W6–W9 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl164

BLAST: improvements for better sequence analysis

Jian Ye, 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 10, 2006; Revised February 22, 2006; Accepted March 20, 2006

ABSTRACT

Basic local alignment search tool (BLAST) is a sequence similarity search program. The National Center for Biotechnology Information (NCBI) maintains a BLAST server with a home page at http://www.ncbi.nlm.nih.gov/BLAST/. We report here on recent enhancements to the results produced by the BLAST server at the NCBI. These include features to highlight mismatches between similar sequences, show where the query was masked for low-complexity sequence, and integrate information about the database sequences from the NCBI Entrez system into the BLAST display. Changes to how the database sequences are fetched have also improved the speed of the report generator.

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 to compare a user's query to a database of sequences (1,2). Several variants of BLAST 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 about an alignment; this is the 'expect' value, or false-positive rate.

The National Center for Biotechnology Information (NCBI) maintains a BLAST server with a homepage at http://www.ncbi.nlm.nih.gov/BLAST/. On the homepage the different BLAST searches are listed by type: nucleotide, protein, translated and genomes. The 'Program Selection Guide' (http://www.ncbi.nlm.nih.gov/blast/producttable.shtml) provides an introduction to the various programs and database options (3). When a query is submitted to the NCBI server, either as a sequence in FASTA format or as a sequence identifier, e.g. GenBank accession number, 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 search and allows the results to be viewed in several formats, which include the familiar BLAST report, a simplified 'hit table', XML and ASN.1 [(4) 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. Searches sent to the server are handled by a sophisticated queuing system that may spread the search over 10 to 20 machines, making the search much faster than if it were run on one machine. Queries and results are stored in an SQL database. More details are available at ftp://ftp.ncbi.nlm.nih.gov/blast/documents/blast-sc2004.pdf

We report here on new display features that we have implemented. These include highlighting mismatches between similar sequences, showing where the query was masked for low-complexity sequence and integrating information from the NCBI Entrez system (5) into the BLAST display. Additionally the new report generator has been optimized for databases with large sequences.

Custom definition lines

During the past five years many genomes have become searchable and the sequences in those databases are typically long contigs or chromosomes. Additionally many long nucleotide sequences have been added to the BLAST databases as a result of high-throughput genomic projects. Traditionally sequences in the BLAST database have been associated with only one descriptive phrase that is normally the same as the 'definition' in the GenBank flat file. This means that only very generic information is provided for matches to long database sequences, even though such a sequence might have annotations for many genes, coding regions (CDS) and other features. The top line of Figure 1 shows a database sequence definition and merely states that the sequence is part of human chromosome 6 and is about 48 million bases long. This reveals little about the region of the database sequence containing the match. To address this issue, we now provide feature information for

