NCBI News

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Department of Health and Human Services

Spring 2003

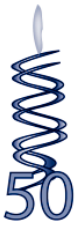
A Field Guide to GenBank® and NCBI Resources: NCBI's Scientific Outreach and Training Program

Biological sequence and structure information are now used in nearly every field of biological research. A working knowledge of these resources and standard computational biology tools are an essential part of every biologist's toolkit. However, keeping up with these databases and tools can be challenging in this period of rapidly changing bioinformatics resources.

In order to help researchers keep abreast of enhancements and the increasing diversity of NCBI molecular biology resources, the NCBI service desk provides a free training program for biologists on the effective use of NCBI databases and tools. The course, called "A Field Guide to GenBank and NCBI Resources", is designed especially for biologists who work at the bench or in the field but use sequence and structure data in their research. All researchers, educators and students who work with biological sequence and structure data should find this to be a useful introduction and survey of the available NCBI tools and databases. Because of the rapid expansion of the resources, even experienced NCBI users will likely learn something new

continued on page 4

First Version of Human Genome Reference Sequence Debuts on DNA's 50th



April 14, 2003 marked the 50th anniversary of the description of the structure of DNA and also saw the release of the first version of the 3 billion base pair reference sequence of the human genome. Annotations to the raw sequence made public on April 14 were released on April 29 when the reference genome, NCBI build 33, appeared in the NCBI Map Viewer.

continued on page 2

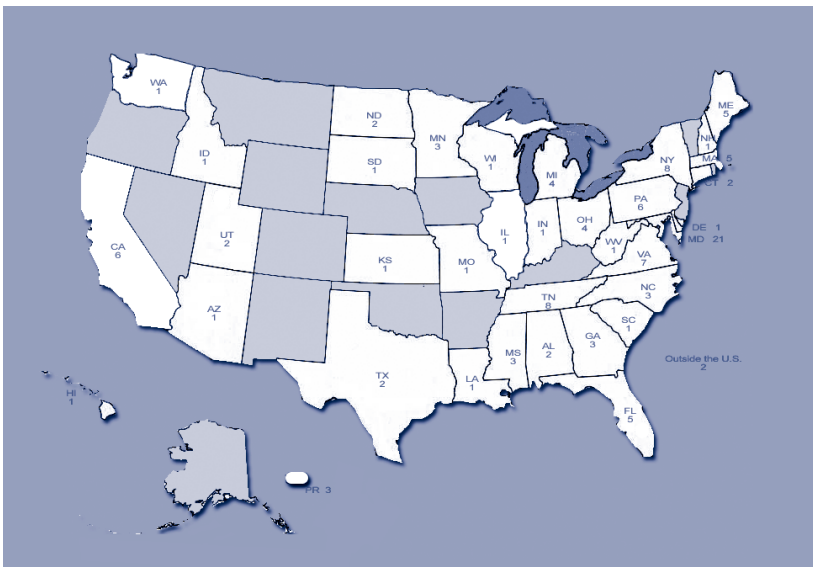


Figure 1: This Field Guide Map shows the number of Field Guides held in each state through April 2003.

In this issue

NCBI News

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

NCBI News
National Library of Medicine
Bldg. 38A, Room 3S308
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 496-2475
Fax: (301) 480-9241
E-mail: info@ncbi.nlm.nih.gov

Editors
Dennis Benson
David Wheeler

Contributors
Medha Bhagwat
Peter Cooper
Eric Sayers

Writers
Vyvy Pham
David Wheeler

Editing and Production
Robert Yates

Graphic Design
Gary Mosteller
Joe Vuthipong

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

NIH Publication No. 03-3272

ISSN 1060-8788
ISSN 1098-8408 (Online Version)

Human Reference Sequence

continued from page 1

The human genome reference sequence is a critical contribution to RefSeq, NCBI's database of reference sequences for genomes, mRNAs, and proteins.

