UTM UNIVERSITI TEKNOLOGI MALAYSIA
SCHOOL OF COMPUTING
Faculty of Engineering

Semester II 2021/2022

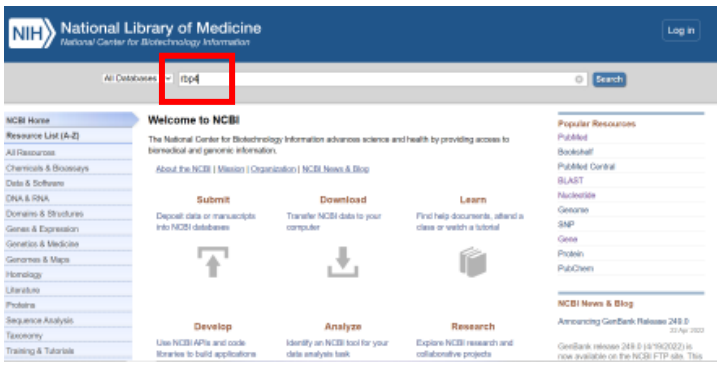
Subject : **Bioinformatics I (SCSB2103)**
Section : **01 – Dr Haslina Hashim**
Topic : **Lab 05 – Advanced Database Searching**

Name : **Gui Yu Xuan**
Matric ID : **A20EC0039**

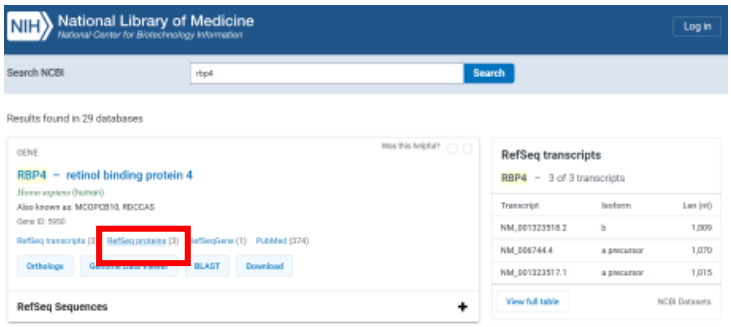
1) Create an artificial protein sequence consisting of human RBP4 followed by the C2 domain of human protein kinase Cα. An example of this is shown in Web Document 5.5 (available in your e-Learning site). Enter this combined sequence into a PSI-BLAST search

   i) In general, are multiple domains always detected by the PSI-BLAST program?

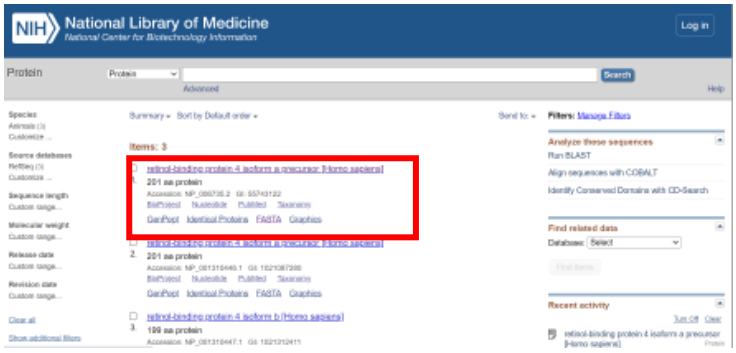
   ii) Do any naturally occurring proteins have both lipocalin and C2 domains?



Go to NCBI website and type rbp4 on the search bar.



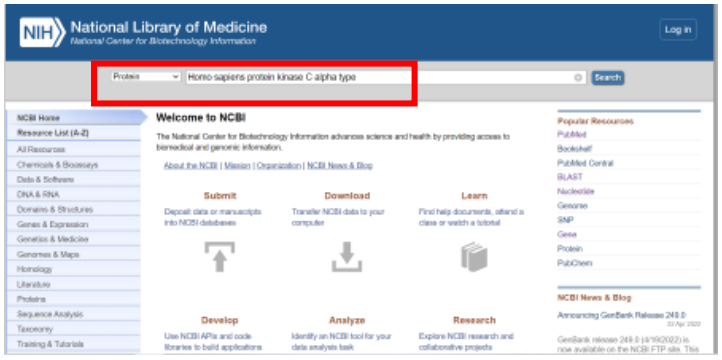
Then, click on the RefSeq proteins.



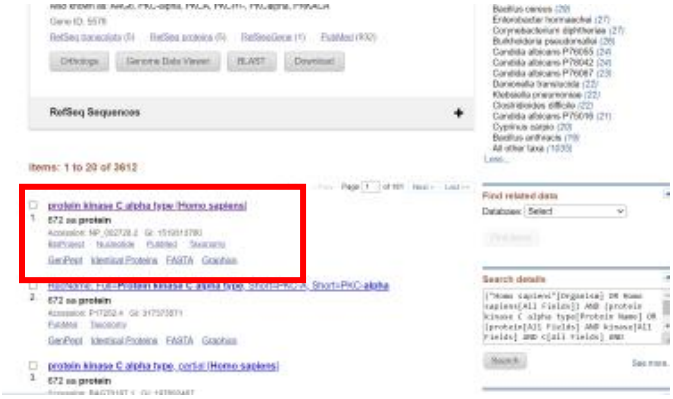
Select the first one.

Click on the FASTA.

Then, the results of the NCBI reference sequence of the RBP4 is shown.

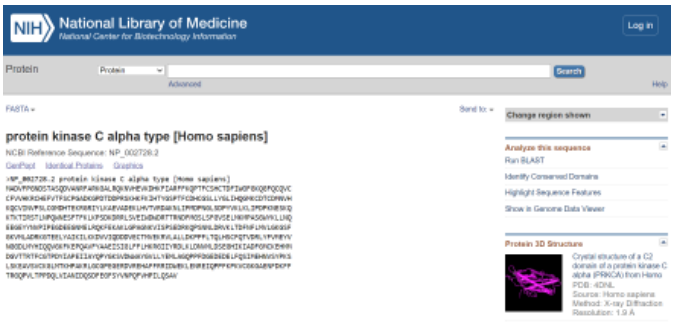


Then, go back to the NCBI home page. Select protein and type Homo sapiens protein kinase C alpha type on the search bar.

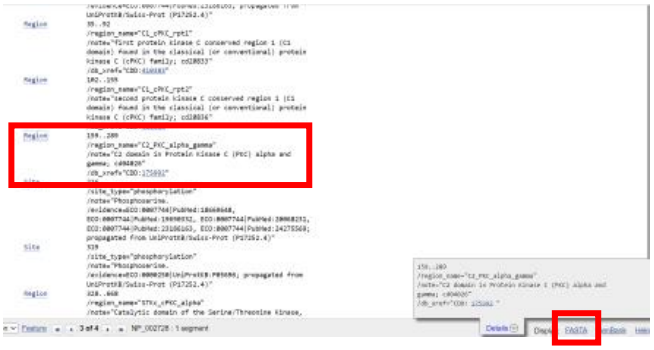


Select the first one.

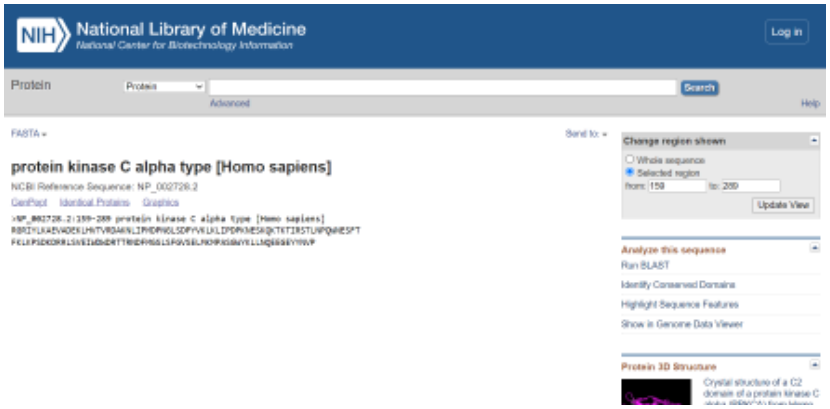
Click on the FASTA link.



The results of the reference sequence of Protein Kinase C alpha type is shown.



Click on the region with name C2_PKC_alpha_gamma and click on the FASTA link at the bottom.



You will get the results.

