Bioinformatics
Explained

Bioinformatics explained: BLAST

March 8, 2007

CLC bio
Gustav Wieds Vej 10 · 8000 Aarhus C · Denmark
Telephone: +45 70 22 55 09 · Fax: +45 70 22 55 19
www.clcbio.com · info@clcbio.com

Bioinformatics explained: BLAST

BLAST (Basic Local Alignment Search Tool) has become the *defacto* standard in search and alignment tools [Altschul et al., 1990]. The BLAST algorithm is still actively being developed and is one of the most cited papers ever written in this field of biology. Many researchers use BLAST as an initial screening of their sequence data from the laboratory and to get an idea of what they are working on. BLAST is far from being basic as the name indicates; it is a highly advanced algorithm which has become very popular due to availability, speed, and accuracy. In short, a BLAST search identifies homologous sequences by searching one or more databases usually hosted by NCBI (`http://www.ncbi.nlm.nih.gov/`), on the query sequence of interest [McGinnis and Madden, 2004].

BLAST is an open source program and anyone can download and change the program code. This has also given rise to a number of BLAST derivatives; WU-BLAST is probably the most commonly used [Altschul and Gish, 1996].

BLAST is highly scalable and comes in a number of different computer platform configurations which makes usage on both small desktop computers and large computer clusters possible.

Examples of BLAST usage

BLAST can be used for a lot of different purposes. A few of them are mentioned below.

- **Looking for species.** If you are sequencing DNA from unknown species, BLAST may help identify the correct species or homologous species.
- **Looking for domains.** If you BLAST a protein sequence (or a translated nucleotide sequence) BLAST will look for known domains in the query sequence.
- **Looking at phylogeny.** You can use the BLAST web pages to generate a phylogenetic tree of the BLAST result.
- **Mapping DNA to a known chromosome.** If you are sequencing a gene from a known species but have no idea of the chromosome location, BLAST can help you. BLAST will show you the position of the query sequence in relation to the hit sequences.
- **Annotations.** BLAST can also be used to map annotations from one organism to another or look for common genes in two related species.

Searching for homology

Most research projects involving sequencing of either DNA or protein have a requirement for obtaining biological information of the newly sequenced and maybe unknown sequence. If the researchers have no prior information of the sequence and biological content, valuable information can often be obtained using BLAST. The BLAST algorithm will search for homologous sequences in predefined and annotated databases of the users choice.

In an easy and fast way the researcher can gain knowledge of gene or protein function and find evolutionary relations between the newly sequenced DNA and well established data.

After the BLAST search the user will receive a report specifying found homologous sequences and their local alignments to the query sequence.

How does BLAST work?

BLAST identifies homologous sequences using a heuristic method which initially finds short matches between two sequences; thus, the method does not take the entire sequence space into account. After initial match, BLAST attempts to start local alignments from these initial matches. This also means that BLAST does not guarantee the optimal alignment, thus some sequence hits may be missed. In order to find optimal alignments, the Smith-Waterman algorithm should be used (see below). In the following, the BLAST algorithm is described in more detail.

Seeding

When finding a match between a query sequence and a hit sequence, the starting point is the *words* that the two sequences have in common. A word is simply defined as a number of letters. For blastp the default word size is 3 *W=3*. If a query sequence has a QWRTG, the searched words are QWR, WRT, RTG. See figure 1 for an illustration of words in a protein sequence.

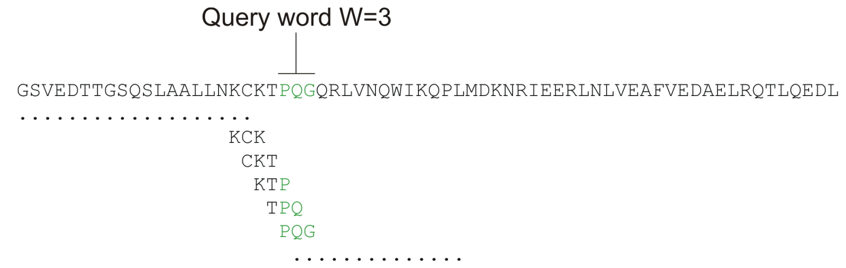

Query word W=3
GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
....
KCK
CKT
KTP
TPQ
PQG
....

