



**UTM**  
UNIVERSITI TEKNOLOGI MALAYSIA

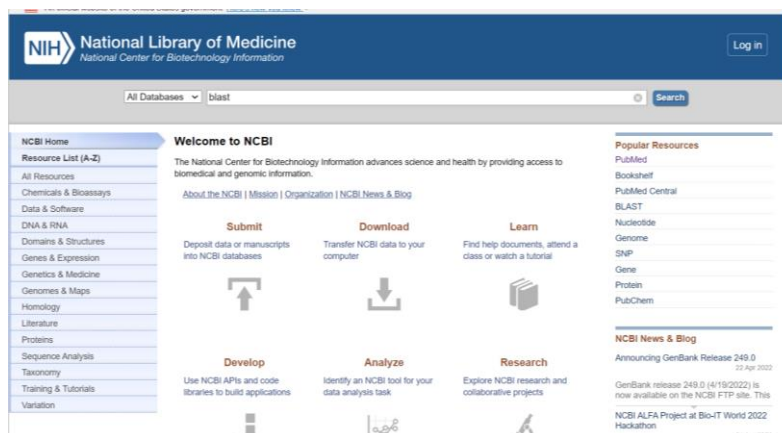
**SCHOOL OF COMPUTING**  
Faculty of Engineering

## 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.

NIH National Library of Medicine  
National Center for Biotechnology Information

Search NCBI

Results found in 32 databases

WEB RESOURCE

**BLAST - Basic Local Alignment Search Tool**

A tool for comparing an amino acid or nucleotide sequence to an entire sequence library, identifying regions of high sequence similarity.

[blast](#) [blastp](#) [blastx](#) [Primer-BLAST](#)

**BLAST**

Use our new Betacoronavirus database for SARS-CoV-2 genome sequence analysis

Literature	Genes	Proteins
Bookshelf <b>2,391</b>	Gene <b>28,496</b>	Conserved Domains <b>74</b>
MeSH <b>32</b>	GEO DataSets <b>16,383</b>	Identical Protein Groups <b>196</b>
NLM Catalog <b>533</b>	GEO Profiles <b>138,189</b>	Protein <b>5,974,751</b>

Then, click on the blastp.

BLAST® » blastp suite Home Recent Results Saved Strategies Help

blastn **blastp** blastx tblastn tblastx **Standard Protein BLAST**

BLASTP programs search protein databases using a protein query more...

Enter Query Sequence:

From  To

Or, upload file  No file chosen

Job Title

☐ Align two or more sequences

Choose Search Set:

Databases ☒ Standard databases (or etc.) ☐ Experimental databases [Try experimental clustered nr database](#)

Compare ☐ Select to compare standard and experimental database

Database

Organism

Exclude ☐ Models (MMXP) ☐ Non-redundant RefSeq proteins (NRP) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm ☐ Quick BLASTP (Accelerated protein protein BLAST)

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

NIH National Library of Medicine  
National Center for Biotechnology Information

BLAST® » blastp suite » RID-69X4JRPR013 Home Recent Results Saved Strategies Help

Format Request Status

unsetting options

Job Title: Protein Sequence

Request ID	69X4JRPR013
Status	Searching
Submitted at	Sun Apr 24 02:56:18 2022
Current time	Sun Apr 24 02:56:21 2022
Time since submission	00:00:02

This page will be automatically updated in 2 seconds

FOLLOW NCBI

[Twitter](#) [Facebook](#) [LinkedIn](#) [RSS](#)

Connect with NLM National Library of Medicine Web Policies Help

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

BLAST® » blastp suite » results for RID-69X4JRPR013

Home Recent Results Saved Strategies Help

[Edit Search](#) [Save Search](#) [Search Summary](#) [How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

**Job Title** Protein Sequence  
**RID** 69X4JRPR013 Search expires on 04-25 14:56 pm [Download All](#)  
**Program** BLASTP [Citation](#)  
**Database** nr [See details](#)  
**Query ID** lcl|Query\_31886  
**Description** unnamed protein product  
**Molecule type** amino acid  
**Query Length** 12  
**Other reports** [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#)

**Filter Results**  
**Organism** only top 20 will appear ☐ exclude  
 Type common name, binomial, taxid or group name  
[Add organism](#)  
**Percent Identity**  to  **E value**  to  **Query Coverage**  to   
[Filter](#) [Reset](#)

Compare these results against the new Clustered nr database [BLAST](#)

**Descriptions** Graphic Summary Alignments Taxonomy

**Sequences producing significant alignments** [Download](#) [Select columns](#) [Show](#) 100

☒ select all 100 sequences selected

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> cytochrome c (Schizosaccharomyces pombe)	Schizosaccharomyces pombe	40.5	40.5	100%	0.007	100.00%	26	AEC04604.1
<input checked="" type="checkbox"/> putine and uridine diphosphatase (Aureobasidium melanogenum)	Aureobasidium melanogenum	40.5	40.5	100%	0.018	100.00%	1487	KAG0692838.1

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

BLAST® » blastp suite » results for RID-69X4JRPR013

Home Recent Results Saved Strategies Help

[Edit Search](#) [Save Search](#) [Search Summary](#) [How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