The reference sequence covers about 99 percent of the human genome's gene-containing , euchromatic, regions, at an accuracy of 99.99 percent. Only the sequence near centromeres, telomeres, and other heterochromatic blocks, along with a small number of unclonable gaps, has yet to be determined. Small updates to the assembly will continue to be made as complex regions are refined and the remaining gaps, about 400 of less than 100 kilobases each, are closed.

The assembly is the end product of several years of collaborative work by the Human Genome Project sequencing centers, NCBI, the University of California at Santa Cruz, and Ensembl, a joint project between EMBL-EBI and the Sanger Institute.

There are about 32,000 genes annotated by NCBI on build 33; of those, almost 18,000 are supported by mRNA alignments and may be considered to be confirmed. The typical confirmed human gene has 12 exons of an average length of 236 base pairs each, separated by introns of an average length of 5,478 base pairs. As a consequence, the average intron length is about twice the average transcript length. Some statistics on NCBI's build 33 are given in Table 1.

To view the genome or download the sequence, start at the "Human Genome Resources" link under "Hot Spots" on the NCBI Home Page. See the upcoming summer edition of the NCBI News for more on the human genome and how to explore it at NCBI.

Gene Identification Method	EST Alignment	Predicted	mRNA Alignment
Number of Genes	3,295	11,972	17,708
Exons/Gene (Avg)	4	5	12
Exon Length (Avg)	440	158	226
Intron Length (Avg)	5,605	10,512	5,478
mRNA Length (Avg)	2,127	820	2,741
Gene Length (Avg)	13,940	43,918	55,147

Table 1: Statistics: First version of the human genome reference sequence, NCBI human genome build 33.

UniGene Expands; *C. elegans* added to LocusLink

With the recent additions of *Ciona intestinalis* (sea squirt), *Oncorhynchus mykiss* (rainbow trout), and *Vitis vinifera* (the wine grape) the UniGene database of non-redundant gene-oriented clusters now covers 15 animals, 12 plants and *Dictyostelium discoideum*, a slime mold. Each UniGene cluster is linked to related information, such as the tissue types in which the gene is expressed, model organism protein similarities, the LocusLink report for the gene, and its map location. Try UniGene by following the "UniGene" link under "Hot Spots" on the NCBI Home Page.

The nematode, *Caenorhabditis elegans*, represented in UniGene with over 19,000 clusters, is now also represented in the LocusLink database of genetic loci with over 22,000 records.

Rat Genome Assembly Deposited into GenBank

With the recent deposit of the draft rat genome assembly into GenBank, researchers can view the rat genome in Entrez and from the Rat Genome Map Viewer. Other resources such as Rat Genomic BLAST and downloadable files of predicted gene and protein sequences complement the navigational and display capabilities of the Viewer.

The Whole Genome Shotgun (WGS) effort that produced the rat genome was managed by the Rat Genome Sequencing Consortium (RGSC), which is led by the Baylor College of Medicine Human Genome Sequencing Center. The current assembly is comprised of over 157,000 WGS contigs fused to produce almost 1,800 supercontigs, or scaffolds.

The rat WGS contigs are assigned accession numbers of the format "AABR########", with the first two digits describing the project version number. Thus, the first version has the accession AABR01000000, and consists of sequences AABR01000001 through AABR01157561. One can either search for AABR01000000 in Entrez, and from its record, click to view the records AABR01000001 through AABR01157561; or, one can search for any of the individual records separately.

The WGS reads are found in the Trace Archive and can also be queried using the BLAST service described below. There are over 38,000,000 rat traces available via the "Hot Spots" "Trace archive" link on the NCBI Home Page.