*To whom correspondence should be addressed. Tel: +1 301 435 5991; Fax: +1 301 480 0814; Email: madden@ncbi.nlm.nih.gov

```
>gi|51465675|ref|NT_007592.14|Hs6_7749 Homo sapiens chromosome 6 genomic contig
Length=48945890

 Features in this part of subject sequence:
   major histocompatibility complex, class I, A precursor


 Score =  257 bits (139),  Expect = 3e-66
 Identities = 142/143 (99%), Gaps = 1/143 (0%)
 Strand=Plus/Plus

Query  99        GGCC-GGTTCTCACACCATCCAGATAATGTATGGCTGCGACGTGGGGTCGGACGGGCGCT  157
                 |||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  20769290  GGCCAGGTTCTCACACCATCCAGATAATGTATGGCTGCGACGTGGGGTCGGACGGGCGCT  20769349

Query  158       TCCTCCGCGGGTACCGGCAGGACGCCTACGACGGCAAGGATTACATCGCCCTGAACGAGG  217
                 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  20769350  TCCTCCGCGGGTACCGGCAGGACGCCTACGACGGCAAGGATTACATCGCCCTGAACGAGG  20769409

Query  218       ACCTGCGCTCTTGGACCGCGGCG  240
                 |||||||||||||||||||||||
Sbjct  20769410  ACCTGCGCTCTTGGACCGCGGCG  20769432


 Features flanking this part of subject sequence:
   58075 bp at 5' side: major histocompatibility complex, class I, G precursor
   53951 bp at 3' side: major histocompatibility complex, class I, A precursor


 Score =  198 bits (107),  Expect = 2e-48
 Identities = 135/148 (91%), Gaps = 3/148 (2%)
 Strand=Plus/Plus

Query  95        GCGAGGCC-GGTTCTCACACCATCCAGATAATGTATGGCTGCGACGTGGGGTCGGACGGG  153
                 ||| |||| ||||||||||||||| ||| | ||||||||||||||||||||| | ||||||
Sbjct  20714485  GCGGGGCCAGGTTCTCACACCATGCAGGTGATGTATGGCTGCGACGTGGGGCCCGACGGG  20714544

Query  154       CGCTTCCTCCGCGGGTACCG-GCAGGACGCCTACGACGGCAAGGATTACATCGCCCTGAA  212
                 |||||||||||||||||  |  ||| ||||||||||||||||||||||||||||||||||
Sbjct  20714545  CGCTTCCTCCGCGGGTA-TGAACAGCACGCCTACGACGGCAAGGATTACATCGCCCTGAA  20714603

Query  213       CGAGGACCTGCGCTCTTGGACCGCGGCG  240
                 |||||||||||||| ||||||||||||
Sbjct  20714604  CGAGGACCTGCGCTCCTGGACCGCGGCG  20714631
```

Figure 1. Excerpt from a BLAST result showing custom-definition lines. The query was bases 241 through 480 of a human MHC A gene nucleotide sequence (NM_002116) in a search against the human genome. The top line of the figure is the traditional sequence definition. Custom definition lines are provided for both of the alignments shown and are relevant to the region matched (first alignment) or nearby regions (second alignment).

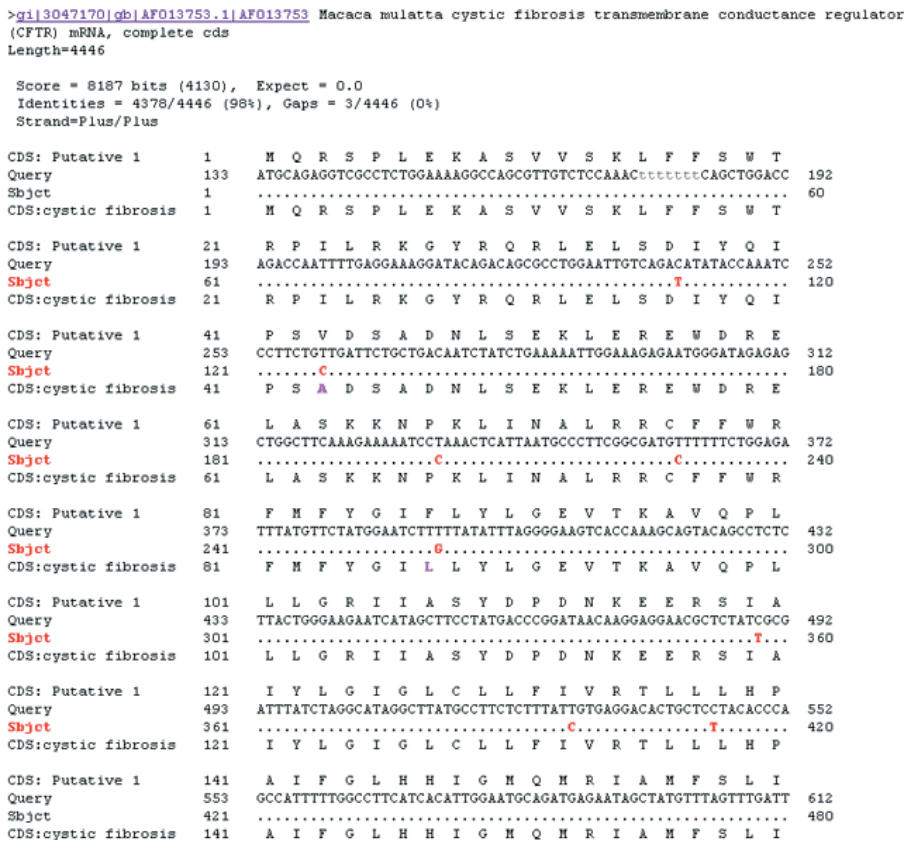
BLAST alignments involving long database sequences (currently defined as larger than 200 kb).

