NCBI News

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

Summer 2000

Conserved Domain Database Debuts with RPS-BLAST Search Interface

Proteins often contain several domains, each with a distinct function. Such domains have evolved as modules that are combined in various arrangements to produce proteins of unique function. Conserved domains are structural modules that have been reused frequently during the process of evolution. NCBI's new Conserved Domain Search (CD-Search) service can be used to identify conserved domains in a protein sequence. The service provides a Web interface for searching NCBI's new Conserved Domain Database (CDD), with the Reverse Position-Specific BLAST program (RPS-BLAST), and retrieving domain alignments including 3-D structures.

The CDD contains domains derived principally from two public protein domain collections, the Simple Modular Architecture Research Tool (SMART)[1] and Pfam,[2] which include collections of multiple sequence alignments for the conserved domains they contain.

To produce the CDD, alignments from SMART and Pfam are processed to provide links from each sequence in the alignment to the protein division of Entrez. Sequences that cannot be found in Entrez databases are either omitted or replaced with a closely related sequence. Whenever possible, non-structurally anchored sequences in the alignment are replaced with closely related sequences that have direct links to 3-D structures.

From the sequences in the alignment for a domain, a representative sequence, preferably with a structure link, is chosen. A PSI-BLAST (Position-Specific Iterated BLAST)[3] type Position-Specific Score Matrix (PSSM) is then calculated from the multiple sequence alignment for the aligned range of the representative sequence. The CD-Search service next uses the RPS-BLAST algorithm to search the resulting databases of PSSMs and identify conserved domains in a protein sequence.

RPS-BLAST is a variant of the PSI-BLAST program. Whereas PSI-BLAST first builds a PSSM that is used as a query for subsequent database searches, RPS-BLAST uses a protein sequence query to search a database of precalculated PSSMs in a single pass. The role of the PSSM has changed from "query" to "subject", hence the term "reverse" in RPS-BLAST.

continued on page 3

Enhanced Access to Taxonomy Database

Several new search and display features enhance the utility of the NCBI Taxonomy database, which has also become a formal component of the Entrez set of integrated databases. Date-bounded queries are now supported, allowing users to identify all species that have been added since a particular date. A search option allows a phonetic spelling, such as "zenopis", of an organism name, then presents a list of candidates from which the correct spelling, "xenopus", can be chosen. Another convenient feature is a direct link from the

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
Steve Bryant
Deanna Church
Alex Lash
Aaron Marchler-Bauer
Renata McCarthy
Scott McGinnis
Jim Ostell
Vyvy Pham
Monica Romiti

Writer
David Wheeler

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

Design Consultants
Gary Mosteller
Tim Cripps

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. 00-3272

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

New Human-Mouse Homology Map

A new version of the Human-Mouse Homology Map is now online. The new map differs from the Davis Human/Mouse Homology Map in that it is produced by integrating orthologs identified at The Jackson Laboratory with putative orthologs identified by sequence homology. The latter step involves the identification of orthologous human/mouse sequence pairs by BLAST comparisons using non-EST mRNA.

To identify human-mouse sequence pairs, two BLAST analyses are performed: one in which the human mRNA is the query and one in which the mouse mRNA is the query. Only pairs exhibiting reciprocal best BLAST scores are included in this analysis. Pairs for which the identification of a reciprocal best hit is not unambiguous, or pairs identified by BLAST that are in conflict with curated homology information are reviewed before being included on the map.

Several new features are offered, including reporting representative STSs associated with the loci on the map, linking human cytogenetic locations to NCBI's MapViewer, providing links to alignments of representative sequences via BLAST2Sequences, and linking gene symbols to LocusLink reports.

The page display style for the new map is determined by the chromosome that is selected as the master. For example, when beginning with a mouse chromosome as the master, genes are ordered based on mouse genetic positions as determined by the Mouse Genome Database (MGD). When a human chromosome is the master map, gene order is determined first by available sequence information, and then by cytogenetic location. Genes with unknown chromosomal positions are grouped at the bottom of the page. In addition, an effort has been made to virtually map genes in which the position of one ortholog is known, but the other is unknown. These genes are indicated by the hatched bars. *—DW*

Catch the Gene Expression Omnibus

To support the public availability of gene expression data, NCBI has launched the Gene Expression Omnibus (GEO), a repository for expression data from any organism or artificial source. Many types of gene expression data from platforms such as spotted microarray, high-density oligonucleotide array (HDA), hybridization filter, and serial analysis of gene expression (SAGE) are now being accepted, assigned accession numbers, and archived as a public data set. A series of precomputed definitions and descriptions of the data, as well as online tools for interactive retrieval and analysis, will follow. Visit the GEO project page for updates and further details at www.ncbi.nlm.nih.gov/geo/. *—AL, DW*

