Livia Songster
osongste@ucsd.edu
A53304057
06 December 2019

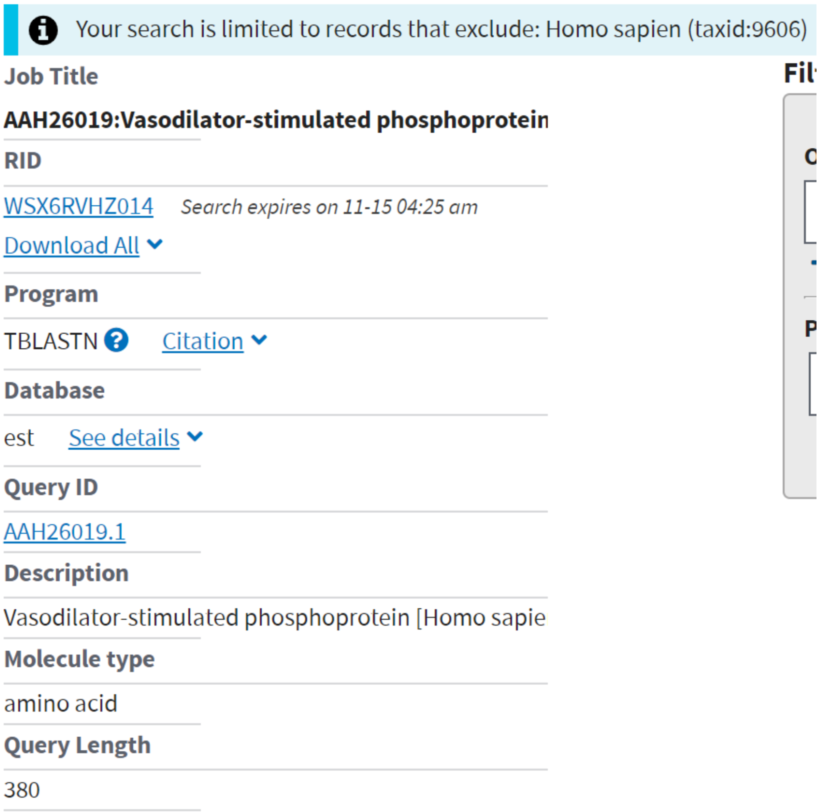
Find-a-gene project assignment [Q1] – [Q4]

[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as its function is known. If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

Name:	Vasodilator-stimulated phosphoprotein (vasp)
Accession:	AAH26019
Species:	Homo sapiens

[Q2] 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 details of the BLAST method used, database searched and any limits applied (e.g. Organism).

BLAST method:	TBLASTN
Database:	Expressed sequence tags (est)
Limits applied:	Organism excludes Homo sapiens (taxID:9606)



Your search is limited to records that exclude: Homo sapien (taxid:9606)

Job Title
AAH26019:Vasodilator-stimulated phosphoprotein

RID
WSX6RVHZ014 *Search expires on 11-15 04:25 am*
Download All

Program
TBLASTN Citation

Database
est See details

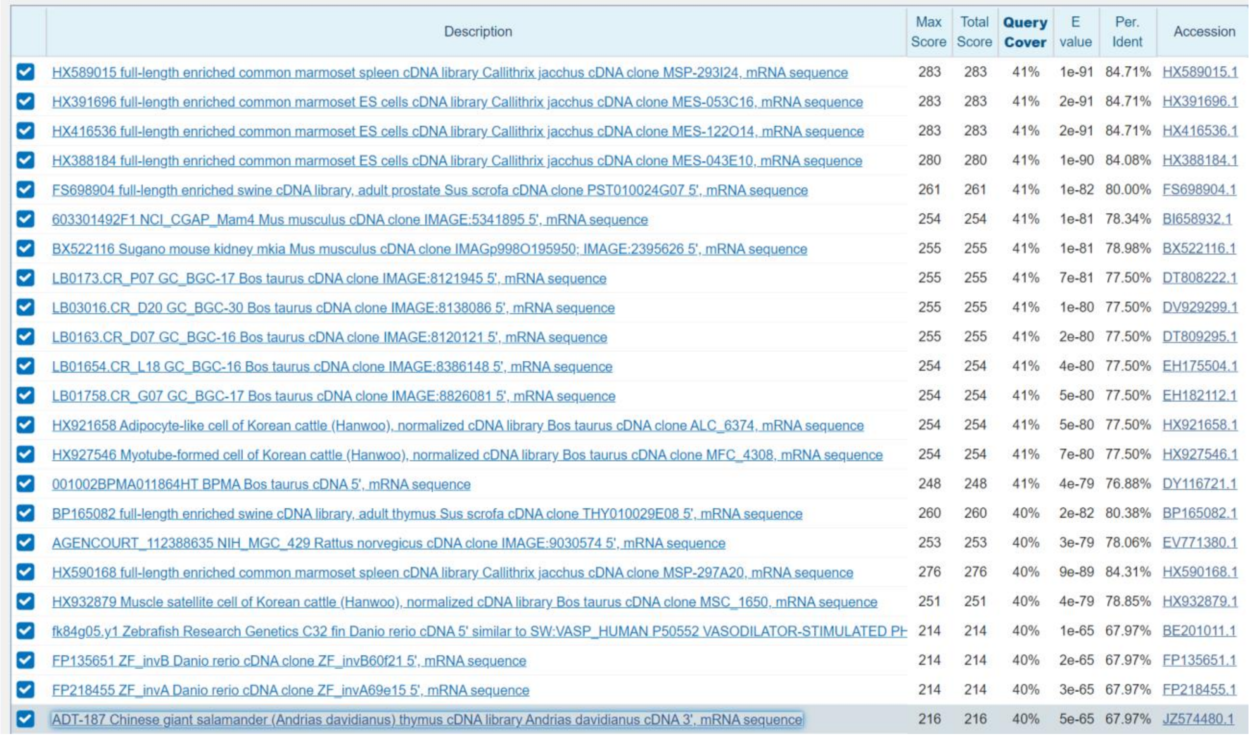
Query ID
AAH26019.1

Description
Vasodilator-stimulated phosphoprotein [Homo sapie

Molecule type
amino acid

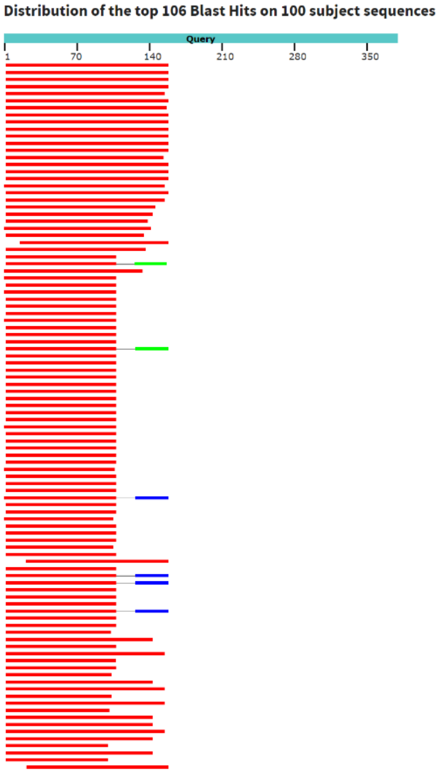
Query Length
380

Also 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 or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called Screen Shot [].png in your Desktop directory). It is not necessary to print out all of the blast results if there are many pages.

