NCBI News

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

Fall 2001

Using TaxPlot to Compare Genomes

TaxPlot is a tool for 3-way comparisons of genomes on the basis of the protein sequences they encode. To use TaxPlot, one selects a reference genome to which two other genomes are compared. Pre-computed BLAST results are then used to plot a point for each predicted protein in the reference genome, based on the best alignment with proteins in each of the two genomes being compared.

Figure 1 shows a TaxPlot in which *E. coli* K12 has been selected as the reference genome for comparison of two strains of *H. pylori*, J99 and 26695. Each point in the figure represents a single *E. coli* protein. The X and Y coordinates represent the BLAST score for the protein's closest match in the two strains of *H. pylori*. There are 217 *E. coli* proteins that are equally similar to proteins in the two *H. pylori* strains, as shown by the points lying on the central diagonal. *E. coli* has 678 proteins with greater similarity to *H. pylori* strain J99 (if only marginally), and 687 proteins with greater similarity to *H. pylori* strain 26695.

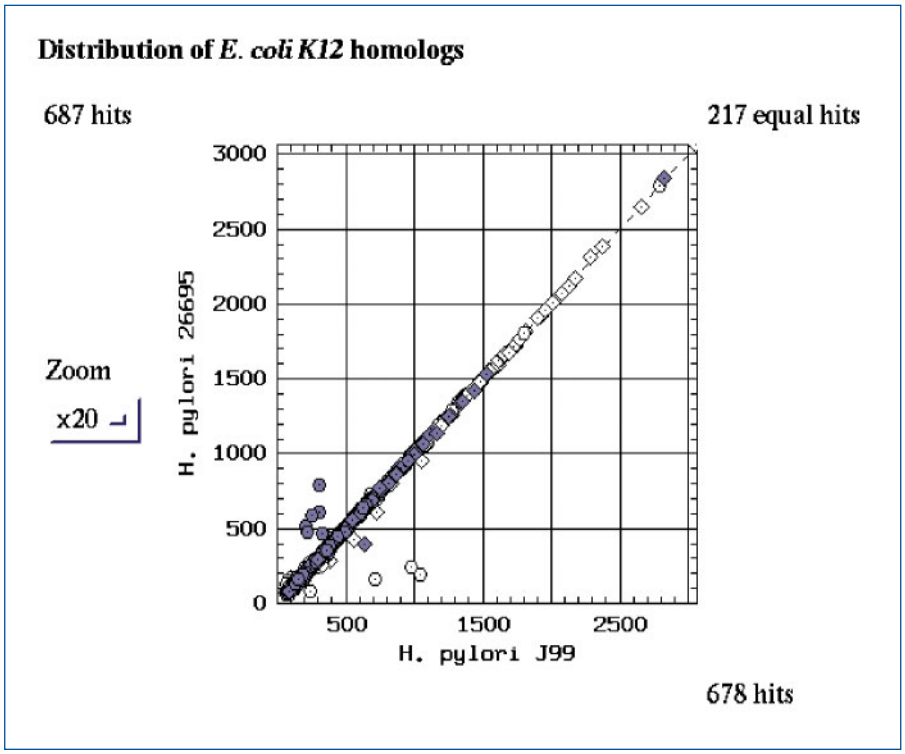

Figure 1: TaxPlot for two strains of *H. pylori* against *E. coli* as the reference genome. Points representing proteins involved in amino acid transport and metabolism are highlighted in blue.

Overall, the proteomes of the two *H. pylori* strains appear to be equally similar to that of *E. coli*. However, a few significant differences between the *H. pylori* strains show up as off-diagonal points toward the left-hand portion of the plot. These points represent proteins in *E. coli* that better match in one strain of *H. pylori* than in the other.

