Tutorial: Using PSI-BLAST to find remote protein homologues

Background:

PSI-BLAST stands for Position-Specific-Iterated BLAST. Position specific iterative BLAST (PSI-BLAST) is a feature of BLAST 2.0 in which a profile (or position specific scoring matrix, PSSM) is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search. The PSSM is generated from the calculated position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform more BLAST searches and the results of each "iteration" are used to refine the profile. This iterative searching strategy results in increased sensitivity. The PSSM can be saved and used to search other databases for a more sensitive remote homolog detection method.

In this example, we will use a protein from the *Cryptoccocus neoformans* genome that has not been characterized. Run a BLASTP search with this protein against the RefSeq proteins, with an E-value cutoff of 0.01. The top 2 matches are to uncharacterized C. neoformans proteins followed by 2 poor matches to another fungal protein.

The goals of this tutorial are:

1. **Save BLAST results in table format so you can compare them between runs**
2. **Execute a PSI-BLAST search & save the PSSM file**
3. **Use that PSSM to search a different database**

Goal 1: Downloading BLAST results in a table format that can be imported into Excel.

- Download the crypto_protein3 from the Exercise 4 homepage.
- Open the BLAST homepage and use the crypto_protein3 sequence as a query against the REFSEQ protein sequences database, with no limits.
- After the results come back, click on the Download link at the top of the page. This should open a menu that will let you download the results in various formats, as shown in Figure 1.

Figure 1: Download options for BLAST results.

- Click on the Hit Table(text) link and save the file to your computer WITH A DESCRIPTIVE NAME such as Crypto_Refseq_hits.txt
- Open Excel and use the File→open command to open the text file you just saved.
- When you open this file in Excel, it should trigger the Text Import Wizard. Make sure the button "Delimited" is checked and then click Next > button at Step 1.
- At Step 2, make sure there is a check-mark to the box to the left of Tab, add a check mark to the left of Comma, then click the Next > button
- At Step 3, click the Finish button and the results should be in an Excel worksheet.

- Note that the 6th row contains the column headings that we want to keep and have just above the data.
- Delete the first 5 rows and then delete the row below the column headings.
- If you expand the columns your data should have a header telling you what is in each row. Save as an Excel file for later use.

Goal 2: Execute a PSI-BLAST search

- Go back to the BLAST homepage select **protein** blast and copy the crypto_protein3 sequence into the text window.
- Choose **NR** protein sequences as the database
- Choose **PSI-BLAST** as the algorithm
- Click the BLAST button. When the first iteration is done, you should have 3 hits with an E-value better than the threshold of 1.
- Under the **Descriptions** section, there is an option that reads "Run PSI-BLAST iteration 2 with max 500" Click the Go button to the right of this option.
- After the 2nd iteration is done, you should see more sequences (~10) listed as producing significant alignments.
- Click on Go button to the right of the Run PSI-BLAST iteration 3.
- I got 395 hits on the 3rd iteration.
- Once it is finished, click on the Download link at the top of the results. One of the options should now read **PssmWithParameters** (ASN.1). Click on this link and save the file to your harddrive with a descriptive name (*i.e.* CryptoProtein3_PSSM3.asn). You may not have the option to give it a name when downloading, so rename it after it has been downloaded.

GOAL 3: Search a new database with the PSSM file.

Now we want to see if we can find remote homologues that are annotated based on this PSSM. Remember that a PSSM represents the most conserved parts of a group of related proteins. It contains a great deal more information about which residues are important (most conserved) and which are less important in the alignment. Thus, we should be able to identify many more proteins using a PSSM than using a single protein sequence.

- Return to the BLAST homepage and click on protein blast again.
- Leave the query text box blank.
- Type in CryptoPSSM into the title box
- Choose **Refseq** as the database
- Further down, select **PSI-BLAST** as the algorithm
- Under the BLAST button, click on the Algorithm parameters link and scroll down to the last section with the header PSI/PHI BLAST.
- You are going to upload the PSSM you just created (and downloaded). Using the Browse button, select the CryptoPSSM.asn file by double-clicking it.
- Execute the PSI-BLAST search by clicking the BLAST button.

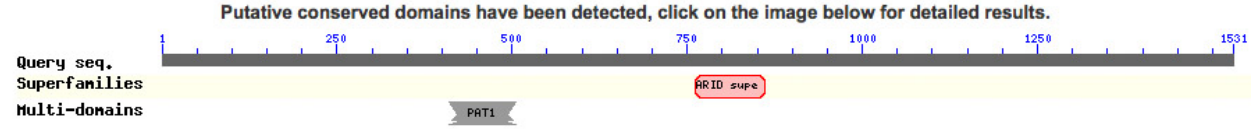
Compare the results from this blast to the first BLASTP you did of the REFSEQ database using just the crypto_protein3 sequence. You should notice that there are many more hits and that the E-values of the top hits are significantly lower than before. Follow the links for the top 1 or 2 hits. What conserved domains do they have? Are they the same as the conserved domains in the crypto protein shown in the figure below?

Try a BL2SEQ between the crypto protein and the top hit from this search that is NOT from *Cryptococcus neoformans*. Change the matrix to blosum45. Could any similarity be detected?

Submit the sequence of the top match that is not a Cryptococcus protein to the Conserved domain search engine. Does it share any domains with the crypto protein?

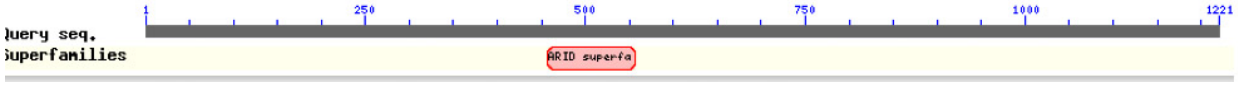
While this method would not be used for genome-wide annotation, it is useful for identifying possible functions for otherwise uncharacterized proteins, identifying remote homologues in other species, and providing evolutionary insights into the relationships between proteins.

Here were the conserved domains identified in Crypto_protein 3:



Here are the conserved domains in the top non-crypto match from the search with the PSSM:

XP_001880760 (predicted protein from Laccaria bicolor):



They do share one of the two domains found in the Crypto protein. It's possible they perform similar functions in their respective organisms.