OK INBRE Bioinformatics

January 2005

University of Oklahoma Health Sciences Center

What is BLAST?

by Edgar Scott, II, M.S.

The fundamental way of learning about a protein or gene is to compare it to well-known proteins or genes. With the large number of sequence databases available, an investigator can search through millions of characterized (and uncharacterized) sequences for one or more sequences that are similar to a query sequence. He or she can search through a list of significant alignments from a sequence similarity search and explore the similarities and differences between the query sequence and the significantly matching sequences. The Basic Local Alignment Search Tool (BLAST) is the main sequence similarity search tool for just such a task. In this month's bulletin, I will introduce you to BLAST: how it can be used, what BLAST algorithms are available, how it works, how to interpret the results and where you can find it.

BLAST can be used to address any number of concerns. For example, an investigator can determine if homologous sequences for a particular protein or gene exists in one or more organisms. Also, given an uncharacterized DNA or protein sequence, an investigator can determine its identity from a list of the sequence alignments returned from a BLAST search. Genome annotation, for instance, relies heavily on the ability to compare uncharacterized sequences discovered in a genome to characterized sequences stored in a database. If a BLAST search that both sequences are significantly similar, then there is a strong possibility that both sequences share a similar function. Other applications of BLAST include discovering new genes or determining what sequence variants have been described for a specific protein sequence.

BLAST is a heuristic algorithm that enables it to align sequences faster than other more rigorous alignment tools. BLAST will first search for short identical matches (called words) between the query sequence and the sequences from the database. If a word match scores above an arbitrary threshold cutoff score, the alignment is lengthened in both directions until the sequence alignment score drops below an arbitrary sequence alignment cutoff score. This is performed for all the sequences in the database. The result is a sorted list of local alignments between the query sequence and several matching database sequences.

It is important to be aware that because BLAST is a heuristic algorithm, some accuracy is sacrificed for speed. The size of the word and the threshold cutoff score can affect the number of false-negative alignment (an alignment that was falsely rejected, but should have been included in the BLAST results). For example, larger word sizes will provide faster searches, but also will result in more false-negatives and a smaller number of BLAST matches. A higher threshold cutoff score will likewise increase the speed of BLAST searches, but will result in more false-negatives and a smaller number of results. Smaller word sizes and lower threshold cut off scores will provide a more accurate search, but also will include more results and make your search slower. Although, there have been major improvements to the BLAST algorithm so that it performs a more thorough search for significant matches, it is still not a perfect tool.

Because DNA sequences can only be compared to other DNA sequences, and likewise for proteins, there are specific BLAST algorithms for comparing each type of sequence. To search a DNA database, one would need to choose the blastn tool. The blastp tool will compare a protein sequence to a protein sequence database. Both are great tools to use when searching for homologous sequences, however, the use of blastp can identify more distantly related sequences because the protein sequence is more conserved than the DNA sequence.

What if an investigator wants to compare the translation of a nucleic acid sequence to a protein database or maybe compare a protein sequence to the translations of the sequences from a DNA database? Instead of translating the sequences by hand, the following translation BLAST algorithms can be used: blastx, tblastn and tblastx. Blastx will compare each of the six-frame translations of a DNA query sequence to a protein sequence database. Tblastn will compare a protein sequence to the six-frame translations of the DNA sequences from a DNA database. Lastly, tblastx will compare the six-frame translations of a DNA query sequence to the six-frame translations of the sequences from a DNA database. An example using one of these translation BLAST tools would involve the comparison of an expressed sequence tag (EST) to a protein sequence database to determine the identity of the cDNA that was originally sequenced. This is where blastx can be used to translate the EST and then use the resulting amino acid sequences as query sequences to be compared to a protein sequence

When searching for homologous sequences with BLAST, one must also be aware of the possibility of false-positives appearing as significant matches. Not only are the alignments from BLAST searches accompanied with percent identity scores, percent similarity scores (if dealing with proteins) and alignment scores, but also with expectation values (E-values). E-values are used to estimate the

The University of Oklahoma Health Sciences Center

Laboratory for Genomics and Bioinformatics

Edgar Scott II, M.S.
Multicampus Bioinformatics Education Specialist

Phone:
(405) 271-2133, Ext. 32511
E-Mail:
edgar-scott@ouhsc.edu

We're on the Web!