**Search Parameters**

Program	blastp
Word size	2
Expect value	200000
Hitlist size	100
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	11
Composition-based stats	0

**Database**

Posted date	Apr 21, 2022 2:37 AM
Number of letters	180,797,797,885
Number of sequences	468,653,756
Entrez query	None

**Karlin-Altschul statistics**

Lambda	0.355191	0.294
K	0.320975	0.11
H	1.89356	0.61
Alpha	0.1938	0.48
Alpha_v	0.161818	2.46392
Sigma		3.18515

**Results Statistics**

**Filter Results**  
**Organism** only top 20 will appear ☐ exclude  
 Type common name, binomial, taxid or group name  
[Add organism](#)  
**Percent Identity**  to  **E value**  to  **Query Coverage**  to   
[Filter](#) [Reset](#)

Compare these results against the new Clustered nr database [BLAST](#)

**Sequences producing significant alignments** [Download](#) [Select columns](#) [Show](#) 100

[GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#) [MSA Viewer](#)

Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Schizosaccharomyces pombe	40.5	40.5	100%	0.007	100.00%	26	AEC04604.1
Aureobasidium melanogenum	40.5	40.5	100%	0.018	100.00%	1487	KAG0692838.1

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?

BLAST® » blastp suite » results for RID-69X4JRPR013

[Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

[< Edit Search](#) [Save Search](#) [Search Summary](#) [How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

**Filter Results**

Organism: only top 20 will appear ☐ exclude  
 Type common name, binomial, taxid or group name  
[+ Add organism](#)

Percent Identity:  to  E value:  to  Query Coverage:  to   
[Filter](#) [Reset](#)

Compare these results against the new Clustered nr database [BLAST](#)

**Descriptions** [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

**Sequences producing significant alignments** [Download](#) [Select columns](#) [Show 100](#)

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
cytochrome C (Schizosaccharomyces)	Schizosaccharomyces	40.5	40.5	100%	0.007	100.00%	26	AEC04604.1
guanine and uridine diphosphatase (Aureobasidium)	Aureobasidium melenogenum	40.5	40.5	100%	0.018	100.00%	1487	KAC3962838.1

Click on the edit search.

Enter Query Sequence:  [number\(s\), gi\(s\), or FASTA sequence\(s\)](#) [Clear](#) [Query subrange](#)

Or, upload file: [Choose File](#) [No file chosen](#)

Job Title:

☐ Align two or more sequences

**Choose Search Set**

Databases: ☒ Standard databases (nr etc.) ☐ Experimental databases [Try experimental clustered nr database](#)

Compare: ☐ Select to compare standard and experimental database

**Standard**

Database:  [?](#)

Organism:  [?](#) [Add organism](#)

Exclude: ☐ Models (AM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

**Program Selection**

Algorithm: ☒ Quick BLASTP (Accelerated protein-protein BLAST) ☐ blastp (protein-protein BLAST)

Change the query sequence to NP\_006735, and organism to Arthropoda (taxid:6656).

Click on the BLAST button.

NIH National Library of Medicine National Center for Biotechnology Information [Log in](#)

BLAST® » blastp suite » RID-69Y12EWS013 [Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

[\[Formatting options\]](#) [Format Request Status](#)

Job Title: NP\_006735:retinol-binding protein 4 isoform...

Request ID	69Y12EWS013
Status	Searching
Submitted at	Sun Apr 24 03:11:30 2022
Current time	Sun Apr 24 03:11:42 2022
Time since submission	00:00:11

This page will be automatically updated in 2 seconds

**FOLLOW NCBI**

[Twitter](#) [Facebook](#) [LinkedIn](#) [RSS](#)