For instance, a number of *E. coli* proteins have low BLAST scores

continued on page 2

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

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)

Using TaxPlot

continued from page 1

against the *H. pylori* J99 strain, yet relatively high BLAST scores against the 26695 strain. These points may represent cases in which selection pressures operating on the orthologs of these *E. coli* proteins in the two *H. pylori* strains are different. To determine if there is a pattern to these differences, one may identify individual points on the plot to learn the function of the *E. coli* proteins indicated.

Subsets of the points plotted can be selected by simply clicking on an area of the graph or by using a menu box to select proteins by functional class. Hyperlinks to the BLAST2 Sequences service provide displays of pairwise alignments. In the figure, those proteins known in *E. coli* to be involved in amino-acid transport and metabolism have been selected and appear blue in the plot. Note that most of the off-diagonal proteins of this type are more similar to proteins from *H. pylori* strain 26695 than J99, suggesting that *H. pylori* strain J99 may be undergoing a restructuring of some aspects of its amino acid processing systems. Such restructuring could represent an important adaptation in the J99 strain of relevance to its pathogenesis.

The TaxPlot tool is accessible from the Entrez Genomes Web page, under Tools and Analysis. In addition to the microbial genome version described here, there is also a TaxPlot service for eukaryotic genomes.

New RefSeq Accession Numbers for Curated Genomic Regions

A new type of RefSeq accession number, of the form "NG_######", now appears in Entrez. These accession numbers are used to designate curated genomic segments. To see all of these records in Entrez, use the query: NG_*[accn]

Examples of curated genomic regions include the human hair keratin gene cluster on chromosome 17 (NG_000018), the human alpha globin region on chromosome 16 (NG_000006), and the human MHC class III complement gene cluster on chromosome 6 (NG_000013). Such gene clusters are difficult to assemble in a purely automated fashion and are therefore maintained manually at NCBI as curated genomic segments.

Currently, all NG_ accession numbers are the result of manual curation and are thus considered to be reviewed RefSeq records. The mRNAs and proteins that these genomic segments encode are also curated and available under distinct RefSeq accession numbers with the prefixes NM_ for nucleotides and NP_ for proteins.

Notice to Subscribers

For your records, there was no Summer 2001 issue of *NCBI News*.

Note that *NCBI News* is also available online. From the NCBI home page, select the About NCBI link.

GenBank *News*

New Genomes in GenBank

The following new complete microbial genomes are now available in GenBank under the accession numbers listed:

Streptococcus pneumoniae: AE005672

Streptococcus pneumoniae strain R6: AE007317

Agrobacterium_tumefaciens: AE007869, circular chromosome; AE006469, linear chromosome; AE00782, plasmid AT; and AE007871, plasmid Ti

Sinorhizobium meliloti: AL591688, main chromosome; AE006469, "megaplasmid" pSymA; and AL591985, "megaplasmid" pSymB

View these complete genomes and learn more about the proteins they encode from Entrez Genomes.

Select the Genomes database from the Entrez home page at: http://www.ncbi.nlm.nih.gov/entrez/

Complete genomes are also available for downloading by FTP from:

ftp://ftp.ncbi.nih.gov/genomes/

GenBank Mirror Sites

To provide alternative sites for downloading GenBank releases and updates, GenBank is now mirrored at two sites:

The San Diego Supercomputer Center (ftp://genbank.sdsc.edu/pub)

Indiana University (ftp://bio-mirror.net/biomirror/genbank).

New High-Throughput cDNA (HTC) Division in GenBank

A new GenBank division has been created for unfinished High-Throughput cDNA sequencing (HTC) data. Sequences in this division may still have 5' and 3' UTRs at their ends, partial coding regions, and introns. Finished HTC sequences will be moved to the appropriate taxonomic GenBank division, in the same manner as High Throughput Genomic (HTG) records are moved into a taxonomic division upon finishing of genomic sequences. Release 126 (October 2001) of GenBank contains 22,002 HTC sequences totaling 27,985,613 bases. The bulk of these HTC sequences (89%) are from *Mus musculus,* with about 10% from *Homo sapiens*. To view these records, use the Entrez query:

"gbdiv_htc"[Properties]

GenBank Release 126

GenBank release 126 (October 2001) contains over 13 million sequences totaling more than 14 billion base pairs. GenBank may be searched using NCBI's Entrez search and retrieval system or may be downloaded from the NCBI FTP site. GenBank flatfiles can be downloaded from the "genbank" directory; the ASN.1 version of GenBank is found in the "ncbi-asn1" directory.

