Biol 381 - Bioinformatics Lab
Advanced BLAST Lab Handout

So far in this course, you've used blastn and blastp to find homologs of a gene or protein. These BLAST programs are great at finding sequences that are close relatives of your query sequence. They also do a decent job of finding sequences that, while somewhat dissimilar to the query, may still be homologs based on a closer analysis of function, structure or other factors. But since detecting homology isn't an exact science, it's possible that some very distant relatives of your query may still exist - relatives too distant to score well in a blastn or blastp search.

Fortunately, tools and techniques more sensitive than standard BLAST are available, making it possible to infer homology in very dissimilar sequences. In today's lab, you'll learn background about some of these tools, such as the hidden Markov model (HMM), a statistical method often used in sequence analysis. Additionally, you'll use tools, conserved domain databases and PSI-BLAST, that make use of this model.

Your write-up for this week will consist of answers to the questions in this handout, as well as the files you generate during the exercise.

Part 1: UGENE BLAST and Conserved Domain Workflow

In this part of today's exercise, you'll continue working with UGENE. This time, you'll design a workflow that BLASTs a sequence, then searches the BLAST results for conserved domains. Using a workflow will enable you to search for conserved domains in several databases at once.

You'll need the following elements:
A data reader
Data writers (one for each output)
Amino Translation
CD Search

Step 1:
Add an element to the scene to BLAST an mRNA sequence. Set the input for this element to a file containing your sequence, or use "Read from Remote Database" to search for it by accession number.

Step 2:
Add an element to translate the mRNA into its amino acid sequence.

Step 3:
Add an element to perform a conserved domain search of three databases. Add a new CD Search element to the scene for each unique database search, and choose the search database from the "Database" parameter drop-down menu.

Note: make sure the "local search" parameter is set to False for each CD Search element!

Step 4:
Finally, write the results of each CD search to a separate file. Connect the elements in a logical order, then validate and run the workflow.

Step 5:
Save your schema by going to Actions → Save Schema As. Submit the .uwl file along with your lab write-up.

Part 2: HMM Background/Pfam Search

In this section, you'll search for conserved domains in your protein. Since conserved-domain search tools often use HMM-based methods to determine homology, you'll first learn some HMM background to get a sense of how these searches work.. By the way, many of the protein secondary structure prediction algorithms we' recently covered in class use HMMs to predict alpha helices and beta-sheets.

Step 1:
Read the paper Eddy2004 for background information on hidden Markov models. Discuss the paper with your group, and answer the following questions.

1. What are two issues in gene finding that HMMs can help solve?

2. What is an example of biological analysis that wouldn't work well with HMM? Why wouldn't it work?

3. HMM can be used for generating sequence alignments. What do you think would represent the states for such a model?

4. Define state path, emission probability, and transition probability in your own words. (You can come up with your definitions as a group.)

5. What calculation can you perform to determine the confidence level in a HMM consensus, such as the position of the 5'-splice site in a DNA sequence?

Step 2:
Go to http://pfam.sanger.ac.uk/search/sequence to search the Pfam conserved domain database for your protein sequence.

Step 3:
Look at the list of conserved domain hits, and make sure that most matches have reasonable E-values. If you got a lot of results with high E-values, you can lower the E-value cutoff and search again.

Step 4:
Click the "Show" button to view the alignment of your sequence with the domain HMM.

1. What do the capital letters in the HMM consensus represent?

2. What is "posterior probability"?

For each conserved domain:

1. What is the domain, and what is its E-value? What is the confidence level of the

alignment?

2. Does the domain have a known function? If so, what is it?

3. What are 3 other proteins that contain this domain, and what are their functions?

Part 3: PSI-BLAST

PSI-BLAST (Position-Specific Iterated BLAST) is a specialized local alignment tool, optimized for finding very distant relatives of a protein sequence. It uses a position-specific scoring matrix (like the one shown in Figure 1) to search for distant homologs. In this part of today's exercise, you'll use PSI-BLAST to find homologs of your protein that are too distant to be returned by a standard blastp search.

