NCBI News

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

February 1998

Speed and Sensitivity: BLAST Version 2.0

Michael Crichton, author of *Jurassic Park* (1990, Knopf, New York), discovered the virtue of BLAST to perform sequence similarity searches when NCBI researcher Mark Boguski informed him that the *Tyrannosaurus rex* sequence Crichton published was 100% contaminated with pBR322 vector.[1] While the scientific community awaits an authentic dinosaur sequence, new features added to NCBI's BLAST service are offering increased speed and sensitivity, providing scientists with an enhanced ability to uncover biologically meaningful relationships among various organisms' sequences.

BLAST (Basic Local Alignment Search Tool) is widely used to perform sequence similarity searching because it can produce valuable results swiftly. Currently, over 5,100 people around the world use NCBI's BLAST server on the World Wide Web daily, and an additional 1,000 are using a server-client version. Together they perform over 38,000 searches each day. The new BLAST version 2.0 programs equip researchers with advanced search strategies that are both fast and convenient. BLAST 2.0 combines the statistical analysis of the original BLAST with the ability to perform gapped alignments (Gapped BLAST) and to construct position-specific score matrices for sequence similarity searches (PSI-BLAST).

Gapped BLAST Is Fast

A traditional BLAST search begins by seeking a "word" in a database sequence that matches a "word" in the query with at least the "threshold" score *T*. Such a "hit" is extended in both directions until the running score drops a certain amount below the best score yet achieved. The alignments produced are evaluated for statistical significance, and any high-scoring segment pairs (HSPs) that meet a user-definable cutoff are reported.

The new Gapped BLAST is considerably faster than the original due to two refinements. First, the original BLAST needed to be very sensitive in detecting weak HSPs because several that involved a single database sequence could, in concert, constitute a significant result. By

Continued on page 2

Protein Families and Genome Evolution: COGs

Evolutionary biologists assume that the genetic constitution of every organism can be traced back to a set of common ancestral genes. This assumption has prompted scientists to perform sequence comparisons between genes from different species to identify the distant and subtle relationships between them. Genes with the same function can often be found in different species. These genes are likely to have evolved from a single ancestral gene and are known as "orthologs." Alternatively, there may be sequences within the same organism that are similar but have different functions; these "paralogs" most likely arose from a gene duplication event and then evolved new functions. The growing number of completely sequenced genomes makes it possible to make unprecedented comprehensive comparisons between major phylogenetic groups or specific organisms and produce an informative outline of these relationships. Such a panoramic perspective will augment our knowledge of the course of evolution and identify protein functions conserved in some organisms but not in others.

In Search of Gene Families from Complete Genomes

Working with the newly sequenced genomes from seven different organisms, three scientists at NCBI, Roman Tatusov, Eugene Koonin,

Continued on page 4

IN THIS ISSUE

NCBI News

NCBI News is distributed two to 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

Editors
Dennis Benson
Barbara Rapp

NCBI Contributors
Renata McCarthy
Ken Katz
Francis Ouellette
Stephen Altschul

Writer
Donna Roscoe

Managing Editor
Roseanne Price

Graphics and Production
Veronica Johnson

Design Consultant
Troy M. Hill

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 to perform research in computational molecular biology.

NIH Publication No. 98-3272

ISSN 1060-8788

BLAST, *continued from page 1*

allowing a single HSP of sufficient score to trigger a gapped extension step, BLAST 2.0 can afford to miss some very weak HSPs in its initial pass. The threshold score *T* can therefore be raised, with an attendant increase in speed. Second, the new program requires the detection of two hits within a short distance of one another on the same diagonal before it invokes an ungapped extension. Even after *T* is adjusted to maintain the same sensitivity, this requirement reduces substantially the number of time-consuming extensions needed. The net result is a program that is not only more sensitive but also three times faster than before. The Gapped BLAST programs blastn and blastp offer fully gapped alignments; blastx and tblastn have "in-frame" gapped alignments and use sum statistics to link alignments from different reading frames. Gapped BLAST is not offered for tblastx searches. (For a description of the BLAST family of programs, see http://www.ncbi.nlm.nih.gov/BLAST/blast_program.html.)