Conserved Domain Database

continued from page 1

A link to the CD-Search tool is found on the main BLAST page. The search form accepts an amino acid query sequence as input. The query sequence is compared, using RPS-BLAST, to the CDD database of PSSMs. Search results may be displayed as pairwise alignments of the query sequence with a representative domain sequence, as shown in Figure 1, for the representative sequence for the aminotransferase class-I domain. The extent of the alignment is illustrated graphically at the top of the output and links to Pfam are given in the RPS-BLAST summary header. From this output page, a multiple sequence alignment may then be generated between the query and other representatives of the aminotransferase class-I domain shown in Figure 2. If a 3-D structure exists for the domain sequence, it may be viewed using Cn3D. In this example, a structure does exist. Cn3D would load the 1B8G_B structure into its structure window and the multiple sequence alignment into its linked sequence window. In this manner, the structural context of the query sequence, implied by the alignment, could be examined in detail.

The source databases used in the CDD are updated several times a year, in roughly bimonthly intervals. NCBI follows these updates and will adjust the CDD with not more than two months' delay. Try a CD-search at www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi. *—DW*

Notes

1. Bateman, A, et. al. *Nucleic Acids Res* 28:263–6, 2000.
2. Schultz, J, et. al. *Nucleic Acids Res* 28:231–4, 2000.
3. Altschul, SF, et. al. *Nucleic Acids Res* 25:3389–402, 1997.

Figure 1: *CD-Search display showing alignment between a query and the representative sequence from CDD for aminotransferase class-I domain defined in the Pfam database.*

Figure 2: *Multiple sequence alignment between a CD-Search query sequence and four representatives of the aminotransferase class-I domain in the CDD database.*

A Pair of Pathogens Added to GenBank

Toxin Toolkit of Vibrio Cholerae *Revealed*

Cholera is a disease arising from contaminated water supplies; in the United States, contaminated shellfish, eaten raw, are the major source of infection. The causative agent, *Vibrio cholerae,* attaches itself to the brush border of the villous absorptive cells of the small intestine of its victim. There it produces the cholera toxin, a multimeric protein including 5 subunits arranged as a doughnut, shown in Figure 1. The toxin is internalized by the intestinal cells where it activates G-proteins leading to massive fluid and electrolyte loss. Control of the production of the cholera toxin is via the ToxR transcription factor, which activates the operon encoding the cholera protein. This operon also contains the ToxT gene, whose product activates several other virulence factors. The complete genome of *V. cholerae,* encoding its toxin toolkit, can be examined now in Entrez Genomes. From www.ncbi.nlm.nih.gov/Entrez, select the Genomes database. *—DW*

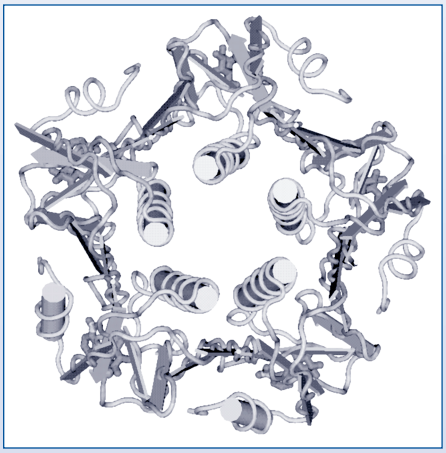

Figure 1: *Cn3D view of cholera toxin B-Pentamer complexed with Metanitrophenyl-Alpha-D-Galactose (MMDB entry 12741, PDB code 1EEI).*

View the Blueprint for a Motile, Hardy Bug

Pseudomonas aeruginosa is an opportunistic pathogen that is extremely resistant to disinfectants and antibiotics used to control it—the bug can live quite happily on a bar of soap. At 6.26 megabases, it is the largest bacterial genome sequenced to date.

P. aeruginosa is also highly motile due to a flagellum that it builds with the help of a couple dozen genes. Using this flagellum, *P. aeruginosa* can travel to places in the body that other microbes do not normally reach, such as the respiratory sinuses, where it can inflict lethal damage. In fact, *P. aeruginosa* causes a wide range of infections in people who are ill or have damaged immune systems, and is particularly dangerous to people with cystic fibrosis or on respirators. About 80% of treated infections are still fatal.

The microbe has 5,565 protein-coding genes, including the blueprint for its flagella, within a six-million-base genome that can now be explored via Entrez Genomes. From www.ncbi.nlm.nih.gov/Entrez, select the Genomes database.