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

Descriptions	Graphic Summary	Alignments	Taxonomy					
Sequences producing significant alignments								
Download Select columns Show 100								
<input checked="" type="checkbox"/> select all 100 sequences selected	GenPept	Graphics	Distance tree of results Multiple alignment MSA Viewer					
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> apolipoprotein D (Bactrocera dorsalis)	Bactrocera dorsalis	55.8	55.8	20%	9e-06	55.81%	193	XP_011201732.2
<input checked="" type="checkbox"/> uncharacterized protein LOC110999209 (Pteris racae)	Pteris racae	55.4	443	32%	2e-05	50.00%	1972	XP_022123845.2
<input checked="" type="checkbox"/> Apolipoprotein D (Eumeta japonica)	Eumeta japonica	54.9	470	40%	2e-05	72.00%	2375	GBPI7935.1
<input checked="" type="checkbox"/> hypothetical protein FF38_04171 (Lucilia cucurina)	Lucilia cucurina	52.8	75.9	27%	7e-05	51.11%	194	KNC26960.1
<input checked="" type="checkbox"/> uncharacterized protein LOC113519703 (Galleria mellonella)	Galleria mellonella	52.4	313	39%	1e-04	40.96%	1253	XP_026760679.2
<input checked="" type="checkbox"/> uncharacterized protein LOC123720757 isoform X2 (Pteris brassicae)	Pteris brassicae	51.1	758	29%	3e-04	55.56%	3544	XP_045533466.1
<input checked="" type="checkbox"/> uncharacterized protein LOC123720757 isoform X1 (Pteris brassicae)	Pteris brassicae	51.1	827	29%	3e-04	55.56%	3897	XP_045533465.1
<input checked="" type="checkbox"/> apolipoprotein D-like (Scaevola lebanonensis)	Scaevola lebanonensis	50.3	74.2	35%	4e-04	40.91%	214	XP_030384604.1
<input checked="" type="checkbox"/> apolipoprotein D-like (Lucilia sericata)	Lucilia sericata	49.4	72.5	25%	8e-04	51.35%	187	XP_037811922.1
<input checked="" type="checkbox"/> PREDICTED: uncharacterized protein LOC106101138 (Paecilomyces)	Paecilomyces	49.0	740	26%	0.001	82.35%	3439	XP_013135715.1
<input checked="" type="checkbox"/> apolipoprotein D-like (Stegomyia fumicola)	Stegomyia fumicola	48.1	70.8	15%	0.002	63.64%	189	XP_035210100.1
<input checked="" type="checkbox"/> hypothetical protein AND_007357 (Anopheles darlingi)	Anopheles darlingi	48.1	48.1	23%	0.002	41.18%	217	ET361000.1
<input checked="" type="checkbox"/> apolipoprotein D-like (Anopheles albimanus)	Anopheles albimanus	48.1	48.1	23%	0.002	41.18%	217	XP_035774281.1
<input checked="" type="checkbox"/> uncharacterized protein LOC118400775 (Anopheles albimanus)	Anopheles albimanus	48.1	117	23%	0.002	41.18%	620	XP_035781272.1
<input checked="" type="checkbox"/> Apolipoprotein D (Paecilomyces)	Paecilomyces	48.1	750	22%	0.002	73.68%	3487	KP092172.1
<input checked="" type="checkbox"/> hypothetical protein DQV91_005312 (Sarcophaga bullata)	Sarcophaga bullata	47.7	47.7	8%	0.003	77.78%	184	IM049608.1
<input checked="" type="checkbox"/> unnamed protein product (Parasitica asolito)	Parasitica asolito	47.7	589	24%	0.003	82.35%	2849	CAG5005077.1
<input checked="" type="checkbox"/> uncharacterized protein LOC106712067 (Paecilomyces)	Paecilomyces	47.7	624	66%	0.003	68.18%	2903	XP_0455337915.1
<input checked="" type="checkbox"/> uncharacterized protein LOC125054729 isoform X3 (Pteris naxos)	Pteris naxos	47.3	322	18%	0.004	66.67%	1439	XP_047512728.1
<input checked="" type="checkbox"/> uncharacterized protein LOC125054729 isoform X2 (Pteris naxos)	Pteris naxos	47.3	358	18%	0.004	66.67%	1613	XP_047512727.1
<input checked="" type="checkbox"/> uncharacterized protein LOC125054729 isoform X1 (Pteris naxos)	Pteris naxos	47.3	394	18%	0.004	66.67%	1787	XP_047512726.1

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.

National Library of Medicine  
National Center for Biotechnology Information

Log in

BLAST® » blastp suite » results for RID-69Y12EWS013

HomeRecent ResultsSaved StrategiesHelp

Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page

Your search is limited to records that include: Arthropoda (taxid:6656)

Job TitleNP\_006735:retinol-binding protein 4 isoform...
RID69Y12EWS013
ProgramBLASTP
Database
Query IDNP\_006735.2
Descriptionretinol-binding protein 4 isoform a precursor [Homo sapiens]
Molecule typeamino acid
Query Length201
Other reportsDistance tree of resultsMultiple alignmentMSA viewer

Filter Results
Organism
Percent Identity
E value
Query Coverage

Compare these results against the new Clustered nr database
BLAST

Descriptions
Graphic Summary
Alignments
Taxonomy

Sequences producing significant alignments
Download
Select columns
Show 100

select all 100 sequences selected
GenPept
Graphics
Distance tree of results
Multiple alignment
MSA Viewer

Select the query ID.

National Library of Medicine  
National Center for Biotechnology Information

Log in

Protein
Advanced
Search
Help

GenPept
Send to
Change region shown
Customize view
Analyze this sequence
Run BLAST
Identify Conserved Domains
Highlight Sequence Features
Find in this Sequence
Show in Genome Data Viewer
Protein 3D Structure

retinol-binding protein 4 isoform a precursor [Homo sapiens]
NCBI Reference Sequence: NP\_006735.2
Identical Proteins
FASTA
Graphics

Go to
LOCUS
DEFINITION
ACCESSION
VERSION
DBSOURCE
KEYWORDS
SOURCE
ORGANISM
REFERENCE
AUTHORS
TITLE

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

The screenshot shows the NCBI protein page for RBP4. The CDS (Coding Sequence) is highlighted in a red box. The FASTA format is selected at the bottom right, also highlighted in a red box.

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

The screenshot shows the NLM page for Homo sapiens retinol binding protein 4 (RBP4), transcript variant 1, mRNA. The accession number NM\_006744.4 is highlighted in a red box.