Figure 1: *Generation of exact BLAST words with a word size of W=3.*

During the initial BLAST seeding, the algorithm finds all common words between the query sequence and the hit sequence(s). Only regions with a word hit will be used to build on an alignment.

BLAST will start out by making words for the entire query sequence (see figure 1). For each word in the query sequence, a compilation of neighborhood words, which exceed the threshold of *T*, is also generated.

A neighborhood word is a word obtaining a score of at least *T* when comparing, using a selected scoring matrix (see figure 2). The default scoring matrix for blastp is BLOSUM62 (for explanation of scoring matrices, see www.clcbio.com/be). The compilation of exact words and neighborhood words is then used to match against the database sequences.

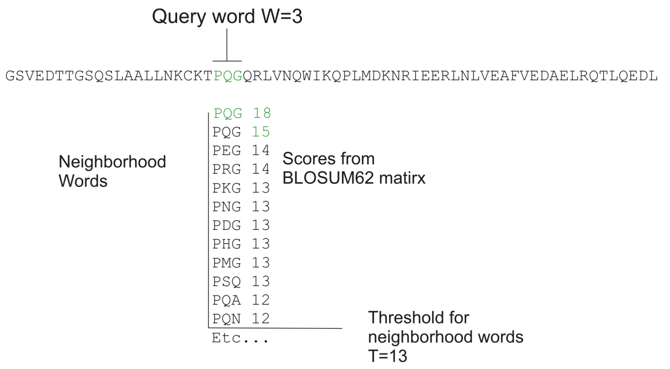

Query word W=3
GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
Neighborhood Words
PQG 18
PQG 15
PEG 14
PRG 14
PKG 13
PNG 13
PDG 13
PHG 13
PMG 13
PSQ 13
PQA 12
PQN 12
Etc...
Scores from BLOSUM62 matirx
Threshold for neighborhood words T=13

Figure 2: *Neighborhood BLAST words based on the BLOSUM62 matrix. Only words where the threshold* T *exceeds 13 are included in the initial seeding.*

After initial finding of words (seeding), the BLAST algorithm will extend the (only 3 residues long) alignment in both directions (see figure 3). Each time the alignment is extended, an alignment score is increases/decreased. When the alignment score drops below a predefined threshold, the extension of the alignment stops. This ensures that the alignment is not extended to regions where only very poor alignment between the query and hit sequence is possible. If the obtained alignment receives a score above a certain threshold, it will be included in the final BLAST result.

```
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA     365
           +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA     330
```

Figure 3: *Blast aligning in both directions. The initial word match is marked green.*

By tweaking the word size *W* and the neighborhood word threshold *T*, it is possible to limit the search space. E.g. by increasing *T*, the number of neighboring words will drop and thus limit the search space as shown in figure 4.

