NCBI News

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

Spring 2000

Dazzling Graphics with Cn3D 3.0

The newly released Cn3D 3.0 provides dazzling graphics through the use of the OpenGL graphics library. Coupled with the graphics enhancements are several new rendering options and color schemes. The Cn3D sequence window has also been refined to allow easier selection of residues, and better depiction of secondary structural elements, protein domains, and hetero-atoms alongside the sequence.

Greater Graphics

Cn3D 3.0 depicts solid objects more realistically, taking advantage of highlighting and shadowing. New rendering options include an alpha-carbon backbone "worm", or "coil" representation at three thickness levels. Alpha-helices may now be represented either as blunt-ended cylinders or cylinders with carboxy-terminal "caps" to indicate the direction of the peptide chain. In addition, ions are depicted as translucent spheres. A new color scheme called Sequence Conservation offers several coloring options when multiple sequences are displayed in the sequence window.

This graphical power can be adjusted to allow for differences in computer speeds, video resolutions, or molecule size using a new Cn3D Quality control tab found under View/Drawing settings. Lower resolutions can be used to facilitate faster rotation speeds for larger molecules. Higher resolutions can be chosen when publication-quality figures are desired.

Enhanced Functionality in the Sequence Window

The Cn3D Sequence window operates in either a single or multiple sequence mode. If a single sequence is displayed, then the window is entitled OneD-Viewer and uses a set of options suitable for operations on a single sequence. If more than one sequence is displayed, the window is entitled DDV (DeuxD-Viewer) and a set of options suitable for multiple sequences is active. Columns of a multiple alignment may be colored according to sequence conservation, for example, with these colors mapped to both the sequence and structure windows.

OneD-Viewer Operation

The Sequence window has been given several new abilities in version 3.0. Secondary structural elements are now represented as 3-D-like cartoon images beneath the sequence. NCBI-assigned protein domains are

continued on page 3

BLAST Now Offers Taxonomic Views of the Output

NCBI's Basic and Advanced BLAST services now offer the option of returning a taxonomically organized report. Clicking on the Taxonomy Reports link on the BLAST results page will generate taxonomy reports in three formats: a Lineage Report, an Organism Report, and a Taxonomy Report. Together, these three reports provide a broad overview of the taxonomic relationships among the records returned from a BLAST search.

The Lineage Report

The Lineage Report gives a simplified view of the relationships between the organisms generating database hits to the query sequence

continued on page 8

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 8N-803
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 496-2475
Fax: (301) 480-9241
E-mail: info@ncbi.nlm.nih.gov

Editors
Dennis Benson
Barbara Rapp

Contributors
Scott Federhen
Donna Maglott
Jim Ostell
Lukas Wagner

Writer
David Wheeler

Editing, Graphics, and Production
Marla Fogelman
Veronica Johnson
Jennifer Vyskocil

Design Consultant
Gary Mosteller

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.

The contents of this newsletter may be reprinted without permission. The mention of trade names, commercial products, or organizations does not imply endorsement by NCBI, NIH, or the U.S. Government.

NIH Publication No. 00-3272

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

HomoloGene: Clusters of Clusters

HomoloGene is a new NCBI database of both curated and calculated orthologs and homologs for the human, mouse, rat, and zebrafish genes represented in UniGene and LocusLink.

Curated orthologs include gene pairs from the Mouse Genome Database (MGD) at the Jackson Laboratory, the Zebrafish Information (ZFIN) database at the University of Oregon, and from published reports. Computed orthologs and homologs are identified from nucleotide sequence comparisons between all UniGene clusters for each pair of organisms. Calculated orthologs and homologs may be considered putative because they are based only on sequence comparison.

Computed similarities are detected using BLAST to compare nucleotide sequences for each pair of organisms, and to identify those sequence pairs that share the greatest degree of nucleotide sequence similarity. The best match for a sequence in one organism to a sequence in a second organism is based on the percentage of identical sequence, called the %ID in the HomoloGene report, for an alignment of a minimum of 100 base pairs. When sequences from two UniGene clusters are reciprocal best matches, the UniGene clusters corresponding to the pair of sequences are considered to represent a putative ortholog pair.