Here is RBP4, obtained via NCBI Gene then RefSeq:

```
>NP_006735.2 retinol-binding protein 4 isoform a precursor [Homo sapiens]
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
```

Here is the entire PKCA protein: C2 domain

```
>NP_002728.2 protein kinase C alpha type [Homo sapiens]
MADVFPGNDSTASQDVANRFARKGALRQKNVHEVKDHKFIARFFKQPTFCSHCTDFIWGFGKQGFQCQVC
CFVVHKRCHEFVTFSCPGADKGPDTDDPRSKHKFKIHTYGSPTFCDHCGSLLYGLIHQGMKCDTCDMNVH
KQCVINVPSLCGMDHTEKRGRIYLKAEVADEKLHVTVRDAKNLIPMDPNGLSDPYVKLKLIPDPKNESKQ
KTKTIRSTLNPQWNESFTFKLKPSDKDRRLSVEIWDWDRTTRNDFMGSLSFGVSELMKMPASGWYKLLNQ
EEGEYYNVPIPEGDEEGNMELRQKFEKAKLGPAGNKVISPSEDRKQPSNNLDRVKLTDFNFLMVLGKGSF
GKVMLADRKGTEELYAIKILKKDVVIQDDDVECTMVEKRVLALLDKPPFLTQLHSCFQTVDRLYFVMEYV
NGGDLMYHIQQVGKFKEPQAVFYAAEISIGLFFLHKRGIIYRDLKLDNVMLDSEGHIKIADFGMCKEHMM
DGVTTRTFCGTPDYIAPEIIAYQPYGKSVDWWAYGVLLYEMLAGQPPFDGEDEDELFQSIMEHNVSYPKS
LSKEAVSVCKGLMTKHPAKRLGCGPEGERDVREHAFFRRIDWEKLENREIQPPFKPKVCGKGAENFDKFF
TRGQPVLTPPDQLVIANIDQSDFEGFSYVNPQFVHPILQSAV
```

Here is the C2 domain of PKCA, obtained by clicking the C2 “region” link from the NCBI Protein page (http://www.ncbi.nlm.nih.gov/protein/NP_002728.1):

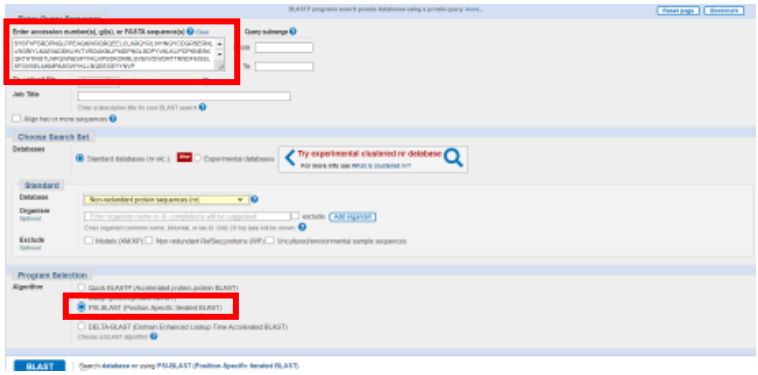
```
>NP_002728.2:159-289 protein kinase C alpha type [Homo sapiens]
RGRIYLKAEVADEKLHVTVRDAKNLIPMDPNGLSDPYVKLKLIPDPKNESKQKTKTIRSTLNPQWNESFT
FKLKPSDKDRRLSVEIWDWDRTTRNDFMGSLSFGVSELMKMPASGWYKLLNQEEGEYYNVP
```

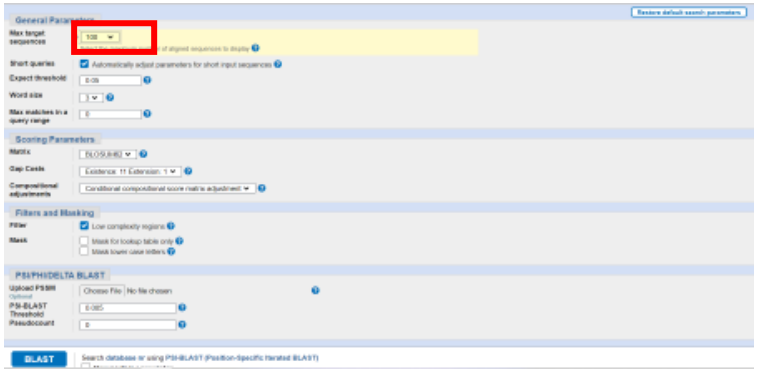
By combining the RBP4 reference sequence and the C2 domain of PKCA obtained by clicking C2 “region” link. Here is a chimeric protein:

```
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDET
GQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQY
SCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLLRGR
IYLKAEVADEKLHVTVRDAKNLIPMDPNGLSDPYVKLKLIPDPKNESKQKTKTIRSTLNPQWNESFTF
KLKPSDKDRRLSVEIWDWDRTTRNDFMGSLSFGVSELMKMPASGWYKLLNQEEGEYYNVP
```

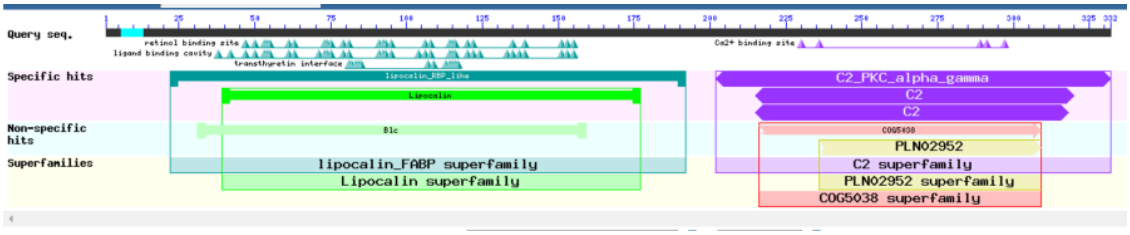
Click on the BLAST on the NCBI Home page.

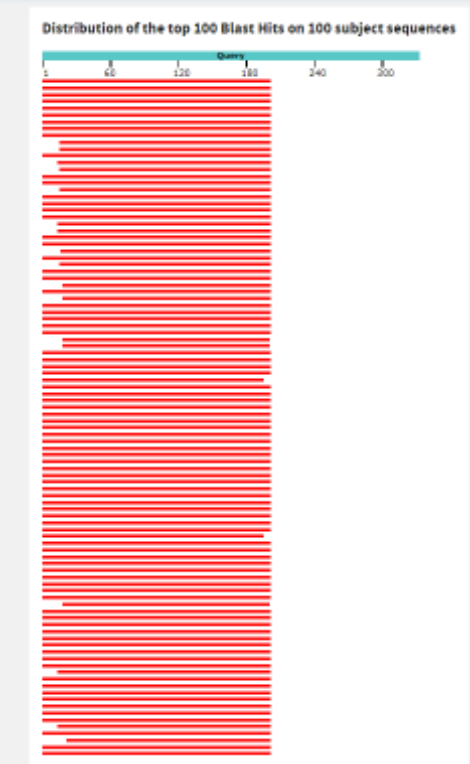
Click on the Protein BLAST.





Copy the chimeric protein and paste it on the query sequence. Change the algorithm to the PSI-BLAST under program selection and change the max target sequences to 100 under general parameters. Then, click on the BLAST button.





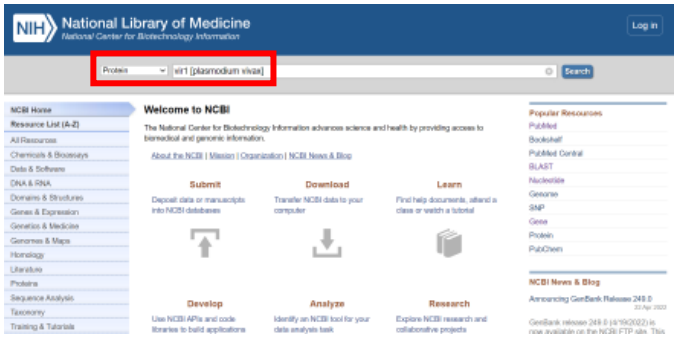
Above are the results. From the graphical summary, we can see that there are some matches to either RBP4 or PKC but not both. The RMP4 portion of the protein is larger so it accumulates higher scores (and lower E values) and those results are listed first. By inspection, there are no proteins with both domains.

2) The purpose of this problem is to compare BLASTP to DELTA-BLAST

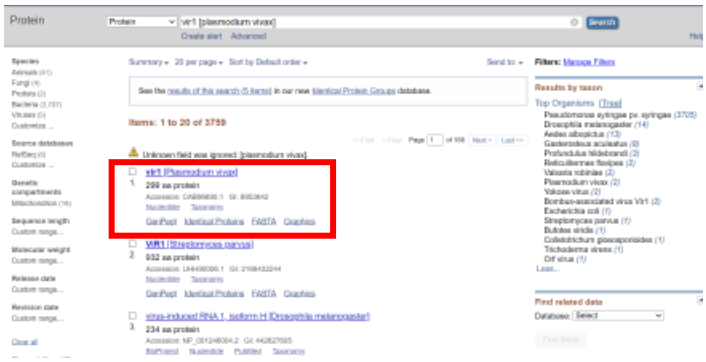
The malarium parasite Plasmodium vivax has a multigene family called vir that is specific to that organism (del Portillo et al., 2001). There are 600–1000 copies of these genes, and they may have a role in causing chronic infection through antigenic variation.

i) Select vir1 and perform a BLASTP search of the nonredundant protein database (restricting the species to Plasmodium vivax).

ii) Then perform a DELTA-BLAST search with the same entry. For each search, approximately how many proteins have an E value less than 1e-10?