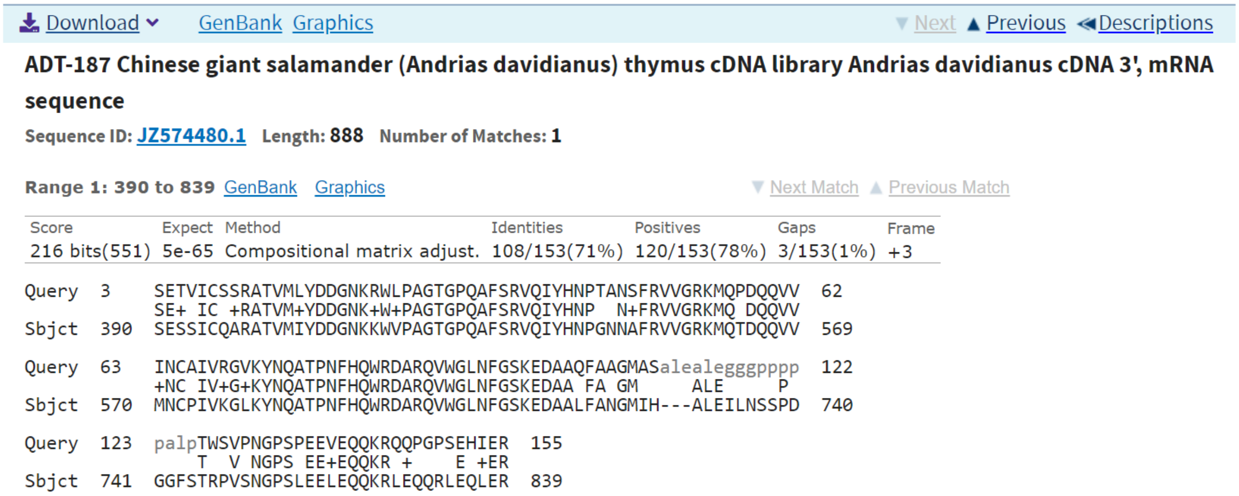


	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	HX589015 full-length enriched common marmoset spleen cDNA library Callithrix jacchus cDNA clone MSP-293I24, mRNA sequence	283	283	41%	1e-91	84.71%	HX589015.1
☑	HX391696 full-length enriched common marmoset ES cells cDNA library Callithrix jacchus cDNA clone MES-053C16, mRNA sequence	283	283	41%	2e-91	84.71%	HX391696.1
☑	HX416536 full-length enriched common marmoset ES cells cDNA library Callithrix jacchus cDNA clone MES-122O14, mRNA sequence	283	283	41%	2e-91	84.71%	HX416536.1
☑	HX388184 full-length enriched common marmoset ES cells cDNA library Callithrix jacchus cDNA clone MES-043E10, mRNA sequence	280	280	41%	1e-90	84.08%	HX388184.1
☑	FS698904 full-length enriched swine cDNA library, adult prostate Sus scrofa cDNA clone PST010024G07 5', mRNA sequence	261	261	41%	1e-82	80.00%	FS698904.1
☑	603301492F1 NCI_CGAP_Mam4 Mus musculus cDNA clone IMAGE:5341895 5', mRNA sequence	254	254	41%	1e-81	78.34%	BI658932.1
☑	BX522116 Sugano mouse kidney mkia Mus musculus cDNA clone IMAGp998O195950; IMAGE:2395626 5', mRNA sequence	255	255	41%	1e-81	78.98%	BX522116.1
☑	LB0173.CR_P07 GC_BGC-17 Bos taurus cDNA clone IMAGE:8121945 5', mRNA sequence	255	255	41%	7e-81	77.50%	DT808222.1
☑	LB03016.CR_D20 GC_BGC-30 Bos taurus cDNA clone IMAGE:8138086 5', mRNA sequence	255	255	41%	1e-80	77.50%	DV929299.1
☑	LB0163.CR_D07 GC_BGC-16 Bos taurus cDNA clone IMAGE:8120121 5', mRNA sequence	255	255	41%	2e-80	77.50%	DT809295.1
☑	LB01654.CR_L18 GC_BGC-16 Bos taurus cDNA clone IMAGE:8386148 5', mRNA sequence	254	254	41%	4e-80	77.50%	EH175504.1
☑	LB01758.CR_G07 GC_BGC-17 Bos taurus cDNA clone IMAGE:8826081 5', mRNA sequence	254	254	41%	5e-80	77.50%	EH182112.1
☑	HX921658 Adipocyte-like cell of Korean cattle (Hanwoo), normalized cDNA library Bos taurus cDNA clone ALC_6374, mRNA sequence	254	254	41%	5e-80	77.50%	HX921658.1
☑	HX927546 Myotube-formed cell of Korean cattle (Hanwoo), normalized cDNA library Bos taurus cDNA clone MFC_4308, mRNA sequence	254	254	41%	7e-80	77.50%	HX927546.1
☑	001002BPMA011864HT BPMA Bos taurus cDNA 5', mRNA sequence	248	248	41%	4e-79	76.88%	DY116721.1
☑	BP165082 full-length enriched swine cDNA library, adult thymus Sus scrofa cDNA clone THY010029E08 5', mRNA sequence	260	260	40%	2e-82	80.38%	BP165082.1
☑	AGENCOURT_112388635 NIH_MGC_429 Rattus norvegicus cDNA clone IMAGE:9030574 5', mRNA sequence	253	253	40%	3e-79	78.06%	EV771380.1
☑	HX590168 full-length enriched common marmoset spleen cDNA library Callithrix jacchus cDNA clone MSP-297A20, mRNA sequence	276	276	40%	9e-89	84.31%	HX590168.1
☑	HX932879 Muscle satellite cell of Korean cattle (Hanwoo), normalized cDNA library Bos taurus cDNA clone MSC_1650, mRNA sequence	251	251	40%	4e-79	78.85%	HX932879.1
☑	fk84g05.y1 Zebrafish Research Genetics C32 fin Danio rerio cDNA 5' similar to SW:VASP_HUMAN P50552 VASODILATOR-STIMULATED PH	214	214	40%	1e-65	67.97%	BE201011.1
☑	FP135651 ZF_invB Danio rerio cDNA clone ZF_invB60f21 5', mRNA sequence	214	214	40%	2e-65	67.97%	FP135651.1
☑	FP218455 ZF_invA Danio rerio cDNA clone ZF_invA69e15 5', mRNA sequence	214	214	40%	3e-65	67.97%	FP218455.1
☑	ADT-187 Chinese giant salamander (Andrias davidianus) thymus cDNA library Andrias davidianus cDNA 3', mRNA sequence	216	216	40%	5e-65	67.97%	JZ574480.1

Alignment Scores ⬛ < 40 🟦 40 - 50 🟩 50 - 80 🟪 80 - 200 🟥 >= 200



On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.



ADT-187 Chinese giant salamander (Andrias davidianus) thymus cDNA library Andrias davidianus cDNA 3', mRNA sequence

Sequence ID: JZ574480.1 Length: 888 Number of Matches: 1

Range 1: 390 to 839 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
216 bits(551)	5e-65	Compositional matrix adjust.	108/153(71%)	120/153(78%)	3/153(1%)	+3

```
Query  3    SETVICSSRATVMLYDDGNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGRKMQPDQQVV  62
            SE+ IC +RATVM+YDDGNK+W+PAGTGPQAFSRVQIYHNP  N+FRVVGRKMQ DQQVV
Sbjct  390  SESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQVV  569

Query  63   INCAIVRGVKYNQATPNFHQWRDARQVWGLNFGSKEDAAQFAAGMASalealegggpppp  122
            +NC IV+G+KYNQATPNFHQWRDARQVWGLNFGSKEDAA FA GM     ALE     P
Sbjct  570  MNCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMIH---ALEILNSSPD  740

Query  123  palpTWSVPNGPSPEEVEQQKRQQPGPSEHIER  155
                T  V NGPS EE+EQQKR +    E +ER
Sbjct  741  GGFSTRPVSNGPSLEELEQQKRLEQQRLEQLER  839
```

```
Chosen match:
> JZ574480.1 ADT-187 Chinese giant salamander (Andrias davidianus) thymus
cDNA library Andrias davidianus cDNA 3', mRNA sequence

Length = 888bp
Score = 216 bits(551), Expect = 5e-65, Method: Compositional matrix adjust.
Identities = 108/153(71%), Positives = 120/153(78%), Gaps = 3/153(1%), Frame
= +3

Query  3    SETVICSSRATVMLYDDGNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGRKMQPDQQVV  62
            SE+ IC +RATVM+YDDGNK+W+PAGTGPQAFSRVQIYHNP  N+FRVVGRKMQ DQQVV
Sbjct  390  SESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQVV  569

Query  63   INCAIVRGVKYNQATPNFHQWRDARQVWGLNFGSKEDAAQFAAGMASalealegggpppp  122
            +NC IV+G+KYNQATPNFHQWRDARQVWGLNFGSKEDAA FA GM     ALE     P
Sbjct  570  MNCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMIH---ALEILNSSPD  740

Query  123  palpTWSVPNGPSPEEVEQQKRQQPGPSEHIER  155
                T  V NGPS EE+EQQKR +    E +ER
Sbjct  741  GGFSTRPVSNGPSLEELEQQKRLEQQRLEQLER  839
```

In general, [Q2] is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of [Q4]), and a non-homologous result. If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

[Q3] 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 [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don’t forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon.

It may not start with a methionine if you don't have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

Chosen sequence:

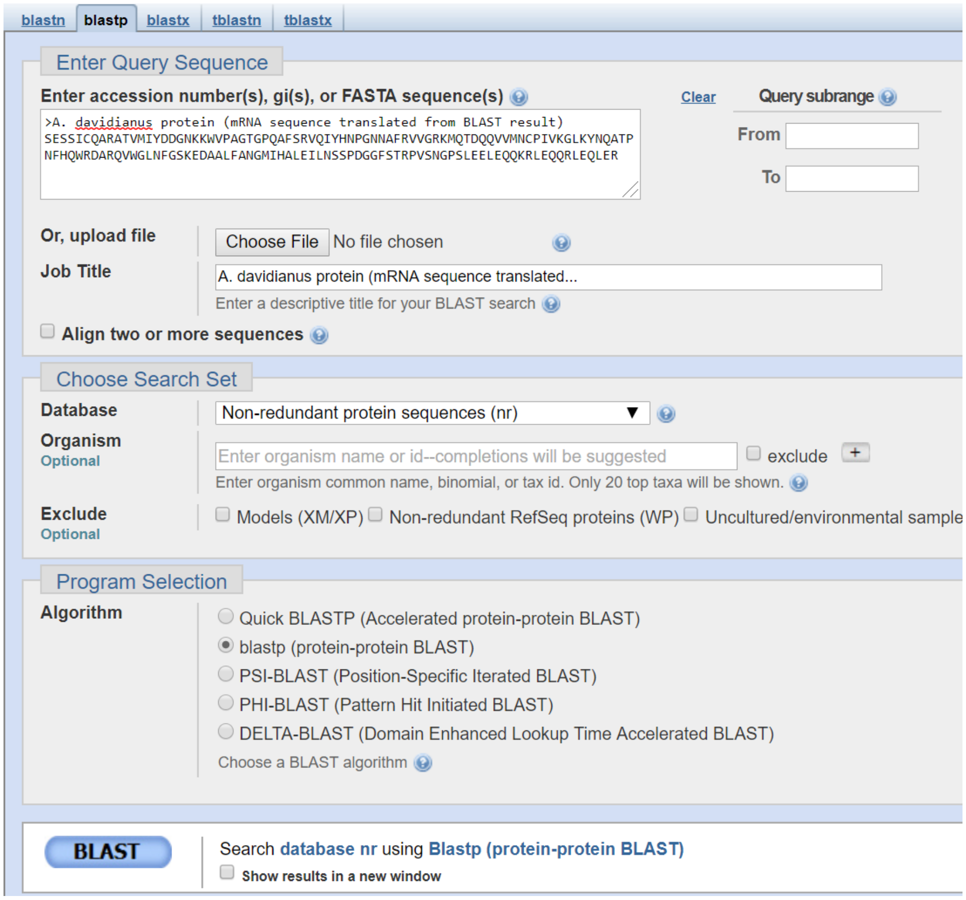
```
>A. davidianus protein (mRNA sequence translated from BLAST result)
SESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQVVMNCPIVKGLKYNQATPN
FHQWRDARQVWGLNFGSKEDAALFANGMIHALEILNSSPDGGFSTRPVSNGPSLEELEQQKRLEQQRLEQLER
```

Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as S. cerevisiae, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

```
Protein name: Vasodilator-stimulated phosphoprotein-like (PREDICTED)
Species: Andrias davidianus (Chinese giant salamander)
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Amphibia; Batrachia; Caudata; Cryptobranchoidea;
            Cryptobranchidae; Andrias
```

[Q4] 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 (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

- If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as "unknown"). Someone has already found and annotated this sequence, and assigned it an accession number.
- If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded.
- If there is a match with 100% identity, but to a different species than the one you started with, then you have likely succeeded in finding a novel gene.
- If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.



blastn | **blastp** | blastx | tblastn | tblastx

Enter Query Sequence

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

```
>A. davidianus protein (mRNA sequence translated from BLAST result)
SESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQVVMNCPIVKGLKYNQATP
NFHQWRDARQVWGLNFGSKEDAALFANGMIHALEILNSSPDGGFSTRPVSNGPSLEELEQQKRLEQQRLEQLER
```

Query subrange

From

To

Or, upload file No file chosen

Job Title A. davidianus protein (mRNA sequence translated...

Enter a descriptive title for your BLAST search

□ **Align two or more sequences**

Choose Search Set

Database Non-redundant protein sequences (nr) ▼

Organism Optional: Enter organism name or id--completions will be suggested □ exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample

Program Selection

Algorithm

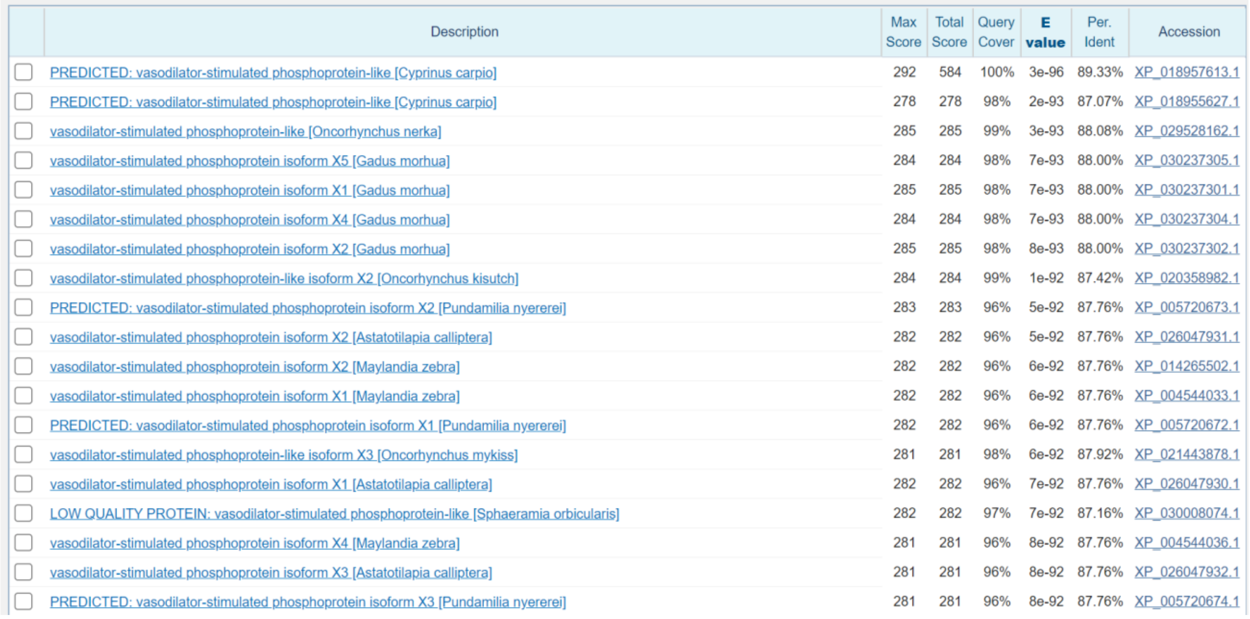
○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

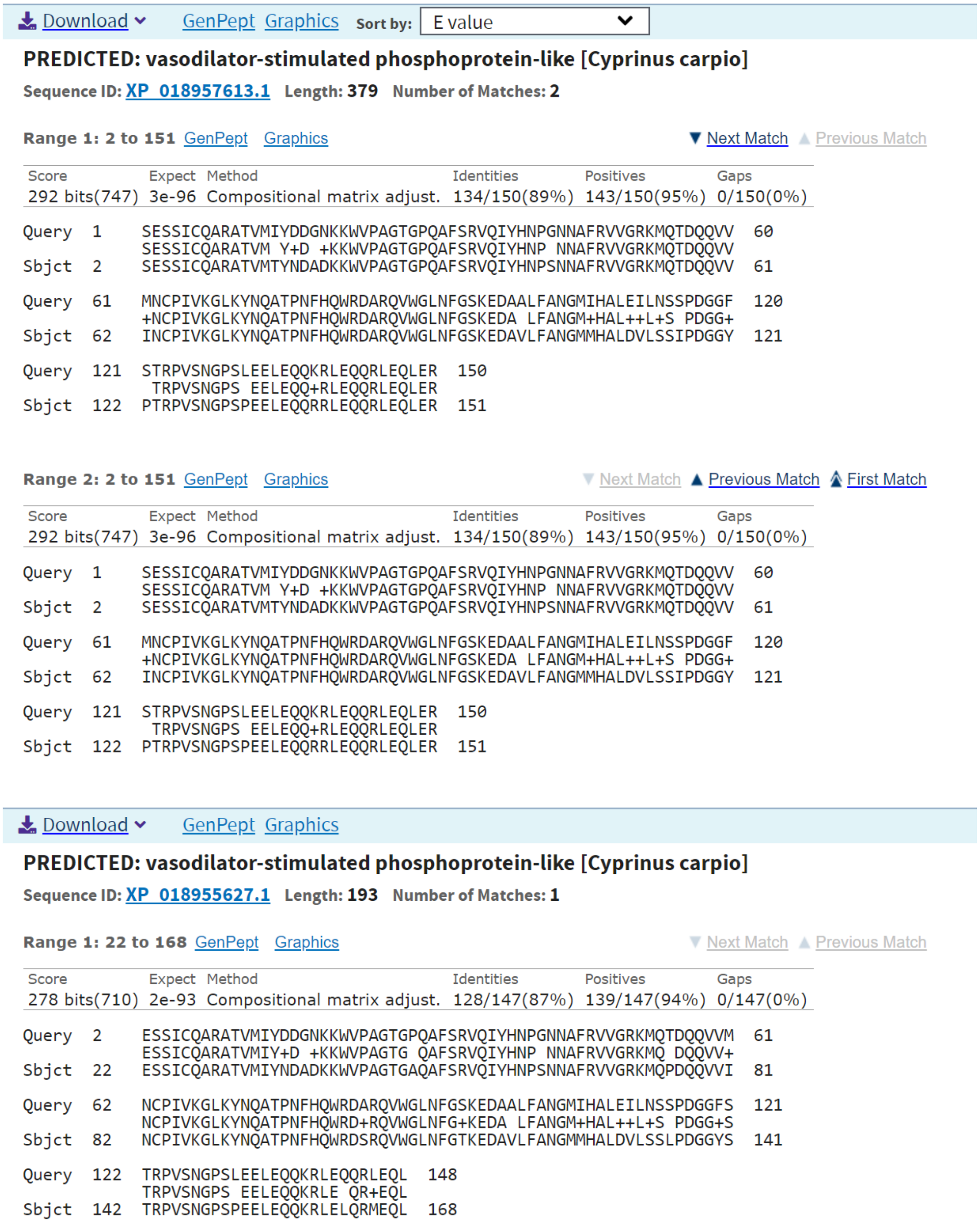
Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

The top result is to a protein from the common (European) carp (Cyprinus carpio). The alignments are on the next page.



