Lecture/Lab: BLAST

Materials last updated June 2007

Joanne Fox, Michael Smith Laboratories, UBC

Key Concepts

- **An introduction to BLAST, the Basic Local Alignment Search Tool**
- **Understanding the BLAST algorithm and BLAST statistics**
- **Understanding the meaning and relevance of BLAST parameters, including scoring matrices**
- **An introduction to the advanced BLAST programs, including PSI-BLAST**

What you will be able to do at end of this section

- **Run BLAST searches through a web interface**
- **Use your understanding of the BLAST algorithm to customize BLAST searches by changing parameters**
- **Apply your knowledge of Entrez queries to BLAST searches**
- **Complete the BLAST assignment**

Introduction

The comparison of sequences is one of the most common bioinformatics analyses carried out today. The premise behind sequence similarity searching is that one sequence by itself is not informative. But combined with the knowledge that exists in sequences databases, you can compare your starting sequence, or query sequence, with other sequences to develop hypotheses concerning relationships and function. The methods used to assess sequence similarity fall into two different classes. The first approach involves the comparative analyses of pairs of sequences to ask questions like, “Do these two proteins have similar functions?” The second approach involves the comparison of sets of sequences to address questions like,

"What are the features common to this family of proteins?" Both of these approaches can provide us with a wealth of information including suggestions regarding functionality of a sequence, evolutionary history, or important residues within a sequence. In this lecture and lab combination, we will take a practical look at the first approach by introducing the BLAST sequence similarity searching tool. The lecture will cover the basics behind BLAST from understanding the algorithm to deciphering the results. The second approach, most often implemented through the generation of multiple sequence alignments, will be covered on Day 4 of the workshop. In the hands-on laboratory, we will focus on web based similarity searching methods, in particular those available on the NCBI BLAST pages. This lecture/lab section will be followed by an assignment where you will be able to apply your skills and carry out some BLAST searches using the example sequences provided.

An Introduction to BLAST

The Basic Local Alignment Search Tool (BLAST) is a powerful way to carry out sequence similarity searching. Other methods such as FASTA and BLAT also exist, but will not be discussed here. Before we go any further, we need to lay down some rules. First, as a bioinformatician, you have an obligation to correctly use the terms: **homology** and **similarity**. Many scientists use these terms interchangeably when they actually mean quite different things. **Similarity** is a measure of how related two sequences are, whereas **homology** is a conclusion about the evolutionary relatedness of two sequences based on an assessment of their similiarity. Two sequences can be said to be 68% *similar* but these same two sequences are either *homologous* or not. There is no degree to homology, two sequences are either related or not. At the next bioinformatics seminar you attend, you can correct the misinformed graduate student who attempts to state that protein X is 23% homologous to protein Y.

BLAST is, of course, the tool that many scientists use to infer relationships of homology based on the degree of sequence similarity between two sequences. BLAST is a local alignment tool which means that it searches for regions of similiarity instead of trying to align the entire length of the sequences. Although global alignment methods result in the most mathematically optimized alignments of full length sequences, they may miss local regions of sequence similiarity. Often, local alignment methods can detect small regions of similarity resulting in a more biologically significant alignment. In addition, global alignment methods are computationally intensive and slow. On the other hand, local alignment tools, like BLAST, are

based on computationally efficient search algorithms which break the large problem of finding similar sequences down into smaller pieces, making the method faster.

Understanding the BLAST Algorithm and BLAST Statistics

As with any sequence similarity searching method, it is important to understand how the method works and what measures or statistics are presented to aid in your evaluation of the results. By understanding the BLAST algorithm and a few key BLAST statistics, you will be better able to interpret BLAST results. Interpreting BLAST results requires you to apply your biological expertise, your understanding of BLAST statistics, and your practical experiences with BLAST.