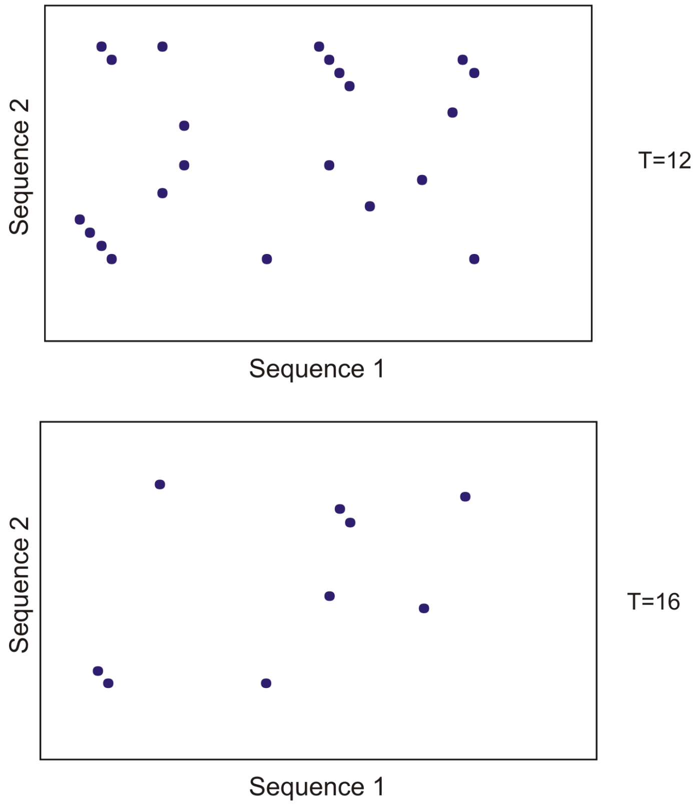


Figure 4: *Each dot represents a word match. Increasing the threshold of* T *limits the search space significantly.*

This will increase the speed of BLAST significantly but may result in loss of sensitivity. Increasing the word size *W* will also increase the speed but again with a loss of sensitivity.

Which BLAST program should I use?

Depending on the nature of the sequence it is possible to use different BLAST programs for the database search. There are five versions of the BLAST program, blastn, blastp, blastx, tblastn, tblastx:

Option	Query Type	DB Type	Comparison	Note
blastn	Nucleotide	Nucleotide	Nucleotide-Nucleotide	
blastp	Protein	Protein	Protein-Protein	
tblastn	Protein	Nucleotide	Protein-Protein	The database is translated into protein
blastx	Nucleotide	Protein	Protein-Protein	The queries are translated into protein
tblastx	Nucleotide	Nucleotide	Protein-Protein	The queries and database are translated into protein

The most commonly used method is to BLAST a nucleotide sequence against a nucleotide database (blastn) or a protein sequence against a protein database (blastp). But often another BLAST program will produce more interesting hits. E.g. if a nucleotide sequence is translated before the search, it is more likely to find better and more accurate hits than just a blastn search. One of the reasons for this is that protein sequences are evolutionarily more conserved than nucleotide sequences. Another good reason for translating the query sequence before the search is that you get protein hits which are likely to be annotated. Thus you can directly see the protein function of the sequenced gene.

Which BLAST options should I change?

The NCBI BLAST web pages and the BLAST command line tool offer a number of different options which can be changed in order to obtain the best possible result. Changing these parameters can have a great impact on the search result. It is not the scope of this document to comment on all of the options available but merely the options which can be changed with a direct impact on the search result.

The E-value

The *expect value*(E-value) can be changed in order to limit the number of hits to the most significant ones. The lower the E-value, the better the hit. The E-value is dependent on the length of the query sequence and the size of the database. For example, an alignment obtaining an E-value of 0.05 means that there is a 5 in 100 chance of occurring by chance alone.

E-values are very dependent on the query sequence length and the database size. Short identical sequence may have a high E-value and may be regarded as "false positive" hits. This is often seen if one searches for short primer regions, small domain regions etc. The default threshold for the E-value on the BLAST web page is 10. Increasing this value will most likely generate more hits. Below are some rules of thumb which can be used as a guide but should be considered with common sense.

- **E-value < 10e-100** Identical sequences. You will get long alignments across the entire query and hit sequence.
- **10e-50 < E-value < 10e-100** Almost identical sequences. A long stretch of the query protein is matched to the database.
- **10e-10 < E-value < 10e-50** Closely related sequences, could be a domain match or similar.
- **1 < E-value < 10e-6** Could be a true homologue but it is a gray area.

- **E-value > 1** Proteins are most likely not related
- **E-value > 10** Hits are most likely junk unless the query sequence is very short.

Gap costs

For blastp it is possible to specify gap cost for the chosen substitution matrix. There is only a limited number of options for these parameters. The *open gap cost* is the price of introducing gaps in the alignment, and *extension gap cost* is the price of every extension past the initial opening gap. Increasing the gap costs will result in alignments with fewer gaps.

