



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

Semester II 2021/2022

Subject : Bioinformatics I (SECB2103)

Section : 01 – Dr Haslina Hashim

Topic : Class Project – Find a gene

Group : Group 4 - Heong Yi Qing (A20EC0043)

Chong Kah Wei (A20EC0027)

Goh Yitian (A20EC0038)

1. Tell me the name of a protein you are interested in. Include the species and the accession number.

Gag-Pol [Human immunodeficiency virus 1] (AAF13061.1)

2. Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC). It is not necessary to print out all of the blast results if there are many pages.

Figure 1 shows the performance of TBLASTN search against homosapiens ESTs using AAF13061.1 as a query. Some of the results are then shown in Figure 2 and the results are arranged from lower E-value to higher E-value. From the results, we choose a match which is IL3-MT0267-261200-410-C08 MT0267 Homo sapiens cDNA, mRNA sequence. For this match, the E-value is 2e-20 and the score is 96.7bits(239) as shown in Figure 3.

NCBI Blast(t) - Protein Sequence

blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST® » tblastn

Translated BLAST: tblastn

blastn blastp blastx **tblastn** tblastx

TBLASTN search translated nucleotide databases using a protein query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange From To

Or, upload file No file chosen

Job Title Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database

Organism ☐ exclude

Exclude ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to ☐ Sequences from type material

Entrez Query

Search database est using tblastn (search translated nucleotide databases using a protein query)

☐ Show results in a new window

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

+ Algorithm parameters

Figure 1: TBLASTN using AAF13061.1 as query

NCBI Blast(t) - Protein Sequence

blast.ncbi.nlm.nih.gov/Blast.cgi#sort_mark

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"/> DKFZ688K0981.1_1.688 (synonym_hlcc3) Homo sapiens cDNA clone DKFZ688K0981.1 mRNA sequence	Homo sapiens	121	121	14%	2e-28	33.65%	736	U09809.1
<input checked="" type="checkbox"/> DA092449 BRACE3 Homo sapiens cDNA clone BRACE3000762.5 mRNA sequence	Homo sapiens	119	119	12%	2e-28	37.93%	582	U09244.1
<input checked="" type="checkbox"/> HES2_35_F01.g1_A035 NIH_MGC_258 Homo sapiens cDNA clone IMAGE 7469715.5 mRNA sequence	Homo sapiens	120	120	11%	9e-28	38.01%	816	U09029.1
<input checked="" type="checkbox"/> ts8e05.x1 NCL CGAP_GC6 Homo sapiens cDNA clone IMAGE 2238464.3 similar to gb.M14123_cds4 RETROVIRU	Homo sapiens	112	112	12%	7e-26	34.44%	631	U03743.1
<input checked="" type="checkbox"/> DA170606 BRAMY2 Homo sapiens cDNA clone BRAMY2032675.5 mRNA sequence	Homo sapiens	112	112	12%	8e-26	35.00%	578	U17060.1
<input checked="" type="checkbox"/> bz72c07.x1 NCL CGAP_Lu24 Homo sapiens cDNA clone IMAGE 3213516.3 similar to gb.M14123_cds4 RETROVIR	Homo sapiens	110	110	12%	4e-25	33.33%	617	U04571.1
<input checked="" type="checkbox"/> DKFZ688B20225.1_1.688 (synonym_hlcc3) Homo sapiens cDNA clone DKFZ688B20225.5 mRNA sequence	Homo sapiens	105	105	11%	3e-23	34.09%	610	U08121.1
<input checked="" type="checkbox"/> x35c11.x1 NCL CGAP_Lu24 Homo sapiens cDNA clone IMAGE 2271572.3 similar to gb.M14123_cds4 RETROVIRU	Homo sapiens	103	103	9%	3e-23	39.42%	503	U08033.1
<input checked="" type="checkbox"/> z28b04.x1 NCL CGAP_GC6 Homo sapiens cDNA clone IMAGE 2242135.3 similar to gb.M14123_cds4 RETROVIRU	Homo sapiens	102	102	11%	2e-22	32.75%	586	U03818.1
<input checked="" type="checkbox"/> bz21e05.x1 NCL CGAP_GC6 Homo sapiens cDNA clone IMAGE 3208640.3 similar to gb.M14123_cds4 RETROVIRU	Homo sapiens	102	102	8%	6e-22	41.80%	643	U06642.1
<input checked="" type="checkbox"/> DB152087 THYMJ3 Homo sapiens cDNA clone THYMJ3029227.5 mRNA sequence	Homo sapiens	97.4	97.4	9%	1e-20	38.81%	595	U15208.1
<input checked="" type="checkbox"/> IL3-MT0267-261200-410-C08 MT0267 Homo sapiens cDNA mRNA sequence	Homo sapiens	96.7	96.7	14%	2e-20	31.88%	595	U05435.1
<input checked="" type="checkbox"/> TK73e04.x1 NCL CGAP_GC6 Homo sapiens cDNA clone IMAGE 3481062.3 similar to TR.O92151.O92151 POLYME	Homo sapiens	95.1	95.1	7%	1e-19	42.59%	649	U06218.1
<input checked="" type="checkbox"/> naa38g04.x1 NCL CGAP_Kid11 Homo sapiens cDNA clone IMAGE 3258799.3 similar to TR.O92151.O92151 POLYME	Homo sapiens	92.8	92.8	9%	1e-19	37.14%	445	U58044.1
<input checked="" type="checkbox"/> np11a02.s1 NCL CGAP_Phe1 Homo sapiens cDNA clone IMAGE 1100330.3 similar to gb.M14123_cds4 RETROVIR	Homo sapiens	90.9	90.9	8%	3e-19	38.46%	388	U08417.1
<input checked="" type="checkbox"/> pc89d02.s1 NCL CGAP_GC61 Homo sapiens cDNA clone IMAGE 1356867.3 similar to gb.M14123_cds4 RETROVIR	Homo sapiens	90.5	90.5	10%	7e-19	36.11%	425	U40931.1
<input checked="" type="checkbox"/> Za39g04.x1 NCL CGAP_GC6 Homo sapiens cDNA clone IMAGE 3220278.3 similar to gb.M14123_cds4 RETROVIRU	Homo sapiens	93.2	93.2	11%	8e-19	31.88%	673	U55063.1
<input checked="" type="checkbox"/> DB299051 BRACE3 Homo sapiens cDNA clone BRACE3011740.3 mRNA sequence	Homo sapiens	89.7	89.7	10%	6e-18	35.42%	584	U29905.1
<input checked="" type="checkbox"/> 602820846F1 NCL CGAP_Skn3 Homo sapiens cDNA clone IMAGE 4746370.5 mRNA sequence	Homo sapiens	90.5	90.5	9%	1e-17	36.96%	769	U06746.1

Figure 2: Part of the results

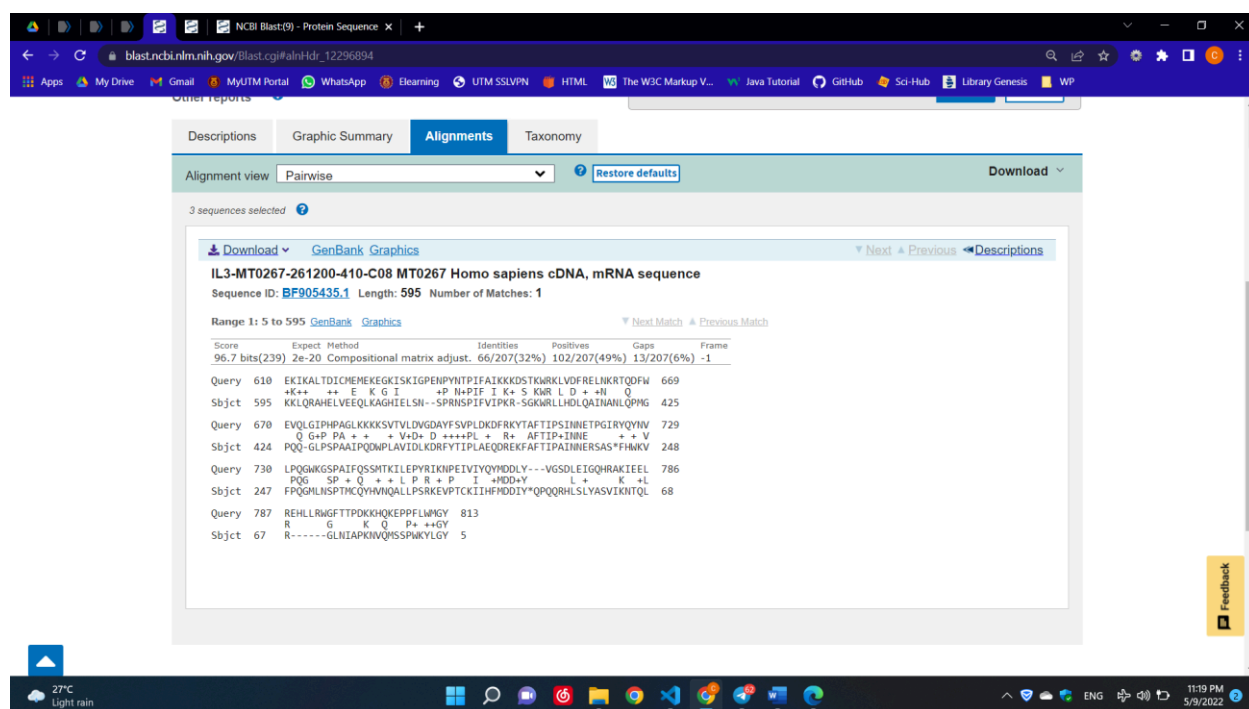


Figure 3: Alignment using IL3-MT0267-261200-410-C08 MT0267 Homo sapiens cDNA, mRNA sequence as subject

>IL3-MT0267-261200-410-C08 MT0267 Homo sapiens cDNA, mRNA sequence

Sequence ID: BF905435.1 Length: 595

Range 1: 5 to 595

Score:96.7 bits(239), Expect:2e-20,

Method:Compositional matrix adjust.,

Identities:66/207(32%), Positives:102/207(49%), Gaps:13/207(6%)

Query 610 EKIKALTDICMEMEKEGKISKIGPENPYNTPIFAIKKKDSTKWRKLVDFRELNKRTQDFW 669
+K++ ++ E K G I +P N+PIF I K+ S KWR L D + +N Q

Sbjct 595 KKLQRAHELVEEQLKAGHIELSN--SPRNSPIFVIPKR-SGKWRLHDLQAINANLQPMG 425

Query 670 EVQLGIPHPAGLKKKKSVTVLDVGDVYFSVPLDKDFRKYTAFTIPSINNETPGIRYQYNV 729
Q G+P PA + + + V+D+ D +++++PL + R+ AFTIP+INNE + + V

Sbjct 424 PQQ-GLPSPAIPQDWPLAVIDLKDRFYTIPLAEQDREKFAFTIPAINNERSAS*FHWKV 248

Query 730 LPQGWKGSPAIFQSSMTKILEPYRIKNPEIVYQYMDDLY---VGSdleIGQHRAKIEEL 786
PQG SP + Q + + L P R + P I +MDD+Y L + K +L

Sbjct 247 FPQGMLNSPTMCQYHVNQALLPSRKEVPTCKIIHFMDDIY*QPQQRHLSLYASVIKNTQL 68

Query 787 REHLLRWGFTTPDKKHQKEPPFLWMGY 813
R G K Q P+ ++GY

Sbjct 67 R-----GLNIAPKNVQMSSPWKYLGY 5

3. Gather information about this “novel” protein. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from step [2]. In some cases, you will be able to do further BLAST searches to obtain even more sequence of your novel gene.