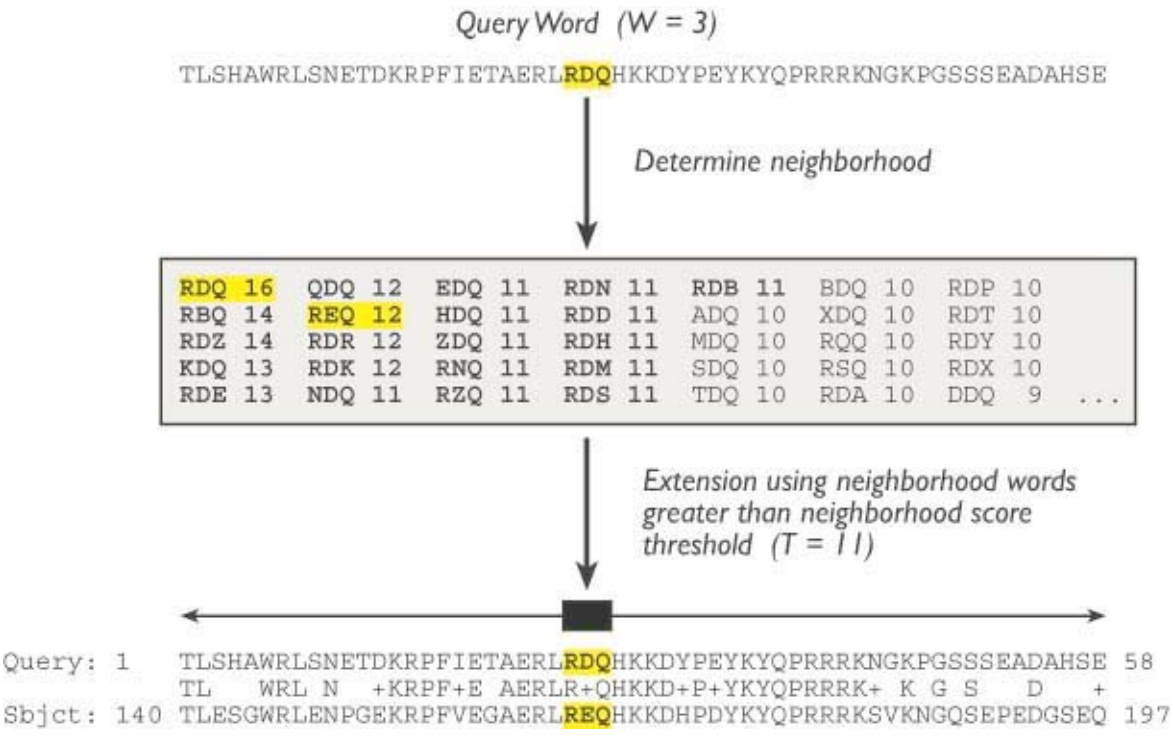
Scoring Matrices

The measure of similarity between two sequences is captured by a scoring scheme in BLAST which is based on scoring matrices. **Scoring matrices** are empirical weighting schemes that are used in comparing sequences and capture information about residue conservation, residue frequency, and evolutionary models. In BLAST, substitution matrices are used for amino acid alignments whereas nucleotide matrices are compared using identity matrices. These matrices are "look-up" tables in which each possible residue is given a score reflecting the probability that it is related to the corresponding residue in the query. The two most commonly used substitution matrices are the BLOSUM and PAM scoring matrices. The **PAM** (point accepted mutation) scoring matrices are based on global alignments of closely related proteins. The PAM1 matrix is calculated by looking at the amino acid substitutions that occur in proteins with no more than 1% divergence (1 change per 100 amino acids). In an effort to model evolutionary changes, the other PAM matrices are extrapolated from PAM1 by matrix multiplication. The **BLOSUM** (blocks substitution matrices) are based on local alignments where the BLOSUM62 matrix is calculated from comparisions of sequences with <62% identity. All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins. An understanding of scoring matrices combined with an understanding of the BLAST algorithm will allow you to more effectively interpret BLAST results.