New FTP Address

Note that NCBI's FTP address has changed to: ftp.ncbi.nih.gov

For Web access, the URL is: ftp://ftp.ncbi.nih.gov

Selected Recent Publications by NCBI Staff

Anantharaman, V, EV Koonin, and **L Aravind.** Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J Mol Biol* 307(5): 1271-92, 2001.

Bhagwat, M and NG Nossal. Bacteriophage T4 RNase H removes both RNA primers and adjacent DNA from the 5' end of lagging strand fragments. *J Biol Chem* 276(30):28516-24, 2001.

Jordan, IK, KS Makarova, JL Spouge, YI Wolf, and **EV Koonin.** Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res* 11(4):555-65, 2001.

Nasidze, I, GM Risch, M Robichaux, **ST Sherry,** MA Batzer, and M Stoneking. Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. *Eur J Hum Genet* 9(4):267-72, 2001.

Ostell, JM, SJ Wheelan, and **JA Kans**. The NCBI data model. *Methods Biochem Anal* 43:19-43, 2001.

Schäffer, AA, L Aravind, TL Madden, S Shavirin, JL 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.

Sreekumar, KR, **L Aravind,** and **EV Koonin.** Computational analysis of human disease-associated genes and their protein products. *Curr Opin Genet Dev* 11(3):247-57, 2001.

Nomura, T, **JM Carlton,** JK Baird, HA del Portillo, DJ Fryauff, D Rathore, DA Fidock, X Su, WE Collins, TF McCutchan, **JC Wootton,** and TE Wellems. Evidence for different mechanisms of chloroquine resistance in 2 Plasmodium species that cause human malaria. *J Infect Dis* 183(11): 1653-61, 2001.

DART Targets Protein Domains

NCBI has strengthened its suite of protein structural analysis tools by introducing the Domain Architecture Retrieval Tool, or DART. Beginning with a protein sequence, DART facilitates searches of the Entrez protein database for protein domain combinations, or architectures.

DART works by first determining the domain architecture of a protein sequence query. It then displays the domain architectures of other proteins that share at least one domain with the query. Figure 1 shows the initial DART report for *E. coli* DNA polymerase I, an enzyme with three distinct domains—a 5'→3' exonuclease domain, a proofreading 3'→5' exonuclease domain, and a DNA polymerase domain. These three domains of the query protein are identified at the top of the report graphic. Below the query architecture graphic is a list of other domain architectures that share at least one domain with the query, with links to the proteins that have these architectures.

Following the graphic output is a list of domain identifiers and descriptions, as shown in Figure 2. One or more domains can be selected for use as a query for a second round of searching. Hence, the first round search seeks any of the domains found in the query, while subsequent rounds can be tailored to find selected domains.

DART displays can also be limited to architectures found in a single organism, taxonomic class, or combination of organisms or classes by using a prunable taxonomic tree.