See us at:

microgen.ouhsc.edu/inbre/

Web sites

1. www.ncbi.nlm.nih.gov/BLAST/
2. www.expasy.uniprot.org
3. www.ebi.uniprot.org
4. www.pir.uniprot.org
5. flybase.bio.indiana.edu/

Bulletin (continued)

significance of an alignment from a BLAST search. An alignment's E-value is an estimate of the number of times the investigator can expect to randomly obtain an alignment that has an alignment score that is equal to or greater than the alignment score of the pairwise sequence alignment in question. For example, if an alignment from a BLAST search has an E-value of 10, then this tells the investigator that he or she can expect to find 10 alignments that score just as well or even better than this alignment, by random chance. The lower an alignment's expectation value, the more significant the alignment found by BLAST. A simple rule of thumb is that an expectation value less than $1x10^{-5}$ is significant.

How can one determine if a query sequence and a BLAST match are homologous? First one would need to determine if the expectation value is significant and if the percent identity score is high enough (greater than 35 percent as a rule of thumb). Second, determine if the two aligned sequences are approximately the same size and if the alignment extends along a large majority of both sequences. Because BLAST returns local alignments, it is possible that the resulting alignment covers a region of high similarity between the two sequences. Two homologous sequences will have significant similarity along the entire length of both sequences. If anything is known about the biochemistry of both sequences being compared, then functional domains and active sites should be readily visible in the alignment with more similarity than other regions. An investigator can also determine if the two proteins share similar 3-D structures, but only if the structures have been determined.

Just about every web site that provides access to a sequence database will also provide BLAST for sequence similarity searching. One very popular web site where an investigator can search for protein or DNA sequences is the National Center for Biotechnology Information (NCBI) web site[1]. This site provides several different DNA and amino acid databases that can be searched, including GenBank, RefSeq, EST databases, structure databases and more. The Universal Protein (UniProt) Database provides well-curated protein sequence databases that can be searched with BLAST and is available through three different web sites: Expert Protein Analysis System (ExPASy)[2], European Bioinformatics Institute (EBI)[3], and Protein Information Resource (PIR)[4] of Georgetown University. There are also a large number of organism specific databases that can be searched with BLAST including FlyBase[5] for the Drosophila genome.

BLAST searching has emerged as an indispensable tool for comparing a DNA or protein sequence to millions of sequences in public sequence databases. With the five BLAST algorithms, one can identify conserved domains between two sequences, the existence of homology, and even identify new genes. Although BLAST is fast, it isn't as accurate as other, more rigorous algorithms such as the Smith-Waterman local alignment algorithm or the Needleman-Wunsch global alignment algorithm. BLAST has become one of the staples of bioinformatics tools in both genomic and proteomic research.

Related bioinformatics papers and links of interest …

- Scott, M.; Madden, T.L.; *BLAST: At The Core of a Powerful and Diverse Set of Sequence Analysis Tools*. Nucleic Acids Res. 2004 July 1; 32(Web Server issue)
- Pertsemlidis, A.; Fondon, J.W.; *Having a BLAST with Bioinformatics (and Avoiding BLASTphemy)* Genome Biol. 2001; 2(10)
- http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.610

About Us…

The OUHSC Laboratory for Genomics and Bioinformatics is a full-service genomics facility offering DNA sequencing (small- and large-scale projects), microarray design and hybridization and other services, including bioinformatics support. Edgar Scott is the INBRE Multicampus Bioinformatics Education Specialist, responsible for fostering the development of bioinformatics education on 14 undergraduate campuses in the state of Oklahoma and coordinating INBRE-related bioinformatics activities with the INBRE Bioinformatics Core. The University of Oklahoma is an equal opportunity institution. This publication is printed and issued by the University of Oklahoma; the cost of $150 was paid by OK INBRE.