Name: Homo sapiens cDNA

Organism: Homo sapiens

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

```
>Homo sapiens cDNA, mRNA sequence (Sequence taken from blast result)
KKLQRAHELVEEQLKAGHIELSN--SPRNSPIFVIPKR-SGKWRLHDLQAINANLQPMG
PQQ-GLPSPA AIPQDWPLAVIDLKDRFYTIPLAEQDREKFAFTIPAINNERSAS*FHWKV
FPQGMLNSPTMCQYHVNQALLPSRKEVPTCKIIHFMDDIY*QPQQRHLSLYASVIKNTQL
R-----GLNIAPKNVQMSSPWKYLGY
```

4. Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Take the protein sequence (step [3]) and use it as a query in a blastp search of the nr database at NCBI.

Figure 4 shows that the BLASTP search is performed using the protein sequence that we obtained earlier from the TBLASTN result as query, which is:

```
>Homo sapiens cDNA, mRNA sequence (Sequence taken from blast result)
KKLQRAHELVEEQLKAGHIELSN--SPRNSPIFVIPKR-SGKWRLHDLQAINANLQPMG
PQQ-GLPSPA AIPQDWPLAVIDLKDRFYTIPLAEQDREKFAFTIPAINNERSAS*FHWKV
FPQGMLNSPTMCQYHVNQALLPSRKEVPTCKIIHFMDDIY*QPQQRHLSLYASVIKNTQL
R-----GLNIAPKNVQMSSPWKYLGY
```

The results are then shown in Figure 5 and the result is arranged from higher percent identity to lower percent identity. Figure 6 shows the top results, which is pol protein from human endogenous retrovirus K.

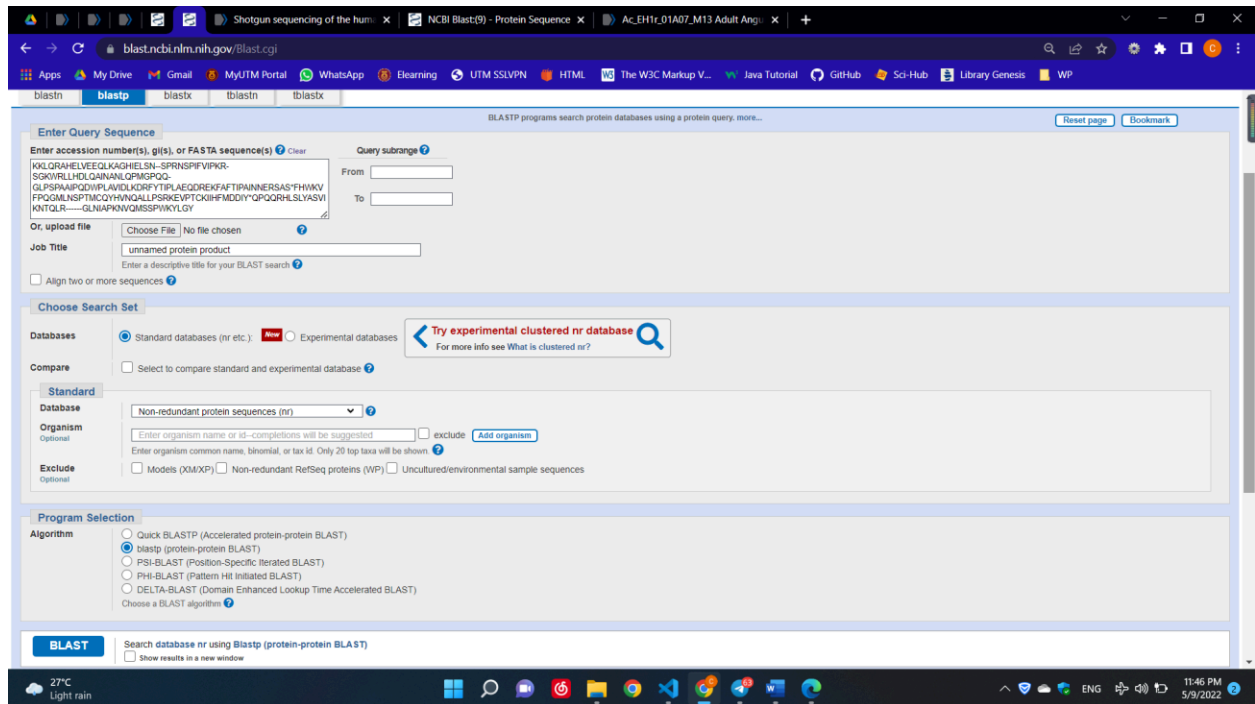


Figure 4: BLASTP search using the protein sequence obtained earlier

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
pol protein [Human endogenous retrovirus K]	Human endogen	213	213	78%	2e-66	66.67%	222	CA571420.1
RecName: Full=Endogenous retrovirus group K member 8 Pol protein; AltName: Full=HERV-K115 Pol protein...	Human sapiens	231	231	99%	1e-66	58.29%	956	PS3133.1
RecName: Full=Endogenous retrovirus group K member 7 Pol protein; AltName: Full=HERV-K(III) Pol protein...	Human sapiens	229	229	99%	2e-65	58.29%	1459	PS3135.1
reverse transcriptase [Homo sapiens]	Homo sapiens	228	228	99%	3e-68	57.79%	573	AB091576.1
polymerase [Homo sapiens]	Homo sapiens	229	229	99%	4e-68	57.79%	596	AAC63291.1
RecName: Full=Endogenous retrovirus group K member 25 Pol protein; AltName: Full=HERV-K_11a22.1 provir...	Homo sapiens	229	229	99%	7e-66	57.79%	954	PS3136.1
polymerase [Homo sapiens]	Homo sapiens	228	228	99%	7e-66	57.79%	956	AAK11553.1
polymerase [Homo sapiens]	Homo sapiens	228	228	99%	8e-66	57.79%	956	AAK11554.1
RecName: Full=Endogenous retrovirus group K member 113 Pol protein; AltName: Full=HERV-K113 Pol protei...	Homo sapiens	228	228	99%	1e-65	57.79%	956	PS3132.1
RecName: Full=Endogenous retrovirus group K member 11 Pol protein; AltName: Full=HERV-K_3a27.2 proviru...	Homo sapiens	228	228	99%	1e-65	57.79%	969	Q8UQ60.2
RecName: Full=Endogenous retrovirus group K member 6 Pol protein; AltName: Full=HERV-K(C7) Pol protei...	Homo sapiens	227	227	99%	3e-65	57.79%	956	Q98XR3.2
Gag-Pro-Pol protein [Homo sapiens]	Homo sapiens	227	227	99%	7e-65	57.79%	1177	AA051791.1
Gag-Pro-Pol-Env protein [Homo sapiens]	Homo sapiens	226	226	99%	1e-64	57.79%	2294	AA051793.1
Gag-Pro-Pol protein [Homo sapiens]	Homo sapiens	226	226	99%	2e-64	57.79%	1879	AA051797.1
polymerase [Homo sapiens]	Homo sapiens	196	196	83%	5e-60	57.74%	198	AA063113.1
polymerase [Homo sapiens]	Homo sapiens	227	227	99%	2e-67	57.29%	597	AAC63290.1
polymerase [Homo sapiens]	Homo sapiens	226	226	99%	3e-67	57.29%	572	AAC63292.1
polymerase [Homo sapiens]	Homo sapiens	225	225	99%	7e-67	57.29%	572	AAC63294.1
pol protein [Human endogenous retrovirus K]	Human endogen	225	225	99%	1e-65	57.29%	740	CA571417.1
RecName: Full=Endogenous retrovirus group K member 18 Pol protein; AltName: Full=HERV-K(C1a) Pol prote...	Homo sapiens	226	226	99%	2e-65	57.29%	812	Q90C97.2
RecName: Full=Endogenous retrovirus group K member 10 Pol protein; AltName: Full=HERV-K10 Pol protei...	Homo sapiens	224	224	99%	4e-64	57.29%	1014	P10266.2
Gag-Pro-Pol protein [Homo sapiens]	Homo sapiens	224	224	99%	1e-63	57.29%	1755	AA051796.1
polymerase [Homo sapiens]	Homo sapiens	194	194	83%	3e-59	57.14%	196	AA063114.1

Figure 5: Top of BLASTP results which are arranged according to the percent Identity

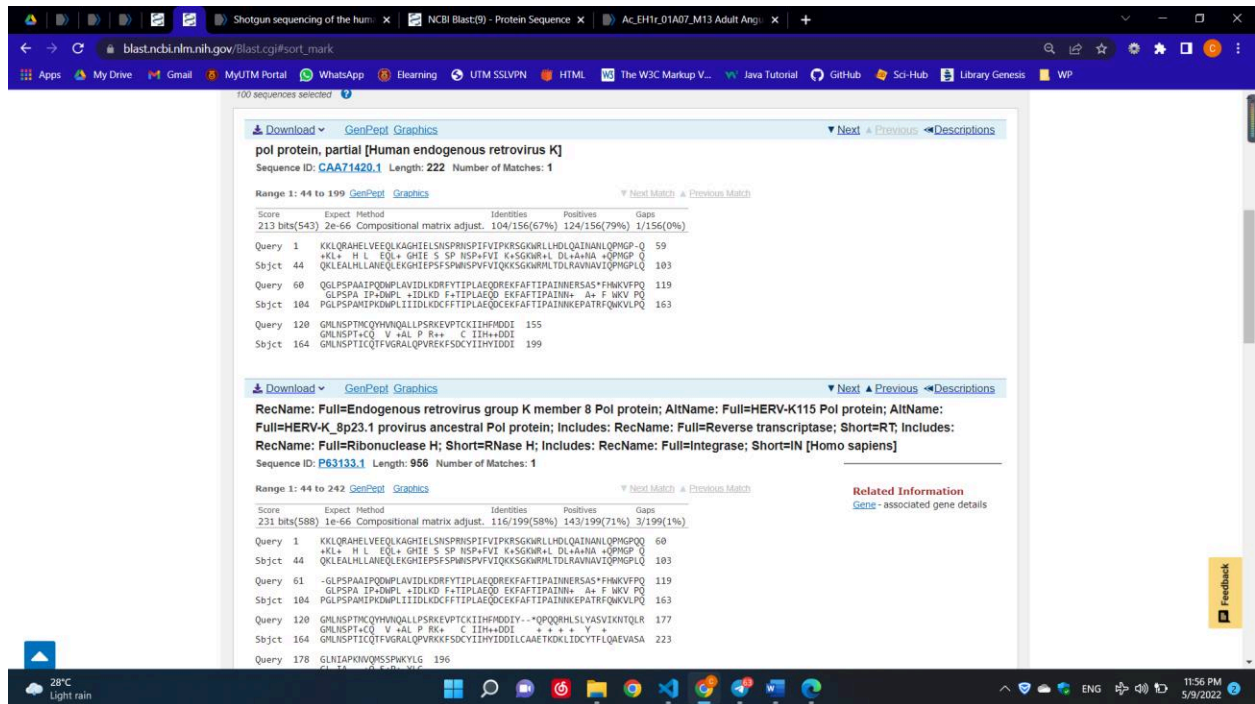


Figure 6: Top result

The top result is only 67% percent identity match to the query and the others are less than that.

Since there is no sequence with 100% percent identity and all sequences are actually less than 100% percent identity matched to the query sequence, there is a possibility that the protein is a novel protein.

5. Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family. A typical number of proteins to use in a multiple sequence alignment is a minimum of 5 or 10 and a maximum 30, although the exact number is up to you.

