NCBI News

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

Winter 2001

COG Database Grows to 3300 Protein Clusters and 44 Complete Genomes

Based on research conducted by NCBI's comparative genomics group, the database of Clusters of Orthologous Groups of proteins (COGs) represents a phylogenetic classification of proteins encoded in complete genomes. The COGs are derived from an "all-against-all" sequence comparison of the encoded proteins. Each COG consists of individual proteins or groups of paralogs from at least three lineages and is therefore considered to correspond to an ancient conserved domain. The database is designed to support research on genome evolution as well as functional annotation of genomes.

At its inception in 1997, the database included 720 clusters from 7 genomes. It now includes more than 3300 COGs from 44 genomes of bacteria, archaea, and the yeast *Saccharomyces cerevisiae*, representing 30 major phylogenetic lineages. In addition, proteins from two eukaryotic genomes, *C. elegans* and *D. melanogaster*, have also been assigned to individual COGs. The COG home page lists the organisms included, number of proteins encoded by each genome, and the portion of those that are included in COGs.

Three general kinds of information can be obtained using the COG database. For functional studies, the COGs have been classified into 18 broad functional categories, including one for uncharacterized COGs. Phylogenetic patterns show the presence or absence of proteins from a given organism in a specific COG and, when used systematically, can identify whether a particular metabolic pathway exists in an organism. Multiple alignments of COG members can be used to identify conserved sequence residues and analyze evolutionary relationships between member proteins.

Individual COG reports contain information on the number of proteins comprising the cluster, their inferred function, a function code from a list of 18 general categories, the phylogenetic pattern for the COG, the unique COG number, and a link to proteins from *C.elegans* and *D.melanogaster* assigned to the COG. If available, the pathway or functional system is also indicated as a functional sub-category. Clicking on the floppy disk icon will generate a FASTA-formatted file of protein sequences for all COG members.

continued on page 2

Plant Genomes Featured in NCBI Services

Plant Genomes Central

A new Plant Genomes Central service provides access to large-scale plant genomic and EST sequencing projects, with links to corresponding taxonomic information. Mapping data for a growing number of plant genomes is also available in the Map Viewer. See http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/PlantList.html.

HomoloGene Expanded to Include Plants

HomoloGene is a homology resource that includes both curated

continued on page 6

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
Colleen Broder
Kathy Kwan
Vyvy Pham
Bart Trawick

Writer
David Wheeler

Editing and Production
Jennifer Carson Vyskocil
Cheryl Richardson

Graphic Design
Tim Cripps
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.

NIH Publication No. 02-3272

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

COG Database

continued from page 1

The COG report also generates a table giving the gene names corresponding to cluster members from each organism. Each gene name is linked to a display of the BLAST output for its encoded protein, which includes both graphical and textual sequence alignments between the COG member and other protein database sequences. A Genomic Context link shows the organization of the genomes of the organisms represented in a COG, centered on the genes coding for the orthologous proteins that comprise the cluster. Finally, a dendrogram, constructed from multiple sequence alignments, displays sequence similarity relationships between the COG members.

A **Phylogenetic Patterns** search tool finds COGs that are shared by any set of organisms. Organisms may be included or excluded from the group using an input table. For closely related organisms belonging to a single clade, pre-computed tables show shared and unique COGs.

The **COGnitor** program is a companion tool that assigns new proteins to pre-existing COGs. COGnitor takes a protein sequence as input for sequence comparison, and suggests inclusion in a COG if there are "best hits" to proteins from at least three lineages. The output shows the COG to which the query protein is predicted to belong, a color-coded BLAST graphic delineating the regions of similarity, and the sequence alignments.

Other useful resources include:

- **List of COGs,** which displays all COGs in the database.
- **Distribution** histograms that show how many COGs contain proteins from a specific number of clades or species.
- **Phylogenetic patterns** table, which organizes the patterns into sets based on the presence or absence of organisms belonging to Archaea, Eukarya or Bacteria.
- **Co-occurrences** table, which shows the number of COGs shared by a particular pair of species or unique to one member.
- **Functional categories** page, summarizing the functions that have been defined, the number of COGs assigned to each category, the number of proteins or domains assigned to each category, and the number of pathways and functional systems associated with each category.