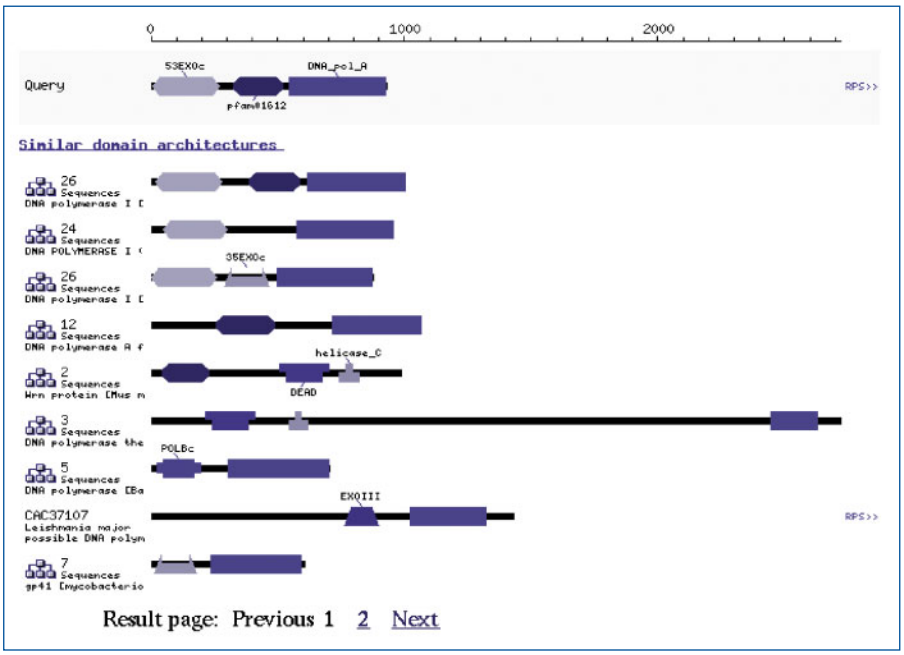

Figure 1: DART display for *E. coli* DNA polymerase I.

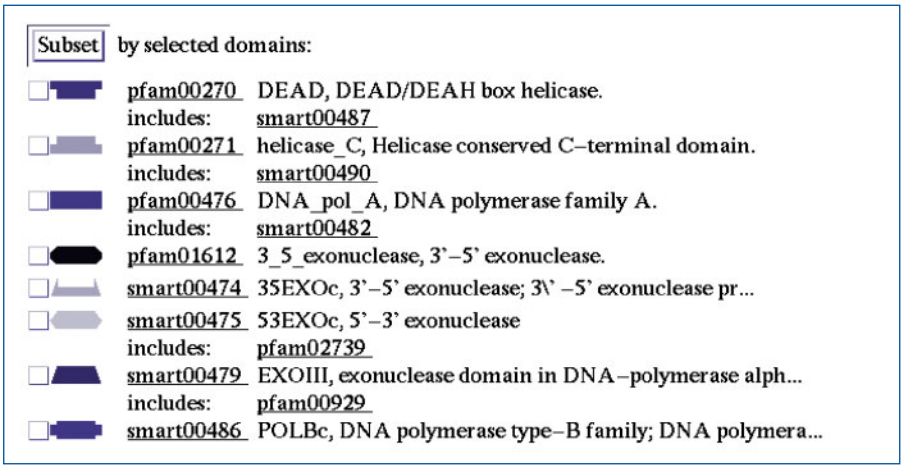

Figure 2: List of domains shown in DART display of Figure 1. Checking the box beside a domain constrains subsequent DART searches to seek this domain.

An interesting observation in the output shown in Figure 1 is that the *E. coli* DNA polymerase I query picks up the human WRN helicase protein due to the presence of a 3'→5' exonuclease domain at the WRN helicase N-terminus. Looking at the structure of the WRN helicase gene using the NCBI Evidence Viewer (see accompanying article, Figure 1), one can see that there is a 5' exon cluster separated from the rest of the gene by an intron. In fact, abstracting the protein sequence coded in the exons of this cluster by selecting and copying from the Evidence Viewer report, then performing a Conserved Domain Search (www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), reveals that this 5' cluster of exons codes for the 3'→5' exonuclease domain module. Hence, the modular structure of the gene parallels the modular structure of the protein.