The complete sequence can be found at ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/Paer/. *—DW*

Protein Molecular Weight Field Now in Entrez

A Molecular Weight search field for proteins is now available in Entrez. This data will be particularly useful to the mass spectrometry research community for which the mass of a protein provides the means of its identification.

As with the Length field in Entrez, the numerical search term must be 6 digits long and must be padded with leading zeros if necessary. For instance, the syntax below can be used to specify a range of molecular weights from 65,000 to 75,000:

```
065000:075000 [Molecular Weight]
```

By default, protein molecular weights included in Entrez are calculated for the whole protein, excluding initial methionine residues. However, if a signal peptide is annotated in the database record, the molecular weight is calculated for the remainder of the protein, after excluding the signal sequence. If cleavage products are annotated, a molecular weight is calculated for each individual cleavage product, rather than for the whole protein. Ambiguous amino acids are treated as one of their possible forms. Molecular weights are not computed for proteins containing unknown amino acids. *—JO, DW*

NCBI acknowledges Dr. Lewis Pannell for his help in developing parameters for protein molecular weight calculation.

OMIM in Entrez: New Searching Power

The Online Mendelian Inheritance in Man (OMIM) database has been integrated into the Entrez suite of databases. This offers more powerful searching and flexibility, and takes advantage of Entrez's extensive cross-referencing to find related records in other Entrez databases. There are links to other related resources as well, such as LocusLink, UniGene, locus-specific databases, and more.

While preserving the functions of the previous OMIM search system, Entrez provides a range of new capabilities, including a Limits function that easily restricts searches by chromosome number, search field, or various record attributes. It is also possible to display multiple records in a number of formats, store selected records in the Clipboard, and combine the results of two or more searches using the History function.

For example, to find OMIM records associated with cancer on chromosome 20, enter the term "cancer" in the search box and check chromosome 20 on the Limits page, accessible from the gray bar under the search box. This search retrieves records containing the search term "cancer" in any field. To retrieve only records containing the term "cancer" in their titles, check the box for the Title field on the Limits page as well.

From the search results page, records can be displayed individually or as a group. In either case, links beside the MIM number lead to related entries in OMIM and other Entrez databases. The LinkOut option leads to associated records in a variety of resources at NCBI, such as LocusLink, and elsewhere. When each entry is displayed individually, a customized blue sidebar is shown, providing a table of contents for that entry. When applicable, colored buttons provide additional links to UniGene, RefSeq, GenBank, and locus-specific databases.

The Help document and FAQ provide additional details, examples, and search tips. Links to related NCBI databases and allied resources are also provided. To try the new search interface, select OMIM from the Entrez home page at www.ncbi.nlm.nih.gov/Entrez. *—RM, BR*

Web Server Software Available for BLAST

NCBI now offers Web server packages that support Advanced BLAST, PHI-BLAST, and PSI-BLAST. Available for Unix and Linux platforms, these packages can be installed in minutes and provide the NCBI BLAST interface, including the popular graphical overview and taxonomically organized outputs. These packages also support batch searches and XML output. Custom BLAST databases, processed with formatdb (included), can be added to the database pull-down menu by performing simple edits to two ASCII files. Download the Web BLAST server package from ftp://ncbi.nlm.nih.gov/blast/server/current_release/. *—DW, SM*

Selected Recent Publications by NCBI Staff

Bogan JA, **DA Natale,** and ML De Pamphilis. Initiation of eukaryotic DNA replication: conservative or liberal? *J Cell Physiol* 184(2):139–50, 2000.

Galperin, MY, and **EV Koonin**. Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol* 18(6):609–13, 2000.

Koonin, EV, L Aravind, and **AS Kondrashov.** The impact of comparative genomics on our understanding of evolution. *Cell* 101(6):573–6, 2000.

Koonin, EV, YI Wolf, and **L Aravind.** Protein fold recognition using sequence profiles and its application in structural genomics. *Adv Protein Chem* 54:245–75, 2000.

Luhn K, **W Makalowski,** and J Brosius. A tRNA pseudogene in the archaeon *Methanococcus jannaschii. DNA Seq* 11(1-2):97-9, 2000.

Pickeral, OK, JZ Li, I Barrow, **MS Boguski, W Makalowski,** and **J Zhang.** Classical oncogenes and tumor suppressor genes: a comparative genomics perspective. *Neoplasia* 2(3): 280–6, 2000.

Ponting, CP, J Schultz, RR Copley, MA Andrade, and P Bork. Evolution of domain families. *Adv Protein Chem* 54:185–244, 2000.

