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 04 – BLAST**

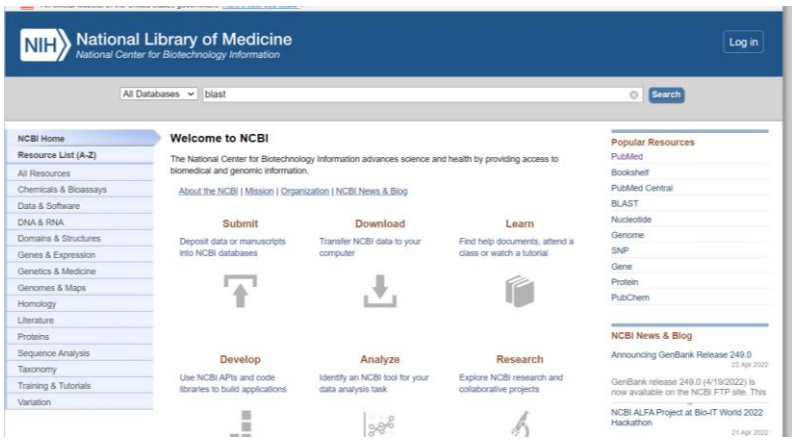
Name : **Gui Yu Xuan**
Matric ID : **A20EC0039**

1) Perform a BLASTP search. In this problem we will explore the effect of a short protein query on the BLASTP parameters

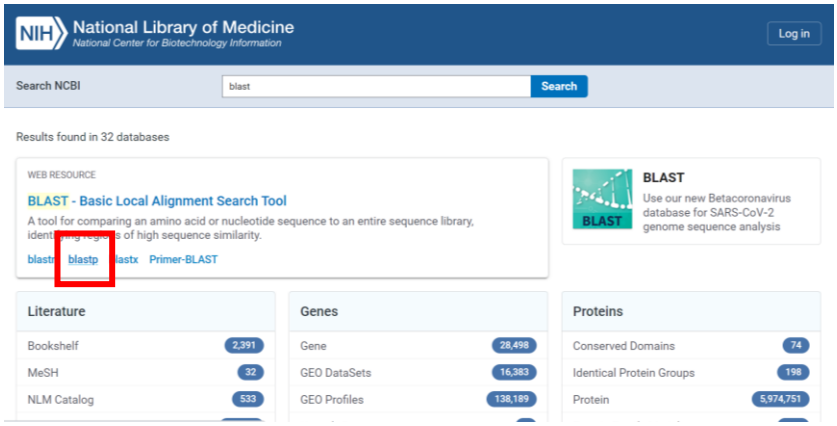
   i) Perform a BLASTP search at NCBI using the following query of just 12 amino acids. By default, the parameters are adjusted for short queries.

      (a) PNLHGLFGRKTG

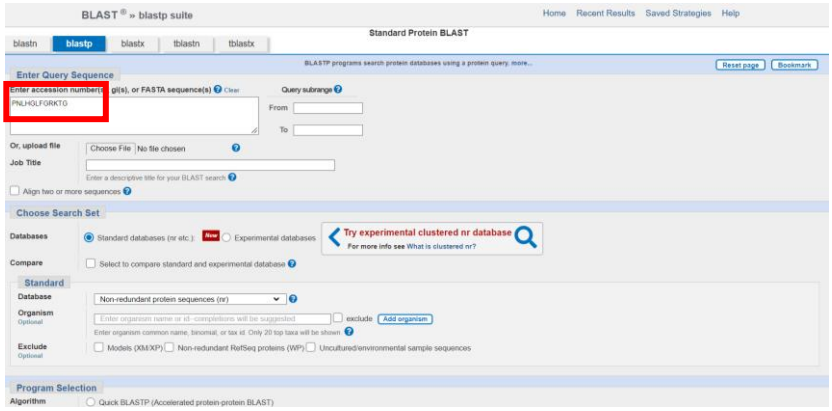
   ii) Inspect the output. What is the E-value cut off? What is the word size? What is the scoring matrix? How do these settings compare to the default parameters?



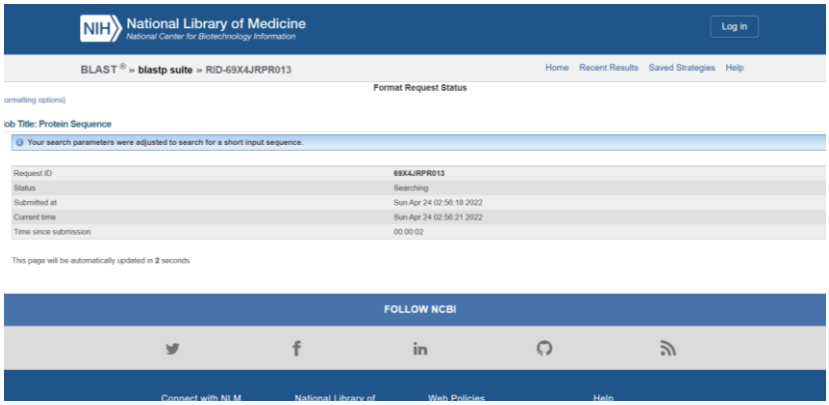
Go to the NCBI website and type “blast” on the search bar.



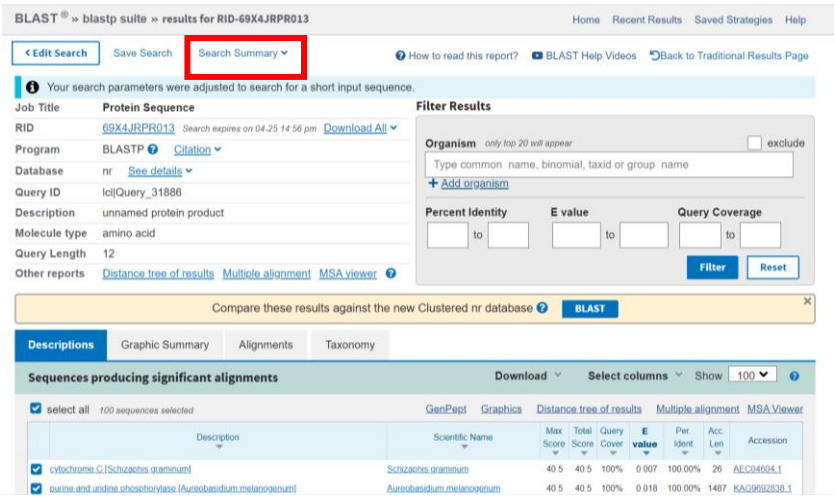
Then, click on the blastp.