Two types of sequence features (CDS and rRNA) are currently supported but this could be expanded to other features. An example is shown in Figure 1 where a custom definition line is displayed for each of the two alignments. According to the custom definition lines the query matches a region inside the human major histocompatability complex (MHC) A gene, as well as a region that is about 54 kb upstream of the MHC A gene and about 58 kb downstream of the MHC G gene, allowing one to quickly draw the conclusion that the query sequence is highly related to the human MHC. This feature is always enabled for reports at the NCBI BLAST web site.

New format options for easier sequence analysis

Frequently alignments are between very similar sequences and it's difficult to identify a few mismatches in the pairwise alignment. To address this issue we recently introduced a new format called 'Pairwise with identities', shown in Figure 2 on an alignment with 98% identity between the query and database sequence. A dot indicates identity between the database sequence and query at that position; mismatches are shown as the database sequence letter in place of the dot and colored red. In addition the word 'Sbjct' (on the left of the figure) is also colored red if there is a mismatch on the line. Enable this option with the 'Alignment View' pull-down menu shown in Figure 3.

The majority of BLAST searches at the NCBI web site are nucleotide queries against nucleotide databases (e.g. BLASTN). Many of these queries are mRNAs or match to sequences with annotated coding regions. The standard BLAST report does not show the amino acid sequence translated from the query or annotated on the database sequence, even though that may be of great interest to the user; furthermore figuring out the positions of the encoded amino acids on the corresponding nucleotide sequence can be challenging, especially if the coding region is long or involves multiple exons. We have introduced a new 'CDS Feature' to display such coding regions. With this option any pre-annotated CDS protein products on the query (if the query is an accession) or the database sequence are fetched from Entrez and shown with the residues aligned to the second base of a codon (Figure 2). For a user-submitted query in FASTA format a putative protein product is calculated using the coding frame of the database sequence as a guide. Mismatched amino acids for the database sequence can also be shown in color. Combined with the 'Pairwise with identities' option discussed above this



```
>gi|3047170|gb|AF013753.1|AF013753 Macaca mulatta cystic fibrosis transmembrane conductance regulator
(CFTR) mRNA, complete cds
Length=4446

 Score = 8187 bits (4130),  Expect = 0.0
 Identities = 4378/4446 (98%), Gaps = 3/4446 (0%)
 Strand=Plus/Plus

CDS: Putative 1       1     M  Q  R  S  P  L  E  K  A  S  V  V  S  K  L  F  F  S  W  T
Query                 133   ATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGTTGTCTCCAAACtttttttCAGCTGGACC  192
Sbjct                 1     ............................................................  60
CDS:cystic fibrosis   1     M  Q  R  S  P  L  E  K  A  S  V  V  S  K  L  F  F  S  W  T

CDS: Putative 1       21    R  P  I  L  R  K  G  Y  R  Q  R  L  E  L  S  D  I  Y  Q  I
Query                 193   AGACCAATTTTGAGGAAAGGATACAGACAGCGCCTGGAATTGTCAGACATATACCAAATC  252
Sbjct                 61    ...........................................T............  120
CDS:cystic fibrosis   21    R  P  I  L  R  K  G  Y  R  Q  R  L  E  L  S  D  I  Y  Q  I

CDS: Putative 1       41    P  S  V  D  S  A  D  N  L  S  E  K  L  E  R  E  W  D  R  E
Query                 253   CCTTCTGTTGATTCTGCTGACAATCTATCTGAAAAATTGGAAAGAGAATGGGATAGAGAG  312
Sbjct                 121   .......C....................................................  180
CDS:cystic fibrosis   41    P  S  A  D  S  A  D  N  L  S  E  K  L  E  R  E  W  D  R  E

CDS: Putative 1       61    L  A  S  K  K  N  P  K  L  I  N  A  L  R  R  C  F  F  W  R
Query                 313   CTGGCTTCAAAGAAAAATCCTAAACTCATTAATGCCCTTCGGCGATGTTTTTTCTGGAGA  372
Sbjct                 181   ..................C...........................C............  240
CDS:cystic fibrosis   61    L  A  S  K  K  N  P  K  L  I  N  A  L  R  R  C  F  F  W  R

CDS: Putative 1       81    F  M  F  Y  G  I  F  L  Y  L  G  E  V  T  K  A  V  Q  P  L
Query                 373   TTTATGTTCTATGGAATCTTTTTATATTTAGGGGAAGTCACCAAAGCAGTACAGCCTCTC  432
Sbjct                 241   ..................G.....................................  300
CDS:cystic fibrosis   81    F  M  F  Y  G  I  L  L  Y  L  G  E  V  T  K  A  V  Q  P  L

CDS: Putative 1       101   L  L  G  R  I  I  A  S  Y  D  P  D  N  K  E  E  R  S  I  A
Query                 433   TTACTGGGAAGAATCATAGCTTCCTATGACCCGGATAACAAGGAGGAACGCTCTATCGCG  492
Sbjct                 301   ...........................................................T...  360
CDS:cystic fibrosis   101   L  L  G  R  I  I  A  S  Y  D  P  D  N  K  E  E  R  S  I  A

CDS: Putative 1       121   I  Y  L  G  I  G  L  C  L  L  F  I  V  R  T  L  L  L  H  P
Query                 493   ATTTATCTAGGCATAGGCTTATGCCTTCTCTTTATTGTGAGGACACTGCTCCTACACCCA  552
Sbjct                 361   ..................................C..............T........  420
CDS:cystic fibrosis   121   I  Y  L  G  I  G  L  C  L  L  F  I  V  R  T  L  L  L  H  P

CDS: Putative 1       141   A  I  F  G  L  H  H  I  G  M  Q  M  R  I  A  M  F  S  L  I
Query                 553   GCCATTTTTGGCCTTCATCACATTGGAATGCAGATGAGAATAGCTATGTTTAGTTTGATT  612
Sbjct                 421   ............................................................  480
CDS:cystic fibrosis   141   A  I  F  G  L  H  H  I  G  M  Q  M  R  I  A  M  F  S  L  I
```

