

Internet On-Ramp

BLAST on the Web

Bruno A. Gaëta, eBioinformatics Inc.,
Sydney, Australia

The last few decades have seen an explosion in the amount of DNA and protein sequence available through public databases. The major nucleotide sequence databases double in size approximately every 14 months, and the number of completely sequenced organism genomes is constantly increasing. It is now impractical for scientists to maintain local copies of the sequence databases and, as a result, the Internet is now the main avenue of access to this information.

Sequence databases are typically searched using keywords or a sequence. Keyword search engines such as the Entrez system provided by the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>) search the annotation section of the sequence database record. The major application of these databases involves sequence similarity searching, where sequences similar to a query sequence supplied by the investigator are identified in a sequence database. In most cases, this operation identifies potential homologues of the query sequence, that is, database sequences sharing a common evolutionary history with the query sequence. Homologous sequences can provide a clue as to the function of the query sequence, its evolutionary history and its structure. For example, if an unknown protein sequence is found to be markedly similar over its whole length to the sequence of a protein of known structure in the database, the two proteins are likely to have similar structures and related functions. Other applications include the identification of coding regions in uncharacterized genomic DNA by looking for DNA regions that can be translated into an amino acid sequence similar to that of a known protein.

BLAST (Basic Local Alignment Search Tool) (1) is the most commonly used suite of programs for sequence database similarity searching. It allows rapid searching for sequences similar to a query sequence. The BLAST programs provide access to a large number of biological sequence databases all over the Internet, including central repositories of public sequence data such as NCBI, the Japanese GenomeNet server (<http://www.blast.genome.ad.jp/>) and EBI, the European Bioinformatics Institute (<http://www.ebi.ac.uk>). BLAST is also used to provide access to genome sequence data provided by large-scale sequencing centers such as the Institute for Genome Research (<http://www.tigr.org>) and the Sanger Centre (<http://www.sanger.ac.uk>). The package is also used as a front-end to smaller collections of sequence data such as organism-specific genome databases, for example, the HIV sequence database at <http://hiv-web.lanl.gov/index.html>. Web-based sequence analysis service providers such as BioNavigator (<http://www.bionavigator.com>) also use the BLAST programs as a central method for accessing their collections of sequence data. A partial list of publicly available BLAST servers is available at http://gifts.univ-mrs.fr/BLASTULA/Blastula_links.html.

All these sites make use of BLAST, but the implementation of the package can vary greatly between different sites. Differences include the version of the program used, the choice of programs available and the optional program parameters that can be modified by the user.

BLAST searches for similar sequences in the database and

creates alignments between the query sequence and the matching sequences. Version 1 of the BLAST package produces ungapped alignments, but later versions create alignments that can contain gaps and provide a better comparison of sequences that diverged through residue substitution and also insertion and deletion mutations. There are two main implementations of this method: the gapped BLAST (2) developed at NCBI, and WU-BLAST (<http://blast.wustl.edu/>), which provides access, for example, to the databases maintained at the EBI. Both implementations produce gapped alignments but use a slightly different approach and set of parameters. The database to be searched will usually dictate the version of BLAST that can be used since most Web sites maintain only one version of the package.

The database to be searched will also dictate which BLAST programs can be used. There are three BLAST programs for searching nucleotide sequence databases (**blastn**, **tblastn** and **tblastx**) and two programs for searching protein sequence databases (**blastp** and **blastx**; the more recently developed PSI-BLAST and PHI-BLAST programs that are variations on **blastp** are not discussed here).