The COGs are also integrated with the Genome division of Entrez. From the COG pages, proteins are linked to the Genome view and Neighbor view. From Entrez Genome, proteins are linked to their respective COGs, and COG data is included in several display options. For example, in the map display of circular genomes, the radial lines corresponding to genes are color-coded according to the functional categories used in the COG system.

The COG service is located at www.ncbi.nlm.nih.gov/COG/. The data is also available by FTP at ftp://ncbi.nlm.nih.gov/pub/COG.
—VP

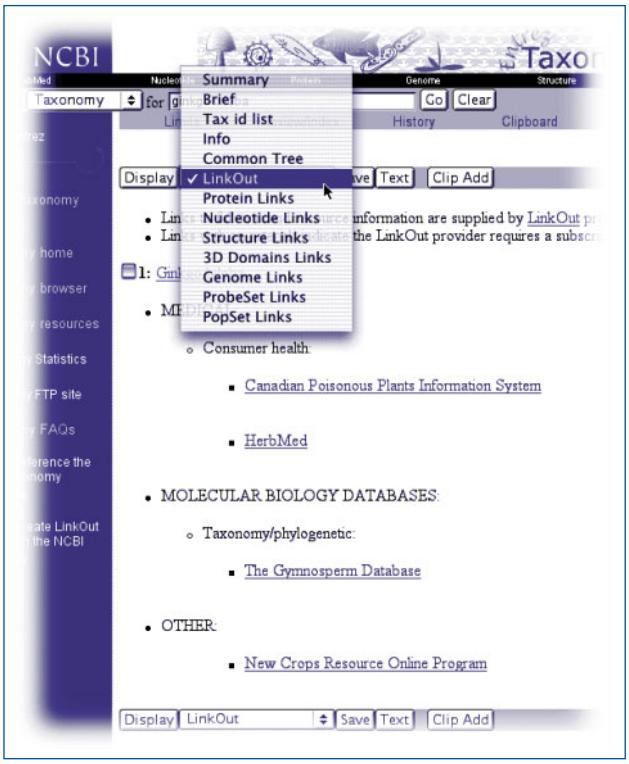
LinkOut—Explore Beyond Entrez

When you are searching Entrez, have you ever gone to another Web site or to your library's bookshelves in order to obtain additional information about what you retrieved? For example, suppose that you are browsing the Entrez Taxonomy database and you come across an entry for *Ginkgo biloba*. You may have heard of this plant many times, but what exactly does it look like? Or maybe you are looking at a *Drosophila melanogaster* sequence within the Entrez Nucleotide database. Wouldn't it be convenient if you could quickly see the corresponding record in a different biological database, such as FlyBase?

You no longer have far to go to retrieve outside information. Just follow the LinkOut hyperlinks located within individual Entrez records and you will be able to explore additional information from sources located around the world.

LinkOut is a feature of Entrez that is designed to provide users with hyperlinks from records in Entrez databases to a wide variety of relevant Web-accessible online resources, including full-text publications, biological databases, research tools, and more. The goal of LinkOut is to facilitate access to relevant online resources beyond the Entrez system in order to extend, clarify, or supplement information found in the Entrez databases.

Let's take a look at the *Ginkgo biloba* example once again. To get to the record, enter *Ginkgo biloba* in the Entrez Taxonomy search box. LinkOut hyperlinks (see Figure) can take you to the Gymnosperm Database where you will find a concise description of this tree, furnished with pictures of the tree, its bark, fruits, and foliage. Other links from this Entrez record tell you even more. The link for Canadian Poisonous Plants Information System provides information relating to the toxic effects of the tree. The New Crops Resource Online Program and HerbMed hyperlinks offer information on *Ginkgo biloba*'s value as a medicinal herb.



LinkOut resources for *Ginkgo biloba*, with menu showing Entrez display options

Links to external resources are listed in the LinkOut display of an Entrez record. Users can access the LinkOut display by selecting **LinkOut** from the drop-down menu located next to the **Display** button. Clicking on the **Display** button will then show the various LinkOut resources for a given record. From the PubMed database, these external LinkOut resources can also be accessed through an icon from the Abstract and Citation formats. Users can customize LinkOut display formats through LinkOut Preferences located in the Cubby.

A tutorial on using LinkOut is available from the LinkOut home page at www.ncbi.nlm.nih.gov/entrez/linkout.