More information on the assembly and other rat genomic resources is available from the NCBI "Genomic Biology" page. Follow the "Rat" link under "Organism-Specific Resources". To download sequence and annotations via FTP, use

ftp.ncbi.nih.gov/genomes

The genome assembly can be displayed using the Rat Genome Map Viewer, as in Figure 1. In the figure, several different maps are shown, including gene sequence, component, contig, GenomeScan, UniGene, and STS. The component map shows the tiling path of the WGS contig accessions used to build the "NW_xxxxxx" scaffolds, whereas the contig map displays the placement of the scaffolds on the chromosomes. Genes that have also been annotated on the genome assembly using the alignment of mRNAs to the contigs are shown on the Genes_seq map, while GenomeScan predictions are included on a separate map. In addition, the Transcript map shows the combinations of exons, or splice variants, that are implied by mRNA alignments. The Rn_UniGene map shows the alignment of rat ESTs to the genome assembly and shows EST-only gene models based on shared introns and alignment to a common position on the genome. The STS map displays the placement, via Electronic PCR, of STSs onto the genomic scaffolds.

Try the rat MapViewer at:

www.ncbi.nlm.nih.gov/mapview

Researchers can also BLAST a query sequence against WGS contigs, WGS traces from the Trace Archive, BAC ends, HTG, RNA, and EST sequences, using the Rat Genome BLAST service. BLAST hits are shown marked in red on the genome assembly. Figure 2 shows the BLAST hit of a *Bos taurus* p53 tumor suppressor phosphoprotein (TP53), mRNA against the Rat genome assembly. The query sequence shows an 82% homology to the rat contig NW_042660.1. The parallel gene sequence map shows the annotated rat gene TP53 that is annotated within the same region, suggesting a homologous relationship between the *Bos taurus* and rat genes.

The Rat BLAST service can be accessed from the BLAST Home Page.

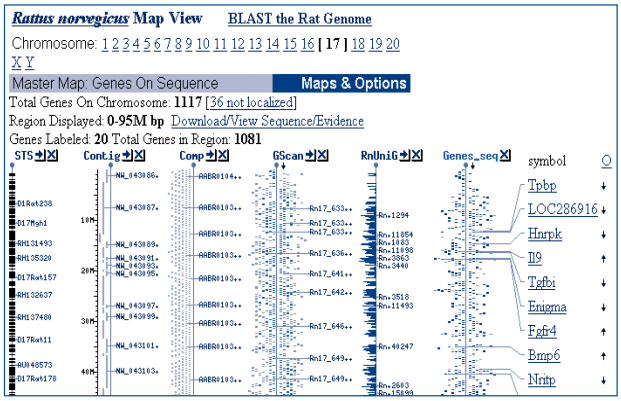


Figure 1: Map Viewer display of rat genome assembly for chromosome 17 showing several parallel maps, including those for predicted and confirmed genes, and the contigs and their components. Click on "Maps and Options" to view all the available maps.

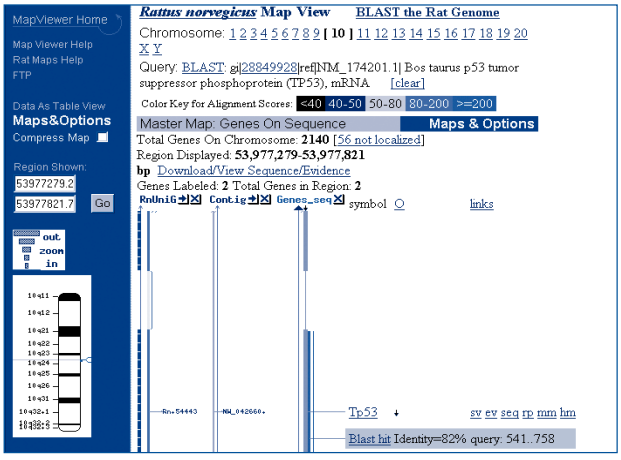


Figure 2: BLAST hit of a *Bos taurus* p53 tumor suppressor phosphoprotein (TP53), mRNA against the Rat genome assembly shown in its genomic context within the MapViewer.

Field Guide to GenBank

continued from page 1

and useful. Five members of the NCBI service desk comprise the Field Guide training staff: Peter Cooper, Susan Dombrowski, Wayne Matten, Rana Morris, and Eric Sayers. All are PhD-level biologists with a broad range of research and teaching experience that make them especially knowledgeable in the diverse interests and needs of the NCBI user community.