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
□	PREDICTED: vasodilator-stimulated phosphoprotein-like [Cyprinus carpio]	292	584	100%	3e-96	89.33%	XP_018957613.1
□	PREDICTED: vasodilator-stimulated phosphoprotein-like [Cyprinus carpio]	278	278	98%	2e-93	87.07%	XP_018955627.1
□	vasodilator-stimulated phosphoprotein-like [Oncorhynchus nerka]	285	285	99%	3e-93	88.08%	XP_029528162.1
□	vasodilator-stimulated phosphoprotein isoform X5 [Gadus morhua]	284	284	98%	7e-93	88.00%	XP_030237305.1
□	vasodilator-stimulated phosphoprotein isoform X1 [Gadus morhua]	285	285	98%	7e-93	88.00%	XP_030237301.1
□	vasodilator-stimulated phosphoprotein isoform X4 [Gadus morhua]	284	284	98%	7e-93	88.00%	XP_030237304.1
□	vasodilator-stimulated phosphoprotein isoform X2 [Gadus morhua]	285	285	98%	8e-93	88.00%	XP_030237302.1
□	vasodilator-stimulated phosphoprotein-like isoform X2 [Oncorhynchus kisutch]	284	284	99%	1e-92	87.42%	XP_020358982.1
□	PREDICTED: vasodilator-stimulated phosphoprotein isoform X2 [Pundamilia nyererei]	283	283	96%	5e-92	87.76%	XP_005720673.1
□	vasodilator-stimulated phosphoprotein isoform X2 [Astatotilapia calliptera]	282	282	96%	5e-92	87.76%	XP_026047931.1
□	vasodilator-stimulated phosphoprotein isoform X2 [Maylandia zebra]	282	282	96%	6e-92	87.76%	XP_014265502.1
□	vasodilator-stimulated phosphoprotein isoform X1 [Maylandia zebra]	282	282	96%	6e-92	87.76%	XP_004544033.1
□	PREDICTED: vasodilator-stimulated phosphoprotein isoform X1 [Pundamilia nyererei]	282	282	96%	6e-92	87.76%	XP_005720672.1
□	vasodilator-stimulated phosphoprotein-like isoform X3 [Oncorhynchus mykiss]	281	281	98%	6e-92	87.92%	XP_021443878.1
□	vasodilator-stimulated phosphoprotein isoform X1 [Astatotilapia calliptera]	282	282	96%	7e-92	87.76%	XP_026047930.1
□	LOW QUALITY PROTEIN: vasodilator-stimulated phosphoprotein-like [Sphaeramia orbicularis]	282	282	97%	7e-92	87.16%	XP_030008074.1
□	vasodilator-stimulated phosphoprotein isoform X4 [Maylandia zebra]	281	281	96%	8e-92	87.76%	XP_004544036.1
□	vasodilator-stimulated phosphoprotein isoform X3 [Astatotilapia calliptera]	281	281	96%	8e-92	87.76%	XP_026047932.1
□	PREDICTED: vasodilator-stimulated phosphoprotein isoform X3 [Pundamilia nyererei]	281	281	96%	8e-92	87.76%	XP_005720674.1



PREDICTED: vasodilator-stimulated phosphoprotein-like [Cyprinus carpio]

Sequence ID: XP_018957613.1 **Length:** 379 **Number of Matches:** 2

Range 1: 2 to 151 GenPept Graphics

```
Score          Expect  Method                        Identities      Positives       Gaps
292 bits(747)  3e-96   Compositional matrix adjust.  134/150(89%)    143/150(95%)    0/150(0%)

Query  1    SESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQVV  60
            SESSICQARATVM Y+D +KKWVPAGTGPQAFSRVQIYHNP NNAFRVVGRKMQTDQQVV
Sbjct  2    SESSICQARATVMTYNDADKKWVPAGTGPQAFSRVQIYHNPSNNAFRVVGRKMQTDQQVV  61

Query  61   MNCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMIHALEILNSSPDGGF  120
            +NCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDA LFANGM+HAL++L+S PDGG+
Sbjct  62   INCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAVLFANGMMHALDVLSSIPDGGY  121

Query  121  STRPVSNGPSLEELEQQKRLEQQRLEQLER  150
             TRPVSNGPS EELEQQ+RLEQQRLEQLER
Sbjct  122  PTRPVSNGPSPEELEQQRRLEQQRLEQLER  151
```

Range 2: 2 to 151 GenPept Graphics

```
Score          Expect  Method                        Identities      Positives       Gaps
292 bits(747)  3e-96   Compositional matrix adjust.  134/150(89%)    143/150(95%)    0/150(0%)

Query  1    SESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQVV  60
            SESSICQARATVM Y+D +KKWVPAGTGPQAFSRVQIYHNP NNAFRVVGRKMQTDQQVV
Sbjct  2    SESSICQARATVMTYNDADKKWVPAGTGPQAFSRVQIYHNPSNNAFRVVGRKMQTDQQVV  61

Query  61   MNCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMIHALEILNSSPDGGF  120
            +NCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDA LFANGM+HAL++L+S PDGG+
Sbjct  62   INCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAVLFANGMMHALDVLSSIPDGGY  121

Query  121  STRPVSNGPSLEELEQQKRLEQQRLEQLER  150
             TRPVSNGPS EELEQQ+RLEQQRLEQLER
Sbjct  122  PTRPVSNGPSPEELEQQRRLEQQRLEQLER  151
```

PREDICTED: vasodilator-stimulated phosphoprotein-like [Cyprinus carpio]

Sequence ID: XP_018955627.1 **Length:** 193 **Number of Matches:** 1

Range 1: 22 to 168 GenPept Graphics

```
Score          Expect  Method                        Identities      Positives       Gaps
278 bits(710)  2e-93   Compositional matrix adjust.  128/147(87%)    139/147(94%)    0/147(0%)

Query  2    ESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQVVM  61
            ESSICQARATVMIY+D +KKWVPAGTG QAFSRVQIYHNP NNAFRVVGRKMQ DQQVV+
Sbjct  22   ESSICQARATVMIYNDADKKWVPAGTGAQAFSRVQIYHNPSNNAFRVVGRKMQPDQQVVI  81

Query  62   NCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMIHALEILNSSPDGGFS  121
            NCPIVKGLKYNQATPNFHQWRD+RQVWGLNFG+KEDA LFANGM+HAL++L+S PDGG+S
Sbjct  82   NCPIVKGLKYNQATPNFHQWRDSRQVWGLNFGTKEDAVLFANGMMHALDVLSSLPDGGYS  141

Query  122  TRPVSNGPSLEELEQQKRLEQQRLEQL  148
            TRPVSNGPS EELEQQKRLE QR+EQL
Sbjct  142  TRPVSNGPSPEELEQQKRLELQRMEQL  168
```

[Q5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width. Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence.

Query protein:

```
>AAH26019.1 Vasodilator-stimulated phosphoprotein [Homo sapiens]
MSSETVICSSRATVMLYDDGNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGRKMQPDQQVVINCAIVRG
VKYNQATPNFHQWRDARQVWGLNFGSKEDAAQFAAGMASALEALEGGGPPPPPALPTWSVPNGPSPEEVE
QQKRQQPGPSEHIERRVSNAGGPPAPPAGGPPPPPGPPPPPGPPPPPGLPPSGVPAAAHGAGGGPPPAPP
LPAAQGPGGGGAGAPGLAAAIAGAKLRKVSKEEASGGPTAPKAESGRSGGGGLMEEMNAMLARRRKATQV
GEKTPKDESANQEEPEARVPAQSESVRRPWEKNSTTLPRMKSSSSVTTSETQPCTPSSSDYSDLQRVKQE
LLEEVKKELQKVKEEIIEAFVQELRKRGSP
```

My novel gene:

```
>Andrias_davidianus protein (mRNA sequence translated from BLAST result)
SESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQVVMNCPIVKGLKYNQATPN
FHQWRDARQVWGLNFGSKEDAALFANGMIHALEILNSSPDGGFSTRPVSNGPSLEELEQQKRLEQQRLEQLER
```

List of proteins from other species (renamed in the alignment)