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

The screenshot shows the BLAST search interface. The accession number NM\_006744.4 is entered in the 'Enter Query Sequence' field. The organism is set to 'Arthropoda (taxid:6656)'. The program selection is set to 'blastn'.

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.

Query ID: [NM\\_006744.4](#)  
 Description: Homo sapiens retinol binding protein 4 (RBP4), transcript ...  
 Molecule type: nucleic acid  
 Query Length: 1070  
 Other reports: [Distance tree of results](#) [MSA viewer](#)

Percent Identity:  to  E value:  to  Query Coverage:  to   
 Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download Select columns Show 100

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> PREDICTED: <i>Frankliniella occidentalis</i> uncharacterized LOC113206112 (LOC113206112) ...	<i>Frankliniella occidentalis</i>	53.6	53.6	6%	0.012	77.78%	1645	XM_026422037.1
<input checked="" type="checkbox"/> <i>Lycaena phlaeas</i> genome assembly, chromosome_8	<i>Lycaena phlaeas</i>	53.6	53.6	5%	0.012	79.66%	19027537	HG995171.1
<input checked="" type="checkbox"/> <i>Pameme fasciana</i> genome assembly, chromosome_2	<i>Pameme fasciana</i>	51.8	51.8	3%	0.041	89.47%	46694440	OU452272.1
<input checked="" type="checkbox"/> <i>Abrostola tripartita</i> genome assembly, chromosome_17	<i>Abrostola tripartita</i>	51.8	51.8	5%	0.041	82.46%	12433443	HG995503.1
<input checked="" type="checkbox"/> PREDICTED: <i>Bombyx mori</i> DNA polymerase epsilon catalytic subunit 1 (LOC101742093) ...	<i>Bombyx mori</i>	51.8	51.8	3%	0.041	89.47%	6901	XM_004923295.4
<input checked="" type="checkbox"/> <i>Bombyx mori</i> mRNA, clone: felid15A03	<i>Bombyx mori</i>	51.8	51.8	3%	0.041	89.47%	6915	AK388416.1

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+.

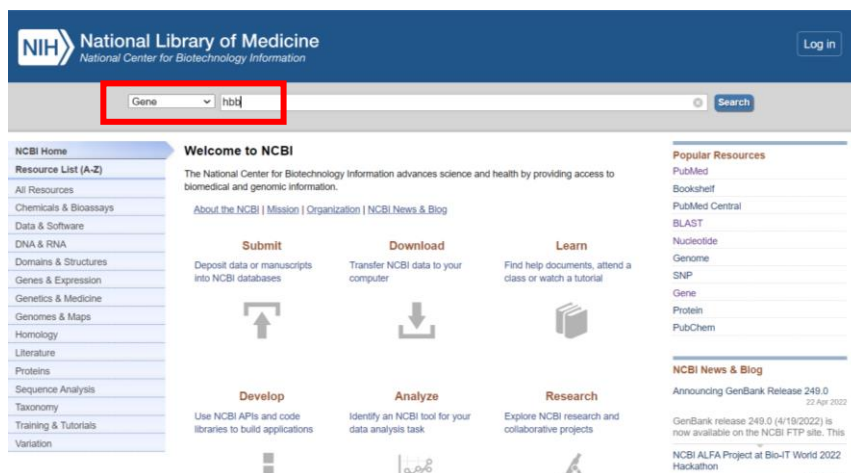
Mosses are plants of the phylum Bryophyta, including the non-seed plant *Physcomitrella patens* that had its genome sequenced (Rensing et al., 2008).

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

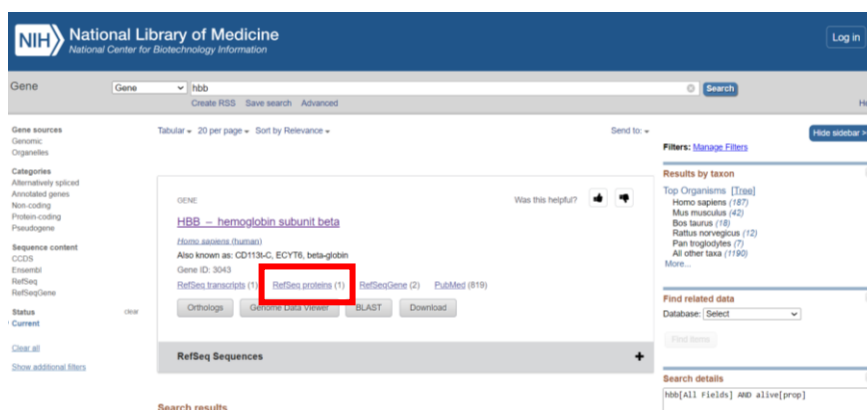
- i) First obtain the accession numbers of all human globins. There are several approaches to doing this, including BLASTP using beta globin and neuroglobin as queries. Other

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.