PSI-BLAST for Motif-Style Searching

"Motif" searches are potentially much more sensitive to distant relationships than are the traditional pairwise similarity searches for which BLAST has been tailored. Position-Specific Iterated BLAST (PSI-BLAST) now brings both speed and ease of operation to motif searching. It can be used to help delineate diverse protein families and to predict function for newly sequenced proteins. PSI-BLAST uses an initial BLAST run to generate a gapped multiple alignment. It then constructs from this alignment a position-specific score matrix, which is employed as a "query" in a subsequent BLAST search. This process can be repeated multiple times to hunt for homologous sequences that would not have been retrieved by the original BLAST algorithm. Currently, PSI-BLAST is limited to protein-protein queries.

NCBI researchers tested the power of PSI-BLAST by applying it to the C-terminal 215 amino acids of the BRCA1 sequence.[2] BRCA1 and other members of the BRCT superfamily typically are involved in DNA damage-responsive cell cycle checkpoints.[3] In multiple iterations, PSI-BLAST automatically identified almost all the previously recognized BRCT proteins and added seven new ones to the roster (see Table 1 and Altschul et al., 1997[4]).

For more information, a BLAST help manual is available on-line.

Continued on page 8

Table 1: Seven New Members of the BRCT Superfamily Identified with PSI-BLAST

Protein	Species	GenBank ID Number	PSI-BLAST iteration	E-value
T10M13.12	*Arabidopsis thaliana*	2104545	1	4e-06
KIAA0259	*Homo sapiens*	1665785	1	0.001
T13F2.3	*Caenorhabditis elegans*	1667334	3	2e-07
SPAC6G9.12	*Schizosaccharomyces pombe*	1644324	7	4e-04
C36A4.8	*Caenorhabditis elegans*	1657667	7	0.010
D90904	*Synechocystis* sp.	1652299	15	0.17
Pescadillo	*Homo sapiens*	2194203	16	0.017

GenBank Submissions: From Deposit to Release

Ever wonder what happens to your cherished sequence once you send it hundreds, even thousands, of electronic miles away to GenBank? Whether you are using the sequence submission tool BankIt on the WWW, the stand-alone program Sequin, or one of the specialized submission procedures for EST, STS, GSS, and HTG sequences, your submission is received by the GenBank staff—a group of highly trained biologists and database specialists who manage the collection and distribution of GenBank data. Currently, over 5,000 sequences arrive each month at GenBank (excluding the specialized submissions). While EST, STS, GSS, and HTG submissions are processed in large numbers using semiautomated systems, all other types of sequence records are processed manually to ensure biological integrity and internal consistency with annotation rules established by the International Nucleotide Sequence Database Collaboration.

Certificate of Deposit: The Accession Number

An NCBI staff member checks that your submission meets minimum requirements and then assigns an accession number to the sequence within 24 hours. The accession number serves as a confirmation that the sequence has been submitted and is a permanent, citable number that will allow your sequence to be referenced in publications by yourself and others. This same number is used to retrieve your sequence from GenBank or from one of the other International Database Collaborators, EMBL and DDBJ.

Accession numbers consist of one letter and five digits, or two letters and six digits, and do not change even if the record or its sequence is updated. GenBank also assigns a unique GenBank identifier, or GI number, to every *sequence* loaded into the GenBank database. The GI numbers for nucleotide and protein sequences are referred to as NIDs and PIDs, respectively. The GI number changes every time the sequence is updated, enabling GenBank to track changes in *sequence* over time.

Checking Accounts: Indexers and Scientists

Under the coordination of Francis Ouellette, a staff of 17 indexers trained in molecular biology and skilled in database production operations annotate, organize, and maintain the 1.7 million database entries. The indexers ensure that all direct submissions receive a systematic quality assurance review. Sequences are screened against GenBank by using BLAST to identify full or partial matches to sequences in the database and then searched to detect vector, yeast, and mitochondrial contamination. Programs 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: the locus name, definition line, taxonomy classification, and journal references. Staff consult with submitters as necessary to add or modify features. Finally, one of 21 senior scientists performs a final review for biological integrity and continuity.

At least four people have reviewed your sequence and its annotations before a draft of the GenBank record is mailed back to you for review. If the record

Continued on page 6

NCBI Data by FTP

The NCBI FTP site contains a variety of directories with publicly available databases and software. The available directories include 'repository,' 'genbank,' 'entrez,' 'toolbox,' 'pub,' and 'sequin.'

The **repository** directory makes a number of molecular biology databases available to the scientific community. This directory includes databases such as PIR, SwissProt, CarbBank, AceDB, and FlyBase.