Schriml, LM, and M Dean. Identification of 18 mouse ABC genes and characterization of the ABC superfamily in *Mus musculus. Genomics* 64(1): 24–31, 2000.

Wang, Y, LY Geer, C Chappey, JA Kans, and **SH Bryant.** Cn3D: sequence and structure views for Entrez. *Trends Biochem Sci* 25(6):300–2, 2000.

Wolf, YI, NV Grishin, and **EV Koonin.** Estimating the number of protein folds and families from complete genome data. *J Mol Biol* 299(4):897–905, 2000.

GenBank Gzipped for Ease of Transfer

GenBank releases and update files are now gzipped rather than Unix-compressed. The gzip format is similar to the common PC Zip format and delivers the greater degree of data compression required for the rapidly growing GenBank files. Gzipped files, which can be recognized by their extension of "gz", can be unzipped using the PC, Unix, or Linux command:

```
gunzip genbank_file.gz
```

The gunzip program is freely available for the Unix, Linux, PC, and Macintosh platforms.

LocusLink Adds Function Information via GeneRIF

LocusLink reports now display a link to a new tool called Gene References into Function, or GeneRIF. GeneRIF is designed to allow researchers to submit information about the functions of genes directly to LocusLink via the Web. Data submitted through GeneRIF must be accompanied by a reference to a publication in PubMed. The GeneRIF submission form contains a field for the PMID of the reference and a free-text field into which functional information about a gene can be entered. For more information about GeneRIF, see www.ncbi.nlm.nih.gov/LocusLink/GeneRIFhelp.html.

Mouse, Rat, Zebrafish Have Their Own Genomes Pages

New special genome resource pages, similar to those for human and *Drosophila,* have been implemented for three more model organisms; the rat, mouse, and zebrafish. Links to all genome resource pages can be found under the Genomic Biology link on NCBI's home page.

Entrez Displays dbEST, dbGSS Records

Searches of dbEST and dbGSS, can now be made using the Entrez search interface. This change allows searchers of dbEST or dbGSS to take advantage of the powerful Entrez Nucleotides search and formatting capabilities. Results are returned using the dbEST/dbGSS display formats, although other formats can be selected in the usual manner within Entrez Nucleotides.

NCBI Conferences

As part of our expanding efforts to provide hands-on education and training to users of our services, NCBI staff are available to answer questions and demonstrate our products at many scientific conferences throughout the year. In addition, we welcome users' suggestions for upcoming meetings that might benefit from NCBI's participation. Please send all comments and questions to info@ncbi.nlm.nih.gov.

The following is NCBI's upcoming exhibit schedule. Please feel free to stop by our booth with questions, comments, and suggestions on how to use and improve our services.

American Society of Tropical Medicine and Hygiene (ASTMH)
Houston, TX
October 29-November 2, 2000

Computational Genomics Conference
Baltimore, MD
November 16-19, 2000

American Society for Cell Biology (ASCB)
San Francisco, CA
December 9-13, 2000
BLAST Workshop: December 11

Microbial Genomes Conference
Co-Sponsored by TIGR and ASM
Monterey, CA
January 28-31, 2001

New BLAST Web Tutorials Now Online

A new suite of BLAST Web tutorials have been created to assist both new and veteran users of BLAST and PSI-BLAST. The three tutorials, entitled Query, BLAST, and PSI-BLAST, offer starting points for users with different backgrounds. A novice should start with the Query tutorial. More experienced users will want to work through the BLAST tutorial before proceeding to the more advanced PSI-BLAST tutorial.

The BLAST tutorials, along with a variety of others, can be found under the Education link on the NCBI home page. You may also navigate directly to www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html.

BLAST 2.0.14 Released

Standalone BLAST version 2.0.14 binaries are now available at ftp://ncbi.nlm.nih.gov/blast/executables/. The source code for BLAST 2.0.14 is available as part of the NCBI Toolbox found at ftp://ncbi.nlm.nih.gov/toolbox/ncbi_tools/.

Mouse Contig BLAST

A new BLAST search form, patterned after the popular human genome BLAST search form, has been implemented for the mouse genome. Mouse genome BLAST can be reached via a link from the Mouse Genome Sequencing page. Like the human genome BLAST form, the mouse version offers blastn or tblastn searches of finished contig data and allows the restriction of searches to data from any or all chromosomes. Try mouse genome BLAST at www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html.