**hemoglobin subunit beta [Homo sapiens]**  
 NCBI Reference Sequence: NP\_000509.1  
 Identical Proteins FASTA Graphics

Go to:

LOCUS NP\_000509 147 aa linear PRI 18-APR-2022  
 DEFINITION hemoglobin subunit beta [Homo sapiens].  
 ACCESSION **NP\_000509**  
 VERSION  
 DISCUSSION  
 REFSEQ: accession M1\_000518.5  
 KEYWORDS RefSeq; NAME Select.  
 SOURCE Homo sapiens (human)  
 ORGANISM Homo sapiens  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominoidea; Homo.  
 1 (residues 1 to 147)  
 AUTHORS Xinh PT, Chuong HQ, Ha HT, Tran HD, Van Dong C, Thanh LVH, Hoa NT, Nghia H, Binh NT, Dang PC and Vu HA.  
 TITLE Spectrum of HBB gene mutations among 696 beta-thalassemia patients and carriers in Southern Vietnam  
 JOURNAL Mol Biol Rep 49 (4), 2681-2686 (2022)  
 PUBMED 35823882

Protein 3D Structure  
 Crystal structure of HbD2 bound to human Hemoglobin PDB: 7CUE  
 Source: Homo sapiens, Streptococcus pyogenes  
 Method: X-ray Diffraction  
 Resolution: 2.75 Å

After you had been redirected to this page, copy the accession number of hbb [NP\_000509].

Enter Query Sequence  
 NP\_000509  
 Or upload file  
 Choose Search Set  
 Databases: Standard databases (in etc.) Experimental databases  
 Standard  
 Database: Reference proteins (refseq\_protein)  
 Organism: Homo sapiens  
 Program Selection  
 Algorithm: DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)  
 BLAST

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.

Sequences producing significant alignments

15 sequences selected

Sequences with E-value BETTER than threshold

☒ select all 15 sequences selected

Description	Scientific Name	Max Score	Positives	Identical	Accession	Select for PSI-BLAST	Used to build PSSM	Newly added
<input checked="" type="checkbox"/> hemoglobin subunit gamma-2 [Homo sapiens]	Homo sapiens	188	188	100%	NP_000175.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> hemoglobin subunit beta [Homo sapiens]	Homo sapiens	186	186	100%	NP_000509.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> hemoglobin subunit delta [Homo sapiens]	Homo sapiens	185	185	100%	NP_000510.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> hemoglobin subunit gamma-1 [Homo sapiens]	Homo sapiens	185	185	100%	NP_000550.2	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> hemoglobin subunit epsilon [Homo sapiens]	Homo sapiens	181	181	100%	NP_005321.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> hemoglobin subunit zeta [Homo sapiens]	Homo sapiens	148	148	97%	NP_005323.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> hemoglobin subunit alpha [Homo sapiens]	Homo sapiens	145	145	97%	NP_000508.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> hemoglobin subunit theta-1 [Homo sapiens]	Homo sapiens	129	129	97%	NP_005322.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> hemoglobin subunit mu [Homo sapiens]	Homo sapiens	129	129	97%	NP_001003938.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> cytoglobin isoform X1 [Homo sapiens]	Homo sapiens	129	129	95%	XP_005257062.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> cytoglobin [Homo sapiens]	Homo sapiens	128	128	95%	NP_599030.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> neuroglobin [Homo sapiens]	Homo sapiens	120	120	92%	NP_067080.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> myoglobin isoform 1 [Homo sapiens]	Homo sapiens	98.6	98.6	97%	NP_001349775.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> cytoglobin isoform X2 [Homo sapiens]	Homo sapiens	89.7	89.7	66%	XP_016875605.1	<input checked="" type="checkbox"/>		
<input checked="" type="checkbox"/> myoglobin isoform 2 [Homo sapiens]	Homo sapiens	55.8	55.8	62%	NP_001369741.1	<input checked="" type="checkbox"/>		

Select all 15 sequences and download it as csv.

You will get 15 accession number.

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

BLAST® » blastp suite » results for RID-6A4Z62K7013

Your search is limited to records that include: mosses (taxid:3208)

Job Title: NP\_000175:hemoglobin subunit gamma-2 [Homo...]  
 RID: 6A4Z62K7013  
 Results for: \*1 refNP\_000175.1 hemoglobin subunit gamma-2 [Homo sapiens] (147aa)

Filter Results: Percent Identity, E value, Query Coverage

No significant similarity found. For reasons why, click here

Select the 13 reference proteins.

NIH National Library of Medicine  
National Center for Biotechnology Information

BLAST® » blastp suite » results for RID-6A4Z62K7013

Log in

Home Recent Results Saved Strategies Help

< Edit Search Save Search Search Summary ▾ How to read this report? BLAST Help Videos Back to Traditional Results Page

ⓘ Your search is limited to records that include: mosses (taxid:3208)

Job Title NP\_000175:hemoglobin subunit gamma-2 [Homo...  
 RID 6A4Z62K7013 Search expires on 04-25 17:09 pm Download All ▾  
 Results for \*13 refNP\_001349775.1 myoglobin isoform 1 [Homo sapiens] 154aa ▾  
 Program Citation ▾  
 Database refseq\_protein See details ▾  
 Query ID NP\_001349775.1  
 Description myoglobin isoform 1 [Homo sapiens]  
 Molecule type amino acid  
 Query Length 154  
 Other reports ⓘ

Filter Results

Percent identity E value Query Coverage

to to to

Filter Reset

No significant similarity found. For reasons why, click here

It shows no significant similarity found.