The **genbank** directory contains files with the latest full release of GenBank, the daily cumulative updates, and the latest release notes.

The **entrez** directory contains the client software for Network Entrez.

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 (sequence similarity search program). Client software for Network BLAST and PowerBlast is also included in this directory.

The **sequin** directory contains the new Sequin submission software for Mac, PC, and UNIX platforms.

Data in these directories can be transferred through the Internet by using the Anonymous FTP program. To connect, type: **ftp ncbi.nlm.nih.gov.** Enter **anonymous** as the login name, and enter your e-mail address as the password. Then change to the appropriate directory. For example, change to the repository directory (cd repository) to download specialized databases.

COGs, *continued from page 1*

and David Lipman, designed a new system for classifying conserved genes and exploring the evolutionary relationships among them. Beginning with a single gene, they looked for the best match to that sequence in every other genome. They continued to perform pairwise sequence comparisons for each protein sequence against every other sequence in all the genomes until nearly 18,000 sequences had been compared. When two genes from different organisms found each other as their best match, they were identified as orthologs. Paralogs in genomes were identified when matches between sequences in genomes were not reciprocal. The NCBI team cataloged the sequences according to their functional similarities into "Clusters of Orthologous Groups," or COGs.[1] A total of 720 unique COGs were identified. Each COG has at least three orthologs from three genomes (Figure 1a) and, in some cases, paralogs from the same lineage (Figure 1b).

The results of this comparison are available on a new NCBI Web page (http://www.ncbi.nlm.nih.gov/COG/). The genomes analyzed include five bacterial genomes, *Escherichia coli, Haemophilus influenzae, Mycoplasma genitalium, Mycoplasma pneumoniae,* and Cyanobacteria *Synechocystis* sp; one archaebacterial genome, *Methanococcus jannaschii;* and one eukaryotic yeast genome, *Saccharomyces cerevisiae*.

COGs Predict Functions

Since orthologs typically have the same function, COGs allow the functions of putative gene products to be predicted from the growing number of newly sequenced genomes. Functions were assigned to the majority of the 720 COGs based on known proteins within the groups or significant similarities to proteins in organisms not included in this study. The COGs were further organized into 15 functional subgroups within 4 major divisions: (1) information storage and processing, (2) cellular processes, (3) metabolism, and (4) poorly characterized. The distribution of proteins from different organisms in the COGs identifies trends in functional diversification. For example, the absence of representative proteins from the pathogenic bacteria (*H. influenzae* and the mycoplasmas) in some metabolic groups was demonstrated.

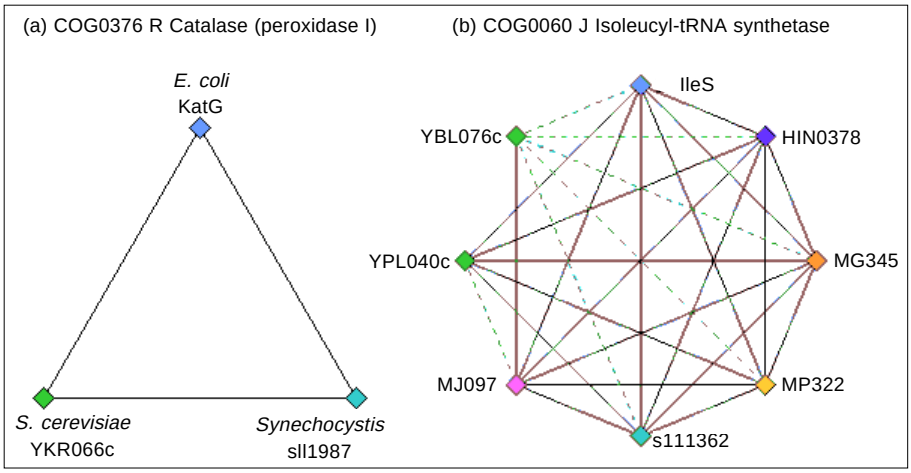

Figure 1. A minimal COG consists of three genes from three different lineages and can be illustrated by a triangle (Figure 1a). COGs were expanded by combining triangles that share sides (Figure 1b). Each of the seven genomes on the COG Web site is represented by a specific color, allowing ready visualization of the orthologous or paralogous relationships between the various genome sequences. Reciprocal best matches between genomes are represented by a solid line. Paralogous relationships, where one genome sequence has a best match to a sequence in another genome but the reverse is not true, are indicated by dashed lines (Figure 1b).