Filters

It is possible to set different filter options before running the BLAST search. Low-complexity regions have a very simple composition compared to the rest of the sequence and may result in problems during the BLAST search [Wootton and Federhen, 1993]. A low complexity region of a protein can for example look like this 'fftfflllsss', which in this case is a region as part of a signal peptide. In the output of the BLAST search, low-complexity regions will be marked in lowercase gray characters (default setting). The low complexity region cannot be thought of as a significant match; thus, disabling the low complexity filter is likely to generate more hits to sequences which are not truly related.

Word size

Change of the word size has a great impact on the seeded sequence space as described above. But one can change the word size to find sequence matches which would otherwise not be found using the default parameters. For instance the word size can be decreased when searching for primers or short nucleotides. For blastn a suitable setting would be to decrease the default word size of 11 to 7, increase the E-value significantly (1000) and turn off the complexity filtering.

For blastp a similar approach can be used. Decrease the word size to 2, increase the E-value and use a more stringent substitution matrix, e.g. a PAM30 matrix.

Fortunately, the optimal search options for finding short, nearly exact matches can already be found on the BLAST web pages `http://www.ncbi.nlm.nih.gov/BLAST/`.

Substitution matrix

For protein BLAST searches, a default substitution matrix is provided. If you are looking at distantly related proteins, you should either choose a high-numbered PAM matrix or a low-numbered BLOSUM matrix. See *Bioinformatics Explained* on scoring matrices on `http://www.clcbio.com/be/`. The default scoring matrix for blastp is BLOSUM62.

Explanation of the BLAST output

The BLAST output comes in different flavors. On the NCBI web page the default output is html, and the following description will use the html output as example. Ordinary text and xml output for easy computational parsing is also available.

The default layout of the NCBI BLAST result is a graphical representation of the hits found, a table of sequence identifiers of the hits together with scoring information, and alignments of the query sequence and the hits.

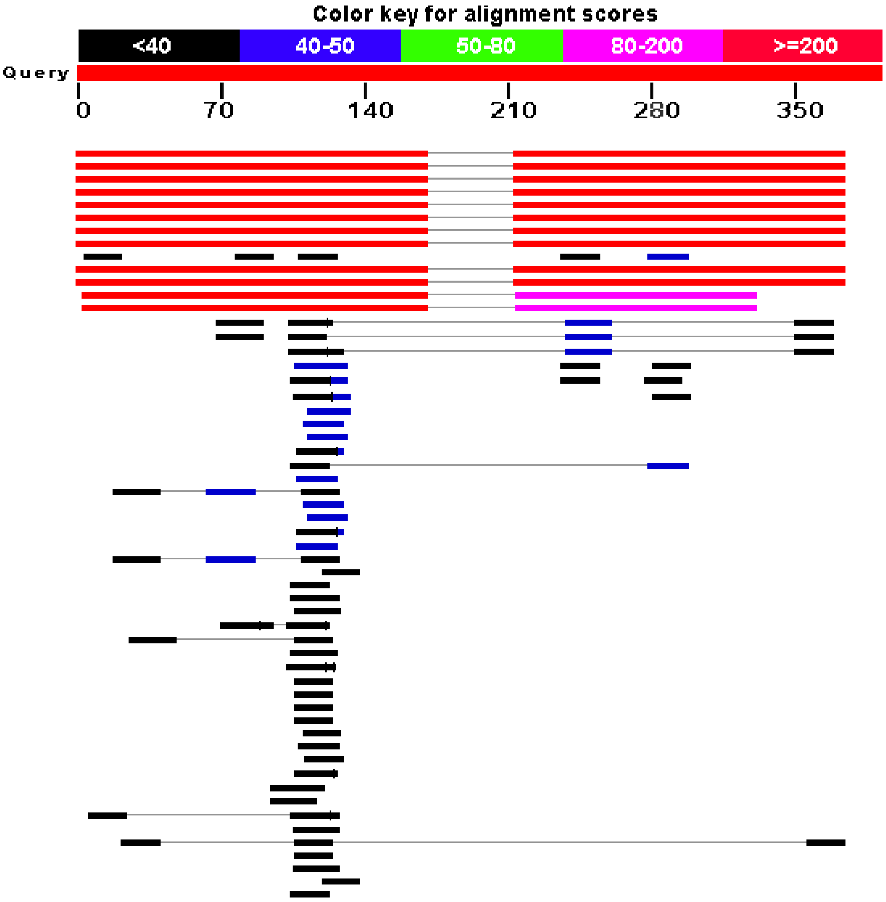
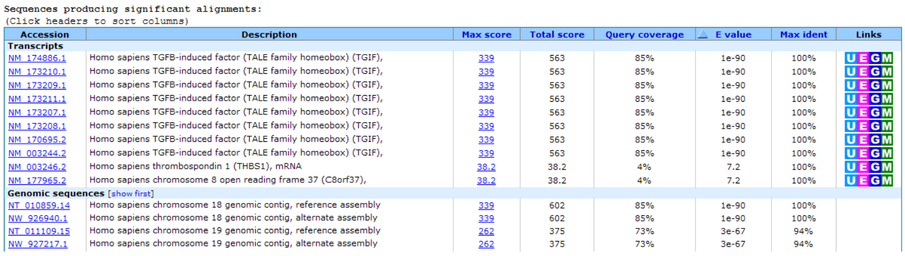