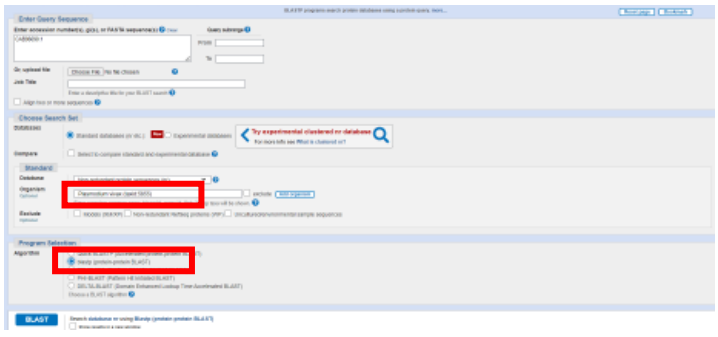


On the NCBI Home page, select protein and type vir1 [plasmodium vivax] on the search bar.

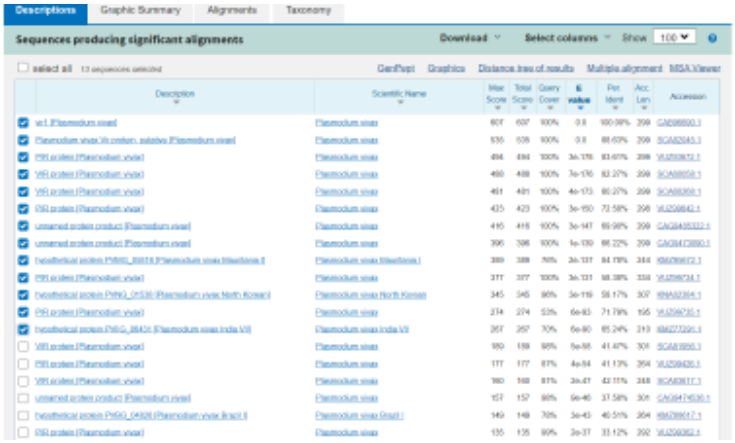


Select the first one.

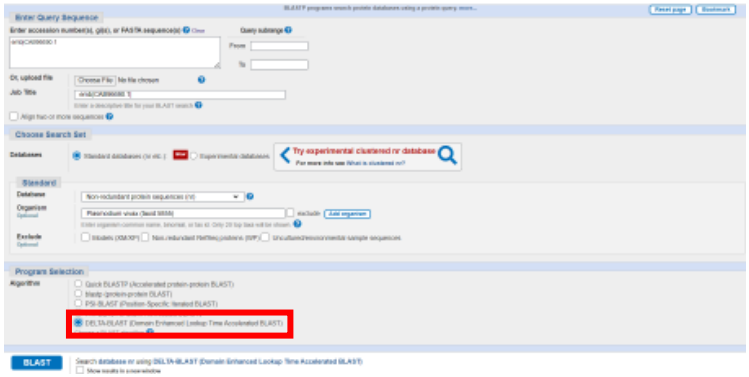
Click on the Run BLAST.



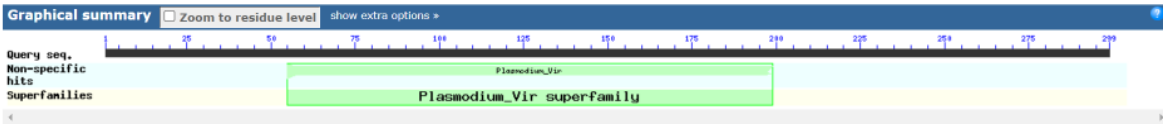
You will be redirected to this page. Type Plasmodium vivax on the organism bar and choose blastp under program selection. Click on the BLAST button and the results is shown as below.

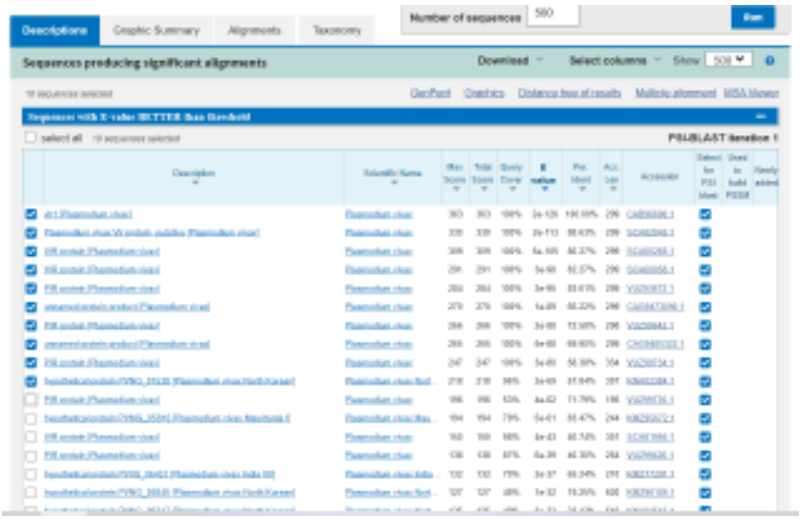


From the results, we can see that there are many matches with significant E values.



Click on the edit search at the top of the page and change the algorithm to DELTA-BLAST. Click on the BLAST button and the result is shown below.





The results of the search are surprising.

BLASTP successfully found a series of matches, including a top 13 (colour coded red) with very low E values form 0.0 to 6e-90.

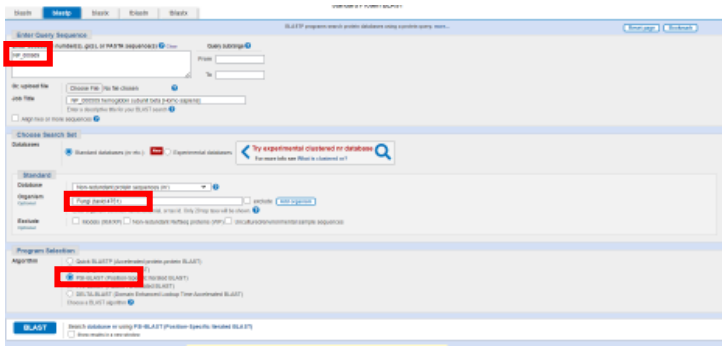
DELTA-BLAST successfully found many more matches by using its PSSM-based approach. However, its best match (2e-126) does not have as low as expect value produced by BLASTP. In many other DELTA-BLAST searches the E values are lower than those found by BLASTP. The results shown here follows from the PSSM approach used by DELTA-BLAST in which a large set of vir1-related proteins were aligned and scored to define the general properties of this family.

3) Are there globins in fungi?

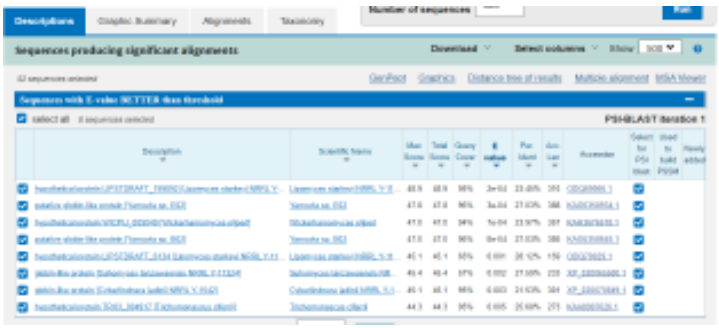
i) Perform a PSI-BLAST search using human beta globin (NP_000509) as a query, restricting the output to sequences from fungi (taxid:4751) in the nr database.

ii) What is the approximate range of lengths of fungal proteins having globin domains? What non-globin domains are often present in fungal globins? Does the presence of these unrelated domains lead to corruption? Why or why not?

iii) In the first iteration there are several hits (with the E values below the 0.005 threshold). After several more iterations there are many dozens of hits including flavohemoproteins that include a globin domain. These fungal proteins have globin domains that are more related to bacterial than vertebrate orthologs. Most of the fungal flavohemoproteins and are quite long (over 400 amino acids and sometimes about 1000 amino acids long), having multiple domains. However, only the globin domain is used for the continued PSI-BLAST iterations.