HomoloGene also contains a set of triplet ortholog clusters in which orthologous clusters in two organisms are also orthologous to the same cluster in a third organism. For the organisms human, mouse, and rat, there are currently over 7,000 of these triplets. For the organisms zebrafish, human, and rodent (mouse or rat), there are currently just over 200 triplets.

HomoloGene Search

To obtain a HomoloGene report for a gene, use the Query box at the top of the HomoloGene page to search using a UniGene ClusterID, LocusLink LocusID, gene symbol, gene name, nucleotide accession number, or any free text appearing in UniGene cluster titles. The HomoloGene report consists of a header section, followed by reports falling within any of three sections entitled Curated Orthologs, Calculated Orthologs, and Mutally Orthologous Pairs.

The header section gives the title of the HomoloGene cluster, followed by a listing of all the possible orthologs contained within it. For each entry, the UniGene cluster ID and the LocusLink ID are given.

The Curated Orthologs section gives pairs of orthologs and, for each pair, a link to the source in which the orthologous relationship is claimed. If the source is a research paper, the link is to a PubMed abstract. In other cases, the source may be MGD or ZFIN, and links are provided to these resources.

The Calculated Orthologs section gives a listing of putative ortholog pairs identified on the basis of sequence similarity. For each pair

continued on page 8

Dazzling Graphics with Cn3D

continued from page 1

indicated using solid 2-D arrows. Hetero-atoms in the features list are indicated in the sequence display as small triangles positioned near the residues that are closest to them in the 3-D structure. As in earlier versions of Cn3D, additional sequences may be aligned to a structurally anchored sequence by importing FASTA-formatted files or by downloading directly from Entrez over the network.

DDV Viewer Operation

When a sequence alignment is shown in the Sequence window, a different set of options is available. The mouse mode may be changed, using the Options menu, from the default of Select to Query in order to use the mouse to identify residue numbers. A Styles panel, also invoked from the Options menu, allows several sequence-alignment display parameters to be adjusted. These include the use of color and the placement of ruler lines.

The enhanced graphical capabilities of Cn3D 3.0 are illustrated in the views of Human Phenylalanine Hydroxylase (PDB code 1PAH) in Figures 1 and 2. The structure window shows an image of the protein in which the NCBI-defined secondary structural elements, helices and strands, are represented as solid cylinders and planks, respectively, with the pointed ends indicating the amino- to carboxy-terminal direction. An iron atom is seen in the center of the structure, represented as a translucent sphere. Using the Annotation panel, the sites of documented mutations, as taken from Online Mendelian Inheritance in Man (OMIM), are marked by showing these amino acid side-chains in a “fat-tube” representation.