If you know of online resources that might be valuable to Entrez users, please contact NCBI or the database provider. We are actively seeking new participants. Generally, high quality, easy to use resources that are directly relevant to particular Entrez records are good candidates. See the LinkOut home page for detailed information on how to provide links for Entrez databases.

Please send your questions or comments to linkout@ncbi.nlm.nih.gov. *—KK, BT*

Investigator Profile

Stephen F. Altschul

Powerful Tools for Identifying Sequence Similarities

The recent pace of whole genome sequencing projects has resulted in an increase—both in volume and in complexity—of available molecular sequence data. Patterns shared by multiple protein and nucleic acid sequences provide valuable insights into genomic organization, molecular structure, and biological function, as well as to unexpected links among diverse biological systems. These previously unknown connections not only speed research progress, but often open new areas for scientific inquiry.

When studying a new gene or protein sequence, researchers often conduct database searches in order to identify similar genes or proteins. This is because the fastest method for identifying the function of a gene or protein is to find a related gene or protein—or an entire family—whose function is already known. The recognition of subtle residue patterns among genes or proteins sometimes relies upon aligning many sequences, a procedure that continues to present a complex and multifaceted problem for research.

Comparison, whether of morphology or protein sequences, lies at the heart of biology.

In 1990, NCBI researchers **Altschul**, **Gish,** and **Lipman**, in collaboration with colleagues **Miller** and **Myers** from Penn State and the University of Arizona, developed and released **BLAST**—the **B**asic **L**ocal **A**lignment **S**earch **T**ool[1]. The BLAST programs implement a set of sequence comparison algorithms that search a database for optimal local alignments to a query sequence. A local alignment represents a possible **homology**, or similarity by descent, between segments from two nucleic acid or protein sequences. The BLAST programs were substantially faster than existing database similarity search programs, and of comparable sensitivity to distant relationships. Of equal importance, using a statistical theory reported the same year by **Karlin** of Stanford University, and **Altschul**[2], the BLAST programs first provided researchers with rigorous guidance for determining which alignments were statistically significant, and therefore worthy of further examination. The ideas underlying BLAST are simple and robust, and can be applied in a variety of contexts, including DNA and protein database searches, gene identification searches, and most recently, sequence motif or profile searches.

BLAST works by breaking the query sequence into short fragments, or "words", and initially seeking very close matches between these words and words from database sequences. Any aligned word pair scoring above a specified threshold is called a "**hit**". Each hit is then extended in both directions in an attempt to generate a local alignment representing statistically significant sequence similarity. The quality of each alignment is represented by a score, defined most simply as the sum of scores for aligning pairs of nucleotides (for DNA) or pairs of amino acids (for proteins).

Because nucleotides or amino acids may be inserted or deleted within a particular sequence during the course of evolution, alignment programs generally allow for the existence of gaps, or spaces introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. Gaps contribute negatively to the overall score of an alignment. The original BLAST programs did not explicitly include gaps within alignments, but rather treated them implicitly by calculating combined statistical assessments of multiple ungapped alignments produced by a single pair of sequences[3].

In 1997, a team of NCBI researchers, including **Altschul**, **Madden**, **Schaffer**, **Zhang,** and **Lipman**, in collaboration with **Zhang** and **Miller** from Penn State, released a set of "gapped BLAST" programs. These new programs not only generated gapped alignments, but also ran several fold faster than the original BLAST programs[4]. This improvement was achieved by incorporating two algorithmic refinements. The first refinement required two hits within a set distance of one another, rather than one, before triggering a search for an ungapped local alignment, or high-scoring segment pair (HSP). The second refinement invoked a gapped extension step whenever an HSP of sufficiently high score was found. Previously, missing a single HSP implicitly involved in a significant alignment match could jeopardize the discovery of the result. Now, by introducing an algorithm for generating gapped alignments, it becomes necessary to find only one HSP, rather than all ungapped alignments subsumed in a significant result. Therefore, careful choice of algorithmic parameters led to increased program sensitivity to distant sequence relationships as well as to increased speed.