Firstly, we use the sequence found in Q3 to do the BLASTP and set the database to reference protein (refseq protein) to generate the protein sequences for multiple alignments.

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastp suite

Standard Protein BLAST

blastn **blastp** blastx tblastn tblastx

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

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange

From To

Or, upload 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 ☐ exclude [Add organism](#)

Exclude ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WPI) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm ☒ blastp (protein-protein BLAST) ☐ PSI-BLAST (Position-Specific Iterated BLAST) ☐ PHI-BLAST (Pattern Hit Initiated BLAST) ☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST Search database refseq_protein using Blastp (protein-protein BLAST) ☐ Show results in a new window

Figure 7: The setting for BLASTP on novel protein

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastp suite » results for RID-AKCB7WU801R

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 [AKCB7WU801R](#) Search expires on 06-16 18:16 pm [Download All](#)

Program **BLASTP** [Citation](#)

Database **refseq_protein** [See details](#)

Query ID **Id|Query_815936**

Description **unnamed protein product**

Molecule type **amino acid**

Query Length **197**

Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#)

Filter Results

Organism ☐ exclude [Add organism](#)

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

☐ select all 0 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input type="checkbox"/>	PREDICTED: putative HERV-K_Xo28 provirus ancestral Pol protein [Ooisthocomus hoazin]	Ooisthocomus...	212	212	100%	1e-64	53.50%	350	XP_009940053.1
<input type="checkbox"/>	PREDICTED: endogenous retrovirus group K member 11 Pol protein-like [Haliaeetus leucocorhalus]	Haliaeetus leu...	216	216	99%	6e-64	55.28%	535	XP_010579894.1
<input type="checkbox"/>	PREDICTED: LOW QUALITY PROTEIN: endogenous retrovirus group K member 9 Pol protein-like [...]	Condylura cris...	214	214	100%	2e-62	53.96%	637	XP_012583238.1
<input type="checkbox"/>	PREDICTED: LOW QUALITY PROTEIN: endogenous retrovirus group K member 18 Pol protein-like [...]	Calidris eyonax	204	204	99%	6e-62	51.76%	302	XP_014809869.1
<input type="checkbox"/>	PREDICTED: LOW QUALITY PROTEIN: putative HERV-K_Xo28 provirus ancestral Pol protein [Taur...	Tauraco erythr...	202	202	100%	8e-62	51.76%	274	XP_009989166.1
<input type="checkbox"/>	PREDICTED: endogenous retrovirus group K member 18 Pol protein-like [Calidris eyonax]	Calidris eyonax	207	207	99%	2e-61	51.76%	473	XP_014809364.1
<input type="checkbox"/>	PREDICTED: endogenous retrovirus group K member 11 Pol protein-like [Colinus striatus]	Colinus striatus	209	209	100%	3e-61	54.27%	570	XP_010202843.1
<input type="checkbox"/>	PREDICTED: uncharacterized protein LOC103583026 [Galeoscoptes variegatus]	Galeoscoptes v...	209	209	99%	5e-61	53.77%	617	XP_008562675.1
<input type="checkbox"/>	PREDICTED: LOW QUALITY PROTEIN: endogenous retrovirus group K member 18 Pol protein-like [...]	Miniopterus na...	209	209	100%	1e-59	56.50%	854	XP_016079066.1

Figure 8: The result of BLASTP on novel protein

From the result, we choose 5 sequences from different family

Descriptions	Graphic Summary	Alignments	Taxonomy					
Sequences producing significant alignments								
Download Select columns Show 100 ?								
<input type="checkbox"/> select all 5 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"/> PREDICTED: putative HERV-K_Xg28 provirus ancestral Pol protein [Opisthocomus hoazin]	Opisthocomus...	212	212	100%	1e-64	53.50%	350	XP_009940053.1
<input checked="" type="checkbox"/> PREDICTED: endogenous retrovirus group K member 11 Pol protein-like [Haliaeetus leucocephalus]	Haliaeetus leu...	216	216	99%	6e-64	55.28%	535	XP_010579894.1
<input type="checkbox"/> PREDICTED: LOW QUALITY PROTEIN: endogenous retrovirus group K member 9 Pol protein-like [...]	Condylura cris...	214	214	100%	2e-62	53.96%	637	XP_012583238.1
<input type="checkbox"/> PREDICTED: LOW QUALITY PROTEIN: endogenous retrovirus group K member 18 Pol protein-like [...]	Calidris pugnax	204	204	99%	6e-62	51.76%	302	XP_014809869.1
<input type="checkbox"/> PREDICTED: LOW QUALITY PROTEIN: putative HERV-K_Xg28 provirus ancestral Pol protein [Taur...	Tauraco erythr...	202	202	100%	8e-62	51.76%	274	XP_009989166.1
<input checked="" type="checkbox"/> PREDICTED: endogenous retrovirus group K member 18 Pol protein-like [Calidris pugnax]	Calidris pugnax	207	207	99%	2e-61	51.76%	473	XP_014809364.1
<input type="checkbox"/> PREDICTED: endogenous retrovirus group K member 11 Pol protein-like [Colius striatus]	Colius striatus	209	209	100%	3e-61	54.27%	570	XP_010202843.1
<input type="checkbox"/> PREDICTED: uncharacterized protein LOC103583026 [Galeopterus variegatus]	Galeopterus v...	209	209	99%	5e-61	53.77%	617	XP_008562675.1
<input type="checkbox"/> PREDICTED: LOW QUALITY PROTEIN: endogenous retrovirus group K member 18 Pol protein-like [...]	Miniopterus na...	209	209	100%	1e-59	56.50%	854	XP_016079066.1
<input type="checkbox"/> PREDICTED: LOW QUALITY PROTEIN: endogenous retrovirus group K member 18 Pol protein-like [...]	Odobenus ros...	199	199	99%	2e-59	49.25%	346	XP_012422424.1
<input type="checkbox"/> PREDICTED: endogenous retrovirus group K member 11 Pol protein-like [Tinamus guttatus]	Tinamus guttatus	202	202	100%	2e-59	49.76%	457	XP_010223024.1
<input checked="" type="checkbox"/> PREDICTED: endogenous retrovirus group K member 25 Pol protein-like [Oryctolagus cuniculus]	Oryctolagus c...	195	195	98%	4e-59	51.01%	257	XP_017200008.1
<input checked="" type="checkbox"/> PREDICTED: endogenous retrovirus group K member 113 Pol protein-like [Tinamus guttatus]	Tinamus guttatus	204	204	99%	4e-59	53.03%	580	XP_010212357.1
<input type="checkbox"/> PREDICTED: endogenous retrovirus group K member 11 Pol protein-like [Tauraco erythrolophus]	Tauraco erythr...	199	199	100%	1e-58	51.76%	418	XP_009976657.1
<input type="checkbox"/> PREDICTED: LOW QUALITY PROTEIN: endogenous retrovirus group K member 18 Pol protein-like [...]	Odobenus ros...	200	200	99%	2e-58	50.75%	489	XP_012422904.1
<input type="checkbox"/> PREDICTED: endogenous retrovirus group K member 18 Pol protein-like [Tauraco erythrolophus]	Tauraco erythr...	196	196	99%	2e-58	51.76%	333	XP_009979359.1
<input type="checkbox"/> PREDICTED: LOW QUALITY PROTEIN: endogenous retrovirus group K member 25 Pol protein-like [...]	Aoteryx mante...	193	193	100%	2e-58	47.00%	247	XP_013799745.1

Figure 9: Choose 5 BLASTP results from different family

Our original query protein:

>AAF13061.1 Gag-Pol [Human immunodeficiency virus 1]

MGARVSVLRGGQLDTWEKIRLRPGGKKKYKMKLLVWASRELERFAVNPGLLTTEGCQQIILEQLQPALKA
 GSEELKSLYNTVATLYCVHQIDVRDTKEALDKLEEIQNKSKQKTQQAANSQVSQNYPIVQNAQGQMVH
 QAISPRTLNAWVKVVEEKAFSPEVIMFTALSEGATPQDLNMMNLIVGGHQAMQMLKDTINEEAAEWDR
 THPIHAGPNPPGQMREPRGSDIAGTTSNLQEQIAWMTGNPPIPVGEIYKRWIVLGLNKIVRMYSFVGILD
 IRQGPKEPFRDYVDRFFKTLRAEQATQEVKNWMTETLLVQNANPDCKTILRALGPGATLEEMMTACQGVG
 GPGHKARVLAEAMSQVQSPNILMQRGNFKGQKRIKCFNCGKEGHLARNCRAPRKKGCWKCGKEGHQMKDC
 TERQANFLRENLAQQREARELSSEQTGAISPTGRELWDKGRNNLLSAAGTEGQGTISSFNFPQITLWQR
 PLVTVRIGGQLIEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVLVG

PTPVNIIGRNMLTQIGCTLNFPISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALTDICMEMEKEGKISK
 IGPENPYNTPIFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVP
 LDKDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPYRIKNPEIVYQYMDDLIV
 GSDLEIGQHRAKIEELREHLLRWGFTTPDKKHQKEPPFLWMGYELHPDKWTVQPIMLPEKDSWTVNDIQK
 LVGKLNWASQIYAGIKVKELCKLLRGAKALTDIVTLTEEALELAENREILKEPVHGVYDPTKDLIAEI
 QKQGQDQWTFQIYQEPFKNLKTGKYQERVAPYDLSITELTEVVQKVTTESIIIWGKTPKFRLP IQRETWE
 AWWMEYWQATWIPEWEFVNTLPLVKLWYQLEKDP IVGAETFYVDGAANRETKLGKAGYVTDGRGRQKVVSL
 TETTNQKTELHAIHLALQDSGSEVNIVTDSQYALGIIQAQPDRSESELVNQIIEKLIGKDKVYLSWVPAH
 KGIGGNEQVDNLVSSGFRKILFLDGLDKAQEEHEKFHSNWRAMASDFNLPPIVAKEIVASCDKCQLKGEA
 MHGQVDCSPGIWQLDCTHLEGKIIILVAVHVASGYIEAEVIPAETGQETAYFILKLAGRWPVKIIHTDNBS
 NFTAACVRAACWWANVTQEFQIPYNPQSQGVVESMNKELKKIIGQVRDQAEHLKTAVQMAVFIHNFKRKG
 GIGGYSAGERIIDIIASDIQTKELQKQITKIQNFRVYYRDSRDPIWKGPALLWKGEAVVIQDNSDIKV
 VPRRKAKILRDYQKQ MAGDDCVAGRQDED

Our novel protein:

>Homo sapiens cDNA, mRNA sequence (Sequence taken from blast result)

KKLQRAHELVEEQLKAGHIELSN--SPRNSPIFVIPKR-SGKWRLLDLQAINANLQPMG
 PQQ-GLPSPAAPQDWPLAVIDLKDRFYTIPLAEQDREKFAFTIPAINNERSAS*FHWKV
 FPQGMLNSPTMCQYHVNQALLPSRKEVPTCKIIHFMDDIY*QPQQRHLSLYASVIKNTQL
 R-----GLNIAPKKNVQMSSPWKYLGY

Below is the list of 5 proteins sequences selected by our group for the Multiple Sequence Alignment:

>XP_009940053.1 PREDICTED: putative HERV-K_Xq28 provirus ancestral Pol protein
 [Opisthocomus hoazin]