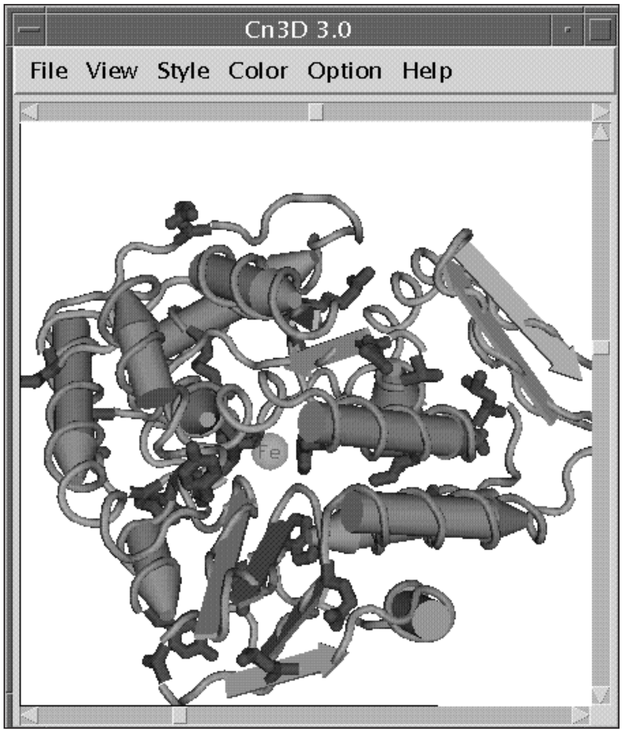
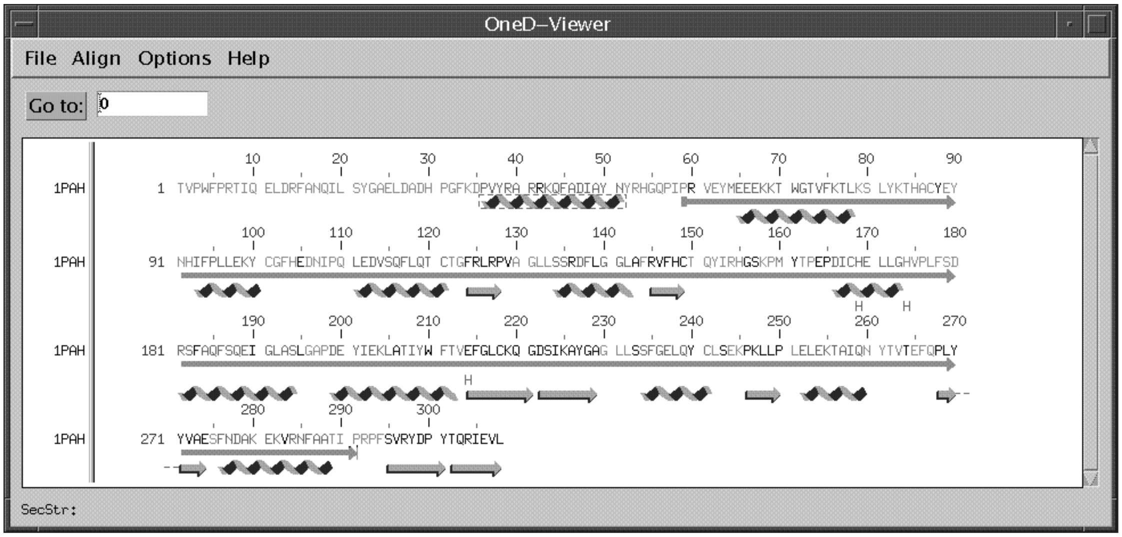


Figure 1: *Cn3D Structure window view of human phenylalanine hydroxylase (PDB code 1PAH).*

The accompanying Sequence window shows these mutable residues in a dark shade (red). Secondary structural elements are shown below the sequence. The single NCBI-derived protein domain for this structure is indicated by the arrow that runs from residue 59 to residue 291. Residues proximal to the iron atom in the structure are marked with an H under the sequence.

For more information, navigate to www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.html.
—*DW*



```
                10        20        30        40        50        60        70        80        90
1PAH     1 TVPWFPRTIQ ELDRFANQIL SYGAELDADH PGFKDPVYRA RRKQFADIAY NYRHGQPIPR VEYMEEEKKT WGTVFKTLKS LYKTHACYEY

               100       110       120       130       140       150       160       170       180
1PAH    91 NHIFPLLEKY CGFHEDNIPQ LEDVSQFLQT CTGFRLRPVA GLLSSRDFLG GLAFRVFHCT QYIRHGSKPM YTPEPDICHE LLGHVPLFSD
                                                                                          H         H
               190       200       210       220       230       240       250       260       270
1PAH   181 RSFAQFSQEI GLASLGAPDE YIEKLATIYW FTVEFGLCKQ GDSIKAYGAG LLSSFGELQY CLSEKPKLLP LELEKTAIQN YTVTEFQPLY
                                          H
               280       290       300
1PAH   271 YVAESFNDAK EKVRNFAATI PRPFSVRYDP YTQRIEVL
```

Figure 2: *Cn3D Sequence window view of human phenylalanine hydroxylase (PDB code 1PAH).*

News Briefs

LocusLink Adds Organisms and Search Fields

Two new organisms have been added to LocusLink: the zebrafish, *Danio rerio,* and the fly, *Drosophila melanogaster.* The data for *Drosophila* include all genes thus far identified in the recent deposition of the *Drosophila* genome. Links to the FlyBase database of *Drosophila* genome annotations are provided for these genes.

