DOI: 10.1093/nar/gkh435

BLAST: at the core of a powerful and diverse set of sequence analysis tools

Scott McGinnis* and Thomas L. Madden

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,
Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received February 20, 2004; Revised April 2, 2004; Accepted April 14, 2004

ABSTRACT

Basic Local Alignment Search Tool (BLAST) is one of the most heavily used sequence analysis tools available in the public domain. There is now a wide choice of BLAST algorithms that can be used to search many different sequence databases via the BLAST web pages (http://www.ncbi.nlm.nih.gov/BLAST/). All the algorithm–database combinations can be executed with default parameters or with customized settings, and the results can be viewed in a variety of ways. A new online resource, the BLAST Program Selection Guide, has been created to assist in the definition of search strategies. This article discusses optimal search strategies and highlights some BLAST features that can make your searches more powerful.

INTRODUCTION

Basic Local Alignment Search Tool (BLAST) is a sequence similarity search program that can be used via a web interface or as a stand-alone tool (1,2). There are several types of BLAST to compare all combinations of nucleotide or protein queries with nucleotide or protein databases. BLAST is a heuristic that finds short matches between two sequences and attempts to start alignments from these 'hot spots'. In addition to performing alignments, BLAST provides statistical information to help decipher the biological significance of the alignment; this is the 'expect' value, or false-positive rate.

The BLAST server at the National Center for Biotechnology Information (NCBI) now has a diverse set of features that can add power to your BLAST searching. The BLAST homepage (http://www.ncbi.nlm.nih.gov/BLAST/) lists the varieties of BLAST searches by type: Nucleotide, Protein, Translated and Genomes. Table 1 documents the default parameters for each link. In the online version of this table (http://www.ncbi.nlm.nih.gov/blast/link_params.html), each cell of the top row and leftmost column of the online version is hyperlinked to a description of that column or row.

The Program Selection Guide can assist in the selection of search type and databases (http://www.ncbi.nlm.nih.gov/BLAST/producttable.shtml). The default nucleotide database used is 'nt', i.e. GenBank without the high-throughput, patent, genomic or sequence tagged site (STS) sequences (see http://www.ncbi.nlm.nih.gov/BLAST/Why.shtml#NUC_ for more details). The default protein database is 'nr': a non-redundant set of all the non-patent sequences; i.e. sequences that are exactly the same over their entire length are merged into one database entry, although information about the sequences that make up the entry is preserved (see http://www.ncbi.nlm.nih.gov/BLAST/Why.shtml#PROT_DB for more details).

When the query is submitted, either as a sequence in FASTA format or as a sequence identifier, e.g. GenBank accession.version, the search is sent to the BLAST server and a 'Request Identifier' (RID) is returned. The query and results are stored in a structured format for up to 24 h after an RID is issued. The RID identifies the query and allows the results to be viewed in several formats, which include the familiar BLAST report, a simplified 'hit table', XML and ASN.1 [(3) and http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.610]. The number of outstanding jobs from one IP address is taken into account when queuing requests, as described at http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.shtml#Queuetime, so that one user does not monopolize the entire service. Infrequently a user will be blocked for an excessive number of queries; typically, this is some user who has become overly enthusiastic with a PERL script rather than a deliberate denial of service attack. Blocking is performed on a case-by-case basis. A user scripting queries to the BLAST server is advised to wait 3 s between queries; not to poll for results for any given RID more often than every 2 min; and to run all scripts between the hours of 9 p.m. and 5 a.m. EST USA. Users anticipating a large volume of searches (several hundred or more) may wish to email blast-help@ncbi.nlm.nih.gov for assistance in formulating their queries and advice on submitting searches.