Evidence Viewer Facilitates Analysis of NCBI Human Gene Models

NCBI has an ongoing program to assemble and annotate the human genome, incorporating updates as new and revised genome data is deposited in public resources. As part of the process, NCBI generates gene models based primarily on alignment of mRNA sequences to the human genomic assembly. These alignments are used as evidence of the intron/exon organization of a gene, as annotated on the contigs. NCBI has developed an Evidence Viewer so that users can see the alignment evidence for the gene models when mRNA is used in this way.

Links to the Evidence Viewer (ev) are now provided in LocusLink and the Human Genome Map Viewer when gene models are presented in an output report. The model sequences are designated with accession numbers beginning with XM_ for nucleotide and XP_ for protein sequences.

In Figure 1, the Evidence Viewer graphic shows a genomic contig from NCBI's human genome assembly (NT_007993, Build 25) aligned to a GenBank mRNA sequence (AF091214) for the human WRN helicase gene involved in Werner Syndrome (OMIM number 277700). Also aligned are an NCBI mRNA Reference Sequence (NM_000553) and an NCBI mRNA Model Sequence (XM_015858). Not included in the figure is the detailed base-by-base ev alignment view, which follows the graphical overview in the report.

The 35 exons of the WRN gene, implied by the mRNA to genomic sequence alignments, are shown

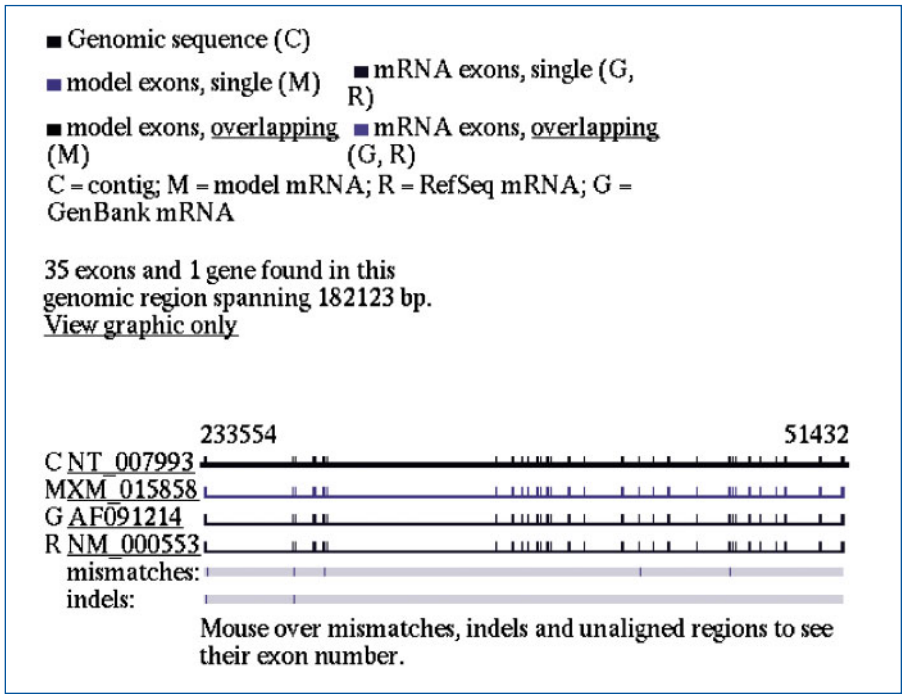

Figure 1: Evidence Viewer Display for human WRN helicase gene.

as vertical tick marks along the gene. Single nucleotide mismatches between the mRNA sequences and the corresponding genomic sequence, as well as insertions and deletions, are marked on the "mismatches" and "indels" scales immediately below the alignments.

In the case of the WRN gene, the NCBI mRNA RefSeq NM_000553 was constructed from the GenBank mRNA record AF91214, then aligned to the genomic sequence to produce the NCBI-generated mRNA Model Sequence XM_015858. The single gene model currently given on the Genes_sequence map of the Map Viewer is based on the mRNA alignment, shown in Figure 1, between the RefSeq and genomic sequence.