The introduction of BLAST and then gapped BLAST rendered it substantially easier for scientists to scan large sequence databases rapidly for relatively weak sequence similarities, and to statistically evaluate the resulting matches. Today, these BLAST programs are widely used tools for searching both protein and nucleic acid databases for sequence similarities, and may compare protein or DNA queries with protein or DNA databases in any combination. However, some of the most interesting similarities are quite subtle and do not rise to statistical significance during a standard BLAST search. Protein database searches using strategies that employ the construction of position-specific score matrices are often better able to detect weak relationships between sequences than are searches using a simple sequence as the query. Yet, employing these methods has not always been simple, frequently involving the use of multiple computer programs as well as a fair amount of scientific expertise.

Altschul and his team have also investigated at least a dozen other potential modifications to the methods used in PSI-BLAST, with the goal of improving overall accuracy in finding true positive matches. Their evaluation resulted in the implementation of a number of refinements to the PSI-BLAST program. Refinements include: the use of more accurately estimated statistical parameters; the filtering of database sequences, as opposed to query sequences, in order to prevent segments with highly restricted or biased amino acid composition from participating in the construction of profiles; and improved treatment of gaps within alignments when estimating position-specific amino acid frequencies[6]. Altschul and his collaborators have many more ideas they would like to implement and evaluate, always striving to provide the biomedical research community with readily accessible and powerful tools for conducting state-of-the-art molecular biology research. —*CB*

The original BLAST paper was the most highly cited paper published in the 1990s and is being supplanted only by the 1997 paper describing the original version of PSI-BLAST.

To overcome this obstacle, the team that developed gapped BLAST incorporated the use of position-specific score matrices into the BLAST protein database search program, extending its capacity to detect weak yet significant sequence similarities. The resulting Position-Specific Iterated BLAST (PSI-BLAST) program features a method for automatically constructing a position-specific score matrix, or profile, from the multiple alignment implicit in the highest scoring matches from an initial BLAST search[4]. Scores are defined for aligning the various amino acids to each profile position. Highly conserved positions yield large positive or negative scores while weakly conserved positions yield scores near zero. The profile is then used to perform a subsequent BLAST search, and the procedure may be iterated, or repeated, further refining the profile. PSI-BLAST runs at approximately the same speed per iteration as gapped BLAST, but in most cases, is far more sensitive to weak yet biologically significant sequence similarities.

A limitation to using PSI-BLAST for large-scale protein analysis has been that on a small percentage of queries, false positives—segments having no direct relationship to the query—enter the list of matches during one iteration and corrupt the profile for subsequent iterations. To mitigate this problem, a team of NCBI investigators headed by Altschul recently improved PSI-BLAST accuracy by incorporating the use of composition-based statistics[5]. Here, the evaluation of an alignment's significance is tuned to a specific profile and the amino acid composition of the sequence to which it is locally aligned. Composition-based statistics have largely suppressed the problem of profile corruption.

[1] Altschul, SF, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. *J Mol Biol* 215(3):403-10, 1990.

[2] Karlin, S and SF Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA* 87(6):2264-68, 1990.

[3] Karlin, S and SF Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci USA* 90(12):5873-7, 1993.

[4] Altschul, SF, TL Madden, AA Schäffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search. *Nucleic Acids Res* 25(17):3389-3402, 1997.

[5] Schäffer, AA, L Aravind, TL Madden, S Shavirin, L Spouge, YI Wolf, EV Koonin, and SF Altschul. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* 29(14):2994-3005, 2001.

[6] Sternberg, MJE, PA Bates, LA Kelley, and RM MacCallum. Progress in protein structure prediction: assessment of CASP3. *Curr Opin Struct Biol* 9(3):368-73, 1999.

The BLAST programs have been widely adopted as standard research tools by the international biomedical community. The advances described above not only improve the accuracy of BLAST searches, but provide scientists worldwide with more powerful methods for characterizing proteins by inferring function from sequence similarity. Using the various versions of BLAST, researchers have assigned many proteins to previously described families, and sometimes have uncovered completely novel families. PSI-BLAST has found relationships that had previously been detectable only with the aid of information about protein three-dimensional structure. This research continues to reveal interesting, and, at times, unexpected truths about evolution.

