ncbi.nlm.nih.gov
Submission Software	authorin@ncbi.nlm.nih.gov

GenBank Internet Access

E-mail Servers	retrieve@ncbi.nlm.nih.gov blast@ncbi.nlm.nih.gov
Network *Entrez*	net-info@ncbi.nlm.nih.gov
Anonymous FTP	ncbi.nlm.nih.gov (130.14.25.1) Directories for GenBank current release and daily updates: genbank (full releases) genbank/daily (cumulative) genbank/daily-nc (noncumulative)
General Information	info@ncbi.nlm.nih.gov

DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service, National Institutes of Health
National Library of Medicine
National Center for Biotechnology Information
Bldg. 38A, Room 8N-803
8600 Rockville Pike
Bethesda, Maryland 20894