BLAST Search database refseq\_protein using Blastp (protein-protein BLAST)

Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ⓘ sign

Algorithm parameters

General Parameters

Max target sequences 100 ⓘ

Short queries ☒ Automatically adjust parameters for short input sequences ⓘ

Expect threshold 20 ⓘ

Word size 3 ⓘ

Max matches in a query range 0 ⓘ

Scoring Parameters

Matrix BLOSUM62 ⓘ

Gap Costs Existence: 11 Extension: 1 ⓘ

Compositional adjustments Conditional compositional score matrix adjustment ⓘ

Filters and Masking

Filter ☐ Low complexity regions ⓘ

Mask ☐ Mask for lookup table only ⓘ  
☐ Mask lower case letters ⓘ

BLAST Search database refseq\_protein using Blastp (protein-protein BLAST)

Show results in a new window

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:

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download ▾ Select columns ▾ Show 100 ⓘ

☒ select all 2 sequences selected

GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> probable pectinesterase 53 [Physcomitrium patens]	Physcomitrium patens	26.6	26.6	15%	10	39.13%	369	XP_024386203.1
<input checked="" type="checkbox"/> probable pectinesterase 53 [Physcomitrium patens]	Physcomitrium patens	25.8	25.8	15%	18	41.67%	418	XP_024372005.1

Descriptions Graphic Summary Alignments Taxonomy

hover to see the title click to show alignments

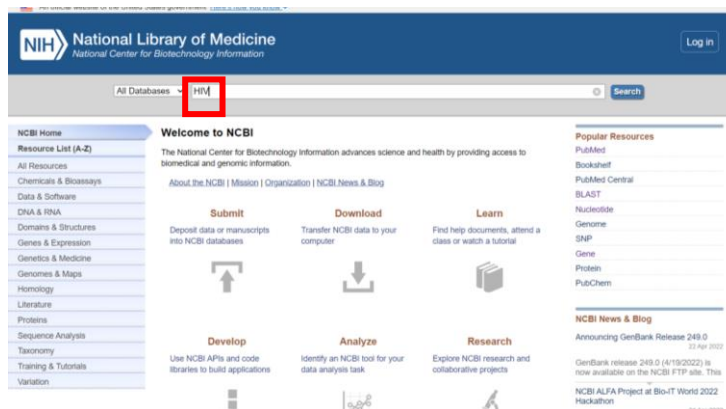
Alignment Scores ■ < 40 ■ 40 - 50 ■ 50 - 80 ■ 80 - 200 ■ >= 200 ⓘ

2 sequences selected ⓘ

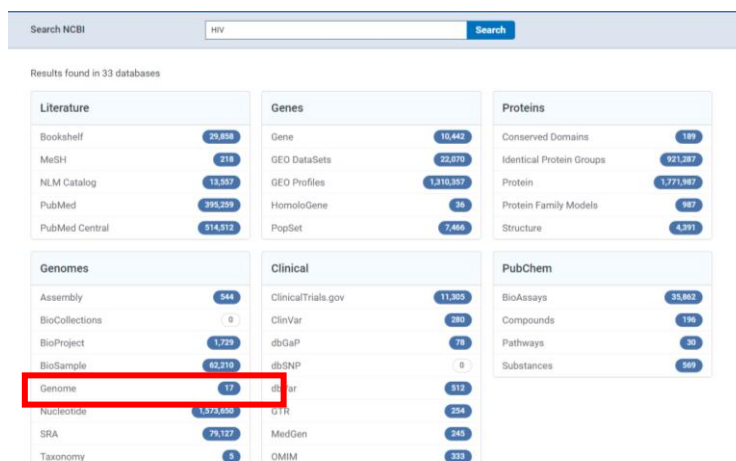
Distribution of the top 2 Blast Hits on 2 subject sequences

- 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.

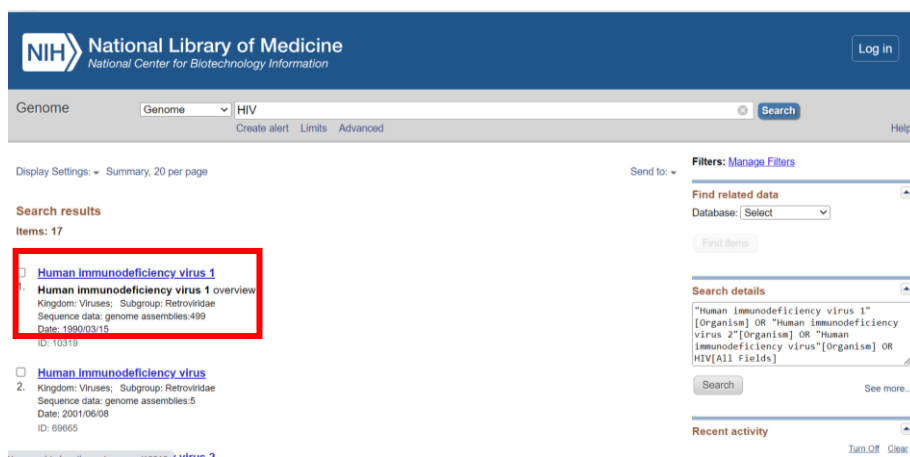
Hint: try the Entrez command “NOT hiv-1[organism]” to focus the search away from HIV-1 matches.