MKNKPEIEVMVESQNQEMLKLKMMIDTGADVTIISAPHWPSHWPTVESFTGVYAFLGAAIEQQPV LKIKWKTNNPFWVD
QWPLTAERLQKISELVEEQ LQVGHIQPSTSPWNTPIFTIPKNSGKWRLLYDLRAVNNVMEDMGALQPGLPSPVMIPENW
TVLVIDLKDCFFTIPLHPDDAERFAFSVPSVNKEEPARRFHWIVLPQGMKNSPTMCQIFVAWAFQPICKKMPQLLIYHY
MDDILIAGQNMDREFVLQEVVREVESRGLNIAPEKIQKQEPWNYLGWVILQGSIKPPKMQLNPEIKTLNDVQKLVGDIQ
WVRTLCDITNDDLAPLVELLGTTSRADDKRTMEP

>XP_010579894.1 PREDICTED: endogenous retrovirus group K member 11 Pol
protein-like [*Haliaeetus leucocephalus*]

MGGENYHGFLTRATVLQGV RQPTLVLTWLTNNPVWVDQWPLPIEKLKALQELVAEQ LAAGHIEPSQSPWNTPVFVIKKK
SGKWRFLHDLQQVNAV MATMGALQPGMPSPAMIPQDWEIIVMDLKDCFFTIPLASQDKEKFAFSVPSINHAEP AKRYQW
RVLPQGMKNSPTICQWFVAQALSPVREKFPTSYCYHYMDDILLASDNKEQLNDMENLARNLLQQYGLVIAPEKVQKIAP
WKYLGMTITSKQVVPQPVKLNLA VKTLNDVQKLMGSLNWIRPYLGLTNSQLQPLDLLKHSNDPTEPRILNKEALNVIH
MVEQCIYKKFVSRIDLSQLVQFFVLIDKTVPF GALVQWNSEWDDPLHILEWMFLSFRPRKTASGLFELIADV IIKTRKR
CVELIGRDPATIVLPVQNWYFEWCLANNYELQVAMAGFQRQISYHLPSHLLLKFAQEIPFGQKYL SQPEPVKGPTVFTD
GSGKTGKAAVVC LLQRALQRAAGTIFLAQIQKGGIVGGCNGSAECLTVREQRDLEQLVVYP

>XP_014809364.1 PREDICTED: endogenous retrovirus group K member 18 Pol
protein-like [*Calidris pugnax*]

MGLFVLPGIIDADFTGEIKIMAWTPSPPCFVPKGQRIALVPLPSVTIPGEGNRKGGFGSTGKPVVLWSKQVSKEQPLL
RCQVHDRHFSGLVDTGADVTIINISDWPEWFLRDPTSAIVGVGGLQKPKQSAKILTFKGPEGQIAHAAPYILPVPCTL
WGCALYIMGPRLVKPVGNLPENKFSIGAIDKQQTLSLTWKSEKTVWVDQWPLKKDKLLHLHDLVQEQLAAGHIVPTTSP
WNTPVFVIPKKGGRWRLI HDLRAINAVIEPMGALQPGLPSPSMLPQNWP IAIIDLKDCFFAIP LHPKDAPRFAFSVPAV
NQEQPTRRYHWTVLPQGM LNSPTICQLTVANALQPV RNANPHVIIYHYMDDILIAAEKDKDLQIVVCQVRLAVQGAGLQ
IAEEKVQREPPWKYL GWKITTHPIQPQALQLALQIKTLNDLQKLLGTINWLRPFLGITTQDLHPLFQLLPLTNRTL CI

>XP_017200008.1 PREDICTED: endogenous retrovirus group K member 25 Pol
protein-like, partial [*Oryctolagus cuniculus*]

SSAEAFSPRAGFFIGAAEEGIPITWKHEDPVWVPQWPLSSDKLVAAQQLIQEQLNLGHIRPSVSPWNTPIFVIKKKSGK
WRLHDLRVINMQMQVMGPVQRGLP LLSALPQGWPIIIIDIKDCFFSIPLHTKDCERFAFTLPACNHEQPDQRYEWVVL
PQGMANSPTMCQLFVGQAIGPLRKRFSSSLKCIHYMDDILLAAKDEFVLDDGFAYLIQLLESKKLFIAPEKVQKGS IATY
LGSCITSTQLFLKGGIAQDS

>XP_010212357.1 PREDICTED: endogenous retrovirus group K member 113 Pol
protein-like [*Tinamus guttatus*]

MQGVGKREAERGSPPHNDTSTGCLERC SAATRG SAGVDVATAVDVMLTDTRVRVIESEL SGPLGQGLSALLLGRSSVSR
QGIFVVPGLIDADYTGVIKIMVYTSAPPVTIPAHSKIAQLIPFKACVPHDALTERCNGEFGSTGPLNMLLAIDIKKGKP

EELVRLQHPPGGQTITLTMVVDTGADVSIIPQHMPRAWPISLAATSMGVGGAQWTVDYSAFLVAAIGVQSTLKLTKWM
ETPVWVDQWPLKQDRHLHIVEQLVQEQLLEAGHIVPSQSPWNTPIFTIPKKSQKWRLLHDLRAINAVMQDMGALQPGLPSP
VMLPEGWDLIIIDLKDCFFTIALHPQDAEKFAFIVPSINKAAPAKRYHWVVLPPQGMKNSPTICQTFVAVWALQPVRAKHP
ELLIYHYMGDILIAGENMCMKSVFQEVGEELGKRGLTIAPEKVQRQGPWNYLGWTIMGSEIRPQKIAIRTEVRTLVVDVQ
RLVGDIQWVRGICGITNDDIAPLMPLLGTSVNASEARELSKEQREAVQAIAEKIAGAYASRIILEKKIWFLIVNSGGSH
QHFMGLIAQILQGLRTLHILEWVFLSY

Simplified original query protein, novel protein and 5 proteins of other family.

>Human immunodeficiency virus 1 AAF13061.1 Gag-Pol

MGARVSVLRGGQLDTWEKIRLRPGGKKKYKMKLLVWASRELERFAVNPGLLETTEGCQQIILEQLQPALKA
GSEELKSLYNTVATLYCVHQKIDVRDTKEALDKLEEIQNKSKQKTQQAANSQVSQNYPIVQNAQGQMVH
QAISPRTLNAWVKVVEEKAFSPEVIMFTALSEGATPQDLNMLNIVGGHQAMQMLKDTINEEAAEWDR
THPIHAGPNPPGQMREPRGSDIAGTTSNLQEQIAWMTGNPPIPVGEIYKRWIVLGLNKIVRMYSVPVGILD
IRQGPKEPFRDYVDRFFKTLRAEQATQEVKNWMTETLLVQANANPDCKTILRALGPGATLEEMMTACQGVG
GPGHKARVLAEAMSQVQSPNILMQRGNFKGQKRIKCFNCGKEGHLARNCRAPRKKGCWKCGKEGHQMKDC
TERQANFLRENLAFFQREARELSSEQTGAISPTGRELWDKGRNNLLSAAGTEGQGTISSFNFPQITLWQR
PLVTVRIGGQLIEALLDTGADDTVLEDINLPKWKPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVLVG
PTPVNIIGRNMLTQIGCTLNFPISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALTDICMEMEKEGKISK
IGPENPYNTPIFAIKKKDSTKWRKLVDFRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLVDVGDAYFSVP
LDKDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPYRIKNPEIVYQYMDDLVY
GSDLEIGQHRAKIEELREHLLRWGFTTPDKKHQKEPPFLWMGYELHPDKWTVQPIMLPEKDSWTVNDIQK
LVGKLNWASQIYAGIKVKELCKLLRGAKALTDIVTLTEEALELAENREILKEPVHGVYDPTKDLIAEI
QKQGQDQWTFQIYQEPFKNLKTGKYQERVAPYDLSITELTEVVQKVTTESIIIWGKTPKFRLPQIRETWE
AWWMEYWQATWIPEWEFVNTLPLVKLWYQLEKDPVGAETFYVDGAANRETKLGKAGYVTDGRGRQKVVS
TETTNQKTELHAIHLALQDSGSEVNIVTDSQYALGIIQAQPDSESELVNQIIEKLIKDKVYLSWVPAH
KGIGGNEQVDNLVSSGFRKILFLDGLDKAQEEHEKFHSNWRAMASDFNLPPIVAKEIVASCDKCQLKGEA
MHGQVDCSPGIWQLDCTHLEGKIIILVAVHVASGYIEAEVIPAETGQETAYFILKLAGRWPVKIIHTDN
NFTSAAVRAACWWANVTQEFGIPYNPQSQGVVESMNKELKKIIGQVRDQAEHLKTAVQMAVFIHNFKRKG
GIGGYSAGERIIDIIASDIQTKELQKQITKIQNFRVYYRDSRDPIWKGPAKLLWKGEAVVIQDNSDIKV

VPRRKAKILRDYGKQMAGDDCVAGRQDED

>(novel)Homo sapiens cDNA, mRNA sequence (Sequence taken from blast result)

KKLQRAHELVEEQLKAGHIELSN--SPRNSPIFVIPKR-SGKWRLLDLQAINANLQPMG

PQQ-GLPSPAIPQDWPLAVIDLKDRFYTIPLAEQDREKFAFTIPAINNERSAS*FHWKV

FPQGMLNSPTMCQYHVNQALLPSRKEVPTCKIIHFMDDIY*QPQQRHLSLYASVIKNTQL

R-----GLNIAPKNVQMSSPWKYLGY

>Opisthocomus hoazin XP_009940053.1 PREDICTED: putative HERV-K_Xq28 provirus
ancestral Pol protein

MKNKPEIEVMVESQNQEMLKLKMMIDTGADVTIISAPHWPSHWPTVESFTGVYAFLGAAIEQQPVLIKWKTNPNFVVD
QWPLTAERLQKISELVEEQLQVGHIQSTSPWNTPIFTIPKNSGKWRLLYDLRAVNNVMEDMGALQPGLPSPVMIPENW
TVLVIDLKDCFFTIPLHPDDAERFAFSVPSVNKEEPARRFWIVLPQGMKNSPTMCQIFVAWAFQPICKKMPQLLIYHY
MDDILIAGQNMDREFVLQEVVREVESRGLNIAPEKIQKQEPWNYLGWVILQGSIKPPKMQLNPEIKTLNDVQKLVGDIQ
WVRTLCDITNDDLAPLVELLGTTSRADDKRTMEP

>Haliaeetus leucocephalus XP_010579894.1 PREDICTED: endogenous retrovirus group
K member 11 Pol protein-like

MGGENYHGFLTRATVLQGVRQPTLVLTWLTNNPVWVDQWPLPIEKLKALQELVAEQLAAGHIEPSQSPWNTPVFVIKKK
SGKWRLHDLQQVNAVMTMGALQPGMPSPAMIPQDWEIIVMDLKDCFFTIPLASQDKEKFAFSVPSINHAEPKRYQW
RVLPQGMKNSPTICQWFVAQALSPVREKFPTSICYHYMDDILLASDNKEQLNDMENLARNLLQQYGLVIAPEKVQKIAP
WKYLGMTITSKQVVPQPVKLNLAVKTLNDVQKLMGSLNWIRPYLGLTNSQLQPLDLLKHSNDPTEPRILNKEALNVIH
MVEQCIYKKFVSRLDLSQLVQFFVLIDKTVPFQALVQWNSEWDDPLHILEWMFLSFRPRKTASGLFELIADVIIKTRKR
CVELIGRDPATIVLPVQNWYFEWCLANNYELQVAMAGFQRQISYHLPSHLLKFAQEIPFGQKYLSQLPEPVKGPTVFTD
GSGKTGKAAVVCLLQRALQRAAGTIFLAQIQKGGIVGGCNGSAECLTVREQRDLEQLVVYP