*To whom correspondence should be addressed. Tel: +1 301 435 5945; Email: mcginnis@ncbi.nlm.nih.gov

Table 1. List of the different links available on the NCBI BLAST home page with the default parameters for each link

	Expect value	Word size	Reward match	Penalty mismatch	Gap existence	Gap extension	Percentage identity	Filtering	Matrix
Nucleotide									
Discontiguous megaBLAST	10	11	1	−2	0	2.5	None	Low complexity; mask for lookup table	–
MegaBLAST	10	28	1	−2	0	2.5	None	Low complexity	–
Standard BLASTN	10	11	1	−3	5	2	–	Low complexity	–
Short Nucleotide Sequences	1000	7	1	−3	5	2	–	None	–
Protein									
Standard blastp	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Psi-BLAST	10	3	–	–	11	1	–	None	BLOSUM62
Phi-BLAST	10	3	–	–	11	1	–	None	BLOSUM62
Short Protein Sequences	20 000	2	–	–	11	1	–	None	PAM30
RPS-BLAST	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Translated									
Blastx	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Tblastn	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Tblastx	10	3	–	–	11	1	–	Low complexity	BLOSUM62
Special Pages									
Genome BLAST pages (blastn)	0.01	28	1	−2	0	0	None	Low complexity; Human repeat	–

An online version of this table (http://www.ncbi.nlm.nih.gov/blast/link_params.html) contains hyperlinks pointing to definitions of the parameters listed.

SEARCH STRATEGIES

Using the default settings for a BLAST search is a sensible approach because they should give the best all-round results. Moving beyond the default settings by changing the type of search and the search parameters requires a strategic approach. Generally, there is a tradeoff between speed and sensitivity and a user should try to use the fastest set of parameters sensitive enough for the task at hand. Use of overly sensitive settings, especially with long queries or very large databases, can mean an excessive wait time for the user or that the job will exceed the CPU resource limit on the server and only an error message will be returned to the user. The BLAST web page encourages optimal parameter setting by offering a number of links for specific purposes, described in Table 1. If the goal is identification of a sequence or an intra-organism comparison, then it is best to use a fast and stringent search. Otherwise, it might be necessary to use more sensitive settings which normally come at a cost in terms of time taken to run the search. In this section we discuss the items in Table 1 under ‘Nucleotide’ and ‘Protein’. We discuss other sections of the table as appropriate in the rest of this article.

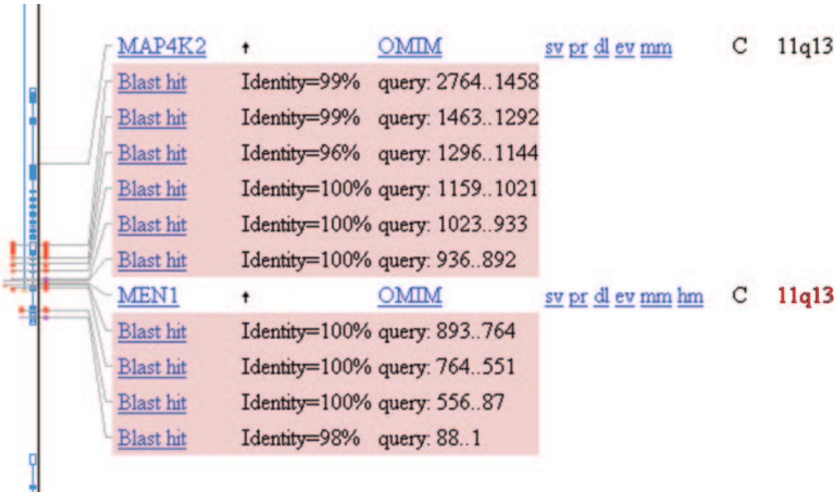
The speed and sensitivity of nucleotide–nucleotide searches varies most dramatically with the word-size and the type of gapped extension. The fastest program is megaBLAST, which defaults to a large word-size (an exact match of 28 bases is required to initiate an extension) and a greedy gapped extension algorithm (4) that has no gap existence cost but merely a gap extension cost, making it ideal for comparing similar sequences (e.g. from the same organism). More sensitive is a search with discontiguous megaBLAST, which uses discontiguous word-matches (not all bases in a word are required to match) to initiate extensions (5,6). It also uses a non-greedy extension, an option that is appropriate for comparisons up to 80% identity. Roughly as sensitive but generally slower (especially for longer queries) is the ‘standard’ BLASTN, which uses an 11-base contiguous word to initiate extensions. Very slow is the option for ‘short’ nucleotide searches. This option is intended only for very short sequences that contain little information and might otherwise not find any hits. Using this option with a query longer than 50 bases will probably exceed the server’s CPU resource limit.

