Biol 381 – Bioinformatics Lab
Lab Handout – Introduction to BLAST

Imagine that you have isolated an unknown protein from an organism you're studying, and you want to learn more about it. You know the protein's sequence, but nothing about its function. How would you go about learning more about it? Alternatively, what if you know the identity of your protein and its function, but want to know how important it is to the survival of the organism?

Often, the most logical first step for problems like this is to perform a similarity search using a tool like BLAST at NCBI. BLAST has access to sequences from many different databases, and can help you find other sequences that are closely related to your sequence. This can be helpful in multiple ways. For problems like the first described above, it may allow you to find sequences that are similar to yours and have a known function, which can give you a clue about the function of your protein. For the second problem, it can tell you something about how conserved your protein is, and therefore how essential it is to the survival of the organism. If your BLAST search returns a large number of very similar sequences, it is likely that your sequence is conserved – and important to the organism's continued survival.

As you can see, BLAST is quite a useful tool for biologists. In today's lab exercise, you'll get a chance to experiment with BLASTing a sequence and seeing what you can learn about the sequence by doing so. Additionally, you'll work with different parameters for your BLAST searches, in order to get an idea of how changing those parameters changes your results, both in terms of the results (or "hits") that you get, and in terms of how closely related (evolutionarily) those matches are considered to be to your sequence. You may find the "Taxonomy Report" useful for this aspect.

Part 1

Your group will BLAST your gene (group/gene assignments will be written on the board) using several different sets of parameters. For each search, fill in the table on the last page of this handout with the query accession number, parameters used, and E-value of the most distant hit. Make sure you also write down observations about the relative sensitivity of each query. Your observations and analysis, along with your answers to the questions in this handout, will be part of this week's write-up.

Step 1:
BLAST your protein using the blastp program and the RefSeq Protein database. Leave all parameters set to their default values.

- Based only on these search results, how conserved is your protein? How can you tell?

Step 2:
Repeat the search, but change the substitution matrix to

a BLOSUM45
b BLOSUM80
c PAM30
d PAM70

See Figure 1 for a rough guide to the relative sensitivity of the PAM and BLOSUM matrices.

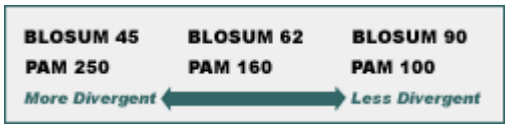

Figure 1: BLOSUM and PAM comparison. Adapted from http://www.ebi.ac.uk/help/matrix.html.

For each search, explain how changing the substitution matrix altered your results.

- Which search is the most sensitive, and which is the least sensitive? (If necessary, you can increase the maximum number of results to get a better idea of how different scoring matrices affect your results.)

Step 3:
Look up the mRNA sequence for your protein, and use it to do a blastn search.

- How do the results compare to the results you obtained from blastp?
- Is blastn better suited to finding closely related genes or distantly related ones?

Step 4:
Change the search program to megablast and run another search using your mRNA sequence.

- What do you notice about your megablast results compared to blastn?
- What happens if you change the word size to 128 or 256? What if you change it to 16?

Step 5:
Predict what will happen to your results if you run your search again using one of the following programs: blastx, tblastn, or tblastx.

- What is your prediction? How do you think the sensitivity of this search will compare to the sensitivity of your other searches? Justify your prediction.

Next, do the search to test your prediction.

- Was your prediction correct?

Step 6:
Now that you know more about how BLAST works, design a search that will be best suited to finding sequences that are close relatives of your protein. To confirm that your search is optimized for closely related sequences, note the E-value of the most distant hit.

Next, design a search that will find distant hits for your protein.

- What is the most distant hit of this search, and do you think it is really homologous to your protein? Why or why not?

Part 2

In Part 1 of this exercise, you saw that you can use BLAST to compare your sequence against an entire database of sequences in order to find sequences that are similar. But what if you have two sequences, and want to use BLAST to see how similar they are to each other? In this part of the exercise, instead of focusing on using BLAST to find homologs of one sequence, you'll use

NCBI's bl2seq feature to look at a pairwise alignment of two sequences at a time in order to closely examine and visualize their relationship to each other.

In addition to the standard scoring and alignment features, bl2seq provides a dot plot as a visual representation of the similarity of your two sequences. Figure 2 is a simplified example of how dot plots work.

- How do you think the dot plot would look if one of the sequences contained an inversion? What about a deletion or insertion?

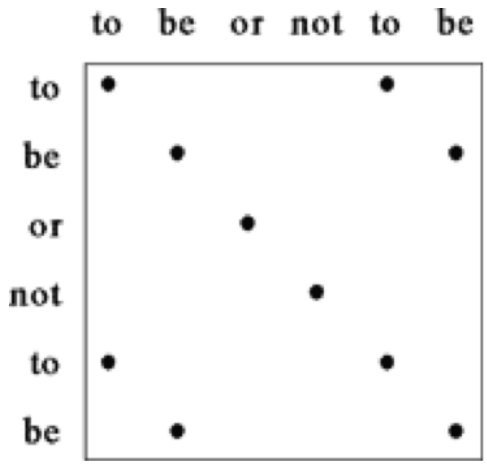

Figure 2: Simplified dot plot example. The "sequence" on the X axis is compared to the sequence on the Y axis. A dot is placed at each position where the two sequences match. Adapted from http://imagebeat.com/dotplot/overview.html.

Each group should download their sequences from Blackboard (Lab 3-Introduction to BLAS/Sequences).
Each group member should follow Steps 1 and 2 with their main sequence (Sequence#) and one of the remaining sequences (Sequence#_[letter]). Make sure everyone has a different pair of sequences to analyze.

Step 1:
Enter Sequence # as a blastn query. Select the "Align two or more sequences" checkbox to get a text box for the subject sequence, and enter Sequence #_A in that box.

Step 2:
When your BLAST results load, look at the "Dot Matrix View" for the alignment. This view is a visual representation of how similar the two sequences are to each other.

- What do you notice about the alignment? Do any patterns stand out? If so, what do the patterns represent?
- Do you think sequences are homologous? Why or why not?

Compare your results with those of the other members of your group.