>Calidris pugnax XP_014809364.1 PREDICTED: endogenous retrovirus group K member
18 Pol protein-like

MGLFVLPGIIDADFTGEIKIMAWTPSPPCFVPKGQRIQLVPLPSVTIPGEGNRKGGFGSTGKPVVLWSKQVSKEQPLL
RCQVHDRHFSGLVDTGADVTIINISDWPEWPLRDPTSAIVGVGGLQKPKQSAKILTFKGPEGQIAHAAPYILPVPCTL
WGCALYIMGPRLVKPVGNLPENKFSIGAIDKQQTLSLTWKSEKTVWVDQWPLKKDKLLHLHDLVQEQLAAGHIVPTTSP
WNTPVFVIPKKGGRWRLIHDLRAINAVIEPMGALQPGLPSPSMLPQNWPPIAIIIDLKDCFFAIPLHPKDAPRFAFSVPAV
NQEQPTRRYHWTVPQGMLNSPTICQLTVANALQPVNRANPHVYIIHYMDDILIAAEKDKDLQIVVCQVRLAVQGAGLQ
IAEEKVQREPPWKYLGWKITHTPIQPQALQLALQIKTLNDLQKLLGTINWLRPFLGITTQDLHPLFQLPLTNRTLCTI

>Oryctolagus cuniculus XP_017200008.1 PREDICTED: endogenous retrovirus group K member 25 Pol protein-like, partial

SSAEAFSPRAGFFIGAAEEGIPITWKHEDPVWVPQWPLSSDKLVAAQQLIQEQLNLGHIRPSVSPWNTPIFVIKKKSGK
WRLLDLRLVINMQMQVMGPVQRGLPLLSALPQGWPIIIIDIKDCFFSIPLHTKDCERFAFTLPACNHEQPDQRYEWWVL
PQGMANSPTMCQLFVGQAIGPLRKRFFSSLKCIHYMDDILLAADKDEFVLDDGFAYLIQLLESKKLFIAPEKVQKGSATY
LGSCITSTQLFLKGGIAQDS

>Tinamus guttatus XP_010212357.1 PREDICTED: endogenous retrovirus group K member 113 Pol protein-like

MQGVGKREAERGSPPHNDTSTGCLERC SAATRGSAAGVDVATAVDVMLTDTRVRVIESELGSLGQGLSALLLGRSSVSR
QGIFVVPGLIDADYTGVIKIMVYTSAPPVTIPAHSKIAQLIPFKACVPHDALTERCNGEFGSTGPLNMLLAIDIKKGP
EELVRLQHPPGQTITLTMVVDTGADVSIIPQHMWPRAWPISLAATSMGVGGAQWTVDYSAFLVAAIGVQSTLKLTKM
ETPVWVDQWPLKQDRHLHIVEQLVQEQLLEAGHIVPSQSPWNTPIFTIPKKSQGWRLLDLRAINAVMQDMGALQPGLPSP
VMLPEGWDLIIIDLKDCFFTIALHPQDAEKFAFIVPSINKAAPAKRYHWVLPQGMKNSPTICQTFVAWALQPVRAKHP
ELLIYHYMGDILIAGENMCMKSVFQEVGEELGKRGLTIAPEKVQRQGPWNYLGWTIMGSEIRPQKIAIRTEVRTLVVDVQ
RLVGDIQWVRGICGITNDDIAPLPLLGTSVNASEARELSKEQREAVQAIAEKIAGAYASRIILEKKIWFIVNSGGSH
QHPMGLIAQILQGLRTLHILEWVFLSY

At EMBL-EBI Multiple Sequence Alignment, MAFFT, paste the 7 sequences into the input and set the output format as ClustalW.

Multiple Sequence Alignment

MAFFT (Multiple Alignment using Fast Fourier Transform) is a high speed multiple sequence alignment program.

We have recently changed the default parameter settings for MAFFT. Alignments should run much more quickly and larger DNA alignments can be carried out by default. Please click the 'More options' button to review the defaults and change them if required.

Important note: This tool can align up to 500 sequences or a maximum file size of 1 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

AUTOMATIC

sequences in any supported format:

PTRRYHWTVLPQGMLNSPTICQLTVANALQPVNRNANPHVIIHYMDDILIAAEKDKDLQIVVCQVRLAVQGAGLQIAEEKVQREPPWKYLGWKITTHPIQPQALQLA
LQIKTLNDLQKLLGTINWLRPFLGITTDLHPLFQLPLTNRTLCI
>Oryctolagus cuniculus XP_017200008.1 PREDICTED: endogenous retrovirus group K member 25 Pol protein-like, partial
SSAEAFSPRAGFFIGAAEEGIPITWKHEDPVVWPQWPLSSDKLVAAQQLIQEQLNLGHIRPSVSPWNTPIFVIKKKSGKWRLLDLRVINMQMQVMGPVQRLPL
LSALPQGWPIIIIDIKDCFFSIPLHTKDCERFAFTLPACNHEQPDQRYEWWVLPQGMANSPTMCQLFVGQAIGPLRKRFSCLKIHYMDDILIAAKDEFVLDDGFAYL
IQLLESKKLFIAPEKVQKGSATYLGSCITSTQLFLKGGIAQDS
>Tinamus guttatus XP_010212357.1 PREDICTED: endogenous retrovirus group K member 113 Pol protein-like
MQGVGKREAERGSPPHNDTSTGCLERC SAATRG SAGVDVATAVDVMLTDTRVRVIESELSGPLGQGLSALLGRSSVSRQGIFVVPGLIDADYTGVIKIMVY TSA

Or upload a file: No file chosen

[Use a example sequence](#) | [Clear sequence](#) | [See more example inputs](#)

STEP 2 - Set your Parameters

OUTPUT FORMAT

ClustalW

Figure 10: Paste the 7 sequences in the MAFFT, output format set to ClustalW

Input form

Web services

Help & Documentation

Bioinformatics Tools FAQ

Results for job mail-120220615-135926-0213-4544388-p1m

Alignments

Result Summary

Guide Tree

Phylogenetic Tree

Results Viewers

Submission Details

Download Alignment File

Hide Colors

CLUSTAL format alignment by MAFFT FFT-NS-i (v7.487)

Human	MGARVSVLRGGQLDTWEKIRLRPGGKKKYKMKLLVWASRELERFAVNPGLLETTEGCQQI
(novel)Homo	-----
Haliaeetus	-----
Opisthocomus	M-----
Tinamus	MQGV-----GK-----REAERGS--PPHNDTSTGC--
Oryctolagus	-----
Calidris	M-----
Human	LEQLQPALKAGSEELKSLYNTVATLYCVHQKIDVRDTKEALDKLE-EIQNKSKQK-TQQA
(novel)Homo	-----
Haliaeetus	-----
Opisthocomus	-----
Tinamus	LERCSAATR-GSAGVD---VAT---AVDVMLTDTRVRIEELSGPLGGQLSALL
Oryctolagus	-----
Calidris	-----
Human	AANSQVSQNYPIVQNAQGGMVHQAIISPRTLNAWVKVVEEKAFSPEVIPMFTALSEGATPQ
(novel)Homo	-----
Haliaeetus	-----
Opisthocomus	-----
Tinamus	LGRSSVS-----RQGFVVPGLIDADYTGVIKIMVYTSAPPVTIP-----
Oryctolagus	-----
Calidris	-----GLFVLPGIIDADFTGEIKIMAWTPSPPCFVP-----
Human	DLNMMLNIVGGHQAAQMLKDTINEEAAEWDRTHPIHAGPNPPGQMRPRGSDIAGTTS-
(novel)Homo	-----
Haliaeetus	-----
Opisthocomus	-----
Tinamus	-----AHSKIAQLI-----PFKAC-VPHDALTEPCNGEFGSTGPL
Oryctolagus	-----
Calidris	-----KGQRIACLV-----PLPSVTIPGEGNRK--GGFGSTGKP
Human	NLQEQIAWMTGNPPIPVGEIYKRNIIVLGLNKIVRMYSVPGILDIRQGPKEPFRDYVDRFF
(novel)Homo	-----
Haliaeetus	-----MGGE-----
Opisthocomus	-----KNKPEIEVMVESQNGEMLKLMIMIDTGADVTIISAPHWPSH-----
Tinamus	MMIATDTYKSGDEELVPIQURGGSTTLTMAVDTGADVETTRQMLMDA

Figure 11: The colour results of the MAFFT

CLUSTAL format alignment by MAFFT FFT-NS-i (v7.487)

Human	MGARVSVLRGGQLDTWEKIRLRPGGKKKYKMKLLVWASRELERFAVNPGLLETTEGCQQI
(novel) Homo	-----
Haliaeetus	-----
Opisthocomus	M-----
Tinamus	MQGV-----GK-----REAERGS--PPHNDTSTGC--
Oryctolagus	-----
Calidris	M-----

Human	LEQLQPALKAGSEELKSLYNTVATLYCVHQIDVRDTKEALDKLE-EIQNKSKQK-TQQA
(novel) Homo	-----
Haliaeetus	-----
Opisthocomus	-----
Tinamus	LERCSAATR-GSAGVD-----VAT-----AVDVMLTDTRVRVIESELGGLGGLSALL
Oryctolagus	-----
Calidris	-----

Human	AANSQVSQNYPIVQNAQGQMVHQAI SPRTLNAWVKVVEEKAFSPEVIPMFTALSEGATPQ
(novel) Homo	-----
Haliaeetus	-----
Opisthocomus	-----
Tinamus	LGRSSVS-----RQGFVVPGLIDADYTGVIKIMVYTSAPPVTIP-----
Oryctolagus	-----
Calidris	-----GLFVLPGIIDADFTGEIKIMAWTPSPPCFVP-----

Human	DLNMMLNIVGGHQAAMQMLKDTINEEAAEWDRTHPIHAGPNPPGQMREPRGSDIAGTTS-
(novel) Homo	-----
Haliaeetus	-----
Opisthocomus	-----
Tinamus	-----AHSKIAQLI-----PFKAC-VPHDALTERCNGEFGSTGPL
Oryctolagus	-----
Calidris	-----KGQRIAQLV-----PLPSVTIPGEGNRK---GGFGSTGKP

Human	NLQEQIAWMTGNPPIPVGEIYKRWIVLGLNKIVRMYSVPVGILDIRQGPKEPFRDYVDRFF
(novel) Homo	-----
Haliaeetus	-----MGGE-----
Opisthocomus	-----KNKPEIEVMVESQNQEMLKLKMMIDTGADVTIISAPHWPSH-----
Tinamus	NMLLAIDIKKGKPEELVRLQHPPGGQTITLTMVVDTGADVSIIPQHMPRA-----
Oryctolagus	-----SSAE-----
Calidris	VVLWSKQVSKEQPLLRCQVHDRH----FSGLVDTGADVTIINISDWPPE-----

Human	KTLRAEQATQEVKNWMTETLLVQNANPDCKTILRALGPGATLEEMMTACQGVGGPGHKAR
(novel) Homo	-----
Haliaeetus	-----
Opisthocomus	-----W-----PTVESFTGVY-----
Tinamus	-----W-----PISLAATSVMGVGG-----
Oryctolagus	-----
Calidris	-----W-----PLRDPTSAIVGVGG-----

Human	VLAEAMSQVQSPNILMQRGNFKGQKRIKCFNCGKEGHLARNCRAPRKKGCWKCGKEGHQM
(novel) Homo	-----
Haliaeetus	-----
Opisthocomus	-----
Tinamus	-----AQ-----
Oryctolagus	-----
Calidris	-----LQKP-----KQSAKILTFK-GPEGQIAH--AAPY-----