The fastest way to identify the function of a protein is to perform a CDD search (7), which uses a database of motifs to characterize ‘conserved-domains’ in a protein sequence. This normally takes just a few seconds and a CDD search is actually performed for every protein–protein search by default. The standard protein–protein search option provides good all-round search parameters. Use of the PSI-BLAST page allows the user to initiate an iterative search (2) that produces a position-specific scoring matrix for further searches. PHI-BLAST searches a database looking only for alignments that include a specified pattern (8). The short sequence option makes use of the PAM30 matrix, which is recommended for short sequences that contain little information (9).

Another consideration is which dataset to search; a database consisting of well-curated sequences will return database matches that are more accurately annotated and contain fewer sequencing errors or vector contamination. Another, more subtle issue, concerns the ‘expect value’ for the matches found. The expect value indicates the validity of the match: the smaller the expect value, the more likely the match is ‘good’ and represents real similarity rather than a chance match (see http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html for more details). The expect value scales roughly with the size of the database; therefore, if it is a database in which 90% of the sequences are not of interest, e.g. they are from the wrong species, then the expect value of all hits is increased by a factor of 10, i.e. the false-positive rate will be higher.

SEARCH AN ENTIRE GENOME

Traditionally, users have chosen nt or nr for their searches. Often this is no longer the optimal choice. nt contained



```
MAP4K2     †                OMIM     sv pr dl ev mm      C  11q13
Blast hit  Identity=99%  query: 2764..1458
Blast hit  Identity=99%  query: 1463..1292
Blast hit  Identity=96%  query: 1296..1144
Blast hit  Identity=100% query: 1159..1021
Blast hit  Identity=100% query: 1023..933
Blast hit  Identity=100% query: 936..892
MEN1       †                OMIM     sv pr dl ev mm hm   C  11q13
Blast hit  Identity=100% query: 893..764
Blast hit  Identity=100% query: 764..551
Blast hit  Identity=100% query: 556..87
Blast hit  Identity=98%  query: 88..1
```

Figure 1. BLAST matches against the human genome presented in the NCBI Map Viewer. The query was the MEN1 mRNA (GenBank accession U93236) from (10). Ten alignments to NCBI RefSeq accession.version NT_033903.6 (a contig on chromosome 11) were found, corresponding to the 10 exons described in (10). The hits are marked as red or pink lines, depending upon the score (red indicates the strongest matches). The MEN1 exons are shown in blue, indicating a 'Confirmed gene model' (see http://www.ncbi.nlm.nih.gov/mapview/static/humansearch.html#genes for more information). The alignments are in reverse order owing to the antisense transcription on the genomic sequence (10).

~10 billion bases as of February 2004, an increase of ~20% from February 2003; nr contained ~540 million residues as of February 2004, an increase of ~25% from a year earlier (C. Camacho, personal communication). Recent announcements (see http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=coffeebrk.chapter.622 and references therein) indicate that this trend will continue. The human genome database contains ~3 billion bases; for users interested only in human sequences, searching the human genome rather than nt is about three times as 'efficient'. The protein subset of the human genome is also only ~12 million residues. Nucleotide–nucleotide searches on the human genome BLAST page include filtering for human repeats using a 'mini-BLAST' search (T. Madden, personal communication) alongside low-complexity filtering by default. The default expect value cutoff on the genome page is set to a more conservative 0.01, rather than the default 10, under the assumption that the search of a genome is very targeted. Results for the search of an mRNA for the MEN1 gene [GenBank accession number U93236; (10)] against the human genome are presented in Figure 1.

LIMIT BY ENTREZ QUERY

The sequence data at NCBI are divided into separate databases for BLAST searching [e.g. expressed sequence tags (ESTs), Trace Archives and whole genomes]. Further restriction can be applied using the 'Limit by Entrez Query' option on the BLAST search pages. For example, the search can be limited to a specific organism or to taxonomic groups, such as Mammals or Archaea. Any query in the Entrez search format (http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers) can be used to restrict the search.