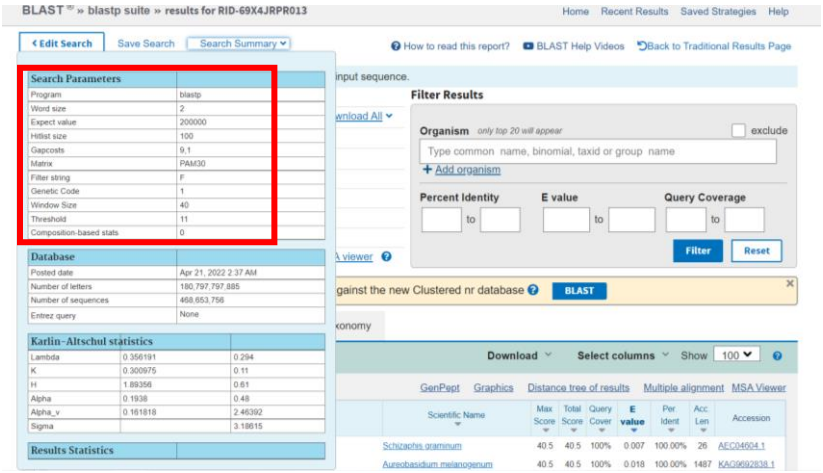
Enter the “PNLHGLFGRKTG” and click on the BLAST button.



You will be redirected to this page. Wait for a while, you will see the page just like the picture below.



Click on "search summary", you will be able to view the results.



Under the search parameters, we can see that the expect value is 200000. Larger value of expect value is because a short query is given, which means that it will be higher probability of occurrence in database. The word size is 2 while the scoring matrix is PAM30. The default expect value is 0, default word size is 3 and the default scoring matrix should be PAM250 or BLOSUM62.

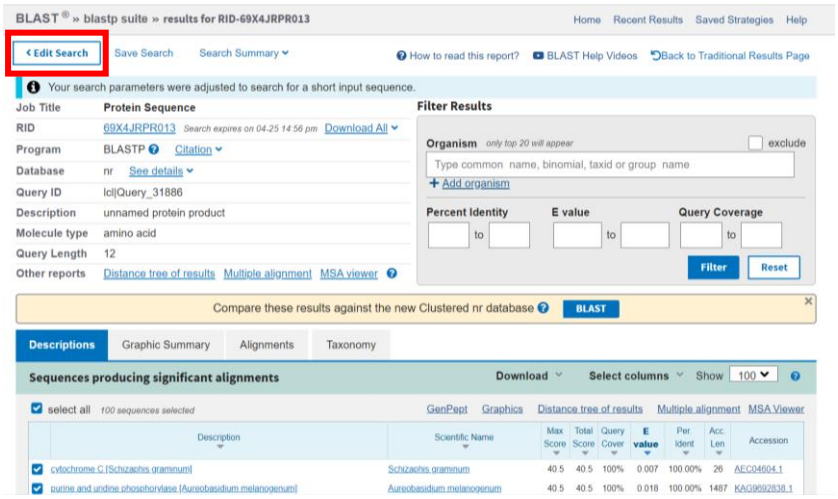
2) Protein searches are usually more informative than DNA searches. Why is this so? Do a BLASTP search

i) Do a BLASTP search using RBP4 (NP_006735), restricting the output to Arthropoda (insects).

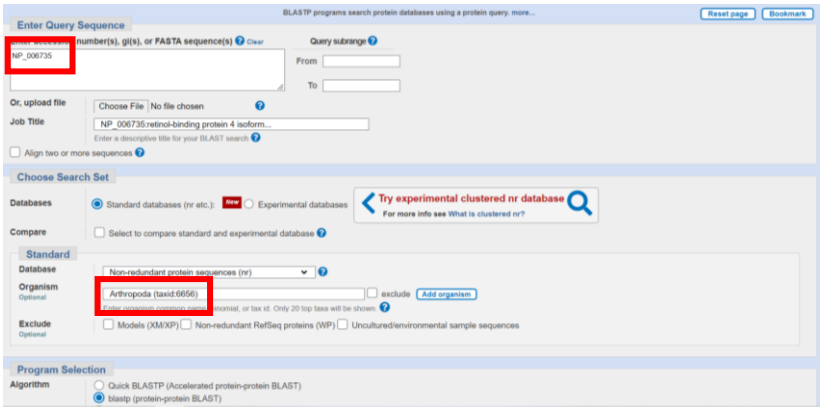
ii) Next, do a BLASTN search using the RBP4 nucleotide sequence (NM_006744). For this query, select only the nucleotides corresponding to the coding region of the DNA.

(To do this visit the NCBI Nucleotide page, follow the link to the coding sequence [CDS], then choose the FASTA format.)

iii) Which search is more informative? How many databases matches have an E value less than 1.0 in each search?

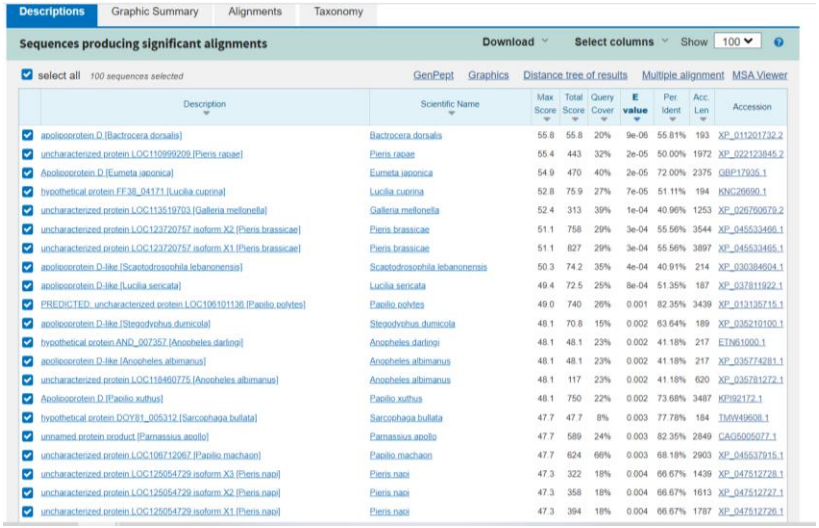


Click on the edit search.