Human	KDCTERQANFLRENLAFAQQREARELSSEQTGAISPTGRELWDKGRNNLLSAAGTEGQGTI
(novel) Homo	-----
Haliaeetus	-----
Opisthocomus	-----
Tinamus	-----W-----
Oryctolagus	-----
Calidris	-----ILPVPCTLW-----GCA

Human	SSFNFPQITLWQRPLVTVRIGGQLIEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIK
(novel) Homo	-----
Haliaeetus	-----NYHGFLTRATVLQGV-----RQPTLVLTWLTN-----
Opisthocomus	-----AFLGAAIEQ-----QPVLKIKWKTN-----
Tinamus	-----TVDYSAFLVAAIGV-----QSTLKLTKWME-----
Oryctolagus	-----AFSPRAGFFIGA-----AEGIPITWKHE-----
Calidris	LYIMGPR-----LVKPVGNLPENKFSIGRID---KQQTLSLTWKSE-----

Human	VRQYDQIPIEICGKKAIGTVLVGPTPVNIIGRNMLTQIGCTLNFPISPIETVPVKLKPGM
(novel) Homo	-----
Haliaeetus	-----NPV-----
Opisthocomus	-----NPF-----
Tinamus	-----TPV-----
Oryctolagus	-----DPV-----
Calidris	-----KTV-----

Human	DGPKVKQWPLTEEEKIKALTDICMEMEKEGKISKIGPENPYNTPIFAIKKKDSTKWRKLVD
(novel) Homo	-----KKLQRAHELVEEQLKAGHIEL--SNSPRNSPIFVIPKR-SGKWRLLED
Haliaeetus	---WVDQWPLPIEKLKALQELVAEQLAAGHIEP--SQSPWNTPVFVIKKK-SGKWRLLED
Opisthocomus	---WVDQWPLTAERLQKISELVEEQLQVGHIQP--STSPWNTPIFTIPKN-SGKWRLLED
Tinamus	---WVDQWPLKQDRLHIVEQLVQEQLAAGHIVP--SQSPWNTPIFTIPKK-SGKWRLLED
Oryctolagus	---WVPQWPLSSDKLVAAQQLIQEQLNLGHIRP--SVSPWNTPIFVIKKK-SGKWRLLED
Calidris	---WVDQWPLKKDKLLHLHDLVQEQLAAGHIVP--TTSPWNTPVFVIPKK-GGRWRLIHD

.:: :: * *: . . * *: *: * * . : *: : *

Human	FRELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVPLDKDFRKYTAFTIPSIN
(novel) Homo	LQAINANLQPMGP-QQGLPSPAIPQDWPLAVIDLKDRFYTIPLAEQDREKFAFTIPAIN
Haliaeetus	LQQVNAVMA TMGALQPGMPSPAMIPQDWEIIVMDLKDCFFTIPLASQDKEKFAFSVPSIN
Opisthocomus	LRAVNNVMEDMGALQPGLPSPVMI PENWTVLVIDLKDCFFTIPLHPDDAERFAFSVPSVN
Tinamus	LRAINAVMQDMGALQPGLPSPVMLPEGWDLIIIDLKDCFFTIALHPQDAEKFAFIVPSIN
Oryctolagus	LRVINMQMQVMGFPVQRGLPLLSALPQGWPPIIIIDIKDCFFSIPLHTKDCERFAFTLPACN
Calidris	LRAINAVIEPMGALQPGLPSPSMLPQNWPPIAIIIDLKDCFFAIPLHPKDAPRFAFSVPAVN

:: : * : * *: : : : :: *: * : : : * : : : * . ** : *: *

Human	NETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPYRIKNPEIVIQYMDDLIVGSDLEIG
(novel) Homo	NERSAS*FHWKVFPQGMLNSPTMCQYHVNQALLPSRKEVPTCKIIHFMDDIY*QPQ---Q
Haliaeetus	HAEPAKRYQWRVLPQGMKNSPTICQWFVAQALSPVREKFPTSYCYHYMDDILLASD-NKE
Opisthocomus	KEEPARRFHWIVLPQGMKNSPTMCQIFVAWAFQPICKMPQLLIYHYMDDILIAGQ--NM
Tinamus	KAAPAKRYHWVLPQGMKNSPTICQTFVAWALQPVRAKHPELLIYHYMGDILIAGE--NM
Oryctolagus	HEQPDQRYEWVLPQGMANSPTMCQLFVGQAIGPLRKRFSCLKIHYMDDILLAAG-DEF
Calidris	QEQPTRRYHWTVPQGMLNSPTICQLTVANALQPVNRNANPHVIIYHYMDDILIAAE-KDK

: . :.: *:*** .***: * : : * . :*:.*: .

Human	QHRAKIEELREHLLRWGFTTPDKKHQKEPPFLWMGYELHPDKWTVQPIMLPEKDSWTVND
(novel) Homo	RHLSLYASVIKNTQLRGLNIAPKNVQMSSPWKYLGY-----
Haliaeetus	QLNDMENLARNLLQQYGLVIAPEKVKIAPWKYLGMTITSKQVVPQPVKLNLAVK-TLND
Opisthocomus	DREFVLQEVRVREVESRGLNIAPEKIQKQEPWNYLGWVILQGSIKPPKMQLNPEIK-TLND
Tinamus	CMKSVFQEVGEELGKRGLTIAPEKVRQGPWNYLGWTIMGSEIRPQKIAIRTEVR-TLVD
Oryctolagus	VLDDGFAYLIQLLESKKLFIAPEKVKGSIATYLGSCITSTQLF-----
Calidris	DLQIVVCQVRLAVQGAGLQIAEEKVQREPPWKYLGWKITTHPIQPQALQLALQIK-TLND

: . :.: * :.*

Human	IQKLVGKLNWASQIYAGIKVKELCKLLRGAKALTDIVTLTEEALELAENREILKEPVHG
(novel) Homo	-----
Haliaeetus	VQKLMGSLNWI-RPYLGLTNSQLQPLL-----DLLKHSNDPT----EPRILNKEALNV
Opisthocomus	VQKLVGDIQWV-RTLCDITNDDLAPLV-----ELLGTTSRAD----DKRTMEP-----
Tinamus	VQRLVGDIQWV-RGICGITNDDIAPLM-----PLLGTSVNAS----EARELSKEQREA
Oryctolagus	---LKG-----GIAQDS-----
Calidris	LQKLLGTINWL-RPFLGITTDLHPLF-----QLLPLTN-----RTL-----

Human	VYYDPTKDLIAEIQKQGQDQWTFQIYQEPFKNLKTGKYQERVAPYDLSITELTEVVQKVT
(novel) Homo	-----
Haliaeetus	IHM--VEQCIY-----KKFVSRI-----DLSQLVQFFV
Opisthocomus	-----
Tinamus	VQA--IAEKIA-----GAYASRI-----ILEKKIWFLI
Oryctolagus	-----

Calidris -----CI-----

Human TESIIIWGKTPKFRLP IQRETWEAWWMEYWQATWIP EWEFVNTLPLVKLWYQLEKDPIVG

(novel) Homo -----

Haliaeetus LID----KTVPF GALVQ-----WNSEWDDPLHILEWMFLSFRP-----

Opisthocomus -----

Tinamus VNS--GGSHQH PMGLIAQ-----I-LQGLRTLHILEWVFLSY-----

Oryctolagus -----

Calidris -----

Human AETFYVDGAANRETKLGKAGYVTD---RGRQKVVS LTETTNQKTELHAIHLALQDSGSEV

(novel) Homo -----

Haliaeetus -----RKTASGLFELIADV IIKTRKRCV-----

Opisthocomus -----

Tinamus -----

Oryctolagus -----

Calidris -----

Human NIVTDSQYALGIIQAQPDRSESELVNQIIEK LIGKDKVYLSWVPAHKGIGGNEQVDNLVS

(novel) Homo -----

Haliaeetus -----ELIGRDPATIV-LPV-----

Opisthocomus -----

Tinamus -----

Oryctolagus	-----
Calidris	-----

Human	SGFRKILFLDGLDKAQEEHEKFHSNWRAMASDFNLPPIVAKEIVASCDKCQLKGEAMHGQ
(novel) Homo	-----
Haliaeetus	-----QNWYFEW-CLANNYEL-----QVAMAGFQRQ
Opisthocomus	-----
Tinamus	-----
Oryctolagus	-----
Calidris	-----

Human	VDCSPGIWQLDCTHLEGKIIILVAVHVASGYIEAEVIPAETGQETAYFILKLAGRWPVK--
(novel) Homo	-----
Haliaeetus	IS-----YHLPSHLLL-----KFAQEIPFGQKYLSQPEPVKGP
Opisthocomus	-----
Tinamus	-----
Oryctolagus	-----
Calidris	-----

Human	IIHTDNGSNFTSAAVRAACWWANVTQEFGIPYNPQSQGVVESMNKELKKIIGQVRDQAEH
(novel) Homo	-----
Haliaeetus	TVFTD-GSGKTGKA-----AVVCLLQRALQRAAG-----
Opisthocomus	-----

Tinamus	-----
Oryctolagus	-----
Calidris	-----

Human	LKTAVQMAVFIHNFKRKGGIGGYSAGERIIDIIASDIQTKELQKQITKIQNFRVYYRDSR
(novel) Homo	-----
Haliaeetus	-----TIFLAQIQKGGIVGGCNGSAECLTV-----REQR
Opisthocomus	-----
Tinamus	-----
Oryctolagus	-----
Calidris	-----

Human	DPIWKGPAKLLWKGEHAVVIQDNSDIKVVPRRKAKILRDYGKQMAGDDCVAGRQDED
(novel) Homo	-----
Haliaeetus	DL-----EQLVVYP-----
Opisthocomus	-----
Tinamus	-----
Oryctolagus	-----
Calidris	-----

6. Create a phylogenetic tree, using either parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use any program such as MEGA, PAUP, or Phylip.