Two new Boolean qualifiers have also been added to LocusLink. The first is termed "disease_known", whereas the second is termed "has_seq". These qualifiers provide the ability to search for records with associated diseases or sequence data. Together, these qualifiers can be used, for example, to determine how many disease genes have been sequenced using the search:

```
disease_known AND has_seq
```

This query returns more than one thousand records.

GenBank Gets Billion Base Boost With Version 118.0

GenBank release 118.0 is now available via FTP from NCBI. Uncompressed, the release 118.0 flat files require roughly 29 megabytes (MB) of disk space. The ASN.1 version of the GenBank sequence data requires roughly 24 MB. Release 118.0 comprises 8,604,221,980 base pairs, 7,077,491 sequences, and represents a growth of 1.23 billion base pairs over release 117.0. Pick up the latest GenBank at ftp://ncbi.nlm.nih.gov/genbank/.

News Briefs

1.5 Million Bases of Bad Bug Sequenced

The complete 1.5 million base pair genomic sequence of the motile, gram-negative bacterium, *Campylobacter jejuni* (Feb 10, *Nature*), has been deposited in GenBank and can be viewed in Entrez Genomes. *C. jejuni* has earned a place in the FDA's "Bad Bug Book" by being the bacterium most frequently associated with diarrhea in the United States. The Bad Bug is linked to 45% of diarrhea cases, leading *Salmonella* (30%), *Shigella* (17%), and *E. coli* (5%). Unchlorinated water and uncooked meat are the most common sources of contamination.

Take a look at the genome on the NCBI Bacterial Genomics page at www.ncbi.nlm.nih.gov/PMGifs/Genomes/eub.html.

BLAST Version 2.0.13 is Released

The latest version of the BLAST suite of programs, version 2.0.13, is now available on the NCBI FTP site at ftp://ncbi.nlm.nih.gov/blast/.

Archives of executables of standalone BLAST are found at this site within the "executable" directory. The BLAST network client is found within the "network" directory.

The major enhancement of this release is that Bl2seq, the standalone version of the Web-based, BLAST2Sequences, can now perform nucleotide-protein (blastx style) comparisons. In this new version of Bl2seq, the "-p" option accepts the arguments "blastn", "blastp", or "blastx" rather than the Boolean "T" or "F" of prior versions.

Fly Genome Deposited in GenBank

The *Drosophila* genome sequence described in *Science* (Mar 2000; 287, 2185–95) is publicly available in GenBank. The genome, sequenced jointly by scientists at Celera Genomics and the Berkeley *Drosophila* Genome Project (BDGP), was deposited in the form of more than 700 "scaffolds" consisting of annotated, ordered sets of contigs having gaps of known length. Approximately 134 scaffolds, representing 98% of the genome, have been mapped to a chromosome and these mappings appear in the GenBank record with a "/chromosome" qualifier in the source feature field.

Many scaffolds contain unsequenced gaps of known length, indicated by the appropriate number of Ns. The sequence spanning the gaps will be filled, primarily by the BDGP, from one end of the genome to the other, in the coming months. As each chromosome is completed, it will be completely reannotated, reviewed, and updated with a complete, consistent set of annotated records. In the meantime, the BDGP plans to provide regular unannotated updates to NCBI of the most recent versions of the scaffold sequences. These updates will be available from the NCBI FTP site. NCBI will also align the annotated GenBank records to the most recent unannotated versions of the scaffolds in order to combine the annotation available from the initial release with the most recent sequence data.—*JO, DW*

Drosophila Finds a Home Page at NCBI

NCBI has combined a gateway to the complete genomic sequence of the fruit fly, an array of links to related *Drosophila* resources such as FlyBase and GadFly, and a mix of pre-computed analyses conducted at NCBI, to produce a special *Drosophila* Web page to serve as a springboard for the analysis of the fly genome sequenced jointly by Celera Genomics and the Berkeley *Drosophila* Genome Project.