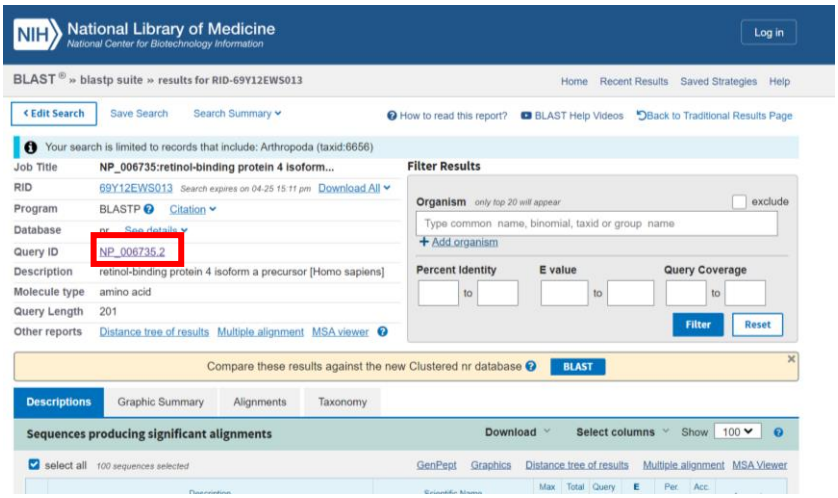


Change the query sequence to NP_006735, and organism to Arthropoda (taxid:6656). Click on the BLAST button.

You will be redirected to this page. Wait for a while.

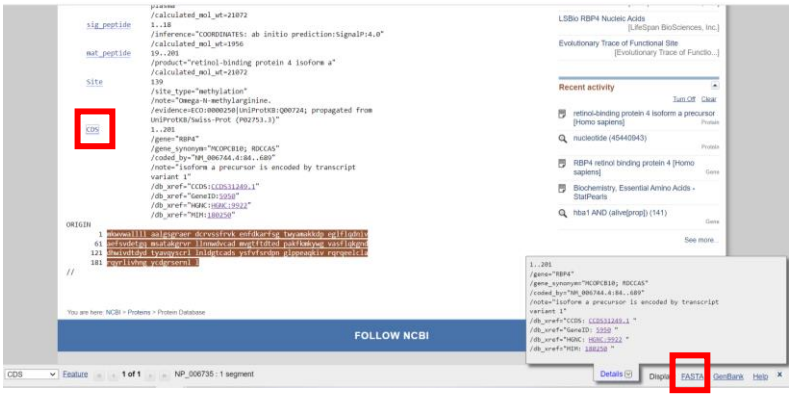


When the result is shown, scroll down to see the descriptions of the sequences producing significant alignments. You will be observed that there are 100 matches having E value less than 1.0.

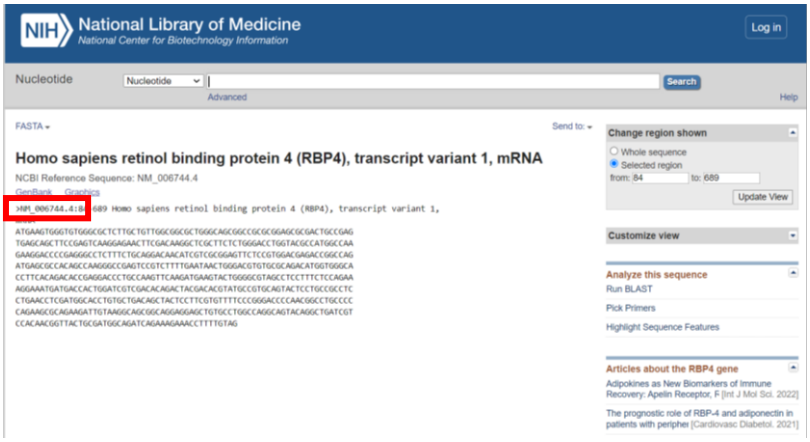


Select the query ID.

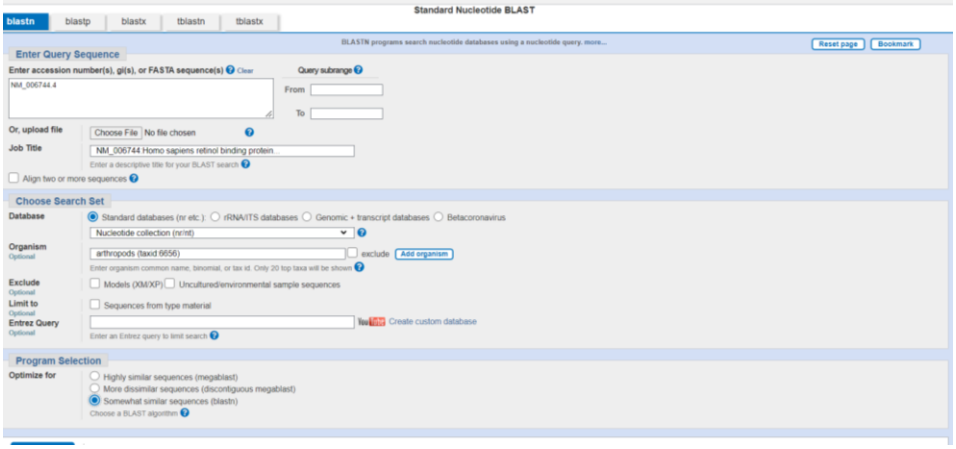
You will be redirected to this page. Scroll down until you reached the CDS in features part.



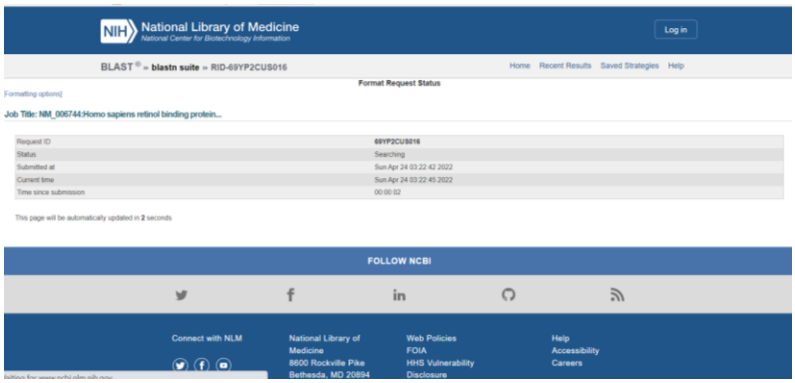
Click on the CDS, the details will be shown. Then, select the FASTA at the bottom of the page.