Figure 12: Input the 7 sequence into MEGA software

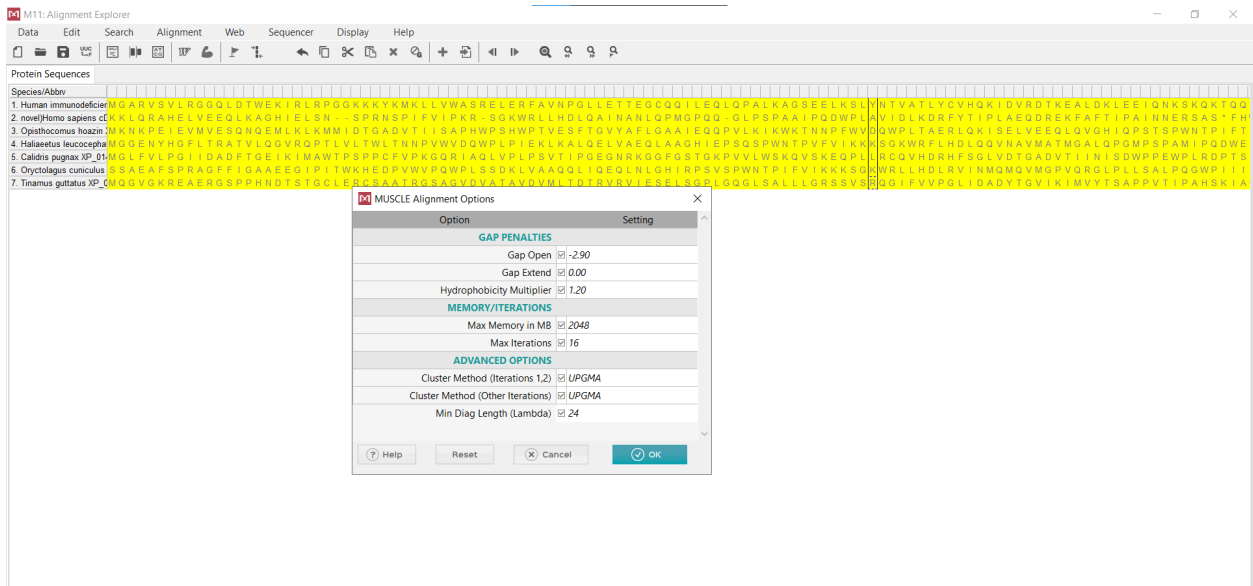


Figure 13: Default MUSCLE Alignment Options

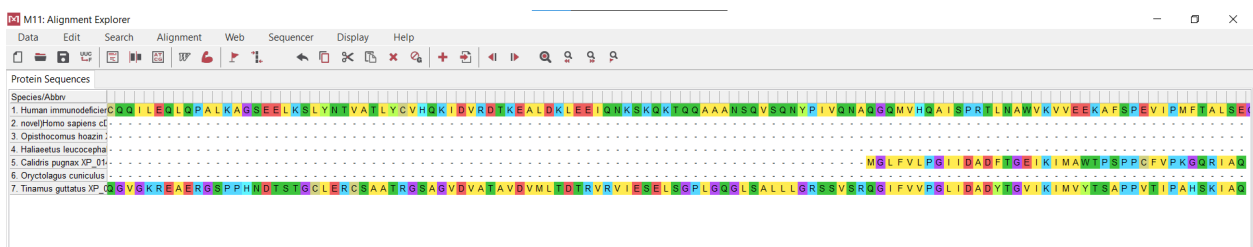


Figure 14: The result of MUSCLE alignments



Figure 15: Export alignment to MEGA format

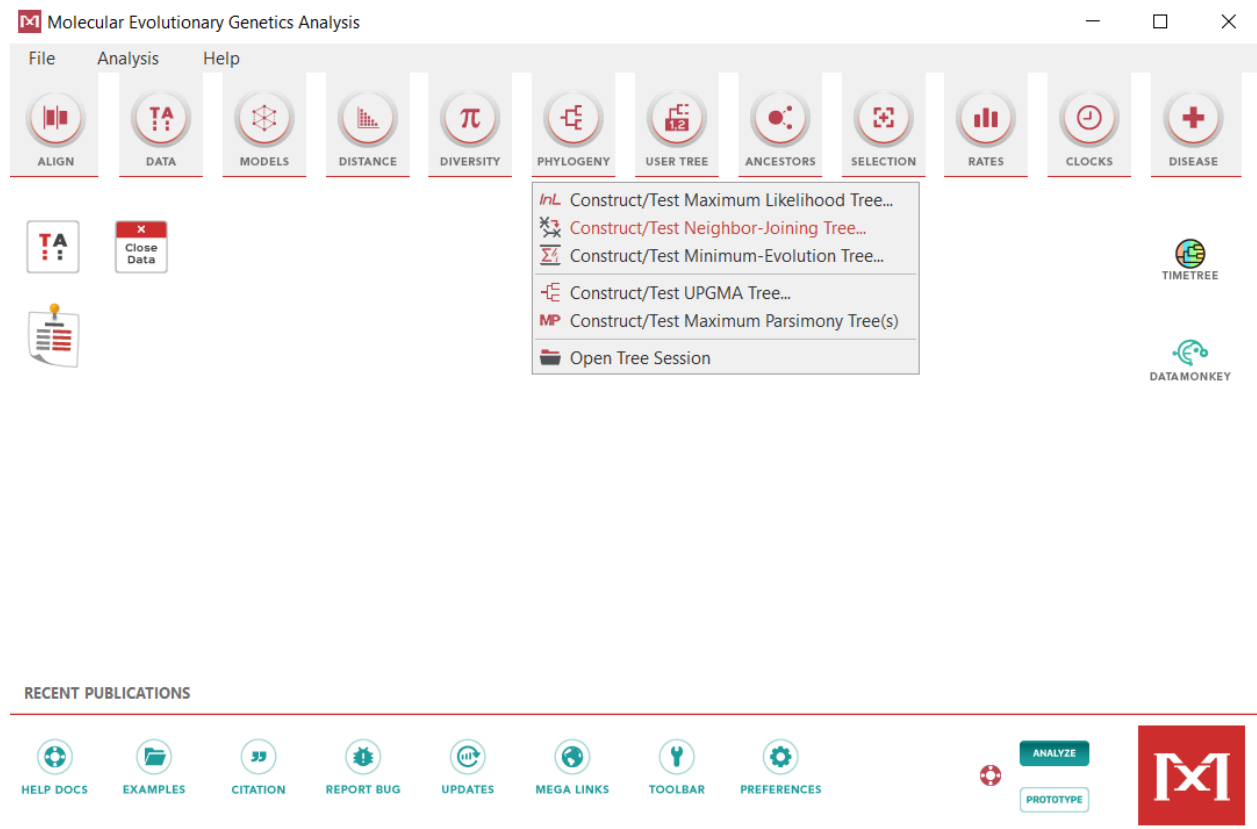


Figure 16: Select Construct/ Test Neighbor-Joining Tree at Phylogeny menu bar

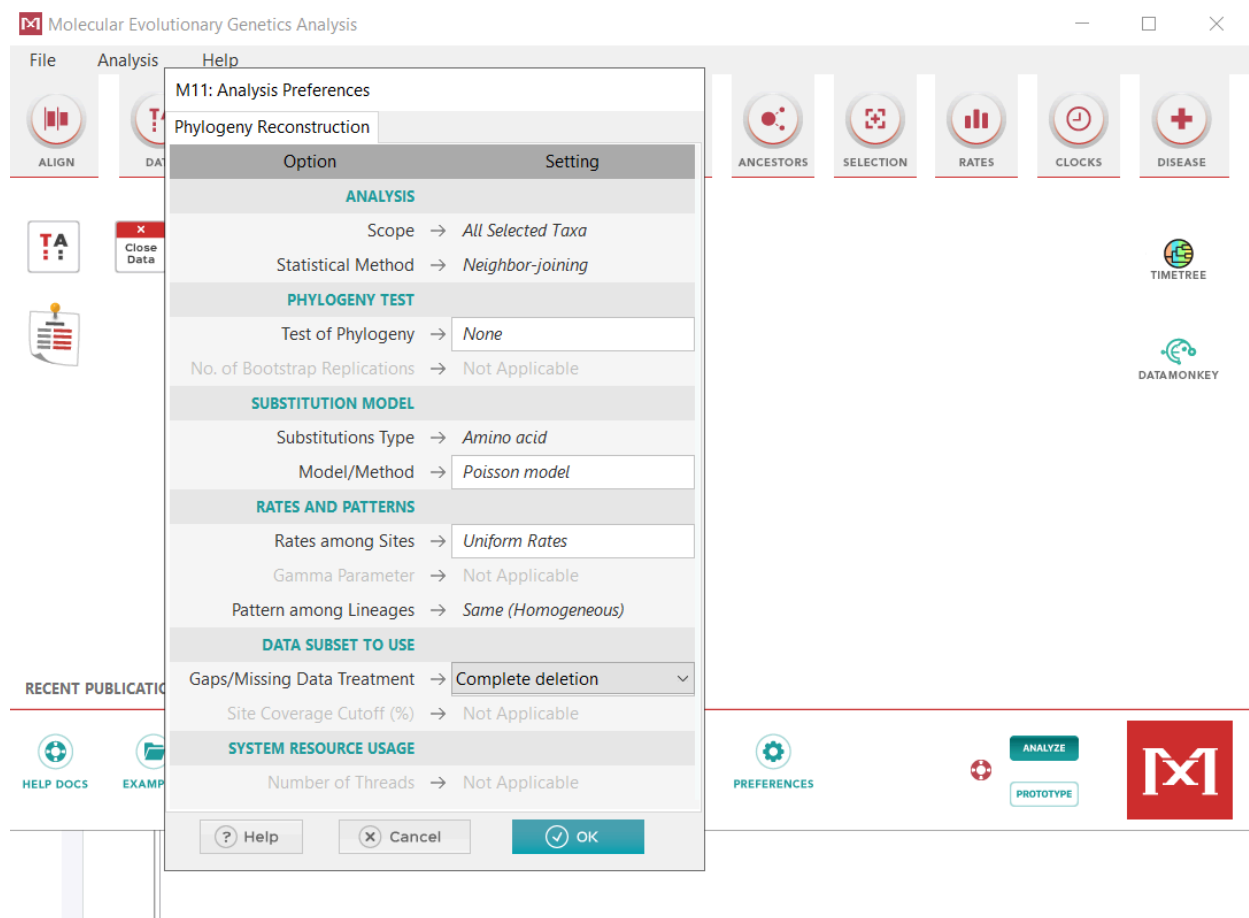


Figure 17: Phylogeny Reconstruction dialog box

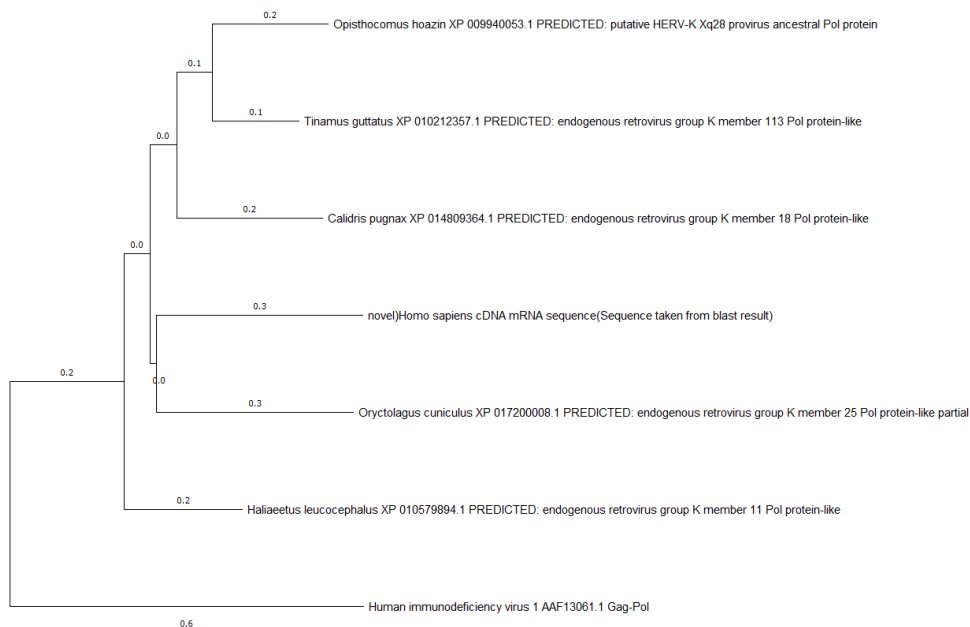


Figure 18: Rectangular Tree Style of Neighbor-Joining Tree with 7 protein sequences.