The Genome Map Viewer

The genome of the fruit fly is accessible through five chromosome links, each of which invokes NCBI's new Genome Map Viewer to provide access to the sequence data. Users can select from five different maps, which run the gamut from physical maps showing chromosomal banding patterns, to sequence maps with links to GenBank records. Using a query box at the top of the main *Drosophila* page, it is possible to perform text searches followed by protein neighbor searches in which the results are graphically mapped onto the *Drosophila* chromosomes.

NCBI Analysis

Pre-computed BLAST comparisons between protein sequences derived from the *Drosophila* sequence and all protein sequences in the BLAST non-redundant database can be accessed through a link on the *Drosophila* page.

In addition, a one can generate a 2-dimensional dot plot showing the similarity of the set of *Drosophila* proteins to the protein sets of any two organisms or taxa. A Related Structures link on the *Drosophila* page leads to a listing of *Drosophila* protein sequences with significant similarity to the sequences of proteins with known structure. Using NCBI's Cn3D macromolecular viewer it is possible to view 3-D images of the *Drosophila* protein sequences superimposed upon protein structures to which they bear sequence similarity.

The *Drosophila* Web page can be reached through a link on NCBI's Genomic Biology Web page at www.ncbi.nlm.nih.gov/Genomes/index.html.—*DW*

A Map Viewer for the Human Genome

NCBI's Human Genome Map Viewer can display up to seven parallel chromosomal maps simultaneously. The maps displayed can be selected from a set of 19, and include cytogenetic maps, such as G-banding chromosomal ideograms, sequence-based maps, such as those showing contig and clone information, and radiation hybrid maps, such as the G3 and GB4 maps used to construct GeneMap '99. To take a broad view of the human genome click on the Map Viewer link on NCBI's Human Genome Resources web page at www.ncbi.nlm.nih.gov/genome/guide/.

Selected Recent Publications by NCBI Staff

Galperin, MY, L Aravind, and **EV Koonin.** Aldolases of the DhnA family: a possible solution to the problem of pentose and hexose biosynthesis in archaea. *FEMS Microbiol Lett* 183(2): 259–64, 2000.

Kirsch, IR, ED Green, R Yonescu, R Strausberg, N Carter, D Bentley, MA Leversha, I Dunham, VV Braden, E Hilgenfeld, **GD Schuler, AE Lash**, GL Shen, M Martelli, WM Kuehl, RD Klausner, and T Ried. A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nat Genet* 24(4):339–40, 2000.

Makarova, KS, **YI Wolf**, O White, K Minton, and MJ Daly. Short repeats and IS elements in the extremely radiation-resistant bacterium *Deinococcus radiodurans* and comparison to other bacterial species. *Res Microbiol* 150 (9-10):711–24, 1999.

Panchenko, AR, A Marchler-Bauer, and **SH Bryant.** Combination of threading potentials and sequence profiles improves fold recognition. *J Mol Biol* 296(5):1319–31, 2000.

Pickeral, OK, W Makalowski, MS Boguski, and JD Boeke. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res* 10(4):411–5, 2000.

Pruitt, KD, KS Katz, H Sicotte, and **DR Maglott.** Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* 16(1):44–7, 2000.

Schaffer, AA, YI Wolf, CP Ponting, EV Koonin, L Aravind, and **SF Altschul.** IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* 15(12):1000–11, 1999.

Sherry, ST, M Ward, and **K Sirotkin.** Use of molecular variation in the NCBI dbSNP database. *Hum Mutat* 15(1):68–75, 2000.

Q & A

Frequently Asked Questions

Q.

A.

In PubMed, how can I restrict my search to an author name?