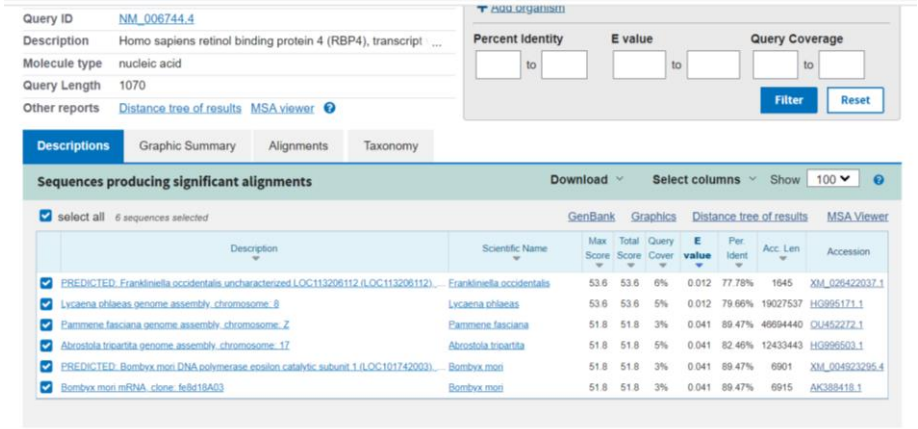
You will be redirected to this page. Then, copy the accession numbers.



Return to the BLAST program and change to the blastn. Enter the accession numbers, “NM_006744.4” and organism, “Arthropoda (taxid:6656)”. Under the program selection, choose the optimise for “somewhat similar sequences (blastn).



You will be redirected to this page. Wait for a while.



When the result is shown, scroll to the description sequences producing significant alignments. You will see there is 6 matches having E value less than 1.0. Hence, we can conclude that protein search is more informative than DNA search. This is because protein sequences are built by 20 amino acids while DNA sequences are built by 4 bases.

3) This problem introduces batch queries. It is possible to search many queries simultaneously, either using the web-based BLAST (as in this problem) or via locally installed BLAST+.

   Mosses are plants of the phylum Bryophyta, including the non-seed plant Physcomitrella patens that had its genome sequenced (Rensing et al., 2008).

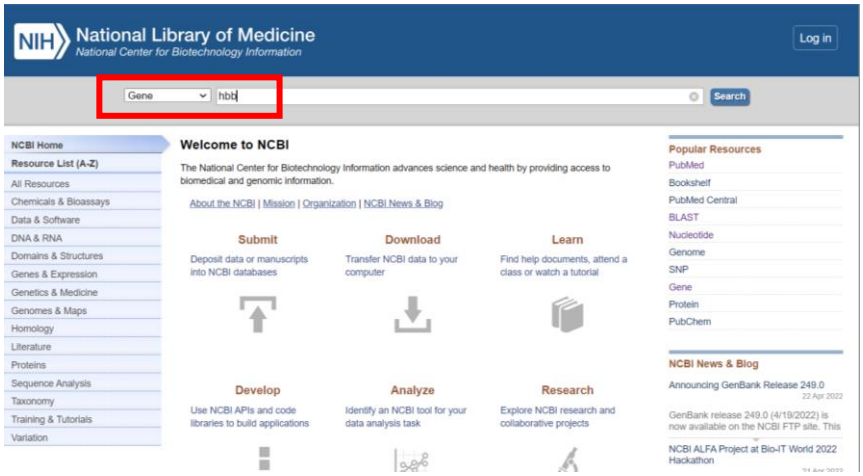
   Do mosses have any globin proteins, and if so, which human globin(s) are they most closely related to?

i) First obtain the accession numbers of all human globins. There are several approaches to doing this, including BLASTP using beta globin and neuroglobin as queries. Other

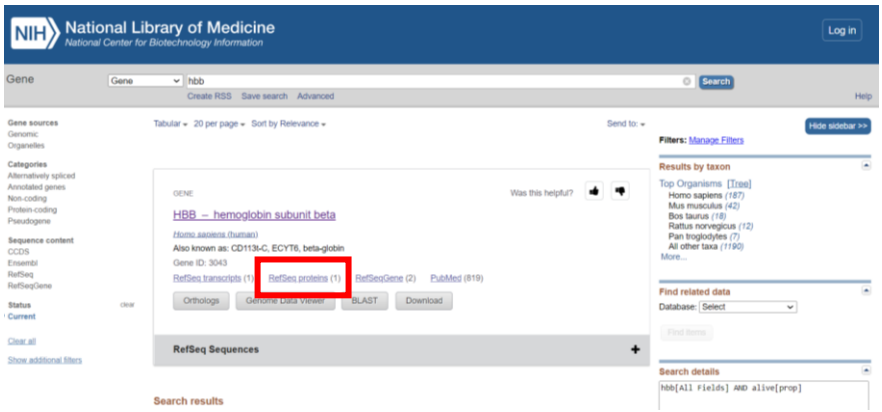
approaches involve DELTA-BLAST (Chapter 5), or Pfam (Chapter 6). These accession numbers are provided in Web Document 4.7.

ii) Perform a BLASTP search using all accession numbers as queries, entering them into the query box. Restrict the output to RefSeq proteins of the mosses.