Consider the sequence from GenPept accession.version AAH04246.1, described as the human homolog of an *Escherichia coli* DNA mismatch repair protein. When searching against the nr database with no restriction by organism or other criteria and using the default display limit of 100 database sequences, no hits to *E.coli* are found. However, if the search is limited to *E.coli* by selecting it from the pull-down list of organisms, there are 47 matches to *E.coli*. The top-scoring match is a 343-residue alignment between the query and a sequence entitled 'mismatch repair protein' (GenPept accession number AAM82372) with an expect value of 1.0e-50. To limit the results still further, use an additional Entrez query. For example, to search only *E.coli* sequences from the Reference Sequence collection [(11) and http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.681], use the Entrez query

''srcdb refseq''[prop]

and limit the search to *E.coli*. The result is seven matches, with the top-scoring match being a 343-residue alignment to NCBI RefSeq protein accession.version NP_755173.1 (again a mismatch repair protein). Other possibilities for a general Entrez query are to restrict the search by date or to only mRNA sequences.

USE THE PROGRAM SELECTION GUIDE

The number of BLAST programs and databases now available can make choosing a search strategy a daunting task. To address this, a new tool called the 'Program Selection Guide' (http://www.ncbi.nlm.nih.gov/BLAST/producttable.shtml) has been designed to assist users. It has been organized on three basic characteristics: (i) the nature of the query sequence, (ii) the purpose of the search and (iii) the dataset intended as the target of the search. An example of how to use the Guide with a nucleotide query is shown in Figure 2.

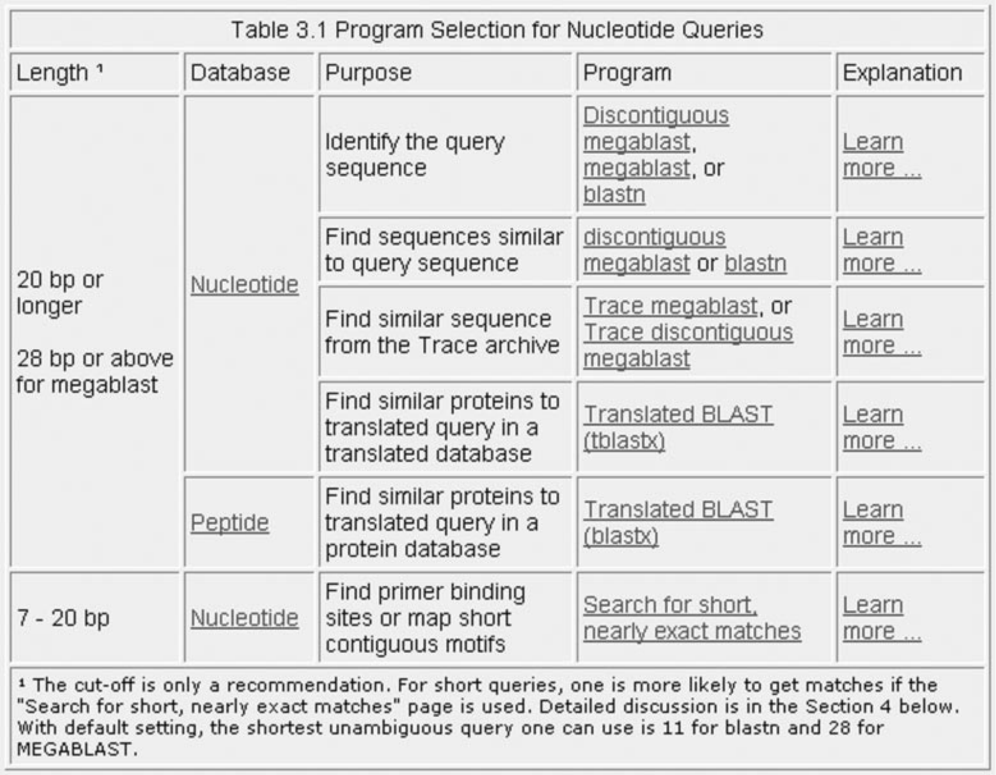


Table 3.1 Program Selection for Nucleotide Queries

Length ¹	Database	Purpose	Program	Explanation
20 bp or longer 28 bp or above for megablast	Nucleotide	Identify the query sequence	Discontiguous megablast, megablast, or blastn	Learn more ...
		Find sequences similar to query sequence	discontiguous megablast or blastn	Learn more ...
		Find similar sequence from the Trace archive	Trace megablast, or Trace discontiguous megablast	Learn more ...
		Find similar proteins to translated query in a translated database	Translated BLAST (tblastx)	Learn more ...
	Peptide	Find similar proteins to translated query in a protein database	Translated BLAST (blastx)	Learn more ...
7 - 20 bp	Nucleotide	Find primer binding sites or map short contiguous motifs	Search for short, nearly exact matches	Learn more ...