Click on the Preview/Index link in the grey area beneath the query box. Choose Author Name from the Index list box and type the author name into the text box to the right. Click on the AND button to add the author name to your query.

Why do I get the message "ERROR: Blast: No valid letters to be indexed"?

You may have accidentally entered an accession number in the search box without changing the input selection from Sequence in FASTA format to Accession or gi. You will also see this error message if your query contains invalid characters or a large number of ambiguity codes (R,Y,K,W,N, etc. for nucleotide). Although BLAST allows ambiguity codes, be aware that these will always contribute a negative value to nucleic acid alignment scores.

How are symbols and gene names selected for LocusLink and RefSeq records?

Both LocusLink and RefSeq use gene symbols and gene names established by the nomenclature committee appropriate for the genome.

How can I see the sequence alignments in GenBank?

From the Entrez Nucleotides search screen, click on the Limits button, and then select Properties from the search field list box. Type in "seqalign present" in the search box and press Go. To view an alignment, follow the Popset link to the right of the Entrez summary.

Is there a way to download the proteome of an organism from GenBank?

Yes, if the organism is represented in Entrez Genomes. Navigate to the Entrez Genomes summary page for the organism in question and click on the Protein Coding Genes link in the Feature table. A table will be displayed giving the protein coding regions of the genome and links to corresponding protein sequences. This table, as well as the corresponding protein sequences, can be saved to a local file. Entrez Genomes is found at www.ncbi.nlm.nih.gov/Entrez/Genome/main_genomes.html.

Where can I find tutorials on the use of the NCBI Web services?

A set of NCBI tutorials as well as other educational material can be found under the Education link on the NCBI home page.

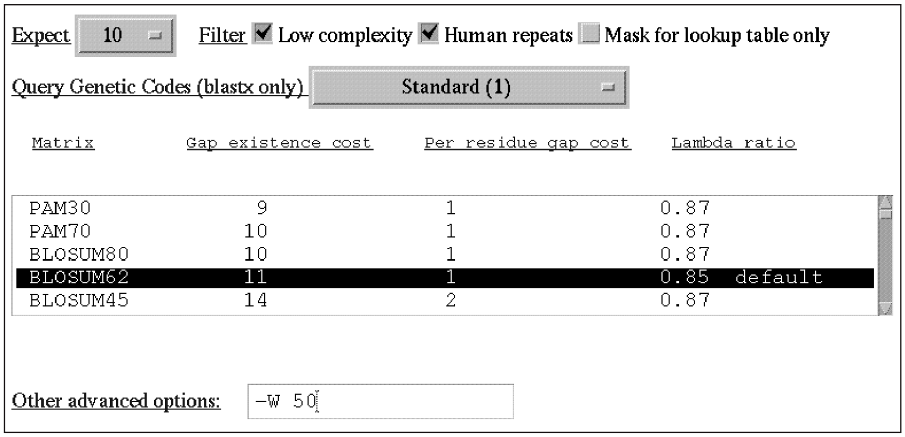
BLAST Lab

```
caaatccgttcttgatcgtacatagcgcatgtcagncaaatccgttcttg
|||||| ||||||||||||||| || |||||||| ||||||| ||||||
caaatccattcttgatcgtacatggcacatgtcagtcaaatccattcttg
```

How to BLAST Using Very Large Nucleotide Queries

As an era of prolific genome sequencing begins, the need arises to run BLAST searches using very large query sequences that often reach hundreds of kilobases in length. Because of its speed and flexibility, the BLAST algorithm can rise to the occasion and perform these searches quickly.

One of the contributors to the speed and flexibility of BLAST is the fact that BLAST matches multi-character "words" between the query and database sequences rather than single characters. By default, these matches need not be perfect, although a scoring threshold for reporting matches can be adjusted. This search strategy offers a tradeoff between speed and sensitivity; smaller word-sizes result in greater sensitivity at the expense of speed while larger word-sizes optimize BLAST for speed.