Expanding COGs into Superfamilies

The COGs represent ancient, conserved protein families with relevant cellular functions because they are from organisms representing the major phylogenetic groups that are estimated to be over 1 billion years old. Conserved sequence motifs within the proteins reflect distinct biochemical activities employed by a variety of proteins to perform their designated role in the cell. The NCBI team also employed motif-style searching by using PSI-BLAST to identify protein superfamilies. Protein superfamilies represent a higher level of protein classification than the COGs alone and can be used to classify highly evolved proteins not assigned to any COG. The largest superfamily contained ATP-ase and GTP-ase motifs broadly distributed in a variety of cellular mechanisms.

Phylogenetic Patterns in COGs

Like pieces of a mosaic that reveal an image when viewed together, COGs can be used to conceptualize genetic evolution. The presence or absence of a representative gene from an organism in a COG can be studied to reveal "patterns" of gene conservation or loss for that particular COG function. Tatusov, Koonin, and Lipman compiled a list of phylogenetic patterns gleaned from the 720 COGs. A single letter of the alphabet was assigned to

Continued on page 8

High Throughput Sequencing Gives Rise to New GenBank Division

Generation-megabase scientists interested in accessing information hot off the sequencer will be pleased to know that large-scale sequencing centers involved in the eukaryotic genome projects are making copious amounts of sequence data available to the public prior to completion. GenBank, in concert with the other members of the International Nucleotide Sequence Database Collaboration, DDBJ and EMBL, has created the High Throughput Genome (HTG) division to handle the evolving assemblage of genomic data. To date, the high throughput sequencing projects include *Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, and *Mus musculus*.

HTG Record Evolution

Genome sequencing centers generate preliminary sequence information from a single genomic clone and deposit random sequence fragments greater than 2 KB into the GenBank HTG division. GenBank assigns a single accession number to the sequence data derived from each clone and indicates the status of the HTG record as it passes through several stages toward completion. Phase 1 records contain sequences that are unordered, unoriented, and contain gaps. In Phase 2, the order and orientation of the sequences have been determined, but gaps remain. In Phase 3, once the sequencing is complete and the error rate is less than 10^{-4}, records are considered finished. Phase 3 records are transferred to the appropriate organism division of GenBank, such as the Primate (PRI) division for human sequences, or the Invertebrate (INV) division for *C. elegans*. Sequences submitted to the HTG division are automatically searched against the various databases using the BLAST programs, and the records are annotated to show the significant matches. This sequence similarity information is valuable for positional cloning and gene hunting.

Accessing HTG records

HTG records can be retrieved in Entrez by selecting the organism and specifying "HTG" in the Keywords field. The Genomes database in Entrez, which offers graphical displays of nucleotide and protein sequences, provides a visual framework for the HTG sequences with links to additional DNA, protein, and bibliographic records. Unfinished HTGs (Phase 1 or 2) are also available for BLAST searching by selecting the "htgs" database, or the "month" database for the latest entries; finished records (Phase 3) are available in the "nr" and "month" BLAST databases.

HTG Information on the Web

The new HTG Web site at http://www.ncbi.nlm.nih.gov/HTGS describes the HTG division and gives more detailed instructions for sequencing centers interested in submitting HTG sequences. ■

HTG: phase 1

HTG DIVISION

```
LOCUS        HSAC000003     120000 bp    DNA          HTG      20-SEP-1996
DEFINITION   *** SEQUENCING IN PROGRESS *** Chromosome 17 genomic sequence; HTGS
             phase 1, 6 unordered pieces.
ACCESSION    AC000003
KEYWORDS     HTG; HTGS_PHASE1.
...
COMMENT      ***                                                        ***
             *** WARNING: Phase 1 High Throughput Genome Sequence       ***
             ***                                                        ***
             * This sequence is unfinished. It consists of 6 contigs for
             * which the order is not known; their order in this record is
             * arbitrary. In some cases, the exact lengths of the gaps
             * between the contigs are also unknown; these gaps are  presented
             * as runs of N as a convenience only. When sequencing is complete,
             * the sequence data presented in this record will be replaced
             *by a single finished sequence with the same accession number.
             * 1       22526:   contig of 22526 bp in length
             * 22527   23035:   gap of unknown length
             * 23036   33919:   contig of 10884 bp in length
             * 33920   34427:   gap of unknown length
             * 34428   61877:   contig of 27450 bp in length
. . .
//
```