The course consists of two main components: a three-hour lecture providing detailed background on NCBI databases and a two-hour hands-on practicum featuring biologically interesting examples that highlight key features of the NCBI resources.

The Field Guide content is flexible and changes often to reflect shifts in emphasis and content of the NCBI Web resources. The focus of the Field Guide can also be tailored to the specific interests of the audience. Special Field Guides have been presented that emphasize resources for marine organisms, insects, and agricultural plants. Components of the Field Guide have also been offered as parts of specialized short courses and workshops, notably the Cold Spring Harbor Laboratory's "Computational Genomics" course, the Jackson Laboratory's "Experimental Genetics of the Laboratory Mouse in Cancer Research", the "MR4 Malaria Bioinformatics Workshop" at the ATCC, and "the Biology of Parasitism" summer course at Wood's Hole Oceanographic Institution.

In addition to the Field Guide, the training staff also offers more detailed modular courses that focus on a particular subset of the NCBI resources. A BLAST module is available that provides more extensive hands-on training using the Web BLAST services and practical experience setting up and using standalone BLAST. A separate module providing more in-depth training in the use of NCBI structure databases and tools will premiere this summer at the University of Pennsylvania. Modules will soon be available that focus separately on Entrez and genomic resources.

The Field Guide continues to grow in popularity. Since the initial offering at Yale University in November of 1998, the course has been presented 140 times in 37 States and Puerto Rico to approximately 8,000 participants. The map on the newsletter front page shows the geographical distribution of Field Guide host institutions to date. At the time of this writing, there are 35 courses scheduled for the remainder of 2003. The course is updated each time it is given and provides the most current view of the NCBI resources available anywhere. For this reason, many institutions offer the Field Guide on a regular schedule, including Vanderbilt University, the University of Michigan, the University of Virginia, the Mayo Clinic, and the University of Pennsylvania. A schedule of upcoming and past Field Guides is available from the FieldGuide Home Page referenced below. Anyone who would like to attend one of the courses listed there should contact the host institution through the links provided.

NCBI provides this training free of charge at educational institutions and government laboratories and pays all travel expenses for two NCBI instructors to conduct the course. To host a course, an institution must have the appropriate facilities: a lecture hall and a teaching computer classroom. The host institution must also be able to attract a minimum audience of 50 people and must manage course registration. The audience size requirement is rarely a barrier; at larger university settings, audience size often exceeds 200 people. When practical, we also ask that the host institution open registration to include people outside their institute in the training. In many cases, participants have traveled hundreds of miles to attend Field Guide courses at host institutions.

For more information about course content and links to the most recent lecture materials and hands-on exercises, see The Field Guide Home Page at:

www.ncbi.nlm.nih.gov/Class/FieldGuide

Additional details about hosting the NCBI courses are linked to the Field Guide Home Page. Courses are now being scheduled for late fall of 2003 and early 2004.

Anyone interested in scheduling a course should contact Peter Cooper at: **cooper@ncbi.nlm.nih.gov**

Topics covered in the course include:

Sequence Databases
- GenBank Database: description and scope
- The NCBI Derivative Databases: RefSeqs

Database Searching using Entrez
- Entrez databases
- Neighboring and Links
- Entrez searching

The NCBI Structures Database
- The Molecular Modeling Database
- Structural Alignments
- Viewing Structures and Structural Alignments with Cn3D

Similarity Searching using NCBI BLAST
- Local Alignment Statistics
- Scoring Systems
- Using BLAST 2.2.6 Web services
- PSI-BLAST
- RPS-BLAST and the Conserved Domain Database
- Genomic BLAST pages

Genomic Resources at NCBI
- Complete Microbial Genomes in Entrez
- Higher Genome Resources
  - RefSeq and LocusLink
  - UniGene
  - Variation Data
  - The Reference Human Genome
  - The Map Viewer
  - Mouse, Rat and Other Genomes

Taxonomy Browser Links to More Resources

Figure 1: Taxonomy report for zebrafish *Danio rerio* showing a table of links to other Entrez databases, links to Genome views of zebrafish chromosomes and LinkOuts to external sources of information.

