Basic Local Alignment Search Tool (BLAST)

BLAST is a popular user-friendly tool for searching all the major sequence databases. BLAST is a heuristic method to find the highest scoring locally optimal alignments between a query sequence and a database sequence. BLAST programs were designed for fast database searching, with minimal sacrifice of sensitivity to distant related sequences. BLAST is used to find sequence homologs to predict the identity, function, 3D structure of the query sequence. BLAST shows better results for protein sequences than nucleotide sequences.

Need for Local Alignments
Local alignments is particularly important in case of new DNA sequences from higher eukaryotes. The most significant alignments are of translations of the new DNA sequence against protein databases, to take advantage of the larger protein alphabet and avoid the 25% random alignment characteristics of nucleic acid alignments. However, genes in the new DNA sequences are then distributed among Exons and a given long DNA sequence may possibly contain more than one gene. Further, some Introns contain genes themselves, e.g. nucleases; thus the new DNA can contain “genes within genes”

Each of these features demands for Local alignments, which would yield alignments for each gene and, perhaps more importantly, for each Exon.

The BLAST programs solve these concerns

Local similarity......no Gaps permitted.
Mathematically rigorous: distribution of scores is an Extreme Value Distribution.
Algorithm uses Lookup Tables and extensions of short “word” matches.

Some of the salient features of BLAST are
Local alignments: Blast tries to find patches of regional similarity, rather than trying for global fit between the query and the database sequence.

Ungapped alignments: BLAST programs work on statistics of ungapped sequence alignments, but theoretically this reduces sensitivity of search. However, output shows multiple local alignments between query and a database sequence that can be used to anticipate the gaps between them. Only identities and conservative replacements are taken into account.

BLAST Sequence search Tools

<table>
<thead>
<tr>
<th>Program</th>
<th>Query Sequence</th>
<th>Database</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTp</td>
<td>Protein</td>
<td>Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>BLASTn</td>
<td>DNA</td>
<td>DNA</td>
<td>DNA</td>
</tr>
<tr>
<td>BLASTx</td>
<td>DNA</td>
<td>Protein</td>
<td>Protein</td>
</tr>
<tr>
<td>TBLASTn</td>
<td>Protein</td>
<td>DNA</td>
<td>Protein</td>
</tr>
<tr>
<td>TBLASTx</td>
<td>DNA</td>
<td>DNA</td>
<td>Protein</td>
</tr>
</tbody>
</table>

Table 1. The five BLAST programs described here perform the following tasks
BLASTp: This program compares an amino acid query sequence against a protein sequence database.
BLASTn: This program compares a nucleotide query sequence against a nucleotide sequence database.
BLASTx: This program searches the six-frame translation products of a nucleotide sequence against a protein database.
TBLASTn: This program searches a protein sequence against translated nucleotide sequences in the databases.
TBLASTx: Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. This program is similar to BLASTx and TBLASTn program.

Major parameters
DATALIB: Database, or group of databases, chosen for the search.
MATRIX: Distance Matrix used: BLOSUM62 (default), PAM40, PAM120, PAM250 etc.
CUTOFF: Score S, determined from EXPECT by default
EXPECT: Number of Random Hits expected to be found (default = 10)
FILTER: Option to use one of two “filters”, to ignore highly repeated sequences or sequences of “low information content”.

The BLAST algorithm works in the following steps

Preprocessing of the query
The first step is to quickly locate ungapped similarity regions between the query sequences and sequences from the database. Similarly, all the words of length w (tuples or words) of the query are compared with those of all the database sequences.
Blast uses a more efficient query are compared with those of all the database words of length w formed with the alphabet of the sequences are generated (for example, with amino acid sequences, if w=2 there are 20^2 =400 possible words, and 20^3 =8000 if w=3) and each word of the query is compared with each word of this exhaustive set and a threshold T for the similarity between words is set. Each position of the query sequence is associated with a list of words that score more than T when compared with the word of the query starting at this position. The similar words are also called neighbors.

Generation of bits
Let D be a sequence of the database, and Q be the query sequence. After the first step, query sequence Q is now represented by lists of neighbors, one list at each position of the query. Comparing Q with D yet consists in looking for identities between the neighbours at each position of Q and the words of D. So each position of Q is compared with each word of D, and it one of the neighbor words at that position of Q is identical to the word of D, a hit is recorded. A hit is made with one or several successive (overlapping) pairs of similar words, and characterized by its
position in each of the two sequences. All the possible hits between the query sequence and sequences from the database are calculated in that way.