The first program, **blastn**, searches a nucleotide sequence database using a nucleotide sequence. It is the only BLAST program to produce nucleotide sequence alignments. Typical uses of **blastn** include checking whether a newly obtained DNA sequence has already been published, identifying coding regions in genomic DNA by searching for encoded (or related) mRNA sequences and extending sequence fragments by searching collections of partially characterized sequences such as EST (Expressed Sequence Tags) databases. **Blastp** compares a protein sequence to a protein sequence database. This program is typically used to identify new protein sequences, retrieve homologues of a given protein, identify the domain structure of a protein or determine the structure of a protein by identifying related proteins of known structure. **Blastx** translates a nucleotide sequence into protein and searches a protein sequence database. It is particularly useful for identifying potential coding regions in newly sequenced DNA. **Tblastn** uses a protein query sequence to search a translated nucleotide sequence database and is useful for detecting unidentified proteins encoded in a nucleotide sequence database. Finally, **tblastx** compares a nucleotide sequence with a nucleotide sequence database, after translating both sequences into proteins. This approach greatly increases the sensitivity of the search when dealing with protein-coding regions but is substantially slower than the direct comparison performed by **blastn**.

Internet BLAST Access

Different sites on the Internet provide access to different databases and to different sets of BLAST programs. For example, the EBI provides access to **blastp**, **blastn** and **blastx**, but not to **tblastn** and **tblastx**. The Sanger Centre only offers access to nucleotide databases and only to **blastn**, **tblastn** and **tblastx**. NCBI maintains both nucleotide and protein sequence databases and so provides access to all five programs. However, to limit the load on the server computer, NCBI does not provide Web access to its large nonredundant nucleotide sequence database using the computer-intensive **tblastx**. An alternative is provided by private sequence analysis service providers such as BioNavigator, which allows **tblastx** searching of its nonredundant nucleotide database.

Internet On-Ramp

When planning a database search, first decide which database to search. For most common applications such as checking whether a sequence is new, identifying the function of an unknown sequence or collecting homologous sequences for further analysis, the best starting point is a nonredundant database that brings together the majority of the available data. These databases are maintained at a number of Web sites around the world including NCBI, EBI, GenomeNet and BioNavigator. Which site to choose depends on personal preferences regarding the user interface and the available output options. For example, the NCBI BLAST server includes a useful graphical overview of the results in the output page, and the BioNavigator BLAST output viewer makes it easy to save the database sequences identified by the BLAST search for further analysis.

When searching specialized databases and database subsets, the choice of BLAST servers is frequently more limited. Users will often have only one option available on the Web if they want to access a specific set of data (e.g., the TIGR gene indices that are available for searching only on the TIGR Web site). The situation is especially complicated when searching for ESTs because there are many proprietary EST collections available only at specific sites on the Web, in addition to the major public domain collection maintained at NCBI.

BLAST Parameter Configuration

The popularity of BLAST has made it one of the most reliable sequence analysis software packages currently available, and one that will perform adequately in most cases without modification of its default parameter configuration. This is why many BLAST Web sites do not allow users to change the program parameters, with the exception of the database to be searched or the BLAST program to be used. In some cases, however, it may be useful to change some of the parameter values to improve the results when the option is available.

The default BLAST configuration works well in general with query sequences of medium length (100–1000 nucleotides or 50–200 amino acids). Searching with longer query sequences can, on occasion, retrieve preferentially longer sequences from the database, which may de-emphasize shorter significant matches. This can present a problem when the aim of the search is to identify significant short regions of similarity in a longer sequence (e.g., to determine the domain structure of a new protein or identify short exons in a comparatively long genomic DNA fragment). For these types of searches, try splitting long query sequences into shorter overlapping fragments (1000 nucleotides or 200 amino acids each) and use each as a separate BLAST query to see if interesting additional hits are identified in the database.

BLAST searches using short query sequences may commonly not report any matches in the database. This is because BLAST reports only significant matches that are unlikely to be the result of random sequence similarity. Since matches with a short sequence are likely to have occurred by chance, the program does not report them. While this is useful when searching for homologous sequences, it prevents using BLAST for searching for potential candidates matching a short protein sequence (such as one obtained by amino-terminal sequencing of an unknown spot on a two dimensional protein gel). A notable application of BLAST with short sequences involves checking whether PCR primers are likely to