The Taxonomy Browser now displays many more links to the diverse data sets within the Entrez system along with links to related information from external datasets through LinkOut. The Entrez database links displayed include those for Nucleotide, Protein, Structure, SNP, UniGene, UniSTS, PopSet, PubMed Central, 3D-Domains, Genomes, and Taxonomy. As an example, clicking on the taxonomy record for *Danio rerio*, the zebrafish, displays standard lineage information along with links to related records from other Entrez databases, arranged in a table, as shown in Figure 1. Links to additional information, such as mapping data and to external LinkOut resources, are also included when available. This functionality is incorporated into each taxonomic node so that users can retrieve related information for taxonomic nodes, such as "*Danio*", that lie above the species level.

Outside groups interested in participating in the LinkOut program in the taxonomy domain should contact us at: **linkout@ncbi.nlm.nih.gov**

Search the NCBI Web and FTP Site

A careful examination of the Entrez Databases available under the database pulldown menu at the top of most NCBI Web pages will reveal a database called "NCBI Web site". Choosing this database allows one to search the entire suite of NCBI Web pages as well as the NCBI FTP site using the flexible Entrez query syntax. Results are returned in one of a variety of format options including Summary, Brief, Home Page, and Related Sites; the latter format shows the NCBI Web site neighbors for a particular Webpage.

The default "Home Page" format displays links to NCBI "Home Pages" residing at the top of the hierarchy of the Web page that was hit by the query. The Home Page format also shows separate links to a listing of all pages within the tree satisfying the query.

Searches may be field-limited using several Entrez "NCBI Web site" fields such as, "category", "description", "Home Page URL", and "Home Page title".

Selected Recent Publications by NCBI Staff

To view the citation for any article listed below, click on the PubMed link on the navigation bar at the top of the NCBI Home Page, enter the PubMed ID number in the search query box, and click go.

Cummings L, **Riley L**, **Black L**, **Souvorov A**, **Resenchuk S**, **Dondoshansky I**, **Tatusova T**. Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes. *FEMS Microbiology Letters*. 2002 Nov 5;216(2):133-8. PMID: 12435493

Geer LY, **Domrachev M**, **Lipman DJ**, **Bryant SH**. CDART: Protein Homology by Domain Architecture Retrieval Tool. *Genome Research*. 2002; 12(10):1619-23. PMID: 12368255

Jordan IK, **Rogozin IB**, Glasko GV, **Koonin EV**. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends in Genetics*. 2003 Feb; 19(2):68-72. PMID: 12547512

Koonin EV, **Wolf YI**, **Karev GP**. The structure of the protein universe and genome evolution. *Nature*. 2002; 420(6912):218-23. PMID: 12432406

Liu Y, Zhou J, **Omelchenko MV**, Beliaev AS, Venkateswaran A, Stair J, Wu L, Thompson DK, Xu D, **Rogozin IB**, Gaidamakova EK, Zhai M, **Makarova KS**, **Koonin EV**, Daly MJ. Transcriptome dynamics of Deinococcus radiodurans recovering from ionizing radiation. *Proc. Natl. Acad. Sci. USA*. 2003 Apr 1; 100(7):4191-6. PMID: 12651953

Marth G, **Schuler G**, Yeh R, Davenport R, **Agarwala R**, **Church D**, Wheelan S, **Baker J**, **Ward M**, **Kholodov M**, **Phan L**, **Czabarka E**, **Murvai J**, Cutler D, Wooding S, **Rogers A**, Chakravarti A, Harpending HC, Kwok PY, **Sherry ST**. Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc. Natl. Acad. Sci. USA*. 2003 Jan 7; 100(1):376-81. PMID: 12502794