Figure 5: *BLAST graphical view. A simple graphical overview of the hits found aligned to the query sequence. The alignments are color coded ranging from black to red as indicated in the color label at the top.*

The graphical output (shown in figure 5) gives a quick overview of the query sequence and the resulting hit sequences. The hits are colored according to the obtained alignment scores.



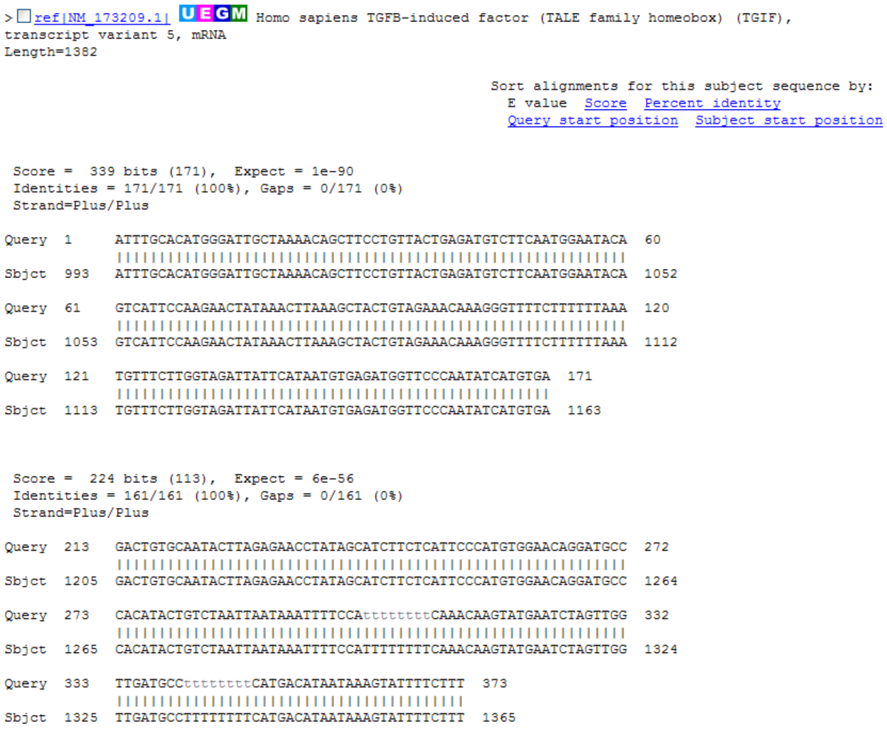
Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_174886.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	UEGM
NM_173210.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	UEGM
NM_173209.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	UEGM
NM_173211.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	UEGM
NM_173207.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	UEGM
NM_173208.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	UEGM
NM_170695.2	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	UEGM
NM_003244.2	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	UEGM
NM_003246.2	Homo sapiens thrombospondin 1 (THBS1), mRNA	38.2	38.2	4%	7.2	100%	UEGM
NM_177965.2	Homo sapiens chromosome 8 open reading frame 37 (C8orf37),	38.2	38.2	4%	7.2	100%	UEGM
Genomic sequences [show first]							
NT_010859.14	Homo sapiens chromosome 18 genomic contig, reference assembly	339	602	85%	1e-90	100%	
NW_926940.1	Homo sapiens chromosome 18 genomic contig, alternate assembly	339	602	85%	1e-90	100%	
NT_011109.15	Homo sapiens chromosome 19 genomic contig, reference assembly	262	375	73%	3e-67	94%	
NW_927217.1	Homo sapiens chromosome 19 genomic contig, alternate assembly	262	375	73%	3e-67	94%	