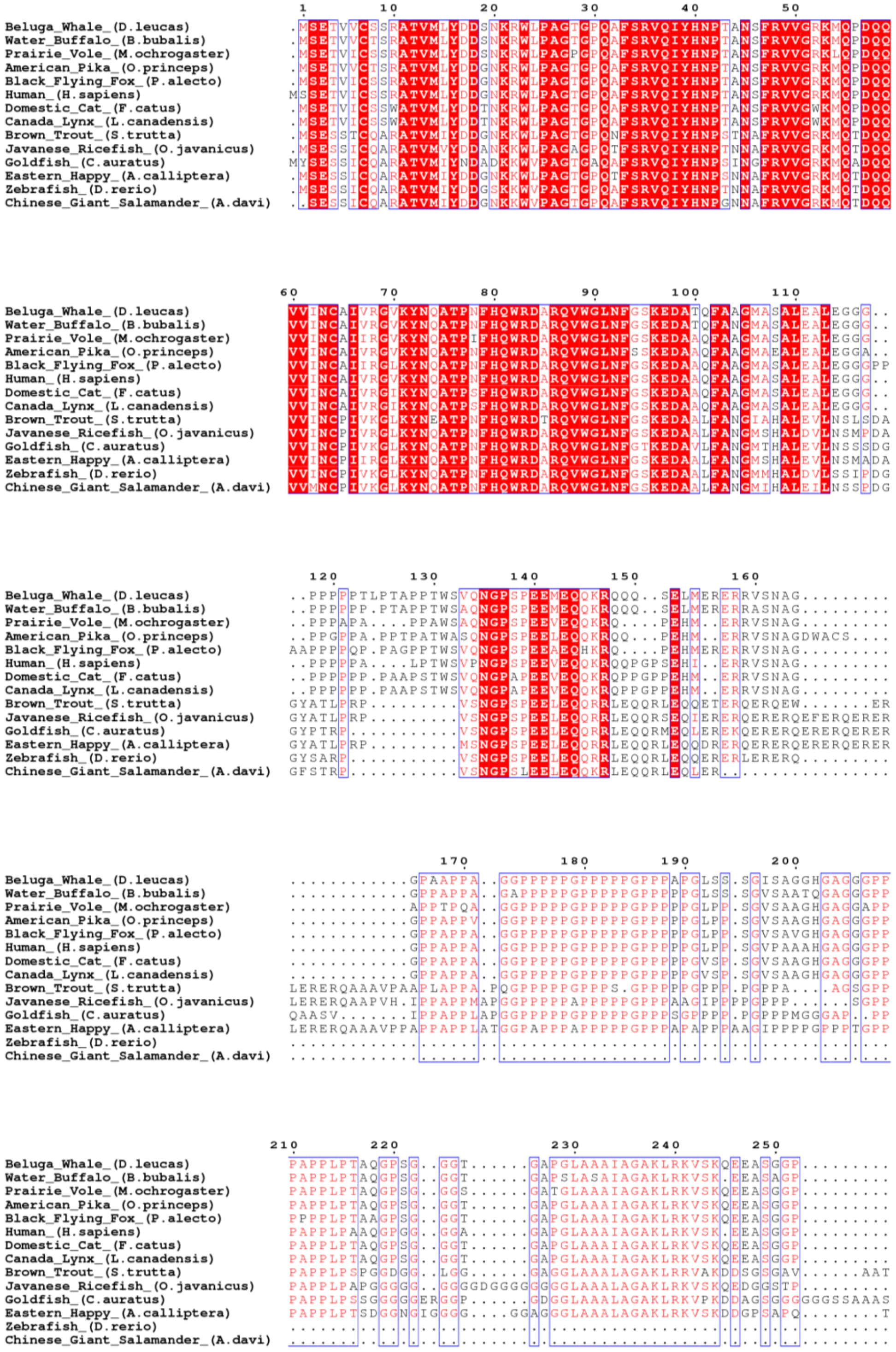
```
>XP_029571362.1 vasodilator-stimulated phosphoprotein-like isoform X2 [Salmo
trutta]

>AAH92890.1 Zgc:110347 [Danio rerio]

>XP_026143669.1 vasodilator-stimulated phosphoprotein-like isoform X3
[Carassius auratus]

>RVE64507.1 hypothetical protein OJAV_G00126530 [Oryzias javanicus]

>XP_026047931.1 vasodilator-stimulated phosphoprotein isoform X2
[Astatotilapia calliptera]

>XP_022441055.1 vasodilator-stimulated phosphoprotein isoform X1
[Delphinapterus leucas]

>XP_025125116.1 vasodilator-stimulated phosphoprotein isoform X2 [Bubalus
bubalis]

>XP_019674903.1 vasodilator-stimulated phosphoprotein isoform X1 [Felis
catus]

>XP_006911864.1 vasodilator-stimulated phosphoprotein isoform X6 [Pteropus
alecto]

>XP_012786028.1 PREDICTED: vasodilator-stimulated phosphoprotein [Ochotona
princeps]
```

```
>XP_005370333.1 vasodilator-stimulated phosphoprotein isoform X1 [Microtus
ochrogaster]
```

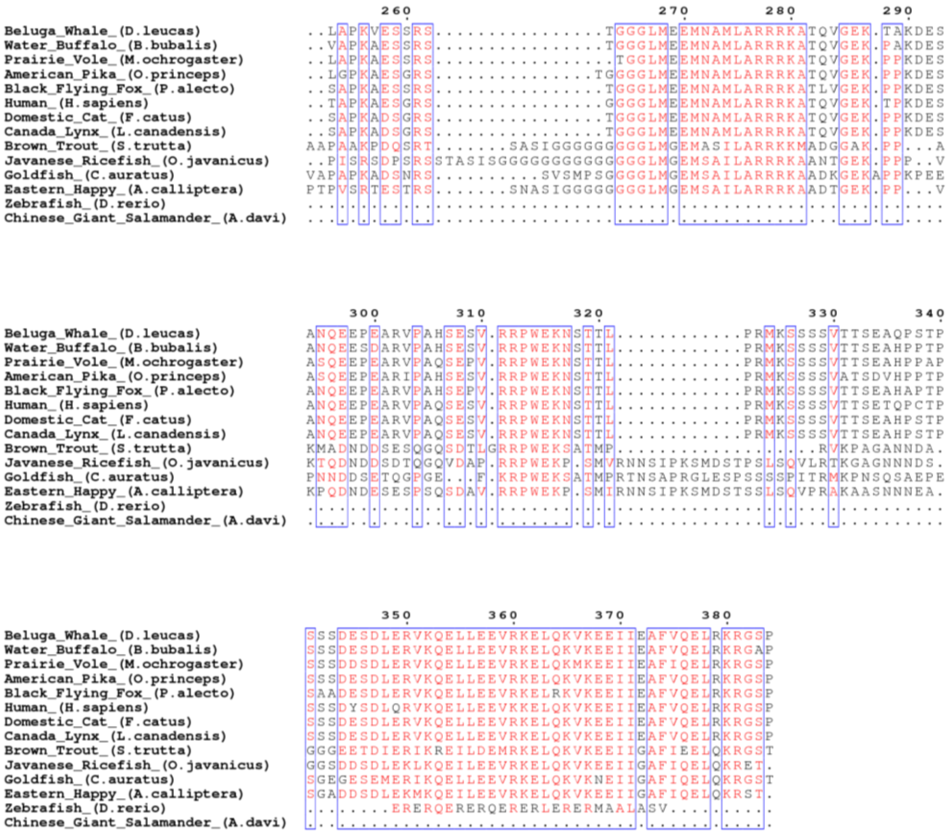
```
>XP_030155687.1 vasodilator-stimulated phosphoprotein isoform X1 [Lynx
canadensis]
```

Alignment:

CLUSTAL multiple sequence alignment by MUSCLE (3.8), then run through ESPript 3.0