A Brief Description of the BLAST Algorithm

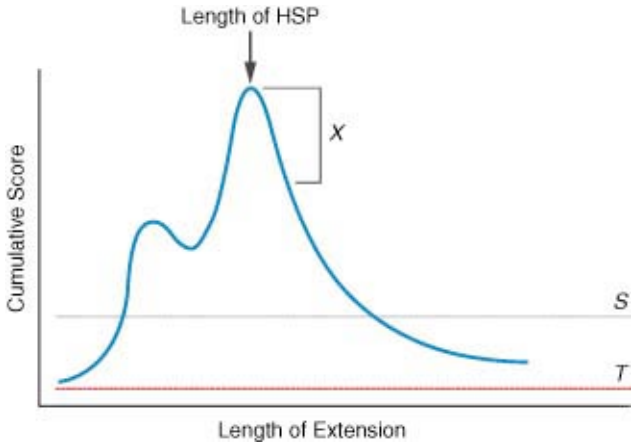
The BLAST algorithm finds regions of local alignments by breaking the query sequence down into smaller chunks of sequence called **words**. These words are then indexed by the computer along with information about where each is found in the intact sequence. The BLAST algorithm then starts by seeding the search with this small subset of letters from the query sequence. These words from your query sequence are then used to scan the database for matches above

a certain threshold. This process of scanning a database with small sequence fragments is far faster than scanning a database with a large sequence.



Pictures used with permission from Chapter 11 of "Bioinformatics: A practical Guide to the Analysis of Genes and Proteins." 3rd Edition A. Baxevanis & B.F.F. Ouellette (eds.) Wiley 2004.

Once the initial word hit has been found, the BLAST algorithm attempts to extend the match in the immediate sequence neighborhood. Extension proceeds and the cumulative score is calculated using the scoring matrices discussed above. As long as positive matches and conservative substitutions outweigh the negative scores for gaps and mismatches, the cumulative score will increase. When the cumulative score starts to drops off, the BLAST algorithm measures the rate of decay and, once past a certain point, stops trying to extend the alignment. The result is an extended sequence alignment that was initially seeded by a word hit.



This is called an **HSP**, or high-scoring segment pair. All HSPs that have a cumulative score above a certain threshold are reported in BLAST reports. Because the BLAST algorithm carries

out these searches using all possible query words, it is possible that more than one HSP may be found for any given pair of sequences.

BLAST Statistics

After an HSP is identified, it is important to determine whether this match is significant or not. Two BLAST statistics, the score (S) and the E-value (E) are particularly helpful in making this interpretation. First, the **score (S)** is a good measure of the quality of an alignment because it is calculated as the sum of substitution and gap scores for each aligned residue. Recall that substitution scores are given by look-up tables (PAM, BLOSUM) whereas gap scores are assigned empirically. Second, the **E-value (E)**, or expectation value is a good measure of the significance of the alignment. The E-value is the number of different alignments, with scores equivalent to or better than S, that are expected to occur in a database search by chance. The lower the E-value, the more significant the alignment result. You can remember this definition by thinking of the E-value as the number of times that you *expect* to see your hit occur in the database due to random chance alone. An E-value of 10 means that if you scrambled your sequences up and searched again in a random set of sequences, you would expect to see a hit as good as the one you're trying to interpret about ten times. Often times, a much more stringent E-value (like 10^{-3} for example) is used as a cut-off for interpreting BLAST results. Understanding BLAST statistics helps you to more effectively interpret BLAST results, but always remember that this is a biological interpretation and whenever possible you should apply your knowledge of biology. Does this alignment cover the interesting regions of your sequence? Do you know any functional information that you can combine with this interpretation? As a sophisticated bioinformatician, it is important that you try not to interpret your BLAST results blindly.