¹ The cut-off is only a recommendation. For short queries, one is more likely to get matches if the "Search for short, nearly exact matches" page is used. Detailed discussion is in the Section 4 below. With default setting, the shortest unambiguous query one can use is 11 for blastn and 28 for MEGABLAST.

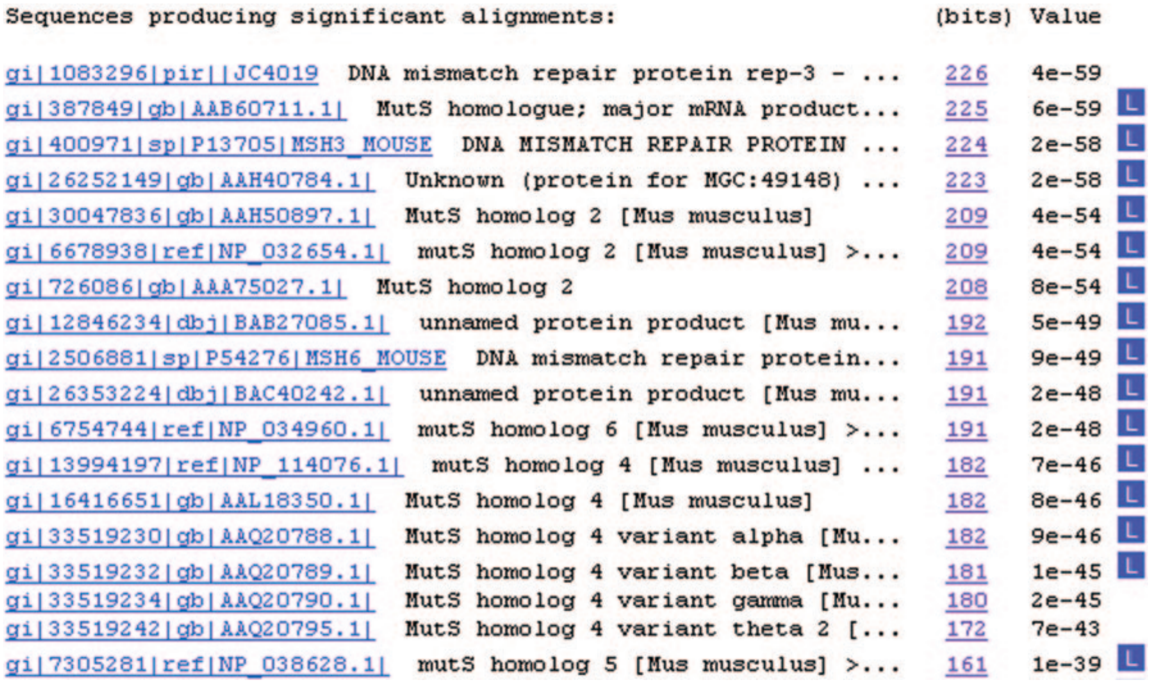
Figure 2. A portion of the third table from the BLAST Program Selection Guide. The focus is on nucleotide queries. Starting from the left side the user chooses the proper row and then moves to the right. Assuming the user has a query >20 bases she would then have the choice between a nucleotide or protein database. For a nucleotide database the user should then pick a choice from the 'purpose' column that best describes her goal and use the corresponding link or links. Links from the BLAST Program column take you directly to the search page. The Learn More links provides a list of available databases.