Figure 6: *BLAST table view. A table view with one row per hit, showing the accession number and description field from the sequence file together with BLAST output scores.*

The table view (shown in figure 6) provides more detailed information on each hit and furthermore acts as a hyperlink to the corresponding sequence in GenBank.

In the alignment view one can manually inspect the individual alignments generated by the BLAST



```
> □ ref|NM_173209.1| UEGM Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),
transcript variant 5, mRNA
Length=1382

                                                                Sort alignments for this subject sequence by:
                                                                  E value   Score   Percent identity
                                                                  Query start position   Subject start position


 Score =  339 bits (171),  Expect = 1e-90
 Identities = 171/171 (100%), Gaps = 0/171 (0%)
 Strand=Plus/Plus

Query  1     ATTTGCACATGGGATTGCTAAAACAGCTTCCTGTTACTGAGATGTCTTCAATGGAATACA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  993   ATTTGCACATGGGATTGCTAAAACAGCTTCCTGTTACTGAGATGTCTTCAATGGAATACA  1052

Query  61    GTCATTCCAAGAACTATAAACTTAAAGCTACTGTAGAAACAAAGGGTTTTCTTTTTTAAA  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1053  GTCATTCCAAGAACTATAAACTTAAAGCTACTGTAGAAACAAAGGGTTTTCTTTTTTAAA  1112

Query  121   TGTTTCTTGGTAGATTATTCATAATGTGAGATGGTTCCCAATATCATGTGA  171
             |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1113  TGTTTCTTGGTAGATTATTCATAATGTGAGATGGTTCCCAATATCATGTGA  1163


 Score =  224 bits (113),  Expect = 6e-56
 Identities = 161/161 (100%), Gaps = 0/161 (0%)
 Strand=Plus/Plus

Query  213   GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  272
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1205  GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC  1264

Query  273   CACATACTGTCTAATTAATAAATTTTCCAttttttttCAAACAAGTATGAATCTAGTTGG  332
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1265  CACATACTGTCTAATTAATAAATTTTCCATTTTTTTTCAAACAAGTATGAATCTAGTTGG  1324

Query  333   TTGATGCCttttttttCATGACATAATAAAGTATTTTCTTT  373
             |||||||||||||||||||||||||||||||||||||||||
Sbjct  1325  TTGATGCCTTTTTTTTCATGACATAATAAAGTATTTTCTTT  1365
```

Figure 7: *Alignment view of BLAST results. Individual alignments are represented together with BLAST scores and more.*

algorithm. This is particularly useful for detailed inspection of the sequence hit found(sbjct) and the corresponding alignment. In the alignment view, all scores are described for each alignment, and the start and stop positions for the query and hit sequence are listed. The strand and orientation for query sequence and hits are also found here.

In most cases, the table view of the results will be easier to interpret than tens of sequence alignments.

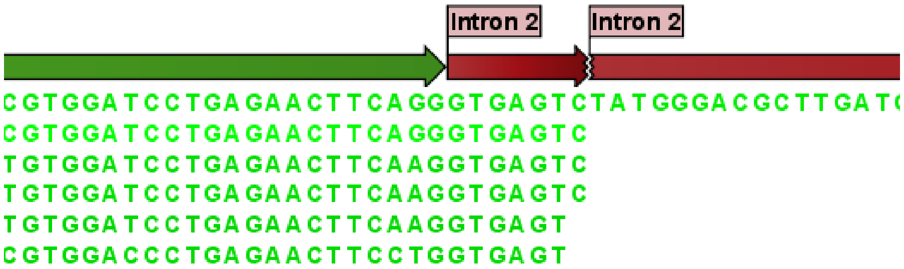
I want to BLAST against my own sequence database, is this possible?