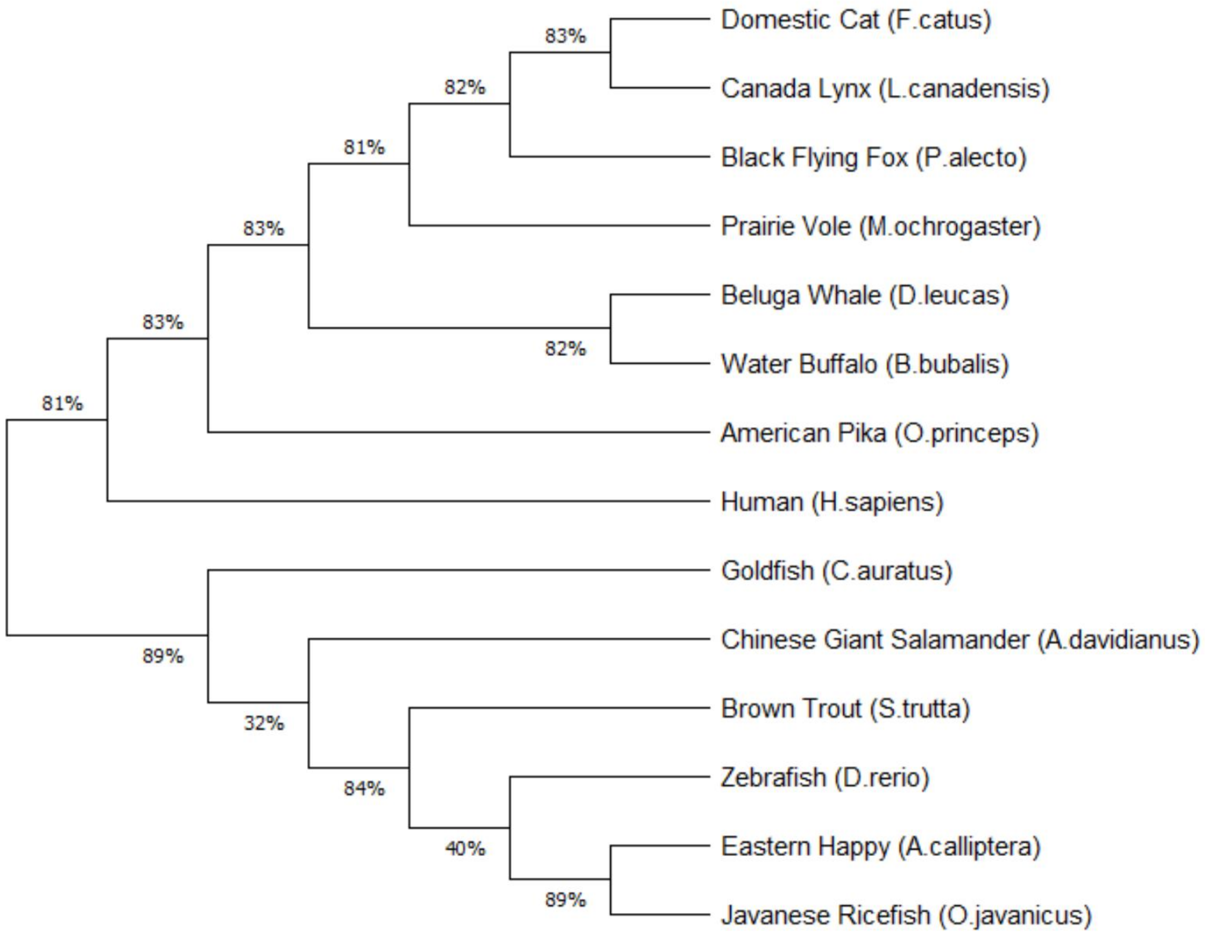
```
                                   1       10        20        30        40        50
Beluga_Whale_(D.leucas)           .MSETVVCSSRATVMLYDDSNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGRKMQPDQQ
Water_Buffalo_(B.bubalis)         .MSETVVCTSRATVMLYDDSNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGRKMQPDQQ
Prairie_Vole_(M.ochrogaster)      .MSETVICSSRATVMLYDDSNKRWLPAGPGPQAFSRVQIYHNPTANSFRVVGRKLQPDQQ
American_Pika_(O.princeps)        .MSETVVCTSRATVMLYDDGNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGRKMQPDQQ
Black_Flying_Fox_(P.alecto)       .MSETVICSSRATVMLYDDGNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGRKMQPDQQ
Human_(H.sapiens)                 MSSETVICSSRATVMLYDDGNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGRKMQPDQQ
Domestic_Cat_(F.catus)            .MSETVICSSWATVMLYDDTNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGWKMQPDQQ
Canada_Lynx_(L.canadensis)        .MSETVICSSWATVMLYDDTNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGWKMQPDQQ
Brown_Trout_(S.trutta)            .MSESSTCQARATVMIYDDGNKKWLPAGTGPQNFSRVQIYHNPSTNAFRVVGRKMQTDQQ
Javanese_Ricefish_(O.javanicus)   .MSESSICQARATVMVYDDANKKWLPAGAGPQTFSRVQIYHNPTNNAFRVVGRKMQTDQQ
Goldfish_(C.auratus)              MYSESSICQARATVMIYNDADKKWVPAGTGAQAFSRVQIYHNPSINGFRVVGRKMQADQQ
Eastern_Happy_(A.calliptera)      .MSESSICQARATVMIYDDGNKKWLPAGTGPQTFSRVQIYHNPSNNAFRVVGRKMQTDQQ
Zebrafish_(D.rerio)               .MSESSICQARATVMIYDDGSKKWVPAGTGPQAFSRVQIYHNPTNNAFRVVGRKMQTDQQ
Chinese_Giant_Salamander_(A.davi) ..SESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQ

                                 60        70        80        90       100       110
Beluga_Whale_(D.leucas)           VVINCAIVRGVKYNQATPNFHQWRDARQVWGLNFGSKEDATQFAAGMASALEALEGGG..
Water_Buffalo_(B.bubalis)         VVINCAIVRGVKYNQATPNFHQWRDARQVWGLNFGSKEDATQFANGMASALEALEGGG..
Prairie_Vole_(M.ochrogaster)      VVINCAIIRGVKYNQATPIFHQWRDARQVWGLNFGSKEDAAQFAAGMASALEALEGGG..
American_Pika_(O.princeps)        VVINCAIVRGVKYNQATPNFHQWRDARQVWGLNFSSKEDAAQFAAGMAEALEALEGGA..
Black_Flying_Fox_(P.alecto)       VVINCAIIRGLKYNQATPNFHQWRDARQVWGLNFGSKEDAAQFAAGMASALEALEGGGPP
Human_(H.sapiens)                 VVINCAIVRGVKYNQATPNFHQWRDARQVWGLNFGSKEDAAQFAAGMASALEALEGGG..
Domestic_Cat_(F.catus)            VVINCAIVRGIKYNQATPSFHQWRDARQVWGLNFGSKEDAAQFAAGMASALEALEGGG..
Canada_Lynx_(L.canadensis)        VVINCAIVRGIKYNQATPSFHQWRDARQVWGLNFGSKEDAAQFAAGMASALEALEGGG..
Brown_Trout_(S.trutta)            VVINCPIVKGLKYNEATPNFHQWRDTRQVWGLNFGSKEDAALFANGIAHALEVLNSLSDA
Javanese_Ricefish_(O.javanicus)   VVINCPIVRGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMSHALDVLNSMPDA
Goldfish_(C.auratus)              VVINCPIVKGLKYNQATPNFHQWRDARQVWGLNFGTKEDAVLFANGMTHALEVLNSSSDG
Eastern_Happy_(A.calliptera)      VVINCPIIRGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMSHALEVLNSMADA
Zebrafish_(D.rerio)               VVINCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMMHALDVLSSIPDG
Chinese_Giant_Salamander_(A.davi) VVMNCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMIHALEILNSSPDG

                                    120       130       140       150         160
Beluga_Whale_(D.leucas)           ..PPPPPTLPTAPPTWSVQNGPSPEEMEQQKRQQQ..SELMERERRVSNAG.........
Water_Buffalo_(B.bubalis)         ..PPPPPP.PTAPPTWSAQNGPSPEEMEQQKRQQQ..SELMERERRASNAG.........
Prairie_Vole_(M.ochrogaster)      ..PPPAPA....PPAWSAQNGPSPEEVEQQKRQ....PEHM..ERRVSNAG.........
American_Pika_(O.princeps)        ..PPGPPA.PPTPATWASQNGPSPEELEQQKRQQ...PEHM..ERRVSNAGDWACS....
Black_Flying_Fox_(P.alecto)       AAPPPPQP.PAGPPTWSVQNGPSPEEAEQHKRQ....PEHMERERRVSNAG.........
Human_(H.sapiens)                 ..PPPPPA....LPTWSVPNGPSPEEVEQQKRQQPGPSEHI..ERRVSNAG.........
Domestic_Cat_(F.catus)            ..PPPPPP.PAAPSTWSVQNGPAPEEVEQQKRQPPGPPEHM..ERRVSNAG.........
Canada_Lynx_(L.canadensis)        ..PPPPPP.PAAPSTWSVQNGPAPEEVEQQKRQPPGPPEHM..ERRVSNAG.........
Brown_Trout_(S.trutta)            GYATLPRP.........VSNGPSPEELEQQRRLEQQRLEQQETERQERQEW.......ER
Javanese_Ricefish_(O.javanicus)   GYATLPRP.........VSNGPSPEELEQQRRLEQQRSEQIERERQERERQEFERQERER
Goldfish_(C.auratus)              GYPTRP...........VSNGPSPEELEQQRRLEQQRMEQLEREKQERERQERERQERER
Eastern_Happy_(A.calliptera)      GYATLPRP.........MSNGPSPEELEQQRRLEQQRLEQQDRERQERERQERERQERER
Zebrafish_(D.rerio)               GYSARP...........VSNGPSPEELEQQRRLEQQRLEQQERERLERERQ.........
Chinese_Giant_Salamander_(A.davi) GFSTRP...........VSNGPSLEELEQQKRLEQQRLEQLER.................

                                                 170         180       190        200
Beluga_Whale_(D.leucas)           ............GPAAPPA..GGPPPPPGPPPPPGPPPAPGLSS.SGISAGGHGAGGGPP
Water_Buffalo_(B.bubalis)         ............GPPAPPA..GAPPPPPGPPPPPGPPPPPGLSS.SGVSAATQGAGGGPP
Prairie_Vole_(M.ochrogaster)      ............APPTPQA..GGPPPPPGPPPPPGPPPPPGLPP.SGVSAAGHGAGGAPP
American_Pika_(O.princeps)        ............GPPAPPV..GGPPPPPGPPPPPGPPPPPGLPP.SGVSAAGHGAGGGPP
Black_Flying_Fox_(P.alecto)       ............GPPAPPA..GGPPPPPGPPPPPGPPPPPGLPP.SGVSAVGHGAGGGPP
Human_(H.sapiens)                 ............GPPAPPA..GGPPPPPGPPPPPGPPPPPGLPP.SGVPAAAHGAGGGPP
Domestic_Cat_(F.catus)            ............GPPAPPA..GGPPPPPGPPPPPGPPPPPGVSP.SGVSAAGHGAGGGPP
Canada_Lynx_(L.canadensis)        ............GPPAPPA..GGPPPPPGPPPPPGPPPPPGVSP.SGVSAAGHGAGGGPP
Brown_Trout_(S.trutta)            LERERQAAAVPAAPLAPPA.PQGPPPPPGPPPS.GPPPPPGPPP.PGPPA....AGSGPP
Javanese_Ricefish_(O.javanicus)   LERERQAAPVH.IPPAPPMAPGGPPPPPAPPPPPGPPPAAGIPPPPGPPP......SGPP
Goldfish_(C.auratus)              QAASV.......IPPAPPLAPGGPPPPPGPPPPPGPPPSGPPPP.PGPPPMGGGAP..PP
Eastern_Happy_(A.calliptera)      LERERQAAAVPPAPPAPPLATGGPAPPPAPPPPPGPPPAPAPPPAAGIPPPPGPPPTGPP
Zebrafish_(D.rerio)               ............................................................
Chinese_Giant_Salamander_(A.davi) ............................................................

                                210       220               230       240       250
Beluga_Whale_(D.leucas)           PAPPLPTAQGPSG..GGT......GAPGLAAAIAGAKLRKVSKQEEASGGP.........
Water_Buffalo_(B.bubalis)         PAPPLPTAQGPSG..GGT......GAPSLASAIAGAKLRKVSK.EEASAGP.........
Prairie_Vole_(M.ochrogaster)      PAPPLPTAQGPSG..GGS......GATGLAAAIAGAKLRKVSKQEEASGGP.........
American_Pika_(O.princeps)        PAPPLPTAQGPSG..GGT......GAPGLAAAIAGAKLRKVSKQEEASGGP.........
Black_Flying_Fox_(P.alecto)       PPPPLPTAAGPSG..GGT......GAPGLAAAIAGAKLRKVSKQEEASGGP.........
Human_(H.sapiens)                 PAPPLPAAQGPGG..GGA......GAPGLAAAIAGAKLRKVSK.EEASGGP.........
Domestic_Cat_(F.catus)            PAPPLPTAQGPSG..GGT......GAPGLAAAIAGAKLRKVSKQEEASGGP.........
Canada_Lynx_(L.canadensis)        PAPPLPTAQGPSG..GGT......GAPGLAAAIAGAKLRKVSKQEEASGGP.........
Brown_Trout_(S.trutta)            PAPPLPSPGGDGG..LGG......GAGGLAAALAGAKLRRVAKDDSGSGAV......AAT
Javanese_Ricefish_(O.javanicus)   PAPPLPAPGGGGG..GGGGDGGGGGGGGLAAALAGAKLRKVSKQEDGGSTP.........
Goldfish_(C.auratus)              PAPPLPSSGGGGGERGGP......GDGGLAAALAGAKLRKVPKDDAGSGGGGGGSSAAAS
Eastern_Happy_(A.calliptera)      PAPPLPTSDGGNGIGGGG....GGAGGGLAAALAGAKLRKVSKDDGPSAPQ........T
Zebrafish_(D.rerio)               ............................................................
Chinese_Giant_Salamander_(A.davi) ............................................................
```



```
                                          260                       270       280       290
                                            .                         .         .         .
Beluga_Whale_(D.leucas)            ..LAPKVESSRS................TGGGLMEEMNAMLARRRKATQVGEK.TAKDES
Water_Buffalo_(B.bubalis)          ..VAPKAESSRS................TGGGLMEEMNAMLARRRKATQVGEK.PAKDES
Prairie_Vole_(M.ochrogaster)       ..LAPKAESSRS.................TGGLMEEMNAMLARRRKATQVGEK.PPKDES
American_Pika_(O.princeps)         ..LGPKAESGRS...............TGGGGLMEEMNAMLARRRKATQVGEK.PPKDES
Black_Flying_Fox_(P.alecto)        ..SAPKAESSRS................TGGGLMEEMNAMLARRRKATLVGEK.PPKDES
Human_(H.sapiens)                  ..TAPKAESGRS................GGGGLMEEMNAMLARRRKATQVGEK.TPKDES
Domestic_Cat_(F.catus)             ..SAPKADSGRS................TGGGLMEEMNAMLARRRKATQVGEK.PPKDES
Canada_Lynx_(L.canadensis)         ..SAPKADSGRS................TGGGLMEEMNAMLARRRKATQVGEK.PPKDES
Brown_Trout_(S.trutta)             AAPAAKPDQSRT.......SASIGGGGGGGGGLMGEMASILARRKKMADGGAK.PP...A
Javanese_Ricefish_(O.javanicus)    ..PISRSDPSRSSTASISGGGGGGGGGGGGGGLMGEMSAILARRRKAANTGEK.PPP..V
Goldfish_(C.auratus)               VAPAPKADSNRS..........SVSMPSGGGGLMGEMSAILARRRKAADKGEKAPPKPEE
Eastern_Happy_(A.calliptera)       PTPVSRTESTRS.......SNASIGGGGGGGGLMGEMSAILARRRKAADTGEK.PP...V
Zebrafish_(D.rerio)                ............................................................
Chinese_Giant_Salamander_(A.davi)  ............................................................

                                       300       310        320                   330       340
                                         .         .          .                     .         .
Beluga_Whale_(D.leucas)            ANQEEPEARVPAHSESV.RRPWEKNSTTL............PRMKSSSSVTTSEAQPSTP
Water_Buffalo_(B.bubalis)          ANQEESDARVPAHSESV.RRPWEKNSTTL............PRMKSSSSVTTSEAHPPTP
Prairie_Vole_(M.ochrogaster)       ASQEEPEARVPAQSEPV.RRPWEKNSTTL............PRMKSSSSVTTSEAHPPAP
American_Pika_(O.princeps)         ASQEEPEARIPAHSESV.RRPWEKNSTTL............PRMKSSSSVATSDVHPPTP
Black_Flying_Fox_(P.alecto)        ANQEEPEARVPAHSEPV.RRPWEKNSTTL............PRMKSSSSVTTSEAHAPTP
Human_(H.sapiens)                  ANQEEPEARVPAQSESV.RRPWEKNSTTL............PRMKSSSSVTTSETQPCTP
Domestic_Cat_(F.catus)             ANQEEPEARVPAQSESV.RRPWEKNSTTL............PRMKSSSSVTTSEAHPSTP
Canada_Lynx_(L.canadensis)         ANQEEPEARVPAQSESV.RRPWEKNSTTL............PRMKSSSSVTTSEAHPSTP
Brown_Trout_(S.trutta)             KMADNDDSESQGQSDTLGRRPWEKSATMP...................RVKPAGANNDA.
Javanese_Ricefish_(O.javanicus)    KTQDNDDSDTQGQVDAP.RRPWEKP.SMVRNNSIPKSMDSTPSLSQVLRTKGAGNNNDS.
Goldfish_(C.auratus)               PNNDDSETQGPGE...F.KRPWEKSATMPRTNSAPRGLESPSSSSPITRMKPNSQSAEPE
Eastern_Happy_(A.calliptera)       KPQDNDESESPSQSDAV.RRPWEKP.SMIRNNSIPKSMDSTSSLSQVPRAKAASNNNEA.
Zebrafish_(D.rerio)                ............................................................
Chinese_Giant_Salamander_(A.davi)  ............................................................

                                          350       360       370       380
                                            .         .         .         .
Beluga_Whale_(D.leucas)            SSSDESDLERVKQELLEEVRKELQKVKEEIIEAFVQELRKRGSP
Water_Buffalo_(B.bubalis)          SSSDESDLERVKQELLEEVRKELQKVKEEIIEAFVQELRKRGAP
Prairie_Vole_(M.ochrogaster)       SSSDDSDLERVKQELLEEVRKELQKMKEEIIEAFIQELRKRGSP
American_Pika_(O.princeps)         SSSDESDLERVKQELLEEVRKELQKVKEEIIEAFVQELRKRGSP
Black_Flying_Fox_(P.alecto)        SAADESDLERVKQELLEEVRKELRKVKEEIIEAFVQELRKRGSP
Human_(H.sapiens)                  SSSDYSDLQRVKQELLEEVKKELQKVKEEIIEAFVQELRKRGSP
Domestic_Cat_(F.catus)             SSSDESDLERVKQELLEEVRKELQKVKEEIIEAFVQELRKRGSP
Canada_Lynx_(L.canadensis)         SSSDESDLERVKQELLEEVRKELQKVKEEIIEAFVQELRKRGSP
Brown_Trout_(S.trutta)             GGGEETDIERIKREILDEMRKELQKVKEEIIGAFIEELQKRGST
Javanese_Ricefish_(O.javanicus)    GGSDDSDLEKLKQEILEEVRKELQKVKEEIIGAFIQELQKRET.
Goldfish_(C.auratus)               SGEGESEMERIKQELLEEVRKELQKVKNEIIGAFIQELQKRGST
Eastern_Happy_(A.calliptera)       SGADDSDLEKMKQEILEEVRKELQKVKEEIIGAFIQELQKRST.
Zebrafish_(D.rerio)                ........ERERQERERQERERLERERMAALASV..........
Chinese_Giant_Salamander_(A.davi)  ............................................
```