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

Genome

Genome

Search

Limits Advanced Help

Human immunodeficiency virus 1

Reference genome: Human immunodeficiency virus 1

Download sequences in FASTA format for genome, protein

Download genome annotation in GFF, GenBank or tabular format

All 499 reference or representative genomes for species:

Browse the list

Download sequence and annotation from RefSeq or GenBank

Try NCBI Datasets - a new way to download genome sequence and annotation we're testing in NCBI Labs

Display Settings: Overview

Send to: ID: 10319

Other genomes for species

Components

Protein

PubMed

Taxonomy

Recent activity

Human immunodeficiency virus 1

HIV (17)

Organism Overview: Genome Assembly and Annotation report [499]

Human immunodeficiency virus 1

Human immunodeficiency virus 1 overview

Lineage: Viruses[24839]; Riboviria[5409]; Paramyovirales[214]; Artverviricota[214]; Revtraviricetes[214]; Ortervirales[192]; Retroviridae[88]; Orthoretrovirinae[83]; Lentivirus[13]; Human immunodeficiency virus 1[1]

plasma

Summary

Sequence data: genome assemblies: 499; sequence reads: 20

Statistics: median total length (Mb): 0.008955

median protein count: 9

median GC%: 41.6

Publications (limited to 20 most recent records)

Click on the protein.

<< First < Prev Page 1 of 267 Next > Last >>

1. [Gag-Pol \[Human immunodeficiency virus 1\]](#)

1435 aa protein

Accession: NP\_057849.4 GI: 268228210

BioProject Nucleotide PubMed Taxonomy

GenPept Identical Proteins FASTA Graphics

2. [Gag-Pol \[Human immunodeficiency virus 1\]](#)

1435 aa protein

Accession: UED13371.1 GI: 2128090702

Nucleotide Taxonomy

GenPept Identical Proteins FASTA Graphics

3. [Gag-Pol \[Human immunodeficiency virus 1\]](#)

1429 aa protein

Accession: AAF13061.1 GI: 6466846

Nucleotide PubMed Taxonomy

GenPept Identical Proteins FASTA Graphics

4. [Gag-Pol, partial \[HIV-1 M.CRF02\\_AG/A3 MAU2140\]](#)

364 aa protein

Accession: CD127924.1 GI: 699030849

Nucleotide Taxonomy

Select Gag-Pol (Human immunodeficiency virus 1).

Protein

Protein

Search

Advanced Help

GenPept

Send to: Change region shown

Customize view

Analyze this sequence

Run BLAST

Protein 3D Structure

HIV-1 M184V reverse transcriptase-DNA complex

PDB: 6UKD

Source: HIV-1 M.B. HXB2R, Human immunodeficiency virus 1

Method: X-ray Diffraction

Resolution: 2.75695 Å

See all 38 structures...

Gag-Pol [Human immunodeficiency virus 1]

NCBI Reference Sequence: NP\_057849.4

Identical Proteins FASTA Graphics

Go to: ☺

LOCUS NP\_057849 1435 aa linear VRL 13-AUG-2018

DEFINITION Gag-Pol [Human immunodeficiency virus 1].

ACCESSION NP\_057849

VERSION NP\_057849.4

DBLINK BioProject: PRJNA485481

DBSOURCE REFSEQ: accession NC\_001802.1

KEYWORDS RefSeq.

SOURCE Human immunodeficiency virus 1 (HIV-1)

ORGANISM Human immunodeficiency virus 1

Viruses; Riboviria; Paramyovirales; Artverviricota; Revtraviricetes; Ortervirales; Retroviridae; Orthoretrovirinae; Lentivirus.

REFERENCE 1 (residues 1 to 1435)

AUTHORS Pettit,S.C., Gulnik,S., Everitt,L. and Kaplan,A.H.

TITLE The dimer interfaces of protease and extra-protease domains influence the activation of protease and the specificity of GagPol cleavage

JOURNAL J. Virol. 77 (1), 366-374 (2003)

PUBMED 12477841

REFERENCE 2 (residues 1 to 1435)

AUTHORS Martozilio,B., Graf,R. and Dobberstein,B.