USE megaBLAST FOR MULTIPLE QUERIES

A common function in high-throughput sequencing projects is to group nucleotides of related function together. A reasonable approach is to first find the very obvious similarities with a fast algorithm (using a nucleotide–nucleotide comparison with a large word-size), and then to use more sensitive algorithms on the sequences that did not have strong matches in the earlier step (e.g. using a smaller word-size or a translating search). As discussed above, megaBLAST was created specifically for the task of efficiently looking for very similar sequences. megaBLAST scans the database once for a large number of queries, making the search very fast. As an example, the 200 *Cyprinus carpio* expressed sequence tag sequences from Savan and Sakai (12) (GenBank accession numbers AU183343–AU183542) were downloaded from the NCBI website and concatenated into one FASTA file; then the file was uploaded into the megaBLAST page using 'Browse/Load query file from disk'. The expect value was changed from its default of 10 to 1.0e−5, and the database was left as nt. By default, megaBLAST returns output in the form of a hit table [(3) and http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.610], which lists the sequence identifiers for each sequence and the start and stop positions for each alignment, as well as the score, expect value and percentage identity matched. The hit table format can also be downloaded in plain text to be analyzed locally. The normal BLAST report, with full alignments and sequence, can also be viewed using the RID. On a run in January 2004, 93 of the 200 queries had at least 1 match to the nt database. This compares well with the analysis by Savan and Sakai (12), who used nucleotide–nucleotide and translating searches to find strong similarities to 129 of the query sequences.

EXPLORE ALL THE HYPERLINKS IN THE BLAST OUTPUT

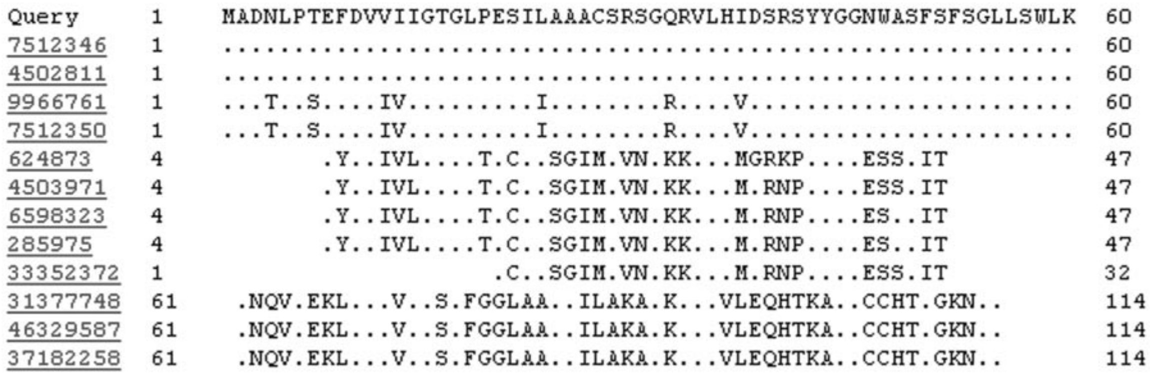
Additional information on the sequences found by BLAST has traditionally been found through links to GenBank or GenPept from the sequence identifiers of the hits. From the GenBank record, it is possible to navigate to other resources on the same sequence; however, this usually involves several steps (or mouse clicks). The LinkOut icons on the BLAST report provide a shortcut to collections of related information, which can be a powerful tool in itself. For example, when a protein–protein comparison of the *E.coli* mismatch repair protein found earlier (GenPept accession number AAM82372) is performed against nr but limited to *Mus musculus*, some very strong hits have uninformative definition lines (Figure 3). Selecting LinkOut is immediately helpful: GenPept accession.version AAH40784.1 is described as a mutS homolog 3, Msh3, and DDBJ GenPept accession.version BAB27085.1 is described as a mutS homolog 6, with the gene symbol Msh6.



```
Sequences producing significant alignments:                    (bits) Value

gi|1083296|pir||JC4019  DNA mismatch repair protein rep-3 - ...   226   4e-59
gi|387849|gb|AAB60711.1|  MutS homologue; major mRNA product...   225   6e-59  L
gi|400971|sp|P13705|MSH3_MOUSE  DNA MISMATCH REPAIR PROTEIN ...   224   2e-58  L
gi|26252149|gb|AAH40784.1|  Unknown (protein for MGC:49148) ...   223   2e-58  L
gi|30047836|gb|AAH50897.1|  MutS homolog 2 [Mus musculus]         209   4e-54  L
gi|6678938|ref|NP_032654.1|  mutS homolog 2 [Mus musculus] >...   209   4e-54  L
gi|726086|gb|AAA75027.1|  MutS homolog 2                          208   8e-54  L
gi|12846234|dbj|BAB27085.1|  unnamed protein product [Mus mu...   192   5e-49  L
gi|2506881|sp|P54276|MSH6_MOUSE  DNA mismatch repair protein...   191   9e-49  L
gi|26353224|dbj|BAC40242.1|  unnamed protein product [Mus mu...   191   2e-48  L
gi|6754744|ref|NP_034960.1|  mutS homolog 6 [Mus musculus] >...   191   2e-48  L
gi|13994197|ref|NP_114076.1|  mutS homolog 4 [Mus musculus] ...   182   7e-46  L
gi|16416651|gb|AAL18350.1|  MutS homolog 4 [Mus musculus]         182   8e-46  L
gi|33519230|gb|AAQ20788.1|  MutS homolog 4 variant alpha [Mu...   182   9e-46  L
gi|33519232|gb|AAQ20789.1|  MutS homolog 4 variant beta [Mus...   181   1e-45  L
gi|33519234|gb|AAQ20790.1|  MutS homolog 4 variant gamma [Mu...   180   2e-45
gi|33519242|gb|AAQ20795.1|  MutS homolog 4 variant theta 2 [...   172   7e-43
gi|7305281|ref|NP_038628.1|  mutS homolog 5 [Mus musculus] >...   161   1e-39  L
```