GenBank *News*

New Genomes in GenBank

The following complete microbial genomes have been deposited recently in GenBank and can be viewed in Entrez Genome:

Brucella melitensis:
AE008917 and AE008918

Encephalitozoon cuniculi:
AL391737 and AL590442-AL590451

Ralstonia solanacearum:
AL646052 and AL646053

Agrobacterium tumefaciens C58:
AE008687–AE008690

Base Quality in HTG Records

The High-Throughput Genomic (HTG) division of GenBank contains unfinished sequence data submitted by large-scale sequencing centers. A typical HTG record might consist of first pass sequence data generated from a single cosmid, BAC, YAC, or P1 clone, and contain one or more gaps.

HTG records contain a status indicator of Phase 0, Phase 1, or Phase 2. Phase 0 sequences are usually one-to-few pass reads of a single clone, so are usually not contigs. Phase 1 sequences are usually unordered, unoriented contigs, with gaps. Phase 2 sequences are usually ordered and oriented contigs, with or without gaps.

In addition to the Phase information, submitters may include a Base Quality score, or PHRAP score, to indicate the quality of the sequence data. This base quality information can be viewed in Entrez. From the Display menu for HTG records, the "Base Quality" view option reports the PHRAP score for each base in a tabular format. The "Graphics" view presents a PHRAP trace for the sequence, with regions of lesser base quality showing up as depressions in the trace.

New Sequin Release Enhances Sequence Updating and Feature Propogation Utilities

Sequin version 3.70 for Macintosh, PC/MS Windows, and Unix Computers was made available from the NCBI FTP site recently. This version features new Update Sequence and Feature Propagation utilities.

The revised Update Sequence utility is based on an alignment indexing function and offers new update possibilities. Users may now "patch" an internal fragment of the current sequence with a corrected sequence. It is also now possible to update with an ASN.1 file containing both sequence and features. This allows overlapping records to be merged.

The revised Feature Propagation utility is found under the Edit menu of the Record Viewer. It was also rewritten with alignment indexing. If you have selected a single feature in the record viewer, only that feature will be propagated. Otherwise, all features from the designated sequence are propagated to all other sequences.

A new version of Sequin will be released shortly that features an updated alignment reader. The new reader uses intelligent parsing to read several text alignment formats, including Phylip, NEXUS, and FASTA+GAP, into ASN.1 Seq-aligns. The reader produces many different informative error messages, so that the user may identify and fix any subtle problems in the text alignment which prevent its successful conversion into a Seq-align.

Plant Genomes

continued from page 1

and calculated orthologs and homologs for nucleotide sequences represented in UniGene and LocusLink. It features pairwise comparisons between 12 plant and animal organisms, including human, mouse, rat, cow, zebrafish, clawed frog, fly, thale cress, barley, wheat, maize, and rice.

The *calculated* orthologs and homologs are the result of nucleotide sequence comparisons between all UniGene clusters and the *Drosophila* genome for each pair of organisms. These orthologs and homologs are considered putative since they are based only on sequence comparisons, and results may change as more full-length mRNA sequences become available. The calculated datasets are available by FTP at ftp://ftp.ncbi.nih.gov/pub/HomoloGene/.

Curated orthologs include ortholog gene pairs reported in the literature, in the Mouse Genome Database at the Jackson Laboratory, and in the Zebrafish Information Network at the University of Oregon.

The HomoloGene service is available at www.ncbi.nlm.nih.gov/HomoloGene/.

Plants and Frog Added to UniGene

The UniGene database of gene-oriented EST clusters now includes its first five plant organisms — thale cress *(Arabidopsis thaliana)*, rice *(Oryza sativa)*, wheat *(Triticum aestivum)*, barley *(Hordeum vulgare)*, and maize *(Zea mays)*. The frog *(Xenopus laevis)* was also added recently, bringing the total number of UniGene databases to 11.

Entrez Searches
Seven Texts on the Bookshelf

The growing list of Entrez databases now includes Books, which contains information from seven textbooks related to molecular biology. To search these texts directly in Entrez, select **Books** from the database menu and enter a text query. For example, try the terms cell cycle control, immunodeficiency, or protein evolution. This will take you to the Bookshelf page, with a report of the number of sections from each book that contain information relevant to your query. Following a link leads to a display of the relevant book sections, including a figure icon for sections that contain figures.

The Books database is also linked to terms in PubMed abstracts. When viewing an abstract, click on the **Books** link to see phrases within the abstract that are hyperlinked to book sections.

The Bookshelf includes the following texts:

C. elegans II.
Riddle, Donald L.; Blumenthal, Thomas; Meyer, Barbara J.; and Priess, James R., editors. Plainview (NY): Cold Spring Harbor Laboratory Press; c1997.