Worked Example: Basic Database Similarity Searching Using BLAST

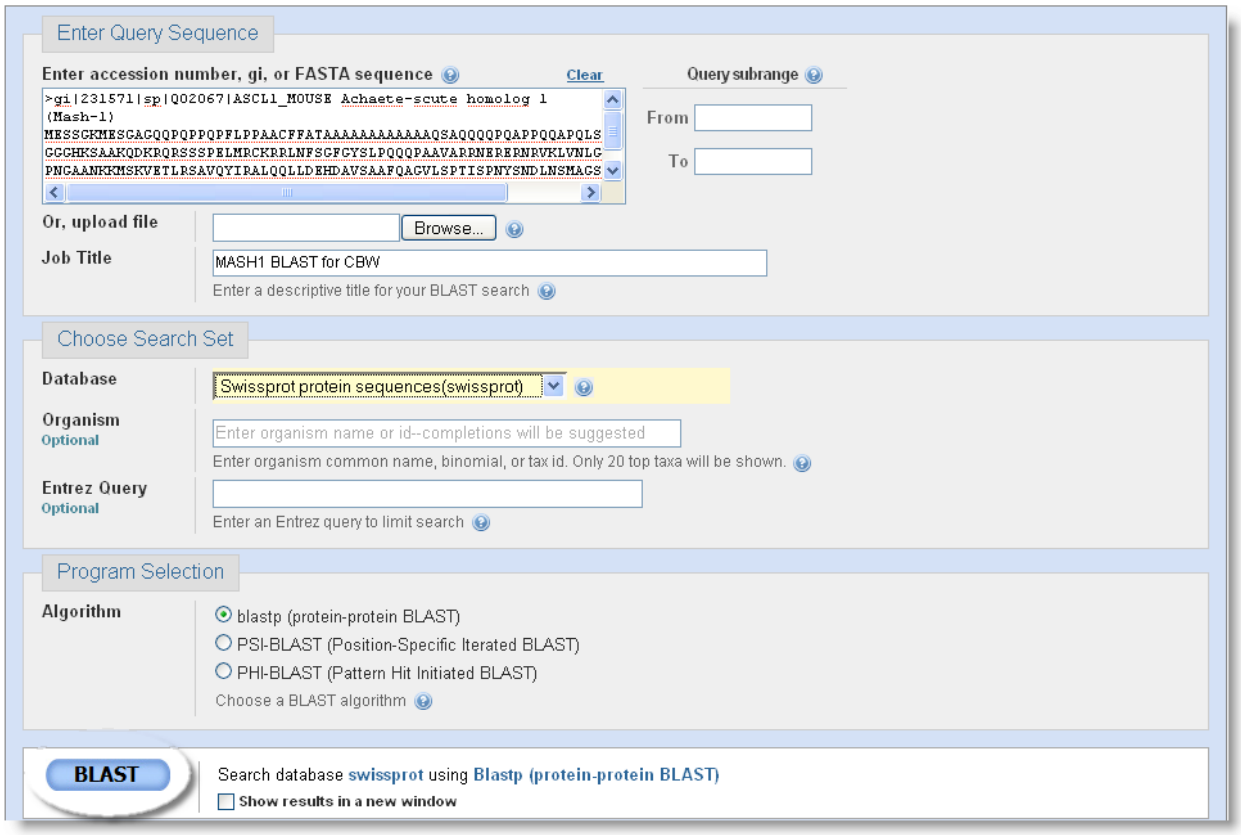
There are many different BLAST programs available, but the ones most commonly used for basic database similarity searching are:

1. blastp: compares a protein sequence against a protein sequence database.
2. blastn: compares a nucleotide sequence against a nucleotide sequence database.
3. blastx: compares a six frame translation of a nucleotide sequence against a protein database

4. tblastn: compares a protein sequence against a six frame translation of a nucleotide database
5. tblastx: compares a six frame translation of a nucleotide sequence against a six frame translation of a nucleotide database.

Choosing the right BLAST program is the first issue that must be considered when preparing a BLAST query. The most meaningful alignments are produced when amino acid sequences are compared. This is a result of the complexity of an amino acid sequence in comparison to a nucleotide sequence. For an amino acid sequence, there are 20 possible letters for each position of the sequence, while for a nucleotide sequence, there are only four possible letters. Therefore, the amino acid sequence contains far richer information than the nucleotide sequence. It is conceivable that two nucleotide sequences with only moderate levels of similarity could encode the same amino acid sequence, and hence the comparison of the two amino acid sequences would yield a far more significant result than a comparison of the nucleotide sequences. In light of this, it is a good practice to always use amino acid sequences in your queries. Alternately, you can use a BLAST program, such as blastx, that will translate a nucleotide sequence for you as it searches through a database. The only exception to this rule would be if you were searching for similarities to a nucleotide sequence that does not code for any protein product.