Wootton JC, Mu J. Malaria parasite diversity and evolution: More challenges for global health. *Global HealthLink*. 2002; 117:12-13. PMID: not in PubMed

Yu H, Hatzivassiloglou V, Friedman C, Rzhetsky A, **Wilbur WJ**. Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2002; 919-23. PMID: 12463959

New Genomes in GenBank

In December 2002, Entrez Genomes reached a milestone of 100 completely sequenced microbial genomes. Recent arrivals include:

Organism	Base Pairs	GenBank Accession Number	RefSeq Accession Number	Journal
***Streptococcus mutans* UA159**	2,030,921 bp	AE014133	NC_004350	Proc. Natl. Acad. Sci. U.S.A. 99 (22), 14434-14439 (2002)
***Shewanella oneidensis* MR-1**	4,969,803 bp,161,613 bp	chromosome: AE014299, megaplasmid: AE014300	chromosome: NC_004347, megaplasmid: NC_004349	Nat. Biotechnol. 99 (2002) In press
Wigglesworthia brevipalpis	697,721 bp	BA000021	NC_004344	Online Publication Nature Genetics 3 September 2002: doi:10.1038/ng986
Mycoplasma penetrans	1,358,633 bp	BA000026	NC_004432	Nucl. Acids Res. 30 (23) 5293-5300 (2002)
***Pseudomonas putida* KT2440**	6,181,863 bp	AE015451	NC_002947	Environ. Microbiol. 4 (12), 799-808 (2002)
***Vibrio vulnificus* CMCP6**	chromosome I: 3,281,945 bp, chromosome II: 1,844,853 bp	chromosome I: AE016795, chromosome II: AE016796	chromosome I: NC_004459, chromosome II: NC_004460	Unpublished
***Bradyrhizobium japonicum* strain USDA110**	9,105,828 bp	BA000040	NC_004463	USDA110 DNA (2002) In press
Buchnera aphidicola (Baizongia pistaciae)	615,980 bp	AE016826	NC_004545	Proc. Natl. Acad. Sci. U.S.A. 100 (2) 581-586 (2003)
***Xylella fastidiosa* Temecula1**	chromosome: 2,519,802 bp, plasmid: 1346 bp	chromosome: AE009442, plasmid: AE009443	chromosome: NC_004556, plasmid: NC_004554	J. Bacteriol. 185 (3), 1018-1026 (2003)
***Clostridium tetani* E88**	2,799,250 bp	AE015927	NC_004557	Proc. Natl. Acad. Sci. U.S.A. 100 (3) 1316-1321 (2003)
***Lactobacillus plantarum* WCFS1**	3,308,274 bp	AL935263	NC_004567	Proc. Natl. Acad. Sci. U.S.A. (2003) In press
***Tropheryma whipplei* str. Twist**	927,303 bp	AE014184	NC_004572	Unpublished
***Pseudomonas syringae* pv. tomato str. DC3000**	6,397,126 bp	AE016853	NC_004578	Unpublished
***Streptococcus pyogenes* SSI-1**	1,894,275 bp	BA000034	NC_004606	Unpublished
Vibrio parahaemolyticus	chromsome I: 3,288,558 bp, chromsome II: 1,877,212 bp	chromosome I: BA000031, chromosome II: BA000032	chromosome I: NC_004603, chromosome II: NC_004605	J. Clin. Microbiol. 38 (6), 2156-2161 (2000)
***Salmonella enterica* subsp. enterica serovar Typhi Ty2**	4,791,961 bp	AE014613	NC_004631	J. Bacteriol. 185 (7), 2330-2337 (2003)
***Bacteroides thetaiotaomicron* VPI-5482**	6,260,361 bp	AE015928	NC_004663	Science 299 (5615), 2074-2076 (2003)
***Enterococcus faecalis* V583**	3,218,031 bp	AE016830	NC_004668	Science 299 (5615), 2071-2074 (2003)

To view the genomes, select the Genomes database from the Entrez Home Page at **www.ncbi.nlm.nih.gov/Entrez**
Complete genomes are also available for downloading by FTP from **ftp.ncbi.nih.gov/genomes**

Entrez Quiz

Q. Find the NCBI contig (eg NT_011520) of which AC005006 is a component.

A. Search Entrez "nucleotides" with AC005006. Select "Assembly" in the Display pulldown and press the "Display" button.

Q. Generate a list of the components comprising NCBI contig NT_011520.

A. Start where you left off in the example above and choose "Components" in the Display pulldown. Then press "Display" to show the components of the contig.