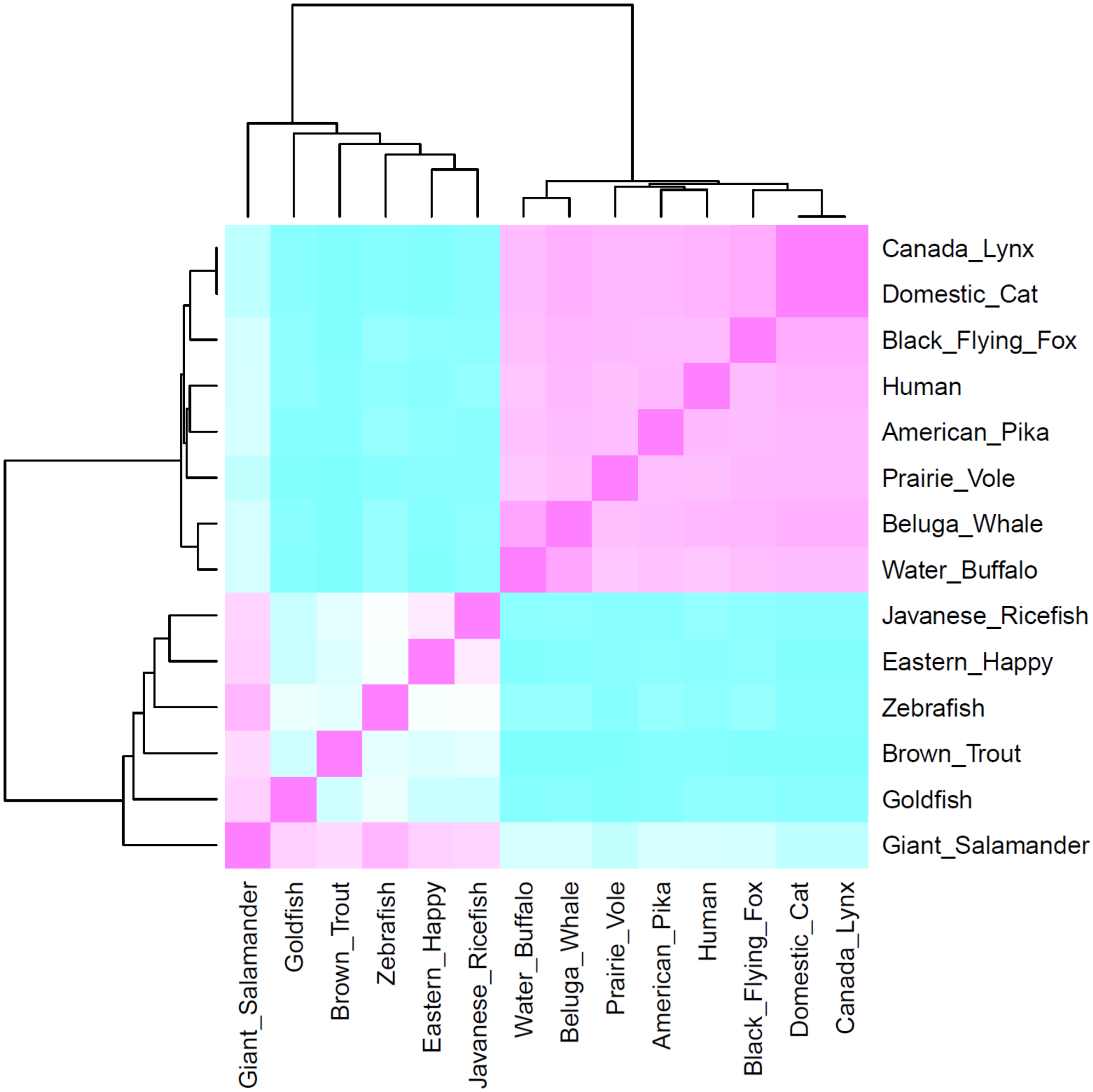
[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use "simple phylogeny" online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