Extension of the bits
Every hit that has been generated is now extended, without gaps, in order to determine whether this hit may be part of a larger segment of similarity. So, each hit is extended in both directions, and in order to make this extension step fast enough, an extension is stopped as soon as the score of the extended hit decreases more than X (the value chosen for X is a parameter of the program) when compared with the best score that has been reached during the extension process.

Types of BLAST
MEGABLAST search: like a Basic BLAST search, but allows you to change certain parameters in order to perform a more specified BLAST search
- Specify an organism or taxonomic class for your search
- Set the E value
- Filter for low complexity or human repeats
- Query Genetic Codes (blastx and tblastx only)
- Change your scoring matrix

There are other advanced BLAST options

PSI-BLAST: Position Specific Iterated-BLAST, this tool can be used when your BLAST search results given you very few matches, the PSI-BLAST will re-iterated the BLAST searches creating a defined profile, upon re-iteration you may reveal alignment matches that are significant that you would not have found using BLAST alone.

PHI-BLAST: Pattern Hit Initiated-BLAST, this tool can be used to search for a specific pattern or motif in your sequence and in the databases the pattern designates the amino acid sequence you are searching for.

BLAST 2 Sequences: this tool produces the alignment of two given sequences using BLAST engine for local alignment.

Steps involved in the BLAST tool:

*How to do a BLAST search:*

Step 1: Go to the BLAST page at NCBI site. All the BLAST related tools/programs are available from NCBI.

Step 2: Paste the sequence in FASTA format in the data entry field.

Step 3: There are several boxes below the sequence entry box. For trial purpose, leave all the fields to default values.

Step 4: The important field is a drop menu that allows the selection of the database to use. The nr (non-redundant databases) is the default setting. Depending on the choice any database can be selected from the list.

Step 5: The only other option that could be changed is the filtering. Uncheck the “Choose filter” check box. In order to get true positive hits, the sequences should
normally be filtered. The filter option ensures that no false positive results are obtained due to short sequences that are very common across biological sequence databases/spectrum.

Step 6: Now click on the “BLAST” button to run the search.
Step 7: A new page will appear showing the ID number of the search and the approximate wait time. Click on the “Format” button and wait for results. The results will be returned when the search is complete.

Interpreting BLAST results:
1. For the more complicated search problems, the length of BLAST hits becomes much more important:
   If the query sequence is short (less than 100 nucleotides or amino acids long), even if there is an exact match the top E-values may be larger than $1 \times 10^{-50}$. We have to check the percentage identity of the top hits, not just the E-values. Hits with low E-values that only have similarities to short regions of the query sequence are more likely to indicate that the sequence have motif or functional domain similarities rather than that they represent related genes or proteins.
   Hits with higher E-values, in the ranges of $1 \times 10^{-50}$ to $1 \times 10^{-5}$ may still indicate that the query and the hits are related. If the hit has at least a 35% identity with the query over at least 80% of its length.

The higher the percentage of hits obtained indicates the more likely of the query related to the matches. But, if the similarity is only in small regions, this likely indicates related functional domains rather than related proteins/gene products.

Terminologies:
Positive/Negative:
It is the label produced by a method, ex: in a dot plot the points are positive and non-points are negative.
True/False:
True points are the ones are correctly assigned as positive (Homologous) or negative (Non-Homologous).
There are four possibilities:

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>P+ True Positive</td>
<td>N+ False Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>P- False Negative</td>
<td>N- True Negative</td>
</tr>
</tbody>
</table>

Statistical Parameters for Sequence Alignment:
For any given alignment, one can calculate a score, which shows the quality of the alignment. Some of the parameters used to determine the statistical significance of the alignment are:
$P$-value: Probability that a random comparison will score at or above a given threshold.
Ex: Probability of a window scoring greater than 100: $P(\text{score}_\text{win} \geq 0.2 \times 10^{-3})$
**E-value:** It provides information about the likelihood that a given sequence alignment is significant. Using larger E-value cut-off in a database search allows more distant matches to be found, but it can also lead to wrong alignments. E-values of 0.01 to 0.001 is generally used for database searching.

**Z-score** (standardized score, standard normal deviate): It is a measure of the significance of an alignment relative to random alignment, similar to BLAST E-value.

\[
Z = \frac{(\text{Obs}_\text{score} - \text{Exp}_\text{score})}{\text{Std\_deviation}}
\]

Rule:
- \(Z < -3\) No evidence of homology
- \(-3 < Z < 6\) Homology possible
- \(6 < Z\) Strong evidence of homology, \((Z > 8)\) better

**References and suggested readings:**