How to Maintain Current Local Databases

Because of the enormous size of the sequence data in GenBank, searching with very large queries or with huge batches of smaller sequences is difficult. The Web interface designed by NCBI to handle general purpose searches cannot be used for batch searching and should not be used for searches that are expected to take a very large amount of CPU time. Although one solution for power-users is to set up standalone BLAST and perform computationally intensive searches locally, the problem of downloading GenBank for local searches then materializes. To help with this problem, NCBI offers a set of standard BLAST databases in FASTA format, and of sizes far smaller than the corresponding full GenBank records, on the NCBI BLAST FTP site.

These BLAST FASTA files are updated as often as NCBI updates the databases used by the BLAST servers (nightly for nr, nt, est, sts, month.na, month.aa, and gss). Nightly transfers of the entire nr or nt database, however, can be quite a burden on users and NCBI systems and is not necessary, especially since most of the nr and nt database does not change on a nightly basis. NCBI makes a program available to merge new entries, from an update file, into the non-redundant databases (nr or nt). This program, "fmerge", can be used to update the nr databases with the following two-step procedure:

1. FTP the newest nr database and run fmerge in "create" mode on it:

```
fmerge -t 1 -n nr -i index.nr
```

2. Periodically FTP the newest month.aa file that contains only the new sequences released by GenBank in the last 30 days, and run fmerge in "update" mode on it:

```
fmerge -t 2 -m month.aa -i
index.nr
```

A similar procedure can be used for the nt database for nucleotide sequences. In that case, the database would be updated using the month.na file, which is analogous to the month.aa file used for the nr database.

The fmerge program does not remove redundancy of new sequences from the database and, since month.na contains est and sts sequences (which nt does not), the local nt or nr will differ slightly from NCBI's version after an update to the appropriate month database. This divergence can be kept to a minimum by refreshing the non-redundant database (nr or nt) and removing the index file once a month.

The fmerge program is available by anonymous FTP from ftp://ncbi.nlm.nih.gov/blast/fmerge/.

The BLAST databases are available at ftp://ncbi.nlm.nih.gov/blast/db/. *—SM, DW*

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.

PSI-BLAST 2.1 Offers Composition-Based Statistics

PSI-BLAST now permits calculated E-values to take into account the amino acid composition of the individual database sequences involved in reported alignments. Such composition-based statistical analysis improves E-value accuracy, thereby reducing the number of false positive results.

The improved statistics are achieved with a scaling procedure[1,2] that employs a slightly different scoring system for each database sequence. As a result, raw BLAST alignment scores will not correspond precisely to those implied by any standard substitution matrix. Furthermore, identical alignments can receive different scores, depending on the compositions of the sequences they involve. The improved statistics are now used by default for all rounds of searching on the Web version of PSI-BLAST, but are not used by Basic or Advanced BLAST. Therefore, if one uses default settings, the results of the first round of PSI-BLAST will be different from those obtained using the same query with Basic or Advanced BLAST.

PSI-BLAST 2.1 is currently available only on the Web, at www.ncbi.nlm.nih.gov/blast/psiblast.cgi. It will be incorporated into the standalone binaries and NCBI toolkit in the near future. *—SM*

Notes

1. Altschul, SF, et al. *Nucleic Acids Res* 25:3389–402, 1997.
2. Schäffer, AA, et al. *Bioinformatics* 15:1000–11, 1999.

NCBI Taxonomy

continued from page 1

Taxonomy home page to organisms commonly used in molecular research projects, such as *H. sapiens*, *A. thaliana*, and *P. falciparum*.

The NCBI taxonomic tree contains more than 79,000 taxa. To simplify the display of the hierarchy, a Common Subtree feature has been implemented, allowing users to build a custom phylogenetic tree for selected taxa. With this tool, users can add or delete taxa, and expand or abbreviate lineages, in the process of creating the tree.

Recent modifications to taxonomic classification are also highlighted, including an explanation and cited sources supporting the change. An example is the recent reorganization of the mosses (*Musci*), offering a more accurate reflection of the current understanding of moss phylogeny.

The Tip of the Day addresses some of the finer points of the Taxonomy database and how it is used in other NCBI services. For example, one tip explains why scientific names are sometimes followed by names or abbreviations, as in "*Homo sapiens L.*" The FAQ section also continues to be a useful source of factual information about the Taxonomy database service. —*MR*

Slight Address Change for NCBI FTP Server

NCBI's FTP server has moved from ncbi.nlm.nih.gov to ftp.ncbi.nih.gov. To access this site with a command line FTP program, use syntax such as: ftp ftp.ncbi.nih.gov. Login as "anonymous" and give your mail address as your password. To access NCBI's FTP site from the Web, click on the FTP site link on our home page at www.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

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