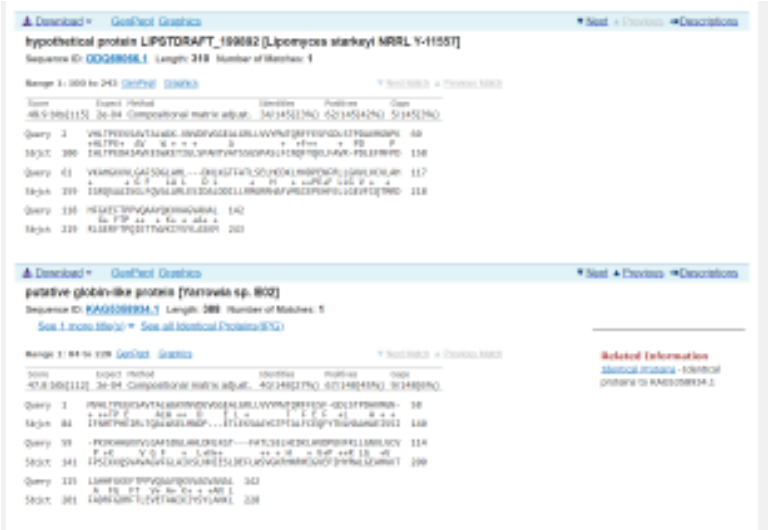


Enter the NP_000509 on the query sequence and type fungi on the organism bar. Change the program selection to the PSI-BLAST.



The description results of the PSI-BLAST sequences with E value BETTER than threshold.

The description results of PSI-BLAST sequences with E value WORSE than threshold.



To see the length of these matches, inspect the pairwise outputs. You can also reformat to view the results as a table.

Most fungal proteins have lengths of 250 to 450 amino acids. Some are more than that.

S globin are often present in fungal proteins. Presence of unrelated domains can lead to corruption which results in the high scores for homologous to the unrelated domains. This corruption can then lead high scores for the assignment and statistical significance to biological-incorrect relationship.

4) Perform HMMER searches

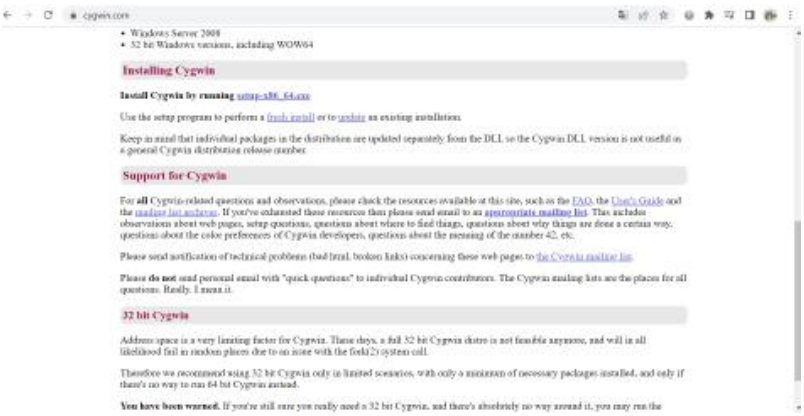
i) First make two different HMMs. You can obtain sets of vertebrate globin and bacterial/fungal/vertebrate globin sequences as web documents 5.6 and 5.7. The multiple sequence alignments that we use as input to HMMER are in these documents.

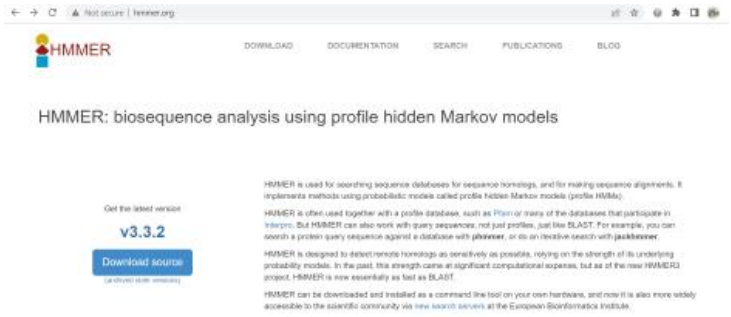
ii) When the profile HMM was built from a multiple sequence alignment of vertebrate alpha and beta globins and used to search the human RefSeq database, there were many database matches, including myoglobin that we could not detect with BLASTP. In contrast, when an alignment of bacterial and fungal globins was used to generate a profile HMM, the output consisted of one result with a non-signficant expect value. Combining several human globins with the bacterial and fungal globins in a multiple sequence alignment resulted in the creation of an HMM that readily detected human globins. Thus, the profile HMM is a model that is sensitive to the choice of sequences that are used as input for the multiple sequence alignment.

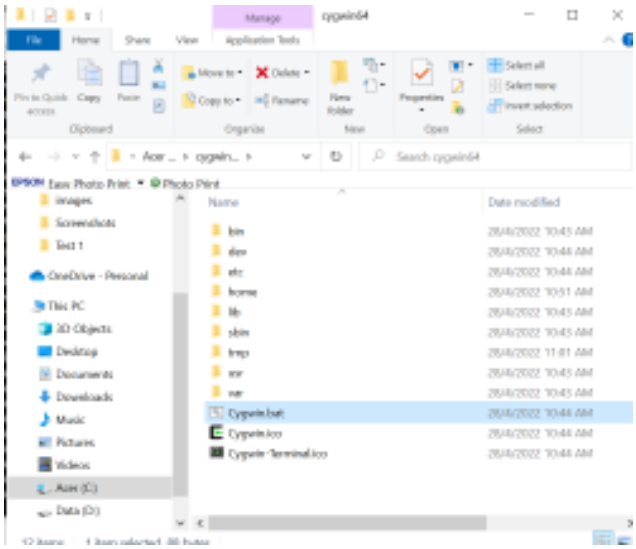
iii) The full results of the HMMER searches for (1) vertebrate, (2) bacterial plus fungal, and (3) bacterial plus fungal plus vertebrate globins are shown in web documents 5.8, 5.9, and 5.10. The HMM match to human myoglobin had a higher score and lower E value in search (3) than in (1).

iv) HMMer searches are run locally. This search was run against all human RefSeq proteins. You can download NCBI databases such as RefSeq by visiting the file transfer protocol (FTP) site from the home page of NCBI or going directly to ►http://www.ncbi.nlm.nih.gov/Ftp/ (WebLink 5.1). Place the downloaded database into the same directory as your input sequences for HMMER.

Follow the YouTube link, https://www.youtube.com/watch?v=MBtbgZ7OmNM install Cygwin and Hmmer.



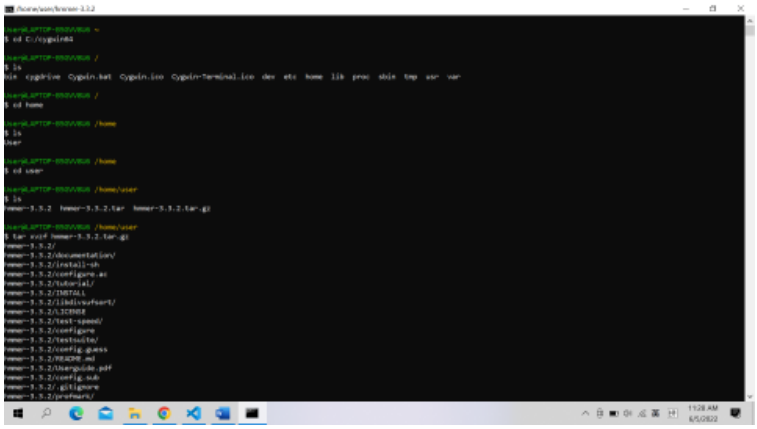


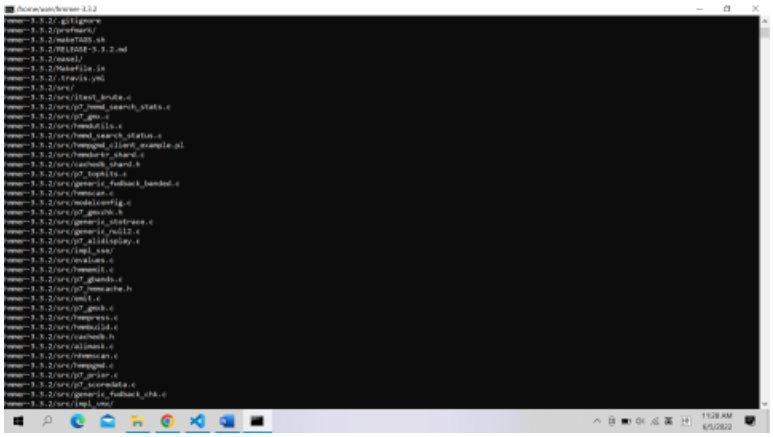


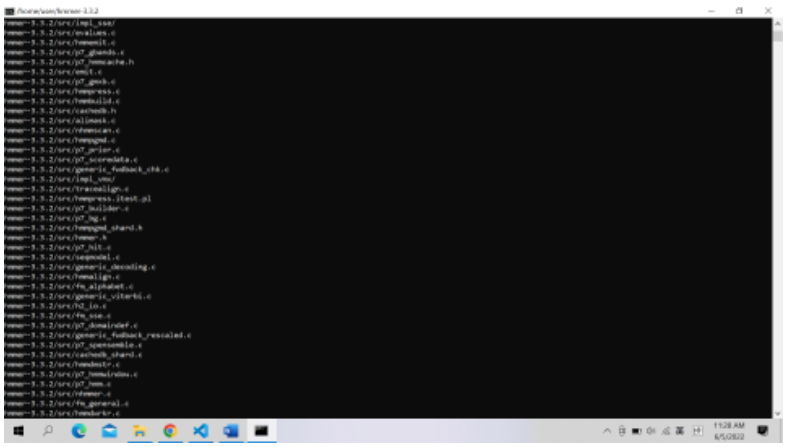
Then, open Cygwin.bat to run the command.