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]
MGRASVLSGGLDRWKIRLPGGKKYKLVHIVASRELERFAVNPGLLETSEGRQILGQLQPSLQT
GSEELRSLVNTATLYCVHQRIEIKDKTEALDKIEEQNKSKKAQAAADGHSNQVSQNPYIVQNIQG
QMVHQASPRTNAMKVVVEKAFSPVIFHSALSEGATPDQNLNTLVGGHQAANMLKETINEEAA
EWDVHPVHAGIAPGQHPREPRGSDIAGTTSTLQEQIWMNTNPPIPVGEIYKRWIILGNKIVRHYSPT
SILDRQGPKEPFRDYDRFYKTLRAEQASQEVKNHITETLLVQANPDCKTILKALGPAATLEENMTAC
QGVGPGHKARVLAEASQVNTSATIMHQGNFRNQRIKVKFCNCGKEGTARNCRAPKKGCHKCGKEG
HQKDCETERQANFLREDLAFLOKAREFSSEQTTRANSPTRELQVWGRDNVSPSEAGADQSTVSFNPQ
VTLWQRLPTVNIIGGQLEALDTGADDTVLEHSLPGRHKPMIIGIGGFIVRQYDQILIEICGMAI
GTVLVGPVNIIGRNLITQIGCTNFPISPIETVPVKLPQGDGPKVQNPLETEEKALVEICTEHEK
EGKISIGPENPYNTPVFAIKKDDSTKWRKLVDFRELNRKQDFHEVQIGIPHPAGLKKKSVTVLDVGD
AYFSVPLDEDFRKYTAFIPISINNETPGIRYQVNLQGWKGSFQSGSNTKILEPFRKQNPDIIVYQY
HDDLTVGSQLEIGQRTKIEELRQHLRWGLTTPDKKHQKEPFLWNGYELHPDKWTVQPIVLPKDSWT
VNDIQKLVGKLNWASQIYVPGIKVRQLCKLLRGTKALTEVIPLTEAELELAENREILKEPVHGVYDPSK
DLIAEQKQSGSQNTYQIQEPFKNLTKGVARHGAHTNDVKQLTEAVQITTESIVWKGTPKFKLP
QKETHETWTEYQATWIPHEFVNTPLVWLVQLEKPIVGAETFYVDSAGANRETKLGKAGVNTNRG
QKQVTLDTTQKTELQATYALQDQSGLEVNIVTDSQVAGIIQAPQDSSESELVHQIIEQLIKKEKVV
AIVPAHKGIGGQEVQDLVLSAGIRKVLFLDSIDKAQDEHEKYNHRAHSDFNLPVVAKEIVASCDIC
QLKGAPHQDQSGISGICQDCTHLEKVLVAVHVASGYEAEVZPAETGQETAYVFLKLAGRPVKTI
HTQMSNFTGATVRAACVQDCTHLEKVLVAVHVASGYEAEVZPAETGQETAYVFLKLAGRPVKTI
NFKRNGSGTGSYAGERIVDIIAQITKELQKQITKIQNFRVYRDSRNPWLKGPALKLWKGEGAVIQD
NSDIKVVPRRAKIRDYGKQWAGDCCVASRQED
```

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.

Descriptions	Graphic Summary	Alignments	Taxonomy
Sequences producing significant alignments			
<div>Download</div> <div>Select columns</div> <div>Show 100</div>			
<input type="checkbox"/> select all 3 sequences selected			
Description	Scientific Name	Max Score	Total Score
<input type="checkbox"/> reverse transcriptase domain-containing protein [Terrabacteria group]	Terrabacteria group	1885	1885
<input type="checkbox"/> reverse transcriptase domain-containing protein [Terrabacteria group]	Terrabacteria group	1748	1748
<input type="checkbox"/> reverse transcriptase domain-containing protein [Terrabacteria group]	Terrabacteria group	1748	1748
<input checked="" type="checkbox"/> Gag-Pol [Simian immunodeficiency virus]	Simian immunodeficiency virus	1687	1687
<input checked="" type="checkbox"/> gag-pol fusion polyprotein [Human immunodeficiency virus 2]	Human immunodeficiency virus 2	1675	1675
<input checked="" type="checkbox"/> pol protein [Simian immunodeficiency virus SIV_mnd 2]	Simian immunodeficiency virus SIV_mnd 2	1377	1377
<input type="checkbox"/> pol protein [Enterococcus faecium]	Enterococcus faecium	1083	1083
<input type="checkbox"/> DDE-type integrase/transposase/recombinase [Escherichia coli]	Escherichia coli	1074	1074
<input type="checkbox"/> pol protein [Enterococcus faecium]	Enterococcus faecium	1047	1047
<input type="checkbox"/> pol protein [Klebsiella pneumoniae]	Klebsiella pneumoniae	1028	1028
<input type="checkbox"/> reverse transcriptase domain-containing protein [Enterococcus faecium]	Enterococcus faecium	999	999
<input type="checkbox"/> reverse transcriptase domain-containing protein [Thalassobius mangrove]	Thalassobius mangrove	996	996
<input type="checkbox"/> pol protein [Listeria monocytogenes]	Listeria monocytogenes	995	995
<input type="checkbox"/> reverse transcriptase domain-containing protein [Escherichia coli]	Escherichia coli	933	933
<input type="checkbox"/> reverse transcriptase domain-containing protein [Helicobacter acinonychis]	Helicobacter acinonychis	919	919
<input type="checkbox"/> pol protein [Escherichia coli]	Escherichia coli	910	910
<input type="checkbox"/> reverse transcriptase domain-containing protein [Acinetobacter baumannii]	Acinetobacter baumannii	884	884

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

You can safely 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.