Figure 2. Demonstration of new format options. FASTA sequence for the human cystic fibrosis trans-membrane conductance regulator sequence (NM_000492) was used as query for a BLASTN search against the nr database using default parameters. Three new display options are shown in this figure. The first is the 'Pairwise with identities' option. Nucleotide matches in the database sequence are shown as dots ('.'), nucleotide mismatches in the database sequence (as well as the database sequence identification) are colored red. The second new option is the presentation of the CDS features, which is shown for both the query and database sequences above and below the BLAST alignment, respectively. The CDS feature annotated on the database sequence was retrieved from Entrez; the putative CDS feature on the query was produced automatically using the CDS of the database sequence as a guide. Mismatches for the amino acid sequence derived from the database sequence are colored pink. Finally the new masking option is shown (see text). Bases 175–181 of the query were masked for low-complexity during the search and are shown in lower-case gray letters.

format makes certain tasks easier, such as analysis for silent and replacement mutations. Owing to the overhead of fetching the CDS feature from Entrez this option is currently not the default. Enable this option by checking the 'CDS feature' box on the BLAST format page as shown in Figure 3.

Low complexity sequences are compositionally biased regions of amino acid or nucleotide sequence, which often result in artificially high scores in sequence similarity searches. Low-complexity filters, such as SEG (6) or DUST, mask these regions and prevent them from overly biasing the results. Traditionally BLAST has replaced the masked regions by Xs or Ns in the BLAST report. The BLAST formatter now can represent these regions by lower-case letters, making them distinct from the (upper-case) non-filtered regions (Figure 2). In addition the user may select from three colors (black, gray, red) to vary the emphasis on these regions (Figure 3). This new display option is now the default, showing the masked regions in gray lower-case.