Figure 3. The one-line descriptions in the BLAST report. The blue 'L' buttons on the right link to the LocusLink resource for each entry.



```
Query     1    MADNLPTEFDVVIIGTGLPESILAAACSRSGQRVLHIDSRSYYGGNWASFSFSGLLSWLK  60
7512346   1    ............................................................  60
4502811   1    ............................................................  60
9966761   1    ...T..S....IV.........I........R....V.......................  60
7512350   1    ...T..S....IV.........I........R....V.......................  60
624873    4       .Y..IVL....T.C..SGIM.VN.KK...MGRKP....ESS.IT               47
4503971   4       .Y..IVL....T.C..SGIM.VN.KK...M.RNP....ESS.IT               47
6598323   4       .Y..IVL....T.C..SGIM.VN.KK...M.RNP....ES..IT               47
285975    4       .Y..IVL....T.C..SGIM.VN.KK...M.RNP....ES..IT               47
33352372  1                   .C..SGIM.VN.KK...M.RNP....ESS.IT               32
31377748  61    .NQV.EKL...V..S.FGGLAA..ILAKA.K...VLEQHTKA..CCHT.GKN..       114
46329587  61    .NQV.EKL...V..S.FGGLAA..ILAKA.K...VLEQHTKA..CCHT.GKN..       114
37182258  61    .NQV.EKL...V..S.FGGLAA..ILAKA.K...VLEQHTKA..CCHT.GKN..       114
```

Figure 4. Query-anchored view of a query (Rab Escort Protein; Swiss-Prot accession P26374) against the human subset of nr. Only the first 60 residues of the P26374 alignment are shown. The top line of sequence represents the query; the other lines are the retrieved database sequences. The identifiers in the leftmost column correspond to the aligned sequence in that row; the numbers are NCBI GI numbers corresponding to the database sequences found. A 'dot' indicates an exact match between the query and database sequence; a letter indicates a substitution. The numbers represent the residue positions. For example, the fifth database sequence (NCBI GI number 624873) is aligned from residues 4–47. The first residue in this alignment is conserved (E), but the second is not (Y rather than F). The signatures of the motif—the bulky hydrophobic residues isoleucine (I), valine (V) and leucine (L)—are conserved, even if I is often switched for V, V for I and L for I. Only NCBI GI number 33352372 does not appear to contain this motif.

Links to structural information (S) and UniGene (U) may also be found on a BLAST report.