It is possible to download the entire BLAST program package and use it on your own computer, institution computer cluster or similar. This is preferred if you want to search in proprietary sequences or sequences unavailable in the public databases stored at NCBI. The downloadable BLAST package can either be installed as a web-based tool or as a command line tool. It is available for a wide range of different operating systems.

The BLAST package can be downloaded free of charge from the following location `http://www.ncbi.nlm.nih.gov/BLAST/download.shtml`

Preformatted databases are available from a dedicated BLAST ftp site `ftp://ftp.ncbi.nlm.nih.gov/blast/db/`. Moreover, it is possible to download programs/scripts from the same site enabling automatic download of changed BLAST databases. Thus it is possible to schedule a nightly update of changed databases and have the updated BLAST database stored locally or on a shared network drive at all times. Most BLAST databases on the NCBI site are updated on a daily basis to include all recent sequence submissions to GenBank.

A few commercial software packages are available for searching your own data. The advantage of using a commercial program is obvious when BLAST is integrated with the existing tools of these programs. Furthermore, they let you perform BLAST searches and retain annotations on the query sequence (see figure 8). It is also much easier to batch download a selection of hit sequences for further inspection.



```
CGTGGATCCTGAGAACTTCAGGGTGAGTCTATGGGACGCTTGATC
CGTGGATCCTGAGAACTTCAGGGTGAGTC
TGTGGATCCTGAGAACTTCAAGGTGAGTC
TGTGGATCCTGAGAACTTCAAGGTGAGTC
TGTGGATCCTGAGAACTTCAAGGTGAGT
CGTGGACCCTGAGAACTTCCTGGTGAGT
```

Figure 8: *Snippet of alignment view of BLAST results from CLC Combined Workbench. Individual alignments are represented directly in a graphical view. The top sequence is the query sequence and is shown with a selection of annotations.*

What you cannot get out of BLAST

Don't expect BLAST to produce the best available alignment. BLAST is a heuristic method which does not guarantee the best results, and therefor you cannot rely on BLAST if you wish to find *all* the hits in the database.

Instead, use the Smith-Waterman algorithm for obtaining the best possible local alignments [Smith and Waterman, 1981].

BLAST only makes local alignments. This means that a great but short hit in another sequence may not at all be related to the query sequence even though the sequences align well in a small region. It may be a domain or similar.

It is always a good idea to be cautious of the material in the database. For instance, the sequences may be wrongly annotated; hypothetical proteins are often simple translations of a found ORF on a sequenced nucleotide sequence and may not represent a true protein.

Don't expect to see the best result using the default settings. As described above, the settings should be adjusted according to the what kind of query sequence is used, and what kind of results you want. It is a good idea to perform the same BLAST search with different settings to get an idea of how they work. There is not a final answer on how to adjust the settings for your particular sequence.

Other useful resources

The BLAST web page hosted at NCBI
`http://www.ncbi.nlm.nih.gov/BLAST`

Download pages for the BLAST programs
`http://www.ncbi.nlm.nih.gov/BLAST/download.shtml`

Download pages for pre-formatted BLAST databases ftp://ftp.ncbi.nlm.nih.gov/blast/db/

O'Reilly book on BLAST
http://www.oreilly.com/catalog/blast/

Explanation of scoring/substitution matrices and more
http://www.clcbio.com/be/

Creative Commons License

References

[Altschul and Gish, 1996] Altschul, S. F. and Gish, W. (1996). Local alignment statistics. *Methods Enzymol*, 266:460–480.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[McGinnis and Madden, 2004] 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(Web Server issue):W20–W25.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.

[Wootton and Federhen, 1993] Wootton, J. C. and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers in Chemistry*, 17:149–163.