Retroviruses.
Coffin, John M.; Hughes, Stephen H.; Varmus, Harold E. Plainview (NY): Cold Spring Harbor Laboratory Press; c1997.

Molecular Biology of the Cell. 3rd ed.
Alberts, Bruce; Bray, Dennis; Lewis, Julian; Raff, Martin; Roberts, Keith; and Watson, James D. New York and London: Garland Publishing; c1994.

Introduction to Genetic Analysis. 7th ed.
Griffiths, Anthony J.F.; Gelbart, William M.; Miller, Jeffrey H.; and Lewontin, Richard C. New York: W H Freeman & Co; c1999.

Modern Genetic Analysis.
Griffiths, Anthony J.F.; Gelbart, William M.; Miller, Jeffrey H.; and Lewontin, Richard C. New York: W H Freeman & Co; c1999.

Molecular Cell Biology. 4th ed.
Lodish, Harvey; Berk, Arnold; Zipursky, S. Lawrence; Matsudaira, Paul; Baltimore, David; and Darnell, James E. New York: W H Freeman & Co; c1999.

Chapter 6, Smallpox and Vaccinia, from *Vaccines. 3rd ed.*
Plotkin, Stanley A.; and Orenstein, Walter A., editors. Philadelphia: W. B. Saunders Company; c1999.

For more information about the Bookshelf, select **Books** from the Entrez home page at www.ncbi.nlm.nih.gov/Entrez.

Selected Recent Publications by NCBI Staff

Kim, W, and **WJ Wilbur.** Amino acid residue environments and predictions of residue type. *Comput Chem* 25(4): 411-22, 2001.

Makarova, KS, VA Ponomarev, and **EV Koonin.** Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol* 2(9): RESEARCH0033, 2001.

Shabalina, SA, AY Ogurtsov, VA Kondrashov, and **AS Kondrashov.** Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet* 17(7):373-6, 2001.

Wolf, YI, FA Kondrashov, and **EV Koonin.** Footprints of primordial introns on the eukaryotic genome: still no clear traces. *Trends Genet* 17(9):499-501, 2001.

Brinkman, FS, and **DD Leipe.** Phylogenetic analysis. *Methods Biochem Anal* 43:323-58, 2001.

Carlton, JM, K Hayton, PV Cravo, and D Walliker. Of mice and malaria mutants: unraveling the genetics of drug resistance using rodent malaria models. *Trends Parasitol* 17(5):236-42, 2001.

Aravind, L, and **EV Koonin.** Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res* 11(8):1365-74, 2001.

Galperin, MY, AN Nikolskaya, and **EV Koonin.** Novel domains of the prokaryotic two-component signal transduction systems. *FEMS Microbiol Lett* 203(1):11-21, 2001.

Mitchell, BD, WC Hsueh, TM King, TI Pollin, J Sorkin, **R Agarwala, AA Schäffer,** and AR Schuldiner. Heritability of life span in the Old Order Amish. *Am J Med Genet* 102(4):346-52, 2001.

BLAST Enhancements in Recent Releases

The last two releases of BLAST (2.2.1 and 2.2.2) introduced several enhancements:

- In both the Web and standalone versions of BLAST and PSI-BLAST, composition-based statistics and edge-correction have been improved.
- The latest standalone version of PSI-BLAST, blastpgp, accepts batch input.
- Formatdb now automatically produces database volumes if a source database is larger than 4 billion letters.
- The tabular output option has been added to blastpgp and rpsblast under the command line switch "-m 8".
- A new -D option for fastacmd will dump an entire BLAST database in FASTA format.
- A new -c option for fastacmd creates separate definition lines for multiple GI numbers (with duplicate sequenc data) that have been merged together in nr.
- Version 4 of the BLAST databases is now fully supported. Use the -A option with formatdb to produce the new database version.

The BLAST binaries for release 2.2.2 are entirely compatible with both the current and the new version 4 of the BLAST databases.

For more information, and to download the latest version, see: ftp://ftp.ncbi.nih.gov/blast/

To receive automatic notification of BLAST version releases, subscribe to the BLASTAnnounce service at: www.ncbi.nlm.nih.gov/BLAST /blastannounce.html.

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