Q. Retrieve the UniGene cluster IDs for a list of 50 human mRNA RefSeqs (eg. NM_130786).

A. Put a list of the accessions, one per line, into a text file. Upload this file using Batch Entrez at:

www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?db=Nucleotide

When the document summary giving links to each accession is returned, use the Display pulldown to select "UniGene Links". Then press the Display button to display the list of UniGene clusters corresponding to the list of RefSeqs.

Q. Beginning with a list of human mRNA RefSeq accessions, generate a list of corresponding protein RefSeq accessions.

A. Upload the mRNA accession list, as above. From the resulting document summary, choose "Protein Links" from the Display pulldown and press Display.

Q. Find all viral 3D structure records which include two DNA chains.

A. Search Entrez Structures with **"viruses[orgn] AND "2"[DNAChain Count]"**. Click on the "Preview/Index" link to see more fields.

Submitting a Segmented Set

A segmented set consists of a number of non-contiguous sequence blocks with a known order and orientation that are grouped together as a set on the basis of their physical proximity. Examples include a set of exon or intron sequences for a gene, or a set of internal transcribed spacers for a ribosomal RNA gene cluster. The submission of a segmented set to GenBank is a fairly straightforward procedure using NCBI's Sequin program.

In a typical scenario, one may have sequenced the six exons of a gene but not the introns. The sequences can be submitted as a segmented set in order to show that the exon sequences belong to one gene. The set can be used to show the relationship between the exons to form a joined mRNA and coding region using the locations of the exons from the individual "parts" entries.

Since GenBank represents each contiguous piece of DNA as one entry, six accession numbers for the "parts" of the set will be issued, however when the set is released into GenBank an additional accession number, beginning with the letters AH, is assigned to the set as a whole. Searches by gene name in Entrez, retrieve both the segmented entry and the six entries consisting of the individual exon "parts".

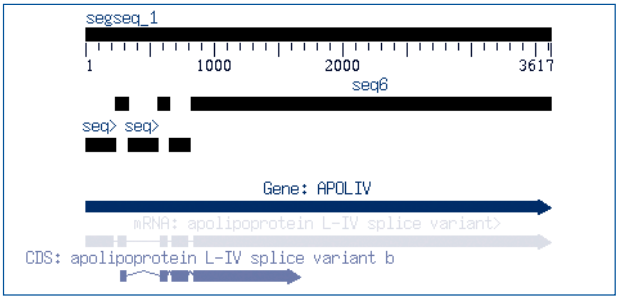


Figure 1: Graphical view of a segmented set in Sequin. The first bar from the top indicates that the entire group of sequences covers 3,617 bases. The next two bars show the positions of the six exon segments of the set. The lower three bars depict the gene, APOLIV, one of its transcripts, and the coding sequence derived from the transcript, respectively. The transcript variant shown lacks exon 3.

The Sequin graphical display of a segmented set of exons for an apolipoprotein L-IV variant which lacks exon 3 view appears in Figure 1. The segmented sequence is made up of six parts depicted in bars 2 and 3 from the top. The gene, shown next, spans all six exons . The mRNA variant, below, consists of the exons 1, 2, 4, 5 and 6. The coding region, shown at the bottom, spans the coding portions of the exons covered by the transcript.

Instructions on how to download and use Sequin are provided at:
www.ncbi.nlm.nih.gov/Sequin/index.html

Three Easy Steps to Submission

You can use Sequin to submit a segmented set comprised of exons in three easy steps.