ALTERNATIVE VIEWS OF BLAST RESULTS

Rather than looking at BLAST results as a series of pair-wise alignments, it is sometimes useful to view the query lined up against a number of retrieved database sequences. This can be particularly useful for finding or observing conserved motifs. The 'query-anchored' alignments provide this view (Figure 4). A dinucleotide-binding motif positioned at bases 11–21 of Rab Escort Protein is characterized by bulky hydrophobic residues followed by a glycine-rich loop (13). The pattern of the conserved motif becomes much clearer when the 'query-anchored' alignment view is selected.

IMPLEMENTATION DETAILS

Searches sent to the BLAST server are handled by a sophisticated system that makes use of a farm of mostly two-CPU machines running LINUX; there are currently about 200 CPUs available, double the number used 2 years ago, For a given query the system splits the database into a number of 'chunks' (typically 10–20) and spreads the calculations across multiple back-end machines. This system also tracks which database chunk has most recently been searched on a given back-end (and is probably still in memory) so it can send another search against the same chunk. The system stores queries, results and various statistics in a pair of machines running Microsoft SQL Server 2000, which can also generate reports on the current state of the system. For example, it is possible to track the number of failed requests organized by any number of criteria such as the database searched, the program used and the back-end machine, allowing, quick diagnosis respectively of BLAST database corruption, issues with a certain part of the algorithm and problems with an individual back-end machine.

FUTURE DIRECTIONS

The number of BLAST queries sent to the server continues to increase, growing from about 100 000 per weekday at the beginning of 2002 to about 140 000 per weekday in early 2004. As described above, the BLAST databases also continue to grow. In order to keep pace with this growth the computing power of the BLAST website will probably double over the course of the next year or two. A new BLAST report formatter is currently being written and to now available on the website. Currently this formatter can present regions masked by filtering as lowercase letters or in different colors. Another enhancement, still at the discussion stage, is on-the-fly title generation for alignments involving very long database sequences that might code for many different genes and typically have uninformative (generic) definition lines. Automatically generated information about the genes or coding regions in the area covered by an such an alignment could be presented to the user. Improvements in web navigability will attempt to steer users to the appropriate settings or link for a given need. This may include links for specialized purposes, e.g. search only mRNAs or a specific taxonomic node.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Richa Agarwala, Stephen Altschul, Kevin Bealer, Christiam Camacho, George Coulouris, Ilya Dondoshansky, Mike Gertz, Yuri Merezhuk, Alexander Morgulis, Jason Papadopoulos, Yan Raytselis, Alejandro Schaffer and Jian Ye for their work, which has made this website possible. The authors would like to thank David Lipman and Jim Ostell for their suggestions and encouragement about the website. The authors would also like to thank Jo McEntyre and Belinda Beck for their editorial suggestions and critical reading of the article.

REFERENCES

1. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
2. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
3. Madden,T.L. (2002) The BLAST sequence analysis tool. In McEntyre,J. (ed.), *The NCBI Handbook* [Internet]. National Library of Medicine (US), National Center for Biotechnology Information, Bethesda, MD.
4. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
5. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
6. Buehler,J. (2001) Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics*, **17**, 419–428.
7. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
8. Zheng,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
9. Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.
10. Chandrasekharappa,S.C., Guru,S.C., Manickam,P., Olufemi,S.E., Collins,F.S., Emmert-Buck,M.R., Debelenko,L.V., Zhuang,Z., Lubensky,I.A., Liotta,L.A. *et al.* (1997) Positional cloning of the gene for multiple endocrine neoplasia-type 1. *Science*, **276**, 404–407.
11. Pruitt,K.D., Tatusova,T. and Ostell,J. (2002) The Reference Sequence (RefSeq) project. In McEntyre,J. (ed.), *The NCBI Handbook* [Internet]. National Library of Medicine (US), National Center for Biotechnology Information, Bethesda, MD.
12. Savan,R. and Sakai,M. (2002) Analysis of expressed sequence tags (EST) obtained from common carp, *Cyprinus carpio* L., head kidney cells after stimulation by two mitogens, lipopolysaccharide and concanavalin-A. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.*, **131**, 71–82.
13. Koonin,E.V. (1996) Human choroideremia protein contains a FAD-binding domain. *Nature Genet.*, **12**, 237–239.