Figure 1 indicates that the WRN gene has a distinct 5' exon cluster separated from the rest of the gene by an intron. It is interesting to speculate that this exon cluster may code for a distinct domain module in the WRN helicase protein. In fact, it does! See the article on DART in this issue to learn the identity of this protein domain.

The simplest case is illustrated here, where a single mRNA aligns to a single place in the genome. However, there can also be multiple or overlapping models. This can be due to a number of reasons, including splice variants, paralogous genes, or inaccuracies in the draft sequence or assembly. The Evidence Viewer is useful in helping researchers to analyze the alternative models presented—e.g., to see where the mismatches are and decide how to interpret the evidence.

Q & A

Frequently Asked Questions

Q.

A.

I have a lot of accession numbers that I'm interested in. Is there any easy way to retrieve them through Entrez, without having to do individual searches for each one?

Yes, Batch Entrez allows you to do this. You can create a file containing the list of accession numbers (or gi numbers), save it on your computer. Click on the Batch Entrez link, which appears in the left-hand sidebar of the Entrez screens. Enter the name of your file in the search box, choose the Nucleotide or Protein database, and press Retrieve. Then proceed as with any Entrez search to display and/or save the results.

How can I download the sequence immediately surrounding a particular gene from the NCBI Human Genome Assembly?

Use the Human Genome Mapviewer. Search for the gene of interest using its gene symbol and click on the link to the gene when the search results are displayed. To retrieve a segment of sequence including the gene, and both upstream and downstream regions, simply click on the "Download/ View Sequence/Evidence" link in the MapViewer. To alter the segment downloaded, change the coordinates for the region and press the "Change Region" button.

Can I use the e-PCR tool to design primers to amplify my gene of interest?

You wouldn't use e-PCR to design the primers per se, but you can use it to see if the UniSTS database already contains a primer pair that you could use to amplify your sequence.

The e-PCR service tool tests a DNA sequence for the presence of sequence tagged sites by comparing the query sequence against the UniSTS database of STS sequences and primer pairs. It looks for STSs in DNA sequences by searching for subsequences that closely match the PCR primers in UniSTS and have the correct order, orientation, and spacing that they could plausibly prime the amplification of a PCR product of the correct molecular weight. The output is a table of links to matching UniSTS records, as well as the primer pairs and PCR conditions used to amplify each STS identified.

caaatccgttcttgatcgtacatagcgcatgtcagncaaatccgttcttg
caaatccattcttgatcgtacatggcacatgtcagtcaaatccattcttg

How to Search Custom Databases in Web-BLAST Using Entrez Queries

A powerful new feature of the BLAST Web interface is the ability to limit BLAST searches to a subset of any database using a standard Entrez query. The limitation by Entrez query gives users a flexibility in database searching hitherto obtainable only by using a local version of BLAST with a locally constructed database. Skillful use of Entrez queries allows the equivalent of on-the-fly construction of databases of exact composition.

Entrez queries are entered into a box in the BLAST formatting area entitled "Limitation by Entrez Query". It is helpful to first construct the query from within Entrez and verify that it returns the desired subset of sequences, before attempting to use it with BLAST. A successful Entrez text query may be pasted into the Entrez Limitation box on the BLAST page. Be sure that the database chosen is compatible with the Entrez limitation used. For example, an Entrez query that picks up ESTs will be of no use when searching the nr database since nr contains no EST sequences. Two examples will illustrate the utility of the Entrez Limitation.

In the simplest case, one might desire to search the nr protein database for matches to a particular type of protein from a particular class of organisms. The following Entrez query defines a search of only viral helicase proteins:

viruses[orgn] AND helicase [protein name]

Note that the limitation above will pick up annotated helicase proteins, but not unannotated proteins. It will, however, ensure that your results contain nothing but viral proteins annotated as helicase or helicase-related proteins.

Another interesting use of the Entrez query limitation is to limit a BLAST search to an arbitrary set of sequences defined by a list of accession numbers.