STEP 1
Save the nucleotide sequences of the "parts" to a file as a set of catenated FASTA formatted sequences. For example:

>seq_first [organism=Homo sapiens] first exon
GAGGTGCTGGGGAGCA....

>seq_last [organism=Homo sapiens] last exon
CCCCTCTTTTCCTGCCCAAG....

STEP 2
Save the amino acid sequence of each of the conceptual translations resulting from all relevant combinations of exons in FASTA format to separate files. For example:

> [protein=apolipoprotein L-IV splice variant a]
MEGAALLKIFVVCIWVQQNHPGWTVAGQFQEKKRFTEEVIEYFQ...

STEP 3
In Sequin, choose "Start New Submission". Fill in the submission and contact information as usual, choose "Segmented sequence" in the "Sequence Format" panel and import the set of nucleotide sequences prepared as above by clicking on "Import Nucleotide FASTA".

To add the coding region information, click on "Annotate–Coding Region and Transcript-CDS". Then use "File-Import Protein FASTA" to import your protein translations one at a time. As each amino acid sequence is imported, Sequin calculates the correct coding region nucleotide locations with respect to the "parts". Add the protein name under "Protein-Name", then click on "Accept".

Add the mRNA feature using "Annotate–Coding Region and Transcript-mRNA". Add the name of the mRNA under "mRNA-Name". Under the "Location" tab, choose the appropriate "SeqId" and "Strand", and fill in the "From" and "To" information for each of the exons. You may wish to create the gene feature, by using "properties–Gene-New". Add the "Gene Symbol", for example APOLIV in this case. Finally, click on "Accept" and "Validate" to check if there are any errors. To submit your segmented set, save the record in Sequin and e-mail it to:
gb-sub@ncbi.nlm.nih.gov

Discontinuation of the GenBank Cumulative Updates

As announced in the Fall/Winter 2002 NCBI Newsletter and in the October and December 2002 GenBank release notes, the GenBank Cumulative Update (GBCU) files listed below are no longer supported:

gbcu.aso.gz	**gbcu.flat.gz**
gbcu.fsa.gz	**gbcu.gnp.gz**
gbcu.qscore.gz	**gbcu.fsa.gz**

These files were formerly available in the GenBank "daily" directory.

As the size of the GBCU has grown to nearly 20% of the total database size, the GBCU files have become increasingly difficult to download and use. The smaller incremental GbUpdate files, which are more practical to download and easier to process and are still found at:

ftp.ncbi.nih.gov/genbank/daily-nc

To facilitate the use of the incremental update files, NCBI furnishes a script for combining the incremental update files into a non-redundant cumulative update file. The script, called "build_gbff_cu.pl", can be downloaded by FTP at:

ftp.ncbi.nih.gov/genbank/tools

Details about the change are available in Section 1.3.2 of the GenBank release notes:

ftp.ncbi.nih.gov/genbank/gbrel.txt

GenBank Release 135: 31 Billion Base Pairs

GenBank release 135 (April 2003) contains over 24 million sequence entries totaling more than 31 billion base pairs. GenBank may be searched using NCBI's Entrez search and retrieval system or may be downloaded from the NCBI FTP site:

ftp.ncbi.nih.gov/genbank

GenBank flatfiles can be downloaded from the "genbank" directory; the ASN.1 version of GenBank is found in the "ncbi-asn1" directory. Uncompressed, the Release 135 flatfiles comprise 114 gigabytes. The ASN.1 version comprises roughly 93 gigabytes. Links to GenBank mirror sites are available on the GenBank FTP site at:

genbank.sdsc.edu/pub
bio-mirror.net/biomirror/genbank

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 3S308
8600 Rockville Pike
Bethesda, Maryland 20894

Official Business
Penalty for Private Use $300

FIRST CLASS MAIL
POSTAGE & FEES PAID
PHS/NIH/NLM
BETHESDA, MD
PERMIT NO. G-816