match many sequences in a genomic DNA preparation and give rise to the wrong PCR products. Using the primer sequence as a query in a **blastn** search of a large nucleotide sequence database could do this, but since the matches would be happening by chance, they would be unlikely to be reported by the program. When a search with a short sequence does not retrieve any (or enough) database hits, a workaround is to raise the value of the Statistical Expectation Threshold parameter (**Expect** on the NCBI page, **MAX.EXP** on the EBI WU-BLAST server). This controls the level of statistical significance that will be reported by the program. The default value for this parameter is 10, corresponding to a cut-off level of similarity where 10 reported database matches are expected to have been found by chance alone. Raising the threshold will cause the program to report more chance matches and may provide an answer in this case. Note that on many BLAST servers, the maximum expectation threshold value that can be selected from the Web form is 1000, which is not enough for some applications involving short query sequences (some servers allow this limit to be bypassed by typing a line command). Raising the statistical expectation is the first thing to do when a search does not report any matches when some were expected. However, remember that most matches with a high statistical expectation score are more likely to be due to chance sequence similarity than to biological relationships.

Another parameter available on most BLAST servers lets the user control whether the query sequence should be filtered before the search. Filtering removes from the query sequence regions of limited amino acid composition or simple repeats. These “low complexity” regions are of concern because they can cause a match to be identified as more significant than it really is or to retrieve many false positives. As a consequence, most BLAST servers filter query sequences by default to mask these regions before starting the search. A low complexity region that has been masked can appear as a string of Ns or Xs (protein BLAST programs) in the query sequence within the sequence alignment region of the BLAST output. In some cases (if the low complexity region is at the beginning or the end of the query sequence), a filtered region may not be shown at all in the BLAST output.

Since filtering the query sequence effectively removes some data from the search, it is often useful to repeat a search with the filtering turned off to obtain better sequence alignments or to check whether the low complexity region adds essential information to the search query. This is particularly important if a search retrieves no matches with filtering turned on.

Most BLAST servers also allow users to change the scoring matrix, or substitution matrix, for protein-protein comparison. This matrix is a table assigning a score to all possible pairings of amino acids in an alignment and controls how the program handles the likelihood that one amino acid can substitute for another between homologous proteins. Most of the modern substitution matrices have been derived empirically, and while the default matrix, BLOSUM62, works best with the majority of sequences, other matrices may perform better when trying to detect distant homologies in particular cases. Basically, the lower the number associated with a BLOSUM matrix, the more distant homologies it can detect, but the more nonrelated sequences it is likely to report. For example, BLOSUM45 is thought to perform better than BLOSUM62 for detecting distant homologies. This rule is reversed for the

Internet On-Ramp

PAM family of matrices: the PAM matrices associated with low numbers, such as PAM30, are more suited to detecting closely related sequences, whereas PAM250 is better at detecting distant homologies (again, with a corresponding increase in false positives). Overall, the BLOSUM matrices that are derived from the BLOCKS database of conserved protein regions (3) are considered better performers than the older PAM matrices, which are derived from a smaller set of aligned globular proteins (4). However, many servers still provide access to PAM matrices because they may perform better in some cases, especially with short query sequences.

While BLAST contains more parameters that can be configured to improve its sensitivity or performance in specific cases, keep in mind that this method was designed first and foremost for speed. It may be useful to tweak some parameters in specific cases and experiment with different values, but for critical searches it is also a good idea to try a slower, more rigorous search method such as the Smith-Waterman algorithm. Implementations of this method include the program **ssearch** in the **fasta** package (5) available on the GeneStream server (<http://vega.crbm.cnrs-mop.fr/>) and BioNavigator, and programs running on specialized hardware (<http://www.irisa.fr/SAMBA>, http://www.ch.embnet.org/software/GMFDF_form.html, <http://www2.ebi.ac.uk/Bic/> and <http://decypher.stanford.edu/expsw.htm>).

REFERENCES

1. **Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman.** 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
2. **Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
3. **Henikoff, S. and J.G. Henikoff.** 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89:10915-10919.
4. **Schwartz, R.M. and M.O. Dayhoff.** 1978. Matrices for detecting distant relationships, p. 353-358. *In* M.O. Dayhoff (Ed.) *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, Natl. Biomed. Res. Found., Washington, DC.
5. **Pearson, W.R. and D.J. Lipman.** 1988. Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. USA* 85:2444-2448.