Imported the sequences into MEGA, aligned with MUSCLE, and created a neighbor-joining tree:


83%
82%
81%
83%
83%
81%
82%
89%
32%
84%
40%
89%
Domestic Cat (F.catus)
Canada Lynx (L.canadensis)
Black Flying Fox (P.alecto)
Prairie Vole (M.ochrogaster)
Beluga Whale (D.leucas)
Water Buffalo (B.bubalis)
American Pika (O.princeps)
Human (H.sapiens)
Goldfish (C.auratus)
Chinese Giant Salamander (A.davidianus)
Brown Trout (S.trutta)
Zebrafish (D.rerio)
Eastern Happy (A.calliptera)
Javanese Ricefish (O.javanicus)

[Q7] Generate a sequence identity based heatmap of your aligned sequences using R. If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and “Save as” FASTA format for example). Read this FASTA format alignment into R with the help of functions in the Bio3D package. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.

Pink = high similarity, Blue = low similarity



[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences. List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimentalTechnique), resolution (resolution), and source organism (source).

HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as Evalue and identity. The results of pdb.annotate() contain the other annotation terms noted above. Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

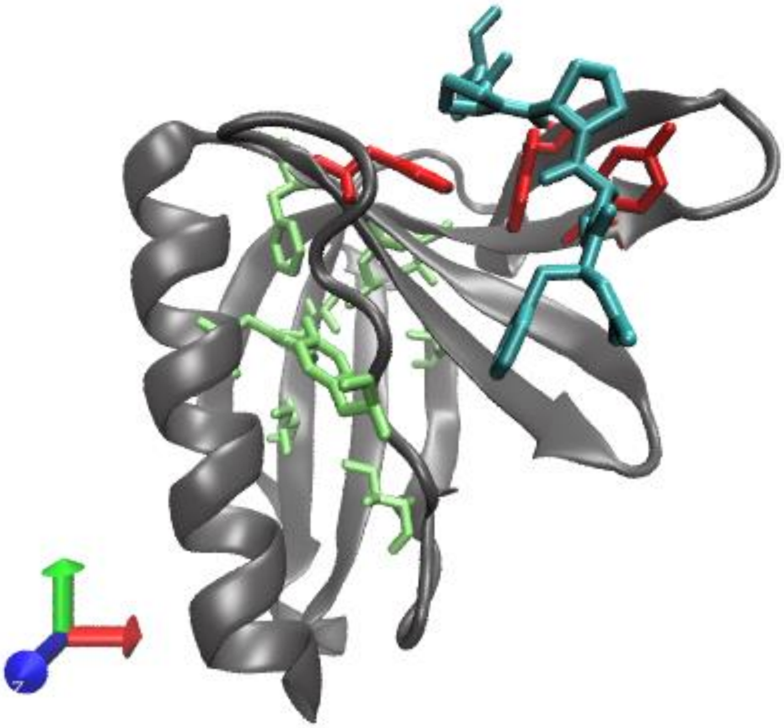
The consensus sequence I generated has >200 gaps, so I calculated the rowSum of the sequence identity matrix. The sequence for the domestic cat had the highest identity to all the others, so I used this as my query for blast.pdb() in R. The sequence is pasted here for reference. There were 22 hits from this search and all of them were for humans except one. I pasted below three hits for three different proteins and multiple organisms (human and mouse).

```
>XP_019674903.1 vasodilator-stimulated phosphoprotein isoform X1 [Felis
catus]
MSETVICSSWATVMLYDDTNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGWKMQPDQQVVINCAIVRGIKYNQATP
SFHQWRDARQVWGLNFGSKEDAAQFAAGMASALEALEGGGPPPPPPPAAPSTWSVQNGPAPEEVEQQKRQPPGPPEH
MERRVSNAGGPPAPPAGGPPPPPGPPPPPGPPPPPGVSPSGVSAAGHGAGGGPPPAPPLPTAQGPSGGGTGAPGLAA
AIAGAKLRKVSKQEEASGGPSAPKADSGRSTGGGLMEEMNAMLARRRKATQVGEKPPKDESANQEEPEARVPAQSES
VRRPWEKNSTTLPRMKSSSSVTTSEAHPSTPSSSDESDLERVKQELLEEVRKELQKVKEEIIEAFVQELRKRGSP
```

Structure ID	***Protein***	***Technique***	***Resolution***	***Source***	***Evalue***	***Identity***
2IYB	TES-MENA complex	X-ray diffraction	2.35	Homo sapiens	1.37e-48	67.257
1EVH	Ena EVH1 domain	X-ray diffraction	1.80 A	Mus musculus	1.28e-48	67.257
3SYX	SPRED1	X-ray diffraction	2.45 A	Homo sapiens	9.57e-13	33.929

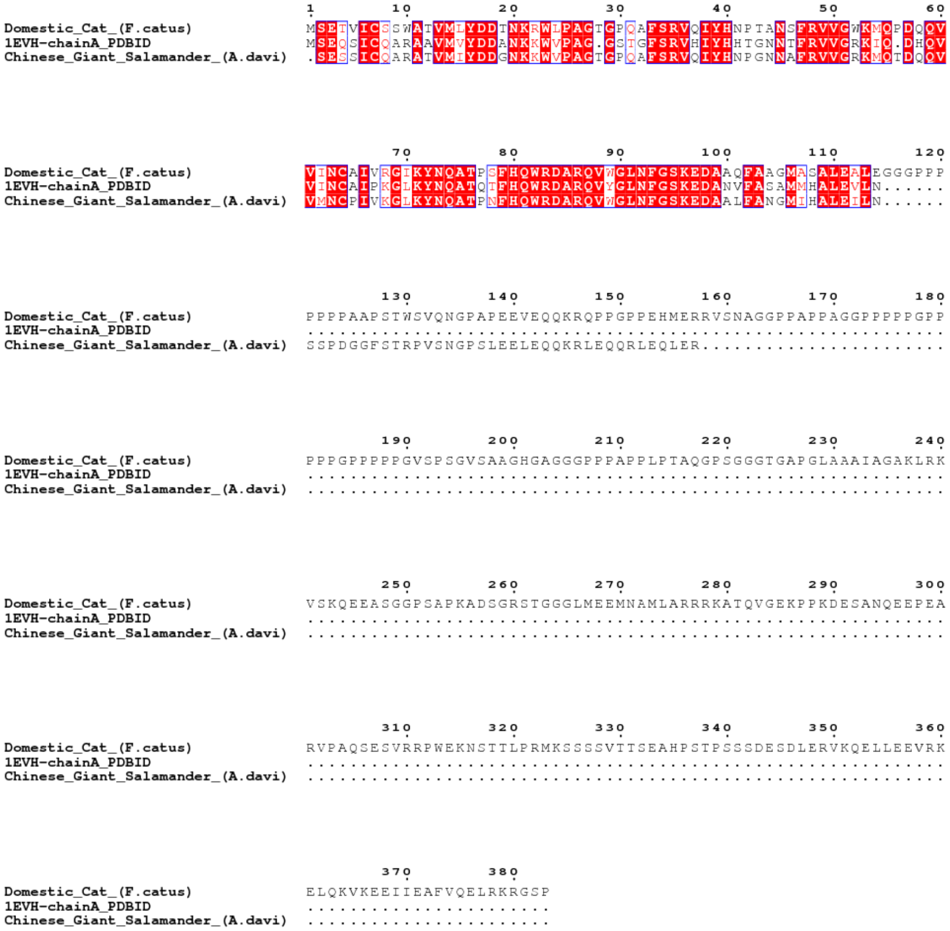
[Q9] Generate a molecular figure of one of your identified PDB structures using VMD. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black). Based on sequence similarity. How likely is this structure to be similar to your "novel" protein?

I used the 1EVH protein. I highlighted conserved residues associated with peptide binding and protein folding. This protein has only 67.257% sequence identity with my query protein from Felis catus.



KEY:
grey cartoon = beta-sheet sandwich + alpha helix
light green licorice = conserved hydrophobic core
red licorice = conserved binding pocket/triad
cyan licorice = peptide ligand in binding pocket

I saved a fasta file with the complete sequence from this PDB crystal structure and did another sequence alignment with the domestic cat sequence (from my original search) and my original novel protein (from the Chinese giant salamander). You can see here in this alignment that this structure only aligns to the EVH1 domain of my novel protein, but this region is very similar. There are only a few differences between this binding domain and my novel salamander gene, so I hypothesize this crystal structure is likely very similar to my novel gene.