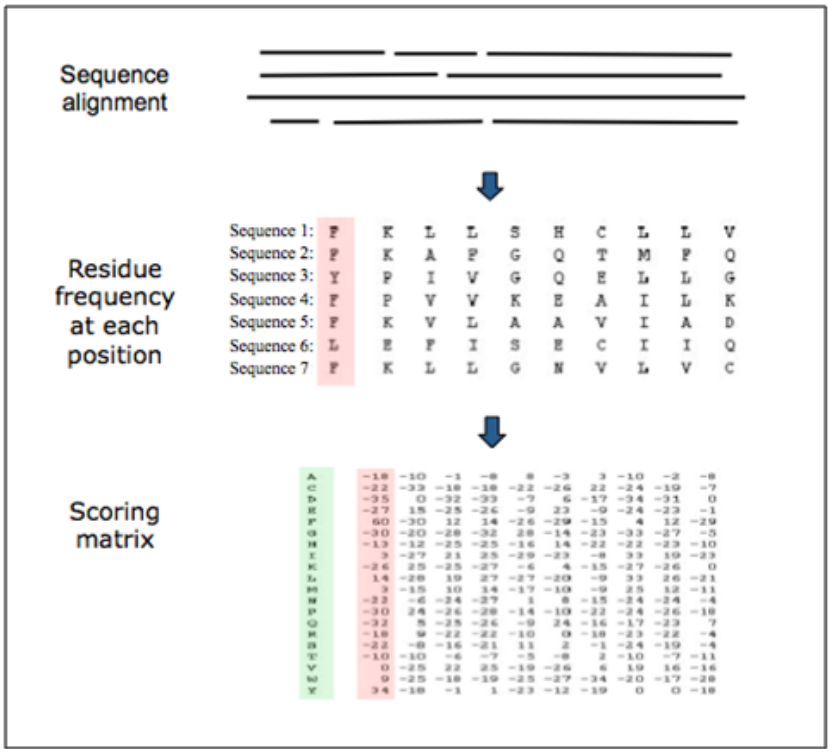

Figure 1: Position-specific scoring matrix example. From http://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/what-are-protein-signatures/signature-types.

Step 1:
First, read through the PSI-BLAST tutorial by Bahgwat and Aravind, then complete Problem 1 under section 2 in the tutorial. Download the human proliferating cell nuclear antigen/*Escherichia coli* DNA polymerase III β-subunit pairwise alignment. Submit the FASTA file with your lab write-up.

Step 3:
Run a blastp search on your protein. In a separate tab, run one PSI-BLAST iteration on your protein. Make sure the maximum number of hits is the same for both searches.

1. Are the results similar for the two BLASTs?

Look at the 3 most distant hits from each search.

1. What is the accession number and E-value for each hit? Are the 3 most distant hits the same for both searches?

Step 4:
Run a second iteration, and find 3 new hits from this iteration (highlighted in yellow). For each of the 3 hits:

1. What is the protein name and accession number?

2. What is the E-value and percent identity?

3. Run a CDD search. Are the conserved domains the same as those of your protein?

4. Based on the results of the above, do you think this hit is homologous to your protein?

Step 5:
Run 3 more iterations. Find 3 new hits from the last (fifth) iteration, and answer Questions 1-4 from Step 4 for each hit.

Step 6:
Run PSI-BLAST on a query you're pretty sure is homologous to your protein - a good bet is to choose one in the same protein family.

1. Why might a PSI-BLAST on a known homolog be a helpful way to analyze very distant relationships? (Hint: Reread section 4 in the PSI-BLAST tutorial.)

2. After 4 iterations, do any of the proteins you analyzed above also appear in the results of the current search?

3. Are any results that *don't* appear in the results of your first search likely homologs? If so, which ones? Explain why you think they are homologs.