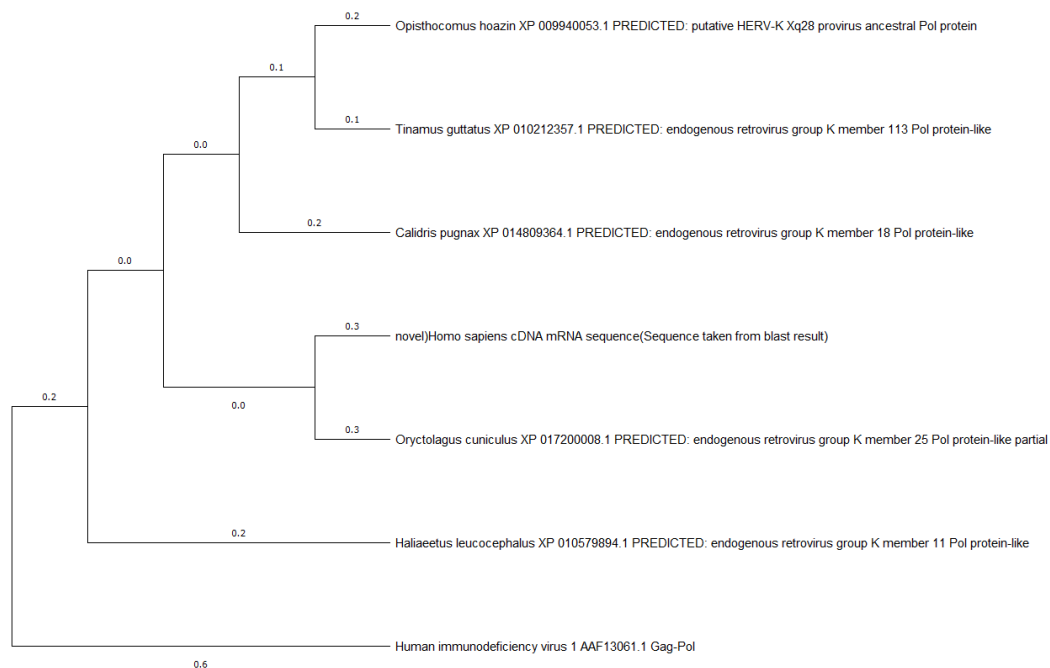


Figure 19 : “Topology Only” Tree Style of Neighbor-Joining Tree with 7 protein sequences.

Bootstrapping with Neighbor-Joining (NJ) Tree

Steps are similar to producing the Neighbor-Joining Tree as mentioned previously, at the dialog box, select Bootstrap method at the Test of Phylogeny and the Number of Bootstrap Replications is set to default which is 500.

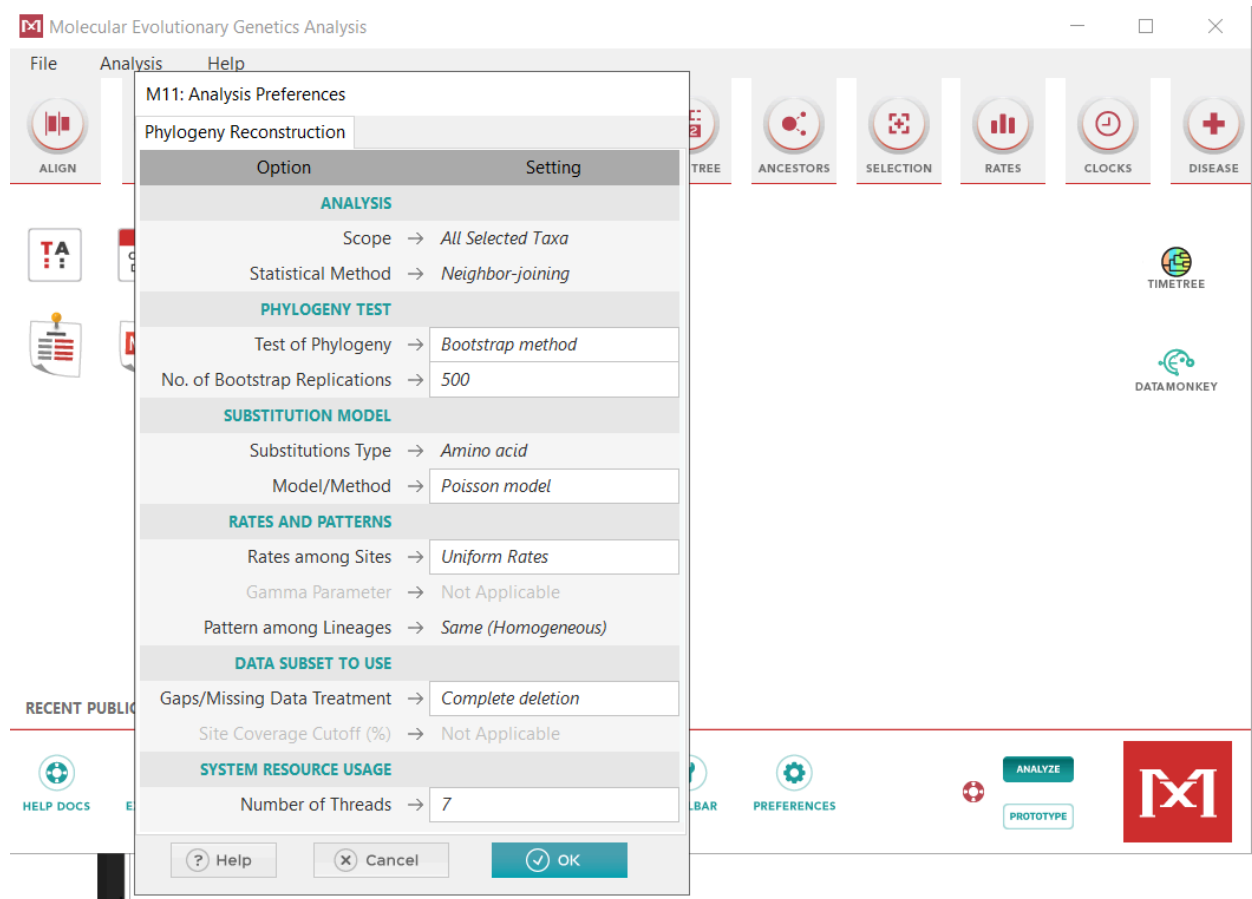


Figure 20: Bootstrap Phylogeny Reconstruction dialog box.

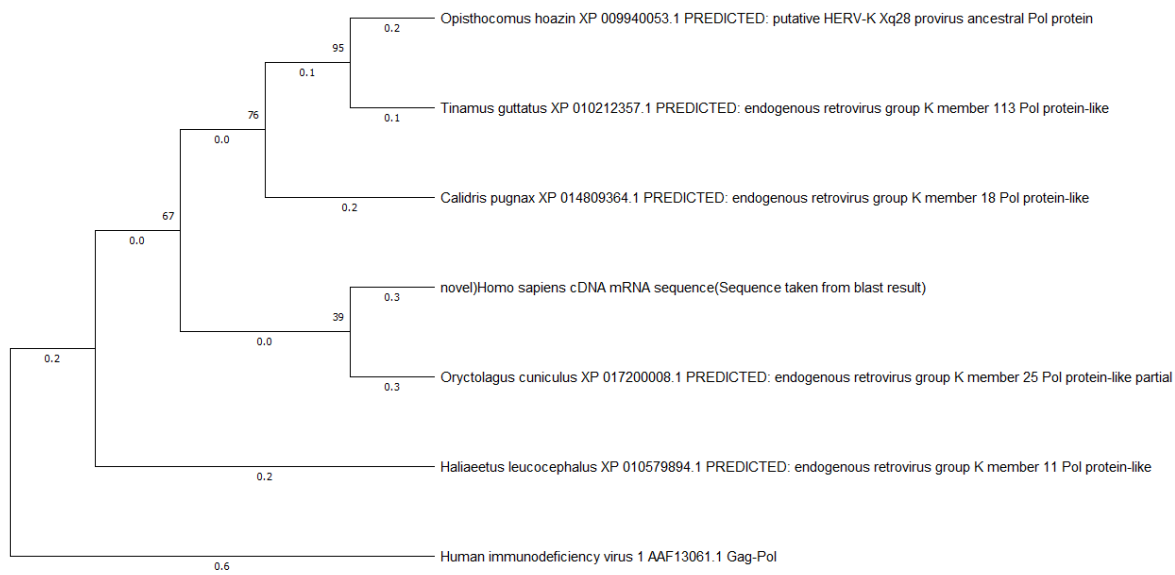


Figure 21: Bootstrap Neighbor-Joining Tree with Poisson Model.

Evolutionary relationships of taxa

The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method [2] and are in the units of the number of amino acid substitutions per site. This analysis involved 7 amino acid sequences. All positions containing gaps and missing data were eliminated (complete deletion option). There were a total of 195 positions in the final dataset. Evolutionary analyses were conducted in MEGA11 [3]

1. Saitou N. and Nei M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.
2. Zuckerkandl E. and Pauling L. (1965). Evolutionary divergence and convergence in proteins. Edited in *Evolving Genes and Proteins* by V. Bryson and H.J. Vogel, pp. 97-166. Academic Press, New York.
3. Tamura K., Stecher G., and Kumar S. (2021). MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* <https://doi.org/10.1093/molbev/msab120>.

Disclaimer: Although utmost care has been taken to ensure the correctness of the caption, the caption text is provided "as is" without any warranty of any kind. Authors advise the user to carefully check the caption prior to its use for any purpose and report any errors or problems to the authors immediately (www.megasoftware.net). In no event shall the authors and their employers be liable for any damages, including but not limited to special, consequential, or other damages. Authors specifically disclaim all other warranties expressed or implied, including but not limited to the determination of suitability of this caption text for a specific purpose, use, or application.

7. Discuss the significance of your novel gene. What have you learned about this gene/protein family?

MT0627 Homo sapiens cDNA is a gene sequence from the cloning products that derived from ORESTES PCR (U.S. Letters Patent application No. 196,716 - Ludwig Institute for Cancer Research) profiles into the pUC 18 vector. The reverse transcription of tissue mRNA and cDNA amplification were performed under low stringency conditions. In this case, the gene is taken from the database of 10,000 sequences from excised human breast tumors in the research. The particular database has significant contribution to the existing public EST databases which consist of mostly the derived sequences from cDNAs and boost the construction of contigs of transcript sequences. On the other hand, as one of the cloning products of ORESTES PCR, the novel gene can become a reference to early identification of important human genes, the decoding of the draft human genome sequence currently being compiled and the shotgun sequencing of the human transcriptome (Neto et al., 2000).

To complete the project, we have applied the knowledge that we have learned from the Bioinformatics lectures. We started our project by identifying a novel gene by performing TBLASTN search using HIV1 gag-pol with the accession number AAF13061.1. Based on the findings, we discovered that the MT0627 Homo sapiens cDNA is a novel gene since the top result of BLASTP searching of the particular protein is only 67% and there is no sequence with 100% matched to the sequence. Then, we did multiple sequence alignment by using the original query protein, the novel protein and a few other members of the family to identify the evolutionary relationships and common patterns among the sequences. By using the same sequences also, we created a phylogenetic tree of the 7 proteins using MEGA software which shows the path from a common ancestor to different descendants through evolutionary time with various approaches.

References

1. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=tblastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
2. <https://www.ncbi.nlm.nih.gov/protein/AAF13061.1/>
3. <https://www.ebi.ac.uk/Tools/msa/mafft/>
4. <https://www.megasoftware.net/>
5. Dias Neto, E., Garcia Correa, R., Verjovski-Almeida, S., Briones, M. R., Nagai, M. A., Da Silva, W., . . . Simpson, A. J. (2000). Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proceedings of the National Academy of Sciences*, 97(7), 3491-3496. doi:10.1073/pnas.97.7.3491