Submitting a good query is the key to getting the most relevant results from your BLAST searches. Let's say that you have the sequence for the mouse MASH-1 protein (sp|Q02067), and want to find out if there are any similar proteins in human. Since we have a protein sequence and we want to search for other protein sequences, we will use blastp. The blastp query page looks like this:



The first and most important thing we have to do to prepare our BLAST search is insert the query sequence into the field provided. The sequence can be input in several ways. First, the sequence in raw or FASTA format can be pasted into the box as shown. Alternately, you could put the accession or GI number of the query sequence into the provided space. Next, we need to choose which database we want to search against. Databases available for each type of BLAST search are shown in a drop down menu. For a complete listing of available databases, click on the context specific help beside the “Database” drop down men. For this example, we will search against SwissProt. Note that a “CD-Search” is run by default, which will allow for the detection of conserved protein domains in our query sequence.

Scrolling down the page we’ll see that there are many options that can be adjusted if desired. We will focus on a few of them, starting with “Entrez Query”. Any entrez query can be input into the provided field in order to limit the results. Limiting results to include only certain organisms is a common task, so an addition box where you can enter organism names has been provided to simplify the process. Since in our example we are interested in proteins from human, we can just type “human (taxid:9606)”. The next option we should consider is filtering. Low complexity sequences and repeat sequences can also affect the outcome of a database similarity search.

For example, a query sequence containing human Alu repeats will align to many other sequences containing Alu repeats, even though they are unlikely to be related. To avoid a search yielding many spurious matches, query sequences should be filtered for low complexity regions and repeats. Filtering is done by default, and we will not turn it off in this case. There are some cases in which we may want to consider turning it off, such as in the case of a query sequence that is low complexity almost throughout its length. Without turning it off, filtering will hide much of the sequence you have to do your comparison, and no results may be found. Another convenient option available is e-value cutoff. You can use this option to allow only the most significant alignments to be reported, by dropping the e-value down from the default value of 10. Further down the query setup page, are options that allow you to select the scoring matrix and gap penalties used. The BLOSUM 62 matrix is the default scoring matrix used. Insertions and deletions occur in sequences over time, therefore introducing some gaps into alignments is acceptable when appropriate. However, you want to avoid adding many gaps leading to aberrant alignments. To avoid this occurrence, the score for an aligned region is penalized for allowing a gap to open, and further penalized if the gap is allowed to widen or extend. Insertions and deletions are rare events, so the penalty to an alignment score is high for opening a gap, but once a gap is open the penalty for extending it is lower as insertions and deletions often involve several neighboring positions in a sequence. For our MASH-1 query, we will leave both of these options at their default values.

Understanding Your BLAST Results

Now that we have prepared our query, we can press the “BLAST!” button that appears on the page. This will take you to the BLAST formatting page. At the top of this page is information on your search, but the first thing you will likely notice is the results from the CD search. Our example contains an HLH domain. To get more information on the domain found, you can click on the picture. The BLAST results page automatically updates itself when your search is results are ready. If your results are not ready, an estimated time to completion will appear on the results page, updating itself periodically until your results are ready. If you don’t have time to wait for your results, or if you’d like to access them multiple times, you can take note of the request ID. On the Recent Results tab, there is an option to retrieve results using a request ID. Each set of query results is kept for approximately 24 hours. With the new BLAST interface, you can now save search results and search strategies. A very handy feature, indeed!

