

Identifying an unknown protein from short peptide sequences

Using

Basic Local Alignment Search Tool (<http://www.ncbi.nlm.nih.gov/blast/>)

Instructions

You are going to run a total of 5 BLAST searches during this exercise. Please follow the instructions carefully to insure that you get the necessary outputs. After you finish these exercises, you can get more information on the BLAST search algorithm by following the link to BLAST course or BLAST Tutorial in the left margin of the main BLAST web page. The questions embedded in the instructions are simply for you to think about while you are doing the exercises. Additional questions at the end are intended for formal responses.

1. Go to the NCBI BLAST search web page (<http://www.ncbi.nlm.nih.gov/blast/>). The NCBI is the National Center for Biotechnology Information. There are a lot of useful programs here, but today, we will be using one called Blastp that searches based on a protein sequence.
2. Under Protein-Blast, select Search for short nearly exact matches. Make sure the database is set to nr (stands for non-redundant and should be the default). This setting will search a collection of databases rather than just a single one. You could limit the search to just a single organism, if you wanted. Leave the rest of the settings in their default mode.
3. First BLAST SEARCH: Enter the sequence of peptide 1 in the window labeled search. Be careful not to make typos. Then press BLAST!.
4. A new page will appear. Jot down the ID number. You can use this number to get at your results later if the search takes a long time because the server is busy. Press format results to continue in a new window.
5. The page of formatted results has four components:
 - a) graphic table at the top
 - b) summary of the search results
 - c) detailed set of pairwise alignments between your query (the peptide you entered) and the database subject
 - d) statistics on the search.

Sections a, b and c rehash the same information in the same order, just in different formats and with progressively greater levels of detail. Save a copy of the output file to disk (for each search you do) or you might have to repeat the search later.

6. Look at the output. What do you see? Did the search find a match for your unknown protein? The E-value is the way the computer has scored your search. Is a good match a large or a small number? If it is not immediately obvious, look at the pair-wise alignments. Find two that differ in the number of identical amino acids and compare their E-values.
7. Second BLAST SEARCH: Try peptide 2. Did it find a match this time?
8. Third and fourth BLAST SEARCHES: For these two searches, use the Standard protein-protein BLAST [blastp] rather than the Search for short nearly exact match settings. What is different about the parameters used for these two types of searches? Enter both peptides together (enter them as Peptide 1 <space> Peptide 2 for search 3 and Peptide 2 <space> Peptide 1 for search 4). One or both of these searches should have found an unambiguous match to your unknown enzyme. Did these two searches provide identical results?
9. Fifth BLAST SEARCH: Now that you know what you are looking for when you have a good match, type in either peptide 1 or 2 (preferably one that provided some type of hit in your original BLAST searches) in reverse order and repeat the search. Did you find a match?
10. Write down the name of the protein. If it is cut off in the table, scroll down to the alignments. Also, try to find the Enzyme Classification number (format X.X.X.X) and the Swiss-Prot accession number (format: sp|PXXXXX|). You will want some of these numbers for the future exercises.
11. Use a hyperlink from the BLAST output file to find the full NCBI entry on your enzyme. At the bottom of the entry, you should find the full amino acid sequence of your protein. Print out this entry to turn in with your problem set.

Credits: This assignment is derived from the work of Dr. Andrew Feig from the University of Indiana.

Questions to help you reflect on the exercise and analyze the output files:

- a) What is the name of your unknown protein?
- b) Why are there multiple hits for what appears to be the same enzyme?
- c) Can you positively identify the organism from which this enzyme derived? If so, what organism is it from? If not, why not?
- d) What are its E.C. number and Swiss-Prot accession number?
- e) When you entered your peptide in backwards, did you get a significant match to anything in the database? What does this teach you about the importance of directionality and orientation in biological macromolecules?
- f) Compare the results from the two searches that used both peptide fragments (Searches 3 and 4). Did you get identical results? This comparison calls into play gaps and gap penalties, but the program has a maximum gap size of 40 amino acids. Explain your results from Searches 3 and 4 in relation to gaps and gap penalties.
- g) Find the peptide fragments you used to identify your unknown protein in the full protein sequence. Highlight them on the print out you are submitting with your problem set. Are the sequences contiguous to one another and does it matter? Please explain.
- h) Think of an example of how you might use a BLAST search for something other than simply identifying a protein for which you have partial sequence information.