PHASE

HTG: phase 3

PRIMATE DIVISION

```
LOCUS        AC000003      122228 bp    DNA          PRI      07-OCT-1997
DEFINITION   Homo sapiens chromosome 17, clone 104H12, complete sequence.
ACCESSION    AC000003
NID          g2204282
KEYWORDS     HTG.
SOURCE       human.
...
COMMENT      The Staden databases, finishing information, and all
             chromatographic files used in the assembly of this clone are
             available from our anonymous ftp site.
             All repeats were identified using RepeatMasker: Smit, A.F.A. &
             Green, P. (1996-1997)
             http://ftp.genome.washington.edu/RM/RepeatMasker.html.
FEATURES             Location/Qualifiers
     source          1..122228
                      /organism="Homo sapiens"
                      /db_xref="taxon:9606"
                     /clone="104H12"
                     /clone_lib="Research Genetics/Cal Tech CITB978SK-B (plates
                     1-194)"
                     /chromosome="17"
     repeat_region          261..370
                     /rpt_family="MLT1B"
```

ACCESSION NUMBER

Selected Recent Publications by NCBI Staff

Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BFF. GenBank. *Nucleic Acids Res* 26:1–7, 1998.

Galperin MY, Koonin EV. A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity. *Protein Sci* 6:2639–43, 1997.

Leipe DD, **Landsman D.** Histone deacetylases, acetoin utilization proteins, and acetylpolyamine amidohydrolases are members of an ancient protein superfamily. *Nucleic Acids Res* 25:3693–7, 1997.

Lipman DJ. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res* 25:3580–3, 1997.

Marchler-Bauer A, Bryant SH. A measure of success in fold recognition. *Trends Biochem Sci* 22:236–40, 1997.

Neuwald AF. An unexpected structural relationship between integral membrane phosphatases and soluble haloperoxidases. *Protein Sci* 6:1764–7, 1997.

Ouellette BFF, Boguski MS. Database divisions and homology search files: a guide for the perplexed. *Genome Res* 7:952–5, 1997.

Pruitt KD. WebWise: navigating the Human Genome Project. *Genome Res* 7:1038–9, 1997.

Schuler GD. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* 75:694–8, 1997.

Sonnhammer EL, Wootton JC. Widespread eukaryotic sequences, highly similar to bacterial DNA polymerase I, looking for functions. *Curr Biol* 7:R463–5, 1997.

Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 278:63 1–7, 1997.

GenBank Reaches One Billion Bases

In 1985, GenBank contained just over 5,700 entries that were obtained principally by scanning the biomedical literature for sequence data. GenBank now contains almost 2 million entries and recently surpassed 1 billion base pairs of genetic information for more than 25,000 organisms. Human sequence data predominate in the database, representing 43% of the billion-plus base pairs. Mouse (*Mus musculus*) and the nematode, *Caenorhabditis elegans*, are second and third, representing 10% and 9%, respectively.

Exponential Growth

Doubling in size every 18 months, GenBank is now built primarily from the direct submission of sequence data from authors and sequencing centers. Currently, more than 70% of the sequence records in the database are ESTs (expressed sequence tags). As EST and genomic sequencing efforts are intensified, the GenBank doubling rate is expected to accelerate. Additional information about GenBank, its various divisions, and its growth statistics can be found in the current release notes (ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt).

GenBank CD-ROM to Be Discontinued

The explosive growth of sequence information is reflected in the GenBank CD-ROM, which expanded from a single disc in 1992 to 12 discs with December 1997 Release 104. Since production costs are escalating and users are opting for the convenience of the Internet over the unwieldy discs, the GenBank CD-ROM will be discontinued following the April 15, 1998, release. GenBank full releases with cumulative and noncumulative update files continue to be available in the genbank/ directory for downloading by Anonymous FTP. Consult the README file in this directory for more details. ■

GenBank Submissions, *continued from page 3*

is not to be held confidential, it is loaded into GenBank after a 5-day review period. A confidential record will not be released into the public database until you have notified GenBank or it is published, whichever comes first. At any time, you may update information in your record. We encourage authors to notify GenBank of publication so that confidential records may be released and public records can be updated in a timely manner. Use the BankIt Update function or send a message with the new information to update@ncbi.nlm.nih.gov; please include your accession number with all correspondence.