General improvements to the BLAST web site

The BLAST graphical overview is a schematic representation of alignments matching the query sequence. It is useful for quickly localizing regions of interest in the query based on it's similarity to other sequences in the database. To reduce the complexity in generating this graphic overview we have now implemented it as HTML tables that use a few small static images (gifs). This design is more robust and also lends itself to future development of a graphical viewer for stand-alone and command-line client BLAST.

The new report generator has improved functionality to fetch part of a database sequence. This can be essential if the database sequence is long, such as a chromosome, and the alignment to be presented only involves a small fraction of the sequence. Previously the entire database sequence was fetched and much of that sequence was not used. This improved functionality has led to a dramatic decrease in formatting time for searches against genomes.

BLAST provides several different modes for viewing BLAST results. The Query-anchored view gives a stacked view of database sequences aligned to the query with indication of insertions and mismatches (3). This provides an easy method to scan alignments and locate things like SNP's and amino acid substitutions among a group of related sequences. Previously the query-anchored views were not fully supported for BLASTX and TBLASTX searches that involved translated

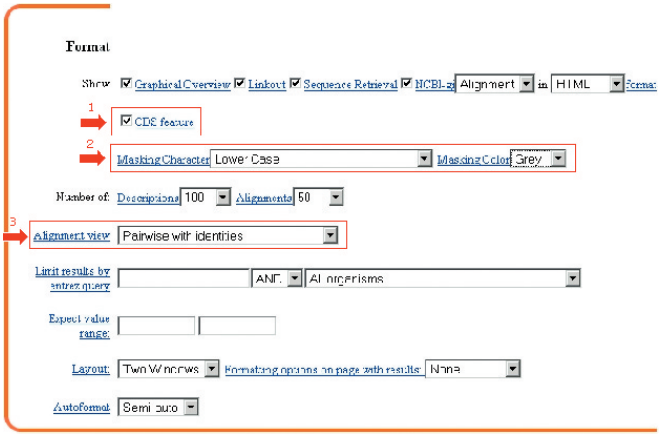


Figure 3. Enabling new features on the BLAST format page. The red arrows point to new report features that may be enabled or modified from this page. The check-box highlighted by arrow 1 enables the CDS feature on a BLASTN or megaBLAST search. The two menus highlighted by arrow 2 change the default behavior for display of masked regions. The menu highlighted by arrow 3 changes how the alignments are displayed in the BLAST report.

sequences. The formatter now supports this format for all these programs. Use the 'Alignment View' pull-down to enable this option (Figure 3).

From the BLAST results it is now possible to select some or all of the database sequences and perform an Entrez query to fetch them. Checking the boxes in the alignment section selects the sequences to download and clicking the 'Get Selected sequences' button takes the user to Entrez, where the sequences can be displayed in various formats, (such as GenBank or FASTA) and saved to a file. The saved file can then be used as input to another program.

Future directions

We are currently redesigning the BLAST web pages to make them more effective tools. Some of the changes will be better organized HTML that makes options apparent to the user, such as making it easier to limit a search or results to a particular organism or subset of the data available. Results will also be made more user-friendly by better organizing the output. Nearing completion is a utility to calculate distances between sequences in the BLAST results and present those as a tree. Finally we are also working on making it possible to save search or formatting strategies for future use.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Richa Agarwala, Stephen Altschul, Kevin Bealer, Christiam Camacho, Peter Cooper, George Coulouris, Susan Dombrowski, Mike Gertz, David Lipman, Wayne Matten, Yuri Merezhuk, Alexander Morgulis, Jim Ostell, Jason Papadopoulos, Yan Raytselis, Eric Sayers, Alejandro Schaffer, Tao Tao, David Wheeler and Irena Zaretskaya, as well as members of the C++ toolkit group at the NCBI, for their work that has made this Web site possible. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health.

Conflict of interest statement. None declared.

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. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
4. 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.
5. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Meth. Enzymol.*, **266**, 141–162.
6. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.*, **266**, 554–571.