Expect 10 Filter ☑ Low complexity ☑ Human repeats ☐ Mask for lookup table only

Query Genetic Codes (blastx only) Standard (1)

Matrix	Gap existence cost	Per residue gap cost	Lambda ratio	
PAM30	9	1	0.87	
PAM70	10	1	0.87	
BLOSUM80	10	1	0.87	
BLOSUM62	11	1	0.85	default
BLOSUM45	14	2	0.87	

Other advanced options: -W 50

Figure 1: *Portion of the Advanced BLAST page demonstrating the use of the "Advanced Options" box and filtering options.*

The default word-size for BLASTn searches is 11, which allows untranslated nucleotide searches to proceed at a rapid pace with a degree of sensitivity appropriate for queries in the range of 1 to 50 kb. A word-size of 11, however, is too small to facilitate rapid searches with queries of 100 kb or more. Fortunately BLAST is flexible enough to allow the word-size to be adjusted upward indefinitely to accommodate ever larger query lengths. For a query of 100 kb, a word-size of 30 to 50 will allow a Web-BLAST search to run to completion and return informative database matches.

Filtering parameters may also be adjusted to facilitate BLAST searches with large queries. Because repetitive sequence within the query leads to a repetitive and relatively uninformative series of database hits, BLAST masks simple, repetitive sequence by default. Repetitious hits waste computational time, and the larger the query sequence, the larger the potential problem—so repeat filtering should always be used with large query sequences. In the case of human sequences, the human repeat filtering option of Advanced BLAST should be used to mask more complex varieties of repeat found in human sequences. Both simple repeat and human repeat filtering are activated using check boxes on the Advanced BLAST Web page. To run a BLASTn search using a word-size of 50 and filtering the query for human repeats, type "-W 50" into the Advanced BLAST Advanced Options box and check the appropriate filtering options as shown in Figure 1.

As an example, a BLASTn search of the default nr nucleotide database with a 400 kb contig from human chromosome 22 will time-out using Advanced BLAST if the default parameters are used. However, if the word-size is changed to 200, and human repeat filtering activated, the search is completed within 15 minutes!

Searches with very large sequences may also be performed using BLAST2Sequences if the word-size is increased sufficiently. Using the default word-size of 11, alignments with query sequences on the order of 150 kb in length will not generally be completed. However, by increasing the word-size to 50, two sequences as large as 250 kb apiece may be aligned. The word-size may be changed on the BLAST2Sequence page using the input box provided.

The BLAST Lab feature is intended to provide detailed technical information on some of the more specialized uses of the BLAST family of programs. Topics are selected from the range of questions received by the BLAST Help Group.

BLAST

continued from page 1

by showing how closely these organisms are related to a “focus organism”, according to the taxonomy database. This focus organism is the organism giving the strongest BLAST hit and this will often be the source organism of the query sequence.

The Organism Report

In the Organism Report, the BLAST results are grouped into blocks by species. Within each species block, the records are sorted by BLAST score. The order of species blocks themselves is based on the BLAST score of the best hit within the block.

The Taxonomy Report

The Taxonomy Report summarizes the relationships among all of the organisms found in the BLAST results. Using this report, it is easy to see how many records are found within broad taxonomic groups such as the mammalia, or the archaea.

For detailed information on the interpretation of the BLAST taxonomy reports, see the Help document at www.ncbi.nlm.nih.gov/blast/taxblasthelp.html. The new taxonomic output format is available via Basic or Advanced BLAST at www. ncbi.nlm.nih.gov/BLAST/.
—SF, DW

HomoloGene

continued from page 2

listed, a %ID score is given as a measure of reliability. If the gene in question is a member of a triplet ortholog cluster, the Mutally Orthologous Pairs section shows the triplet.

FTP Access to Data

The current datasets for the calculated orthologs and homologs and the mutally orthologous pairs are available via FTP at ftp://ftp.ncbi.nlm.nih.gov/pub/HomoloGene/. Link to the HomoloGene Web page from the UniGene page at www.ncbi.nlm.nih.gov/UniGene/.
—LW, DW

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