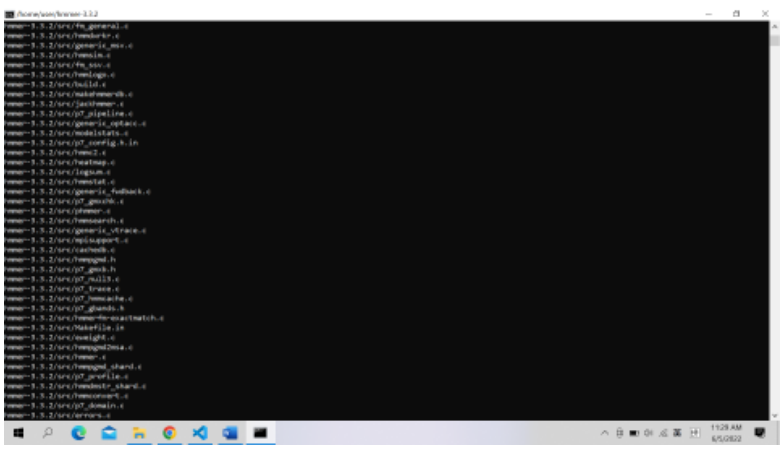
Type:

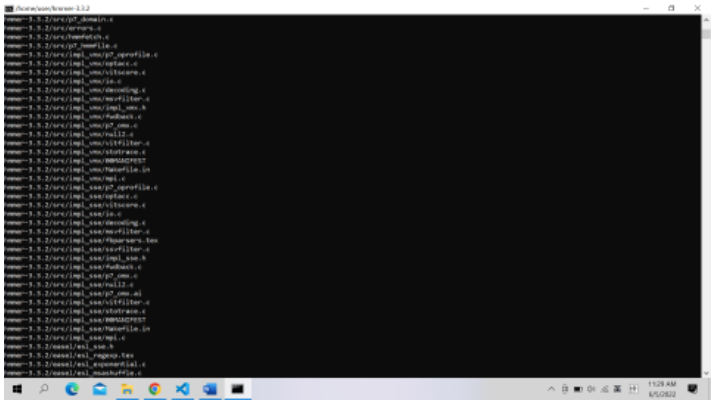
i. cd C:/cygwin64
ii. ls
iii. cd home
iv. ls
v. cd user
vi. ls
vii. tar xvzf hmmer-3.3.2.tar.gz
viii. ls
ix. cd hmmer-3.3.2 (move to new directory)
x. . /configure
xi. Make
xii. make check

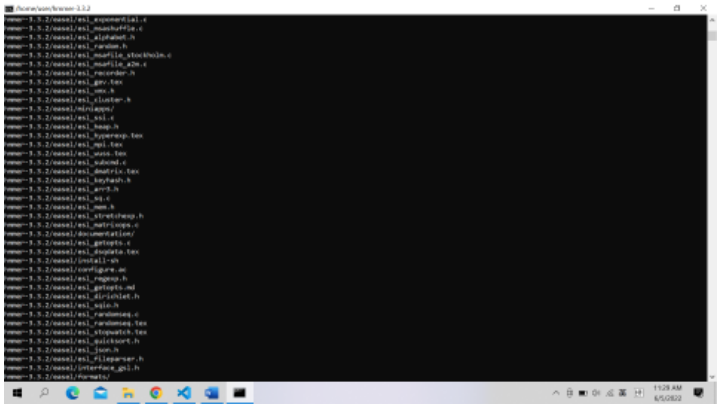


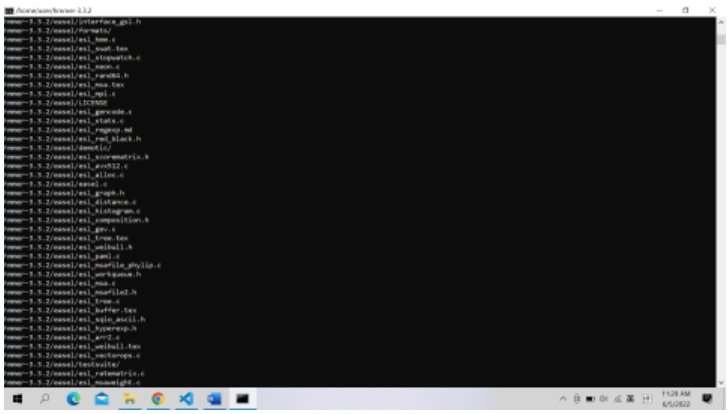


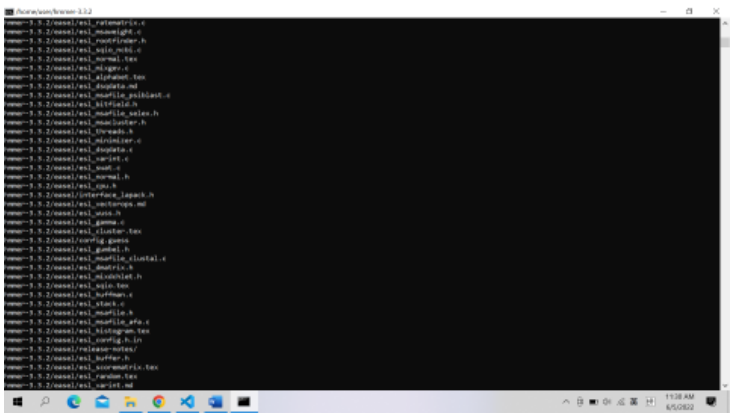


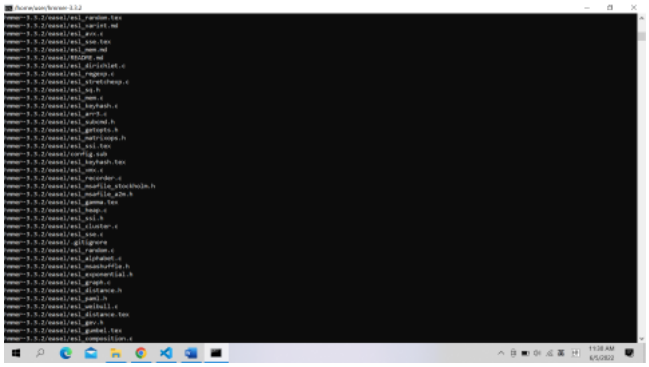


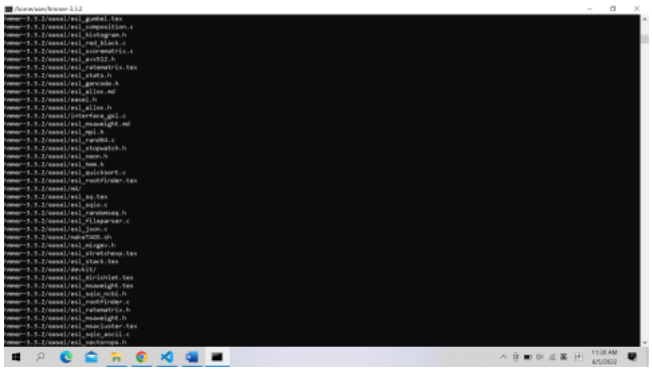


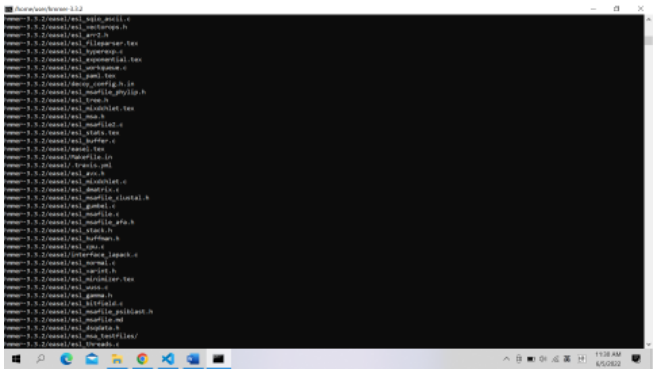


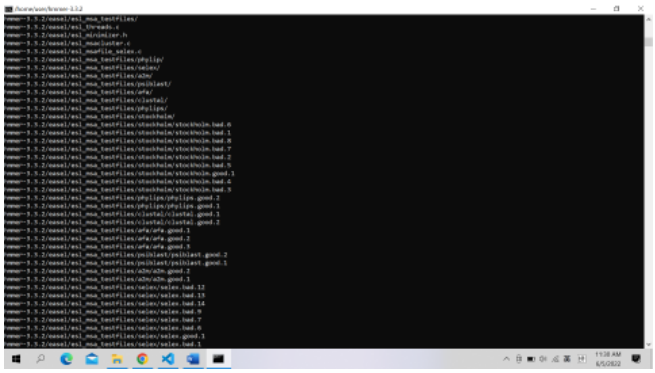


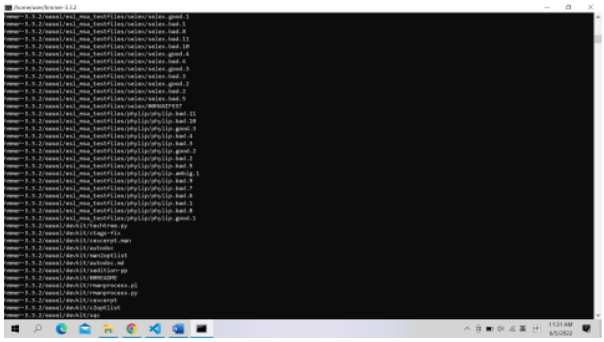


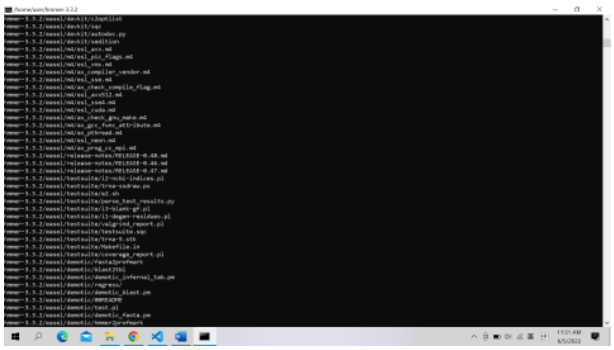


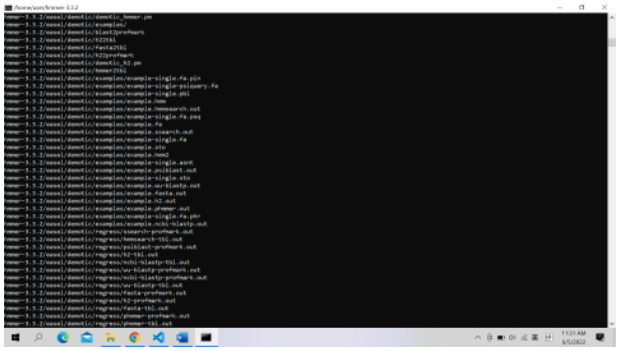


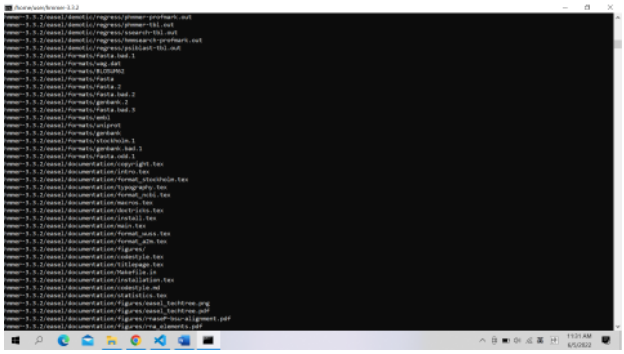


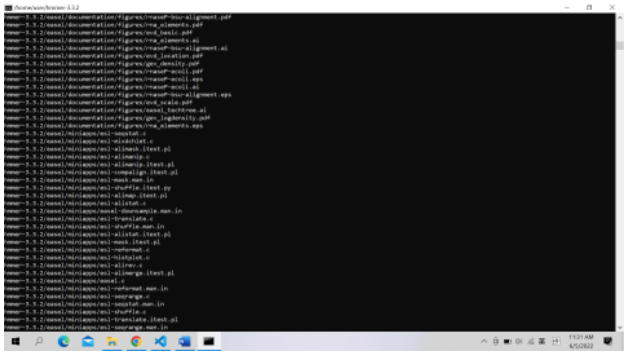


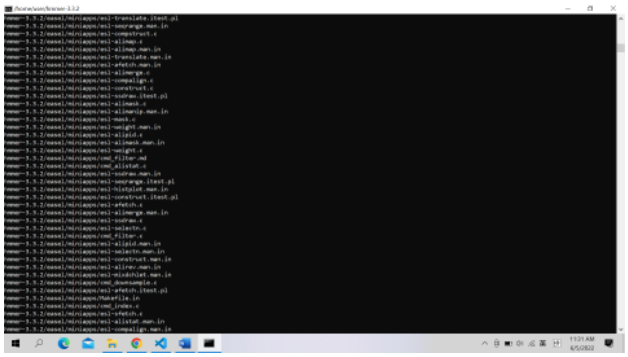


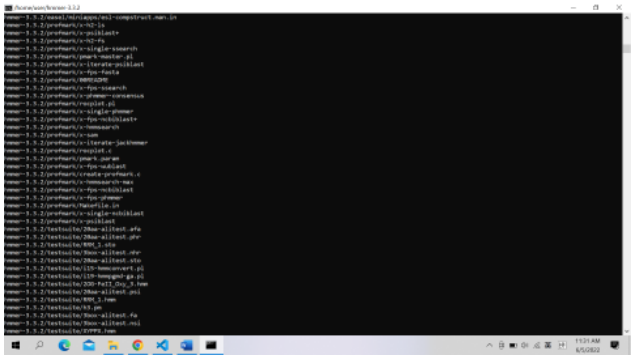


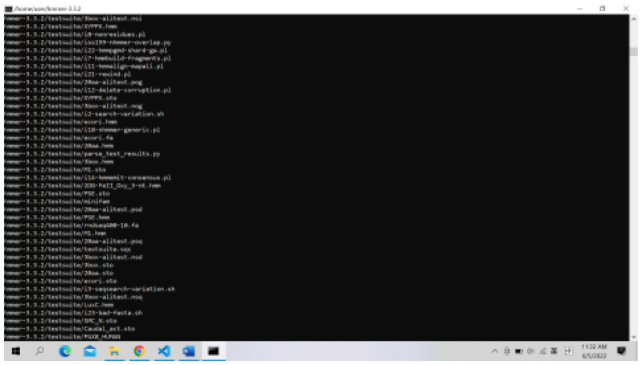


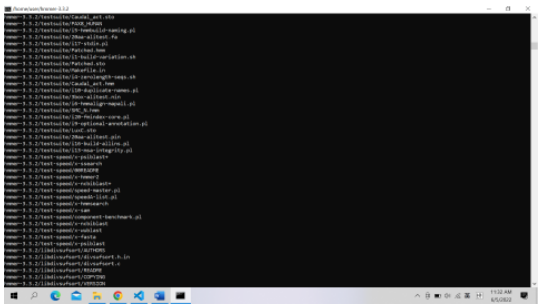


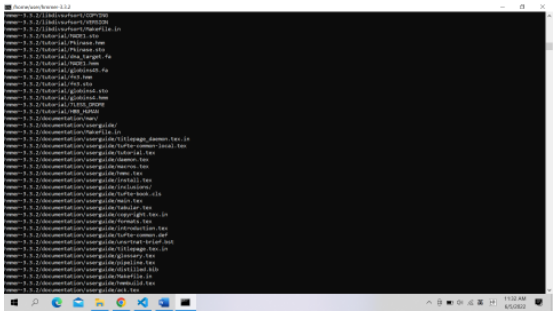


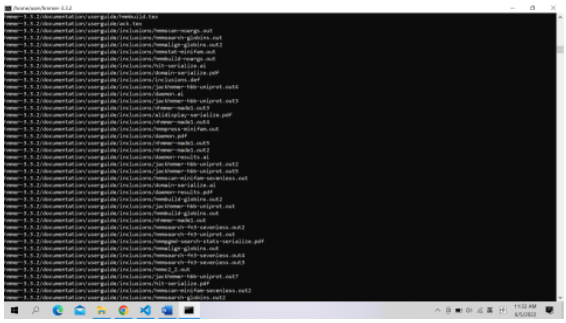


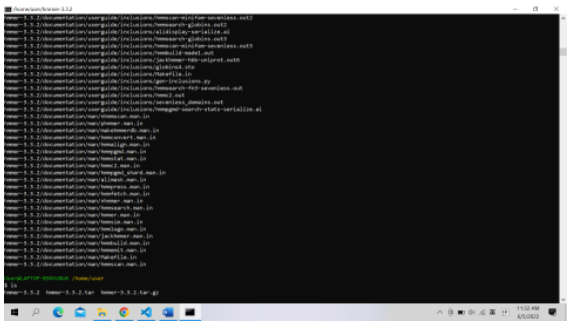


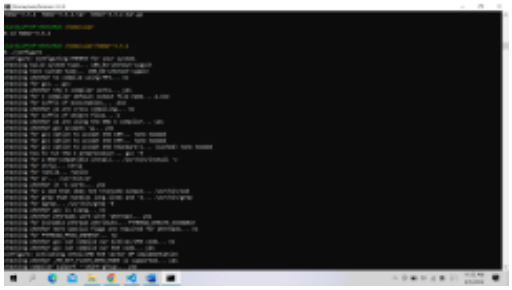


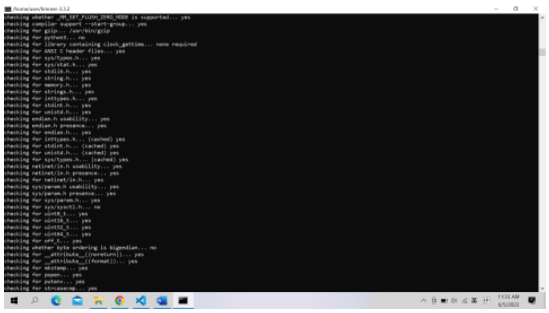


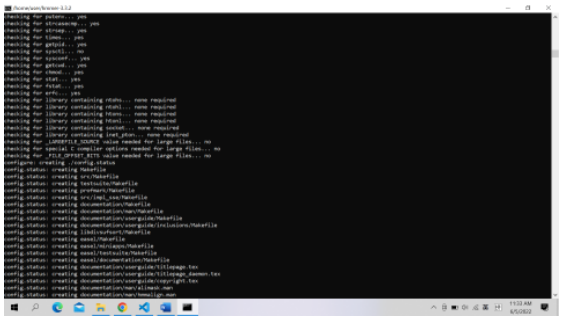


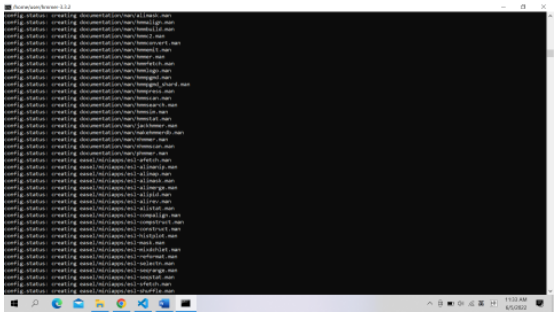


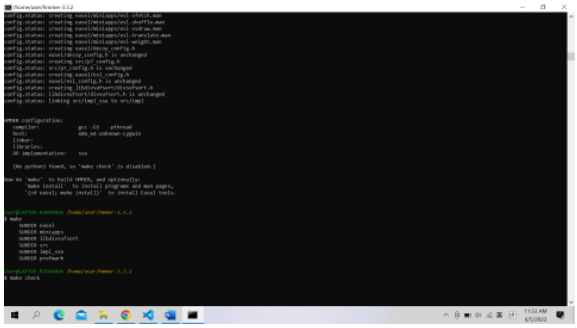












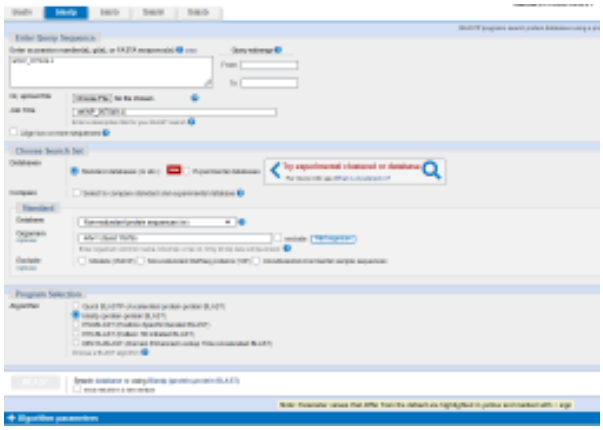
5) We previously performed a series of BLAST searches using HIV-1 pol as a query (NP_057849).

   i) Perform a BLASTP search using this query. Look at the taxonomy report to see which viruses match this query.

   ii) Next, repeat the search using several iterations of PSI-BLAST. Compare this taxonomy report to that of the BLASTP search.

iii) What do you observe? Are there any nonviral sequences detected in the PSI-BLAST search?

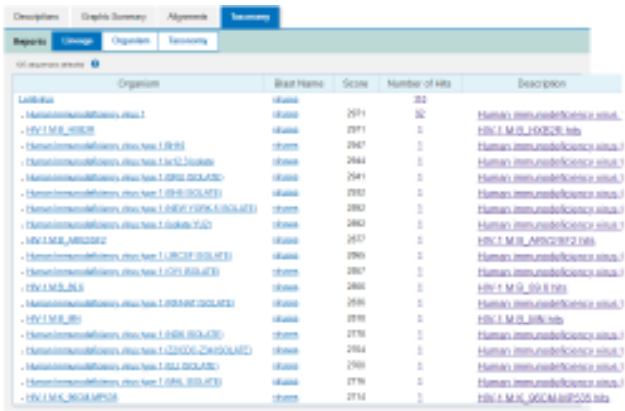
iv) Did you expect to find any?



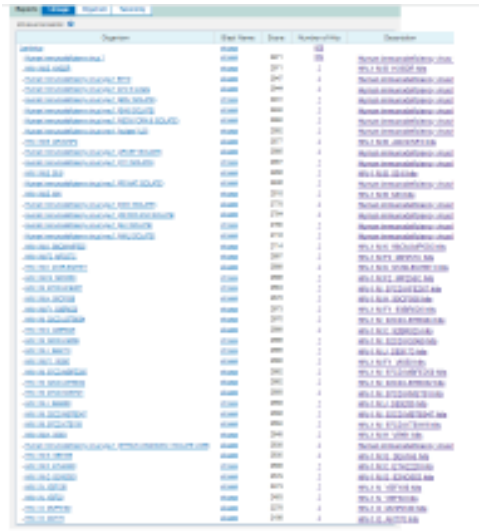
Go to https://www.ncbi.nlm.nih.gov/protein/NP_057849.4 and click on Run BLAST.

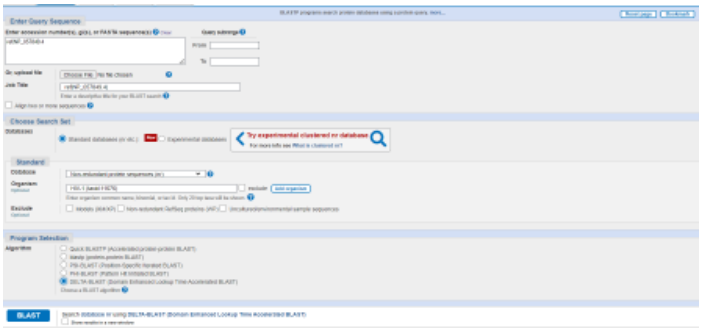
This output shows the lineage and E values (not shown) for different taxonomic groups:

Taxonomy Report for BLASTP:



Taxonomy Report for PSI-BLAST:





Taxonomy Report for DELTA-BLAST:

Based on the outputs, there are more hits with DELTA-BLAST than BLASTP and PSI-BLAST. There are both viral and non-viral detected in the PSI search. This is expected because the PSSM-based approach is far more sensitive than that of standard BLAST.

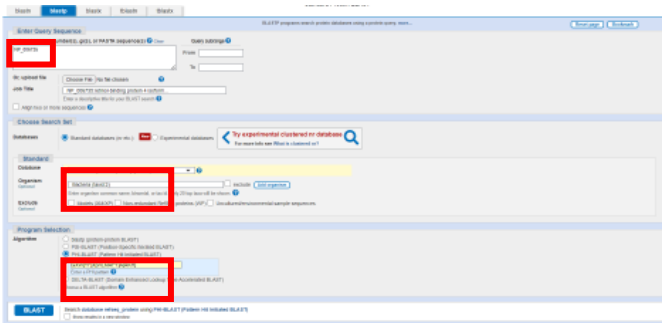
6) Explore PHI-BLAST using human RBP4 (NP_006735) as a query, restricting the output to bacteria and the RefSeq database.

i) Use the PHI pattern GXW[YF]X[VILMAFY]A[RKH]. Perform this search, and save the results.

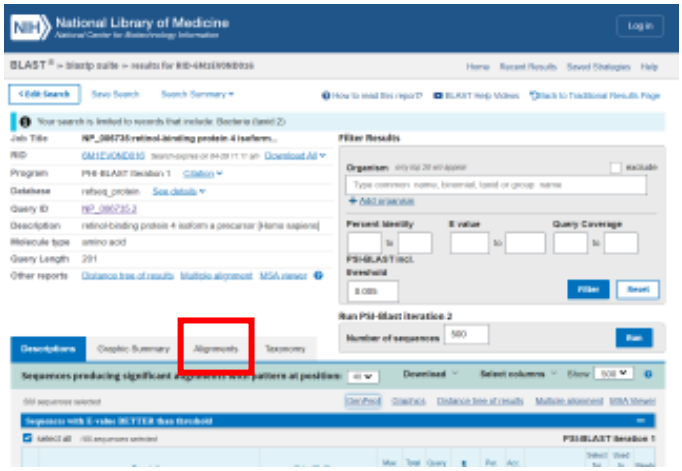
ii) Then repeat the search using the PHI pattern GXW[YF][EA][IVLM].

iii) How do the results differ?

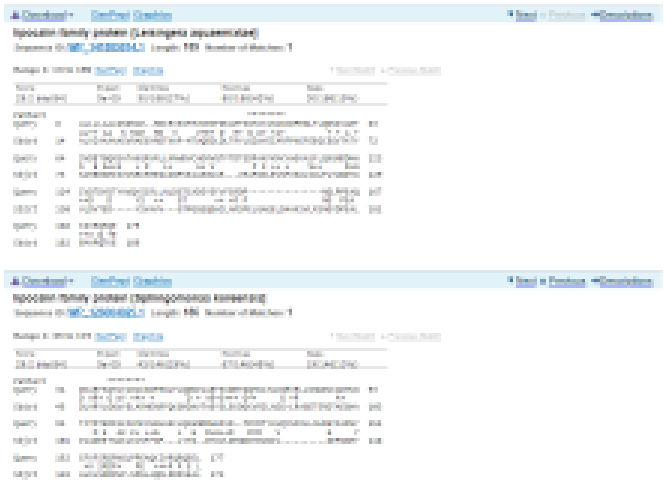
iv) Select one protein that appears as a bacterial protein in a pairwise alignment with the human RBP4 query; what are the E values, and why do they differ?



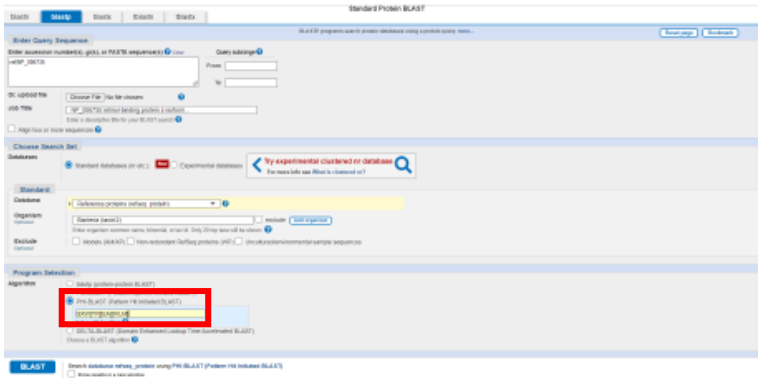
Go to NCBI website to enter the details. After entering the accession number, change the database to the reference proteins and organism to bacteria under Standard section. Then, select PHI-BLAST algorithm and paste the PHI pattern under Program Selection.



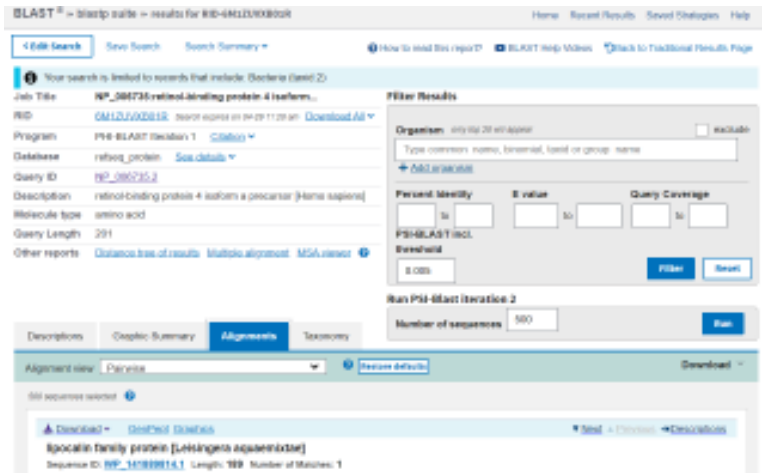
After click on the BLAST button, you will come to this page. Then, click on the alignment.



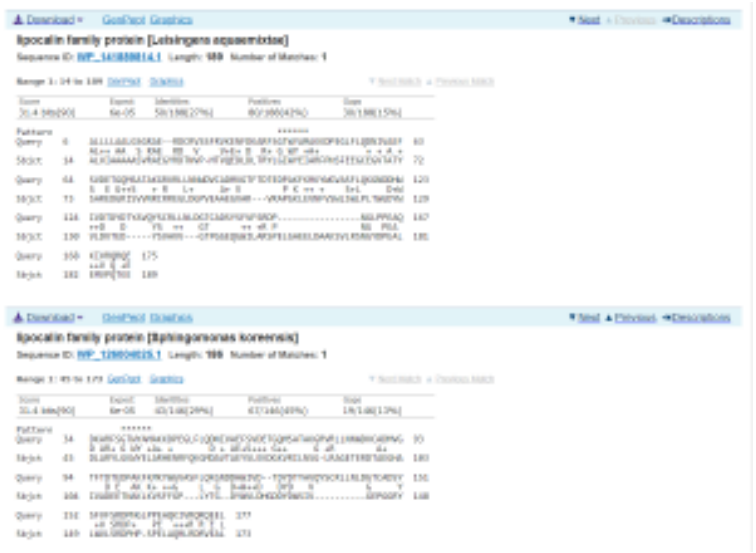
For every database match, the results include asterisks indicating where this pattern matches. Above are the top two results.



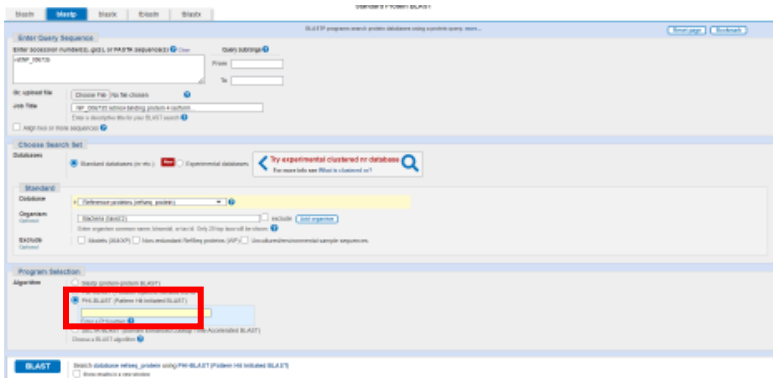
Next, click on edit search, change to the second PHI pattern which is shorter and more restrictive.



Click on the alignment. The results are shown.



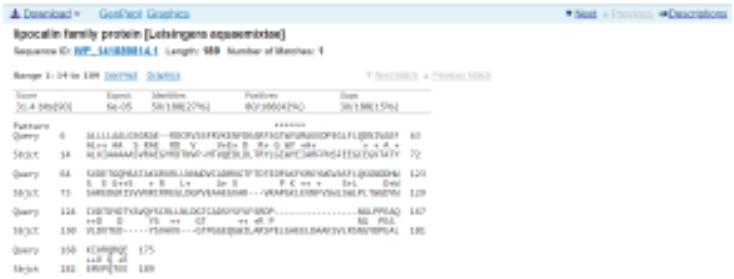
The top two results are identical, except there are 6 asterisks now instead of 8 asterisks previously. Inspecting the results from both searches, the first PHI pattern had more database matches that were significant.



Click on edit search button to back to this page. Empty the PHI pattern and BLAST.

When no PHI pattern is entered, the top 2 results change. The above is the results.

If we instead do a PHI-BLAST search without using a PHI pattern, only Pseudoalteromonas luteoviolacea] is significant.



There are many previously significant, such as Leisingera aquaemixtae, but currently not significant in the absence of PHI pattern. Therefore, a PHI pattern can dramatically improve the sensitivity of a BLAST protein search.