Gaining Interest: Release of Records

The turnaround time from submission to release is anywhere from 1 to 3 weeks, depending on the number of submissions GenBank is processing. Once the record is loaded into the database, the public can see the record the next day by using the Query e-mail server (query@ncbi.nlm.nih.gov), Network Entrez, or WWW Entrez. Entrez provides links to additional sequences, graphic displays, structures, genome records, and PubMed. Still have questions? Write to us at info@ncbi.nlm.nih.gov. ■

Frequently Asked Questions

How do I do a BLAST search with a short DNA sequence?

You will probably need to increase the Expect (E) value, since a short query is more likely to occur by chance in the database. You may also want to turn off the low-complexity filter, since short queries often contain low-complexity sequence. Another parameter that becomes important with a short query is Word size, which is used by BLAST to nucleate regions of similarity. The default Word size is 11 for nucleotides, so if your query sequence falls below this, you may want to decrease Word size (W). For more detail, see the FAQ section of the BLAST Web page.

Is it possible to perform a BLAST search against just human ESTs?

Many separate databases are now available for BLAST searching. Select "Human ESTs" in the Database field pull-down menu when using Gapped BLAST.

Where can I get more information about the Interactive Digital Differential Display (DDD) facility used in the Cancer Genome Anatomy Project (CGAP) Project?

DDD is a computational method for comparing gene frequencies among various cDNA libraries or pools of libraries. It is available from the CGAP Web site at http://www.ncbi.nlm.nih.gov/ncicgap/ddd.html.

How can I obtain the EST clones described in my UniGene search?

Information on clone availability is located in the dbEST record (http://www.ncbi.nlm.nih.gov/dbEST/index.html). Click on "Search dbEST" and enter the GenBank accession number. Individuals interested in obtaining materials can (1) contact the submitter of the sequence, (2) refer to the Source field (if present) for sources providing the clone, (3) refer to the Clone ID and library number located under "Clone Info," which can be used to order a particular clone through the I.M.A.G.E. Consortium. To see a list of distributors participating in the I.M.A.G.E. Consortium, scroll down the dbEST Web page and click on "Distributors."

How can I search for just review articles or specify a certain time period when searching PubMed?

In the Advanced mode, set the Search Field pull-down menu to "Publication Type" and enter the word "review" into the text box. To display a list of available terms for Publication Type, select "List Terms" from the Mode menu and enter a term using "Publication Type" in the Search Field. To search a range of dates, use a colon between the limiting years (e.g., 1966:1976), and set the Search Field to "Publication Date." Search results can also be limited to the last 30 days or another period of time by selecting one of the options under the Publication Date limit menu.

In CGAP, is there a way to tell which libraries are made with tissue from the same donor?

For any tissue, including microdissected tissues, clicking on the link for "Tissue sample" will lead to a list of all libraries made from the same samples.

BLAST, *continued from page 2*

Notes

[1] Boguski MS. A molecular biologist visits *Jurassic Park*. *Biotechniques* 12:668–9, 1992.

[2] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–402, 1997.

[3] Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV. A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J* 11:68–76, 1997.

[4] *Op. cit.* 2. ■

COGs, *continued from page 4*

represent each genome (e.g., "e" for *E. coli*), and a dash was indicated when the organism was not represented in the COG. A COG that has proteins from all seven genomes has a phylogenetic pattern shown as "ehgpcmy." A COG that is missing representative sequences from the pathogenic species *M. genitalium* and *M. pneumoniae* has the pattern "eh_ _cmy." These two patterns are the most frequently occurring patterns, being displayed by 114 and 119 COGs, respectively, and therefore represent conserved patterns. The conserved patterns demonstrated continuity between the genomes, while rare patterns suggest unique functions that need investigating. The addition of more genomes to the COG analyses is expected to illuminate the functional role behind the rare patterns.

Piece by Piece

The NCBI team continues to expand its COG research, analyzing eight more genomes: *Helicobacter pylori*, *Bacillus subtilis*, *Borrelia burgdorferi*, *Treponema pallidum*, *Chlamydia trachomatis*, *Methanobacterium thermoautotrophicum*, *Archaeoglobus fulgidus*, and *Caenorhabditis elegans*. These analyses will be incorporated into the Web site upon completion. Refinements and new additions are expected to build a COG collection that will become a valuable resource for characterizing genomes and comprehending life's blueprint.

Note

[1] Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 278:631-7, 1997. ■

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