Suppose that you have identified a UniGene cluster of interest and wish to BLAST an mRNA from that cluster against the ESTs from the same cluster in order to see how they map to the mRNA. This can be accomplished without downloading a single sequence by using the human EST database in conjunction with an Entrez limitation. Take as an example, UniGene cluster Hs.2. This cluster includes several mRNA sequences, of which we will use one, D90042, as our query. We wish to map the 18 ESTs belonging to this cluster to the mRNA, so we need to BLAST D90042 against these 18 sequences. The UniGene report for Hs.2 lists the accession numbers involved, as shown in

continued on page 8

EST SEQUENCES (18)

BG533459	cDNA clone IMAGE:4072143 liver	5' read	P M
BG568400	cDNA clone IMAGE:4716802 liver	5' read	P M
BG569293	cDNA clone IMAGE:4722596 liver	5' read	P M
BG618195	cDNA clone IMAGE:4767316 liver	5' read	M
BG563731	cDNA clone IMAGE:4712210 liver	5' read	P M
AI792606	cDNA clone IMAGE:1870937 colon	5' read 1.2 kb	P
AI262683	cDNA clone IMAGE:1870937 colon	3' read 1.2 kb	P A
AI733799	cDNA clone IMAGE:1870937 colon	3' read 1.2 kb	P A
BG569272	cDNA clone IMAGE:4722638 liver	5' read	P
BG566307	cDNA clone IMAGE:4712733 liver	5' read	P
AV658623	cDNA clone GLCFOD10	3' read	P
AV684197	cDNA clone GKCFZH06	5' read	
AV658656	cDNA clone GLCFOG07	3' read	P
BG617259	cDNA clone IMAGE:4734378 liver	5' read	
AU099534	cDNA clone HSI08034		
BF126423	cDNA clone IMAGE:4071536 liver	5' read	P
BG204539	cDNA clone (no–name)		
AI460128	cDNA clone IMAGE:2151449 colon	3' read	P A

Figure 1: List of EST sequence in UniGene cluster Hs.2 taken from a portion of the UniGene report.

BLAST Lab
continued from page 7

Figure 1. We can simply select and copy this portion of the report into a text editor and then parse the accessions by hand or by using a script. However we perform the parsing, we hope to wind up with an Entrez limitation of the form:

BG569293 OR BG533459 OR....

In this case, the list of 18 accession numbers are specified explicitly, connected with Boolean OR logic.

The graphical overview for this search, given in Figure 2, shows the alignment of the 18 ESTs in Hs.2 to mRNA D90042 from the same cluster. Such an alignment is a useful way to visualize the distribution of 3' and 5' ESTs within a cluster. The BLAST report itself can be used to distinguish between 5' and 3' ESTs.

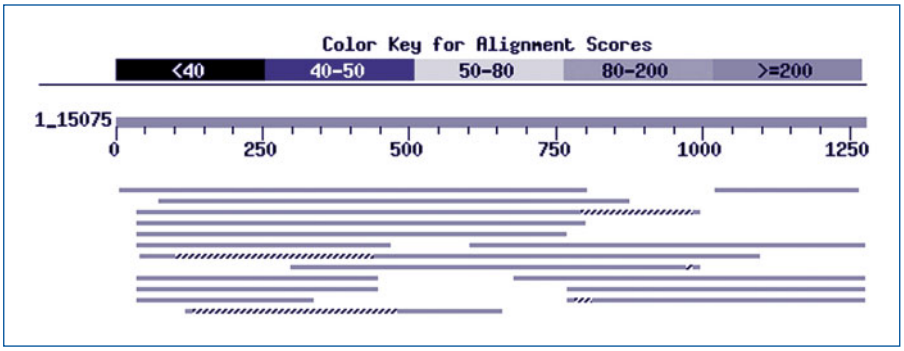

Figure 2: BLAST graphical overview for search with an mRNA from UniGene clusterHs.2 against 18 ESTs from the same cluster.

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.

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