BLAST results contain a large amount of information to examine. So what does it all mean? The first thing you will notice is a graphical representation of your results. At the top of the graph is a linear view of your query sequence, with the bars below indicating where matches to it occur. Each of the bars is colored according to the score the alignment received. Grey areas in the bars represent areas that are not similar to the query sequence that are surrounded by areas of similarity. Mousing over each of the bars will display the identifier of the aligned sequence, while clicking on a bar will take to the alignment of that sequence with the query. Under the graphic representation of the results you will find a list of hits in order of decreasing significance. Information in this hit list includes database identifiers and descriptions for each of the hits, as well as the scores and e-values for each of the alignments. Underneath your list of hits with scores and e-values, you will find the alignments of your query sequence with the database sequences it is similar to. Along with these alignments you will again find score and e-values, but also information regarding the length of the region of similarity, and percent identity within the similar region. Examining these values and the alignment itself is an important step in deciding if your results are significant. To illustrate this, consider a situation where your top hit appears to be the most significant one, but a hit further down the list actually is a more significant hit as its alignment involves a functionally important area of your query sequence, while the top hit does not.

NCBI BLAST – Advanced BLAST Methods

We have already seen that sequence similarity searches using proteins often yield more significant results than those done with nucleic acid sequences. Advanced sequence similarity searching methods using protein sequences are even more powerful. Several advanced methods are available from the NCBI BLAST pages, including **PSI-BLAST** (position specific iterated BLAST). Other advanced methods like PHI-BLAST (pattern hit initiated BLAST), RPS-BLAST (reverse position specific BLAST), and MegaBLAST also exist but will not be covered here.

PSI-BLAST has many advantages over blastp. It is an iterative method with refined statistical methods, and its strength is in its ability to detect weak protein relationships. This is particularly useful for gaining insight into the function of unknown proteins. Since it is still based on BLAST, it is still fast and the NCBI PSI-BLAST user interface is easy to use. PSI-BLAST can be thought about in simplified terms as a number of steps that repeat:

Step 1. blastp is performed using your query sequence to search a database of your choice. A standard scoring matrix such as BLOSUM 62 is used.

Step 2. The most significant hits from this search are used to build a multiple sequence alignment.

Step 3. The multiple sequence alignment is used to generate a position specific scoring matrix (PSSM). This is done by moving along each position in the multiple sequence alignment and counting the presence or absence of different residues. This information is captured in a PSSM where highly conserved residues in the alignment are given high scores, while more weakly conserved residues are given lower scores.

Step 4. Another round of similarity searching is performed, this time using the new PSSM. This PSSM (created from a set of hits found in the previous round) represents a customized scoring matrix. For example, if in the first round your hit was similar to a particular class of protein, let's say DNA binding proteins, then in the second round the PSSM will be more effective at finding sequences that belong to that same class of proteins.

Step 5. Steps 2-4 can be repeated until no new sequences are found in successive rounds of searching. This is known as convergence.

A few words of caution to consider: PSI-BLAST is very sensitive, but not highly specific. That is to say that it is very good at detecting remote sequence similarities within a set of proteins, but it is also likely to pick up false positives. Carefully inspecting and thinking about all of the sequences that will be included in the creation of the new PSSM is an important task. Applying your biological knowledge in order to include or exclude sequences from the multiple sequence alignment step will help to eliminate false positive results.

Appendix

1. Resources

i) Original Papers

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res*. **25**:3389-3402

- Altschul SF, Boguski MS, Gish W, and Wootton JC (1994) “Issues in searching molecular sequence databases.” *Nature Genetics* **6**:119-129

- Henikoff S, Henikoff JG (2000) “Amino acid substitution matrices.” *Adv. Protein Chem*. **54**:73-97

ii) Text books:

- Baxevanis, AD. And Ouellette, BFF. (editors) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. New York: Wiley Interscience, 2005. (Chapter 11)

iii) Web Sites:

- NCBI BLAST homepage: http://www.ncbi.nlm.nih.gov/BLAST/
- NCBI BLAST FAQ: http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.shtml
- NCBI BLAST tutorial: http://www.ncbi.nih.gov/Education/BLASTinfo/tut1.html

iv) Acknowledgements:

- I would like to thank the following individuals for their contributions to these materials: David Wishart, Francis Ouellette, Andy Baxevanis.

2. Demonstration Overheads