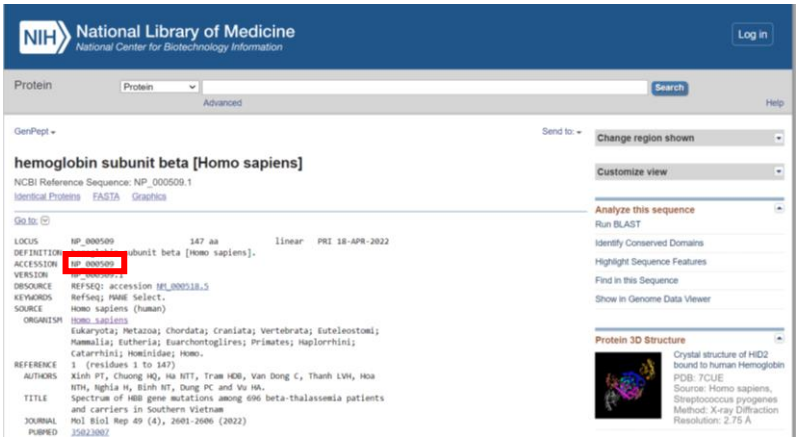
iii) Results for each query are shown (one at a time) via a pull-down menu. Currently there are significant, although distant matches of all human globins to moss proteins except for hemoglobin subunit mu. (See for example the match between human epsilon globin and predicted moss protein XP_001786089.1 with an E value of 0.01. A BLASTP search with that moss protein confirms it is related to many annotated plant globins.) Notably, only one human protein (neuroglobin, NP_001030585.1) has very strong matches to moss proteins such as P. patens predicted protein XP_001764902.1 (E value 2e-10, 27% identity across a span of 138 amino acid residues).



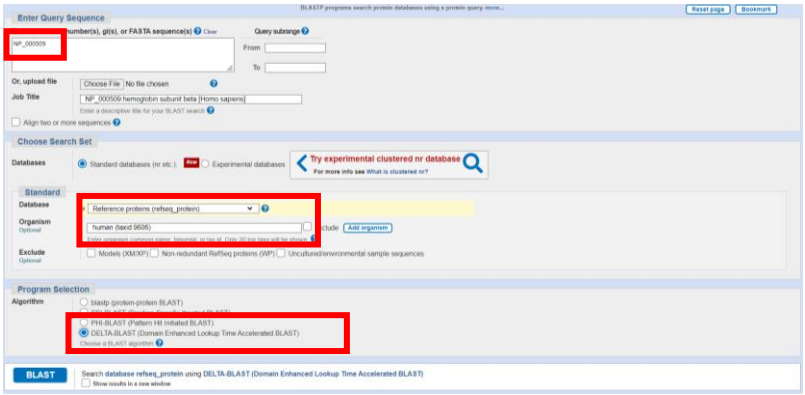
First, go to the NCBI website. Change to the Gene and type hbb on search bar.



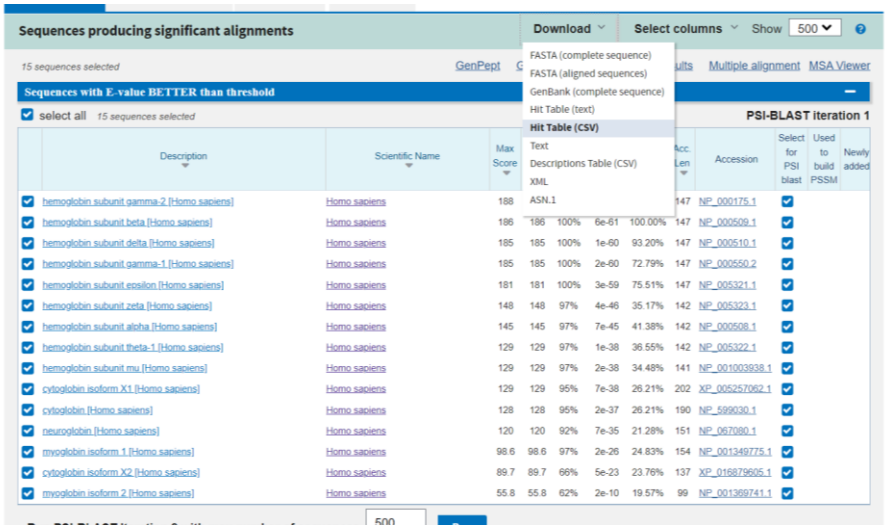
Then, select the RefSeq proteins.



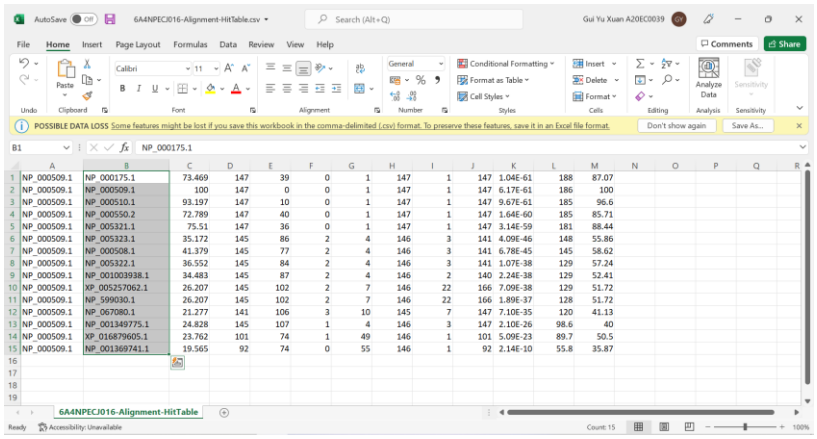
After you had been redirected to this page, copy the accession number of hbb [NP_000509].



Then, go to the blast page, paste the accession number on the query sequence. Under the Standard section, change the database to reference protein and type human on the organism bar. Under program section, choose the delta-blast algorithm. Click on the BLAST button.

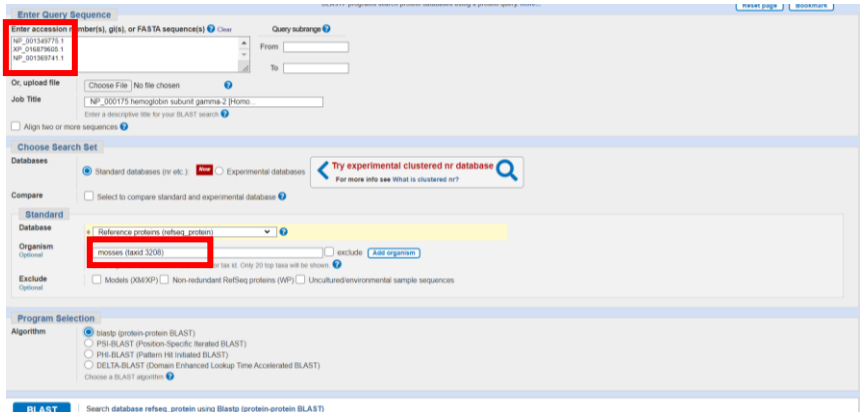


Select all 15 sequences and download it as csv.

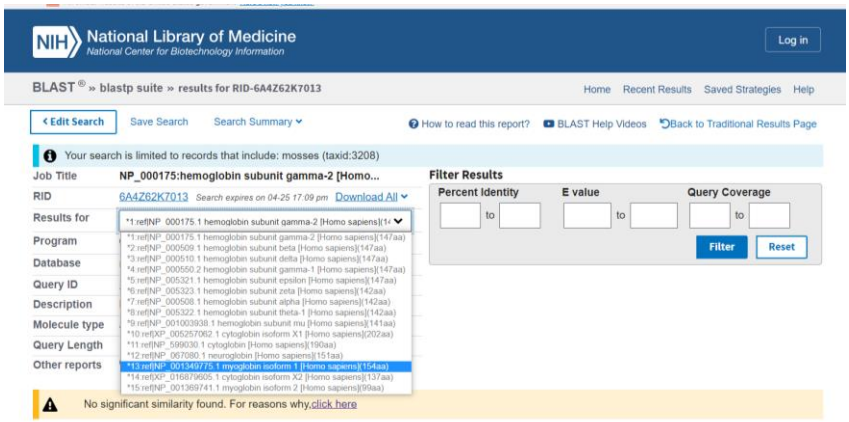


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	NP_000509.1	NP_000175.1	73.469	147	39	0	1	147	1	147	1.04E-61	188	87.07
2	NP_000509.1	NP_000509.1	100	147	0	0	1	147	1	147	6.17E-61	186	100
3	NP_000509.1	NP_000510.1	93.197	147	10	0	1	147	1	147	9.67E-61	185	96.6
4	NP_000509.1	NP_000550.2	72.789	147	40	0	1	147	1	147	1.64E-60	185	85.71
5	NP_000509.1	NP_005321.1	75.51	147	36	0	1	147	1	147	3.14E-59	181	88.44
6	NP_000509.1	NP_005323.1	35.172	145	86	2	4	146	3	141	4.09E-46	148	55.86
7	NP_000509.1	NP_000508.1	41.379	145	77	2	4	146	3	141	6.78E-45	145	58.62
8	NP_000509.1	NP_005322.1	36.552	145	84	2	4	146	3	141	1.07E-38	129	57.24
9	NP_000509.1	NP_001003938.1	34.483	145	87	2	4	146	2	140	2.24E-38	129	52.41
10	NP_000509.1	XP_005257062.1	26.207	145	102	2	7	146	22	166	7.09E-38	129	51.72
11	NP_000509.1	NP_599030.1	26.207	145	102	2	7	146	22	166	1.89E-37	128	51.72
12	NP_000509.1	NP_067080.1	21.277	141	106	3	10	145	7	147	7.10E-35	120	41.13
13	NP_000509.1	NP_001349775.1	24.828	145	107	1	4	146	3	147	2.10E-26	98.6	40
14	NP_000509.1	XP_016879605.1	23.762	101	74	1	49	146	1	101	5.09E-23	89.7	50.5
15	NP_000509.1	NP_001369741.1	19.565	92	74	0	55	146	1	92	2.14E-10	55.8	35.87

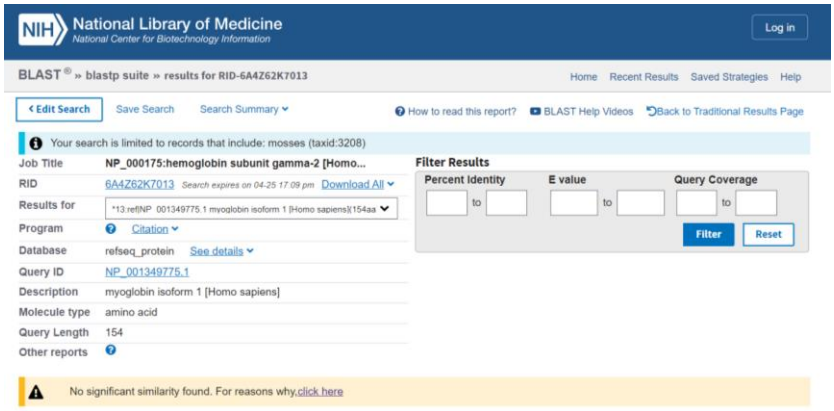
You will get 15 accession number.



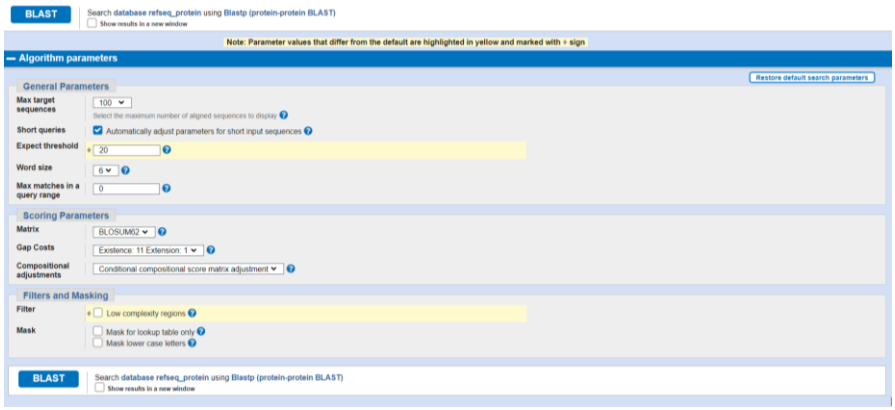
Enter the 15-accession number to the query sequence and change the organism to mosses.



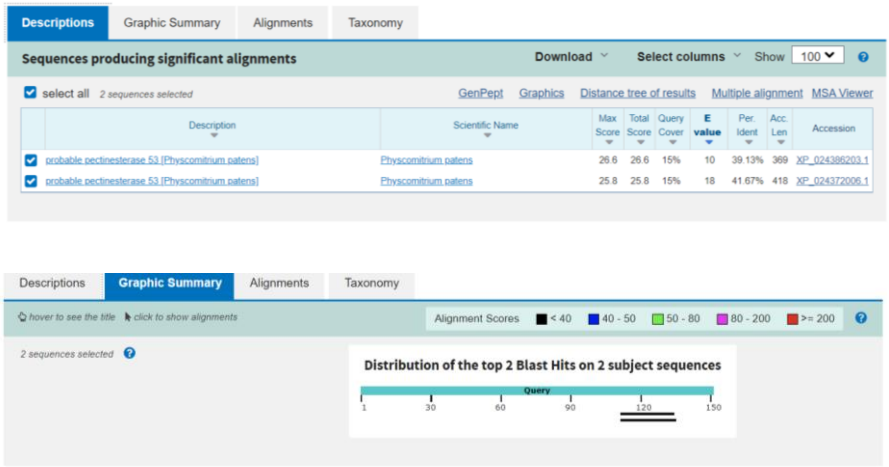
Select the 13 reference proteins.



It shows no significant similarity found.

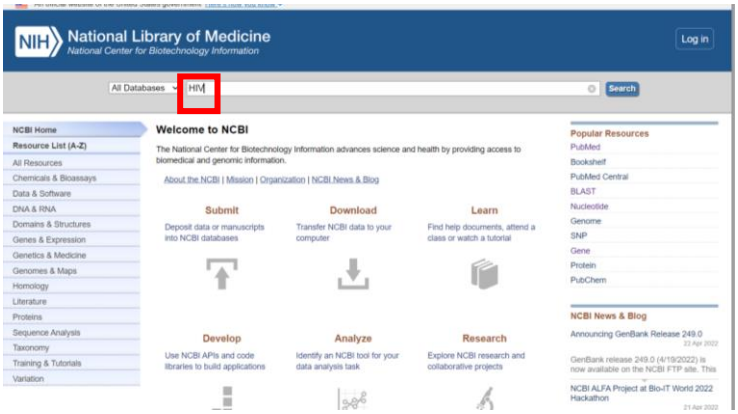


Click on the edit search, you will be redirected to this page. Under the algorithm parameters, change the expect threshold to 20 and click on BLAST button. You will get the results as below:

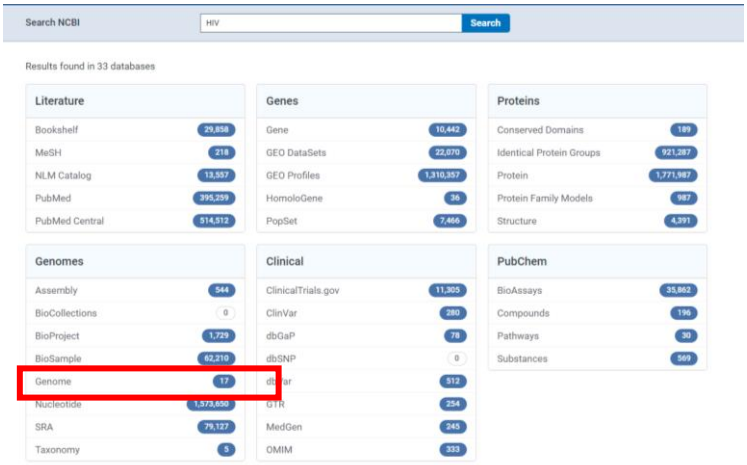


4) Is the pol protein of HIV-1 more closely related to the pol protein of HIV-2 or to the pol protein of simian immunodeficiency virus (SIV)? Use the BLASTP program to decide.

Hint: try the Entrez command "NOT hiv-1[organism]" to focus the search away from HIV-1 matches.

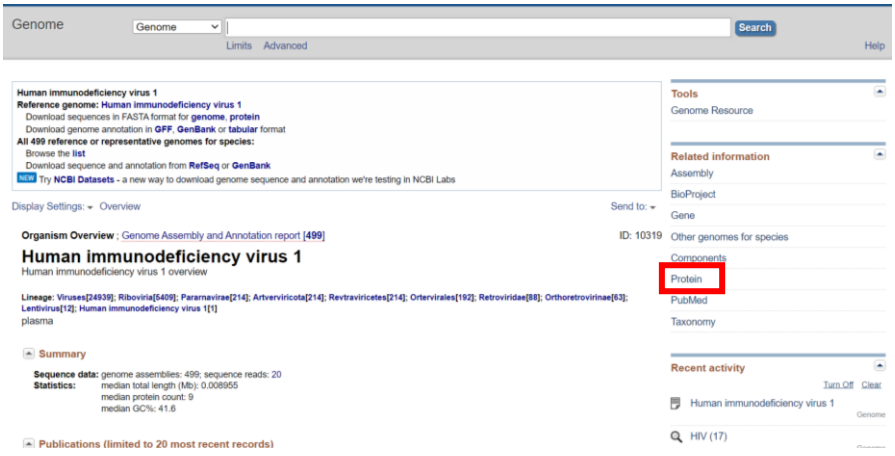


Go to NCHI website and type HIV on the search bar and click on the search button.

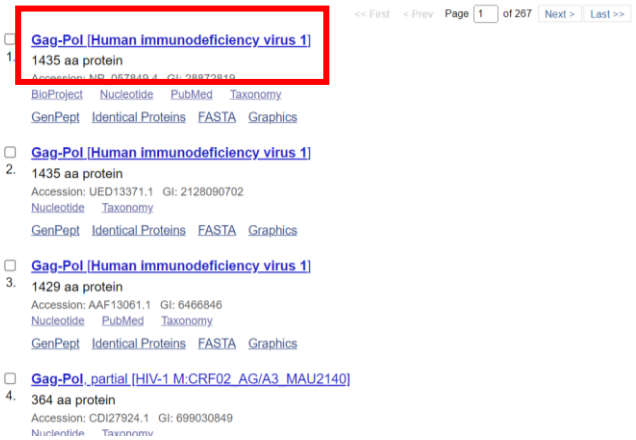


After you had been redirected to this page, click on the genome.

Then, click on the Human immunodeficiency virus 1



Click on the protein.



Select Gag-Pol (Human immunodeficiency virus 1).

Select the FASTA to obtain the sequences of 1435 amino acids. Then, click on the Run BLAST.

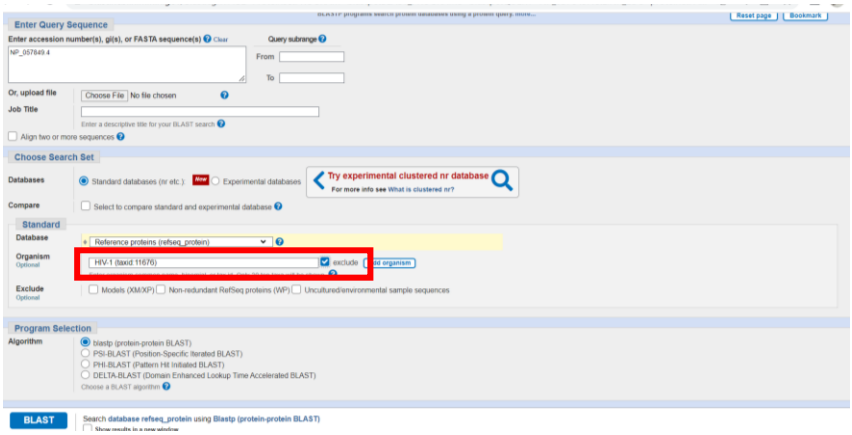
Gag-Pol [Human immunodeficiency virus 1]

NCBI Reference Sequence: NP_057849.4

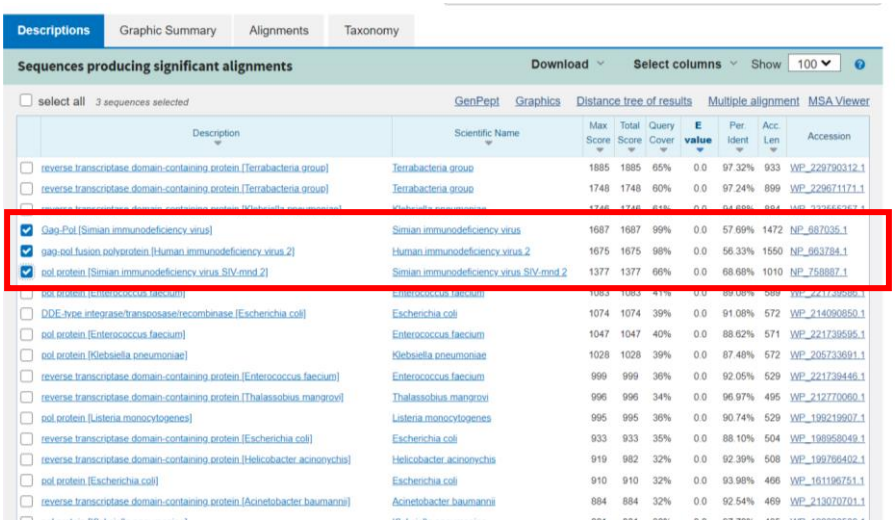
GenPept Identical Proteins Graphics

```
>NP_057849.4 Gag-Pol [Human immunodeficiency virus 1]
MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVNPGLLETSEGCRQILGQLQPSLQT
GSEELRSLYNTVATLYCVHQRIEIKDTKEALDKIEEEQNKSKKKAQQAAADTGHSNQVSQNYPIVQNIQG
QMVHQAISPRTLNAWVKVVEEKAFSPEVIPMFSALSEGATPQDLNTMLNTVGGHQAAMQMLKETINEEAA
EWDRVHPVHAGPIAPGQMREPRGSDIAGTTSTLQEQIGWMTNNPPIPVGEIYKRWIILGLNKIVRMYSPT
SILDIRQGPKEPFRDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTAC
QGVGGPGHKARVLAEAMSQVTNSATIMMQRGNFRNQRKIVKCFNCGKEGHTARNCRAPRKKGCWKCGKEG
HQMKDCTERQANFLREDLAFLQGKAREFSSEQTRANSPTRRELQVWGRDNNSPSEAGADRQGTVSFNFPQ
VTLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAI
GTVLVGPTPVNIIGRNLLTQIGCTLNFPISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEK
EGKISKIGPENPYNTPVFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGD
AYFSVPLDEDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPFRKQNPDIVIYQY
MDDLYVGSDLEIGQHRTKIEELRQHLLRWGLTTPDKKHQKEPPFLWMGYELHPDKWTVQPIVLPEKDSWT
VNDIQKLVGKLNWASQIYPGIKVRQLCKLLRGTKALTEVIPLTEEAELELAENREILKEPVHGVYYDPSK
DLIAEIQKQGQGQWTYQIYQEPFKNLKTGKYARMRGAHTNDVKQLTEAVQKITTESIVIWGKTPKFKLPI
QKETWETWWTEYWQATWIPEWEFVNTPPLVKLWYQLEKEPIVGAETFYVDGAANRETKLGKAGYVTNRGR
QKVVTLTDTTNQKTELQAIYLALQDSGLEVNIVTDSQYALGIIQAQPDQSESELVNQIIEQLIKKEKVYL
AWVPAHKGIGGNEQVDKLVSAGIRKVLFLDGIDKAQDEHEKYHSNWRAMASDFNLPPVVAKEIVASCDKC
QLKGEAMHGQVDCSPGIWQLDCTHLEGKVILVAVHVASGYIEAEVIPAETGQETAYFLLKLAGRWPVKTI
HTDNGSNFTGATVRAACWWAGIKQEFGIPYNPQSQGVVESMNKELKKIIGQVRDQAEHLKTAVQMAVFIH
NFKRKGGIGGYSAGERIVDIIATDIQTKELQKQITKIQNFRVYYRDSRNPLWKGPAKLLWKGEGAVVIQD
NSDIKVVPRRKAKIIRDYGKQMAGDDCVASRQDED
```

Here is the result of FASTA.



After click on Run BLAST, you will be redirected to this page. Type HIV-1 on the organism bar under Standard section and click on the exclude. Then, click on BLAST button.



Pol protein of HIV 1 is compared to the Simian immunodeficiency virus and pol protein of HIV 2. You can see that the both have 0 score E value while SIV has 99% query coverage which is slightly higher than 98% query coverage.

5) You perform a BLAST search, and a result has an E value of about 1×10^{-4}. What does this E value mean? What are some parameters on which an E value depends?

The BLAST E-value is the number of expected hits of similar quality (score) that could be found just by chance. Each E value has an associated score S.

An E value of about 1×10^{-4} means that for the query used, and for the database searched of some particular size, you can expect to obtain a score $\geqslant$ S by chance one time in 10,000.

You can safety reject the null hypothesis which states that the alignment between the query and the database match occurred by chance. Such an E value implies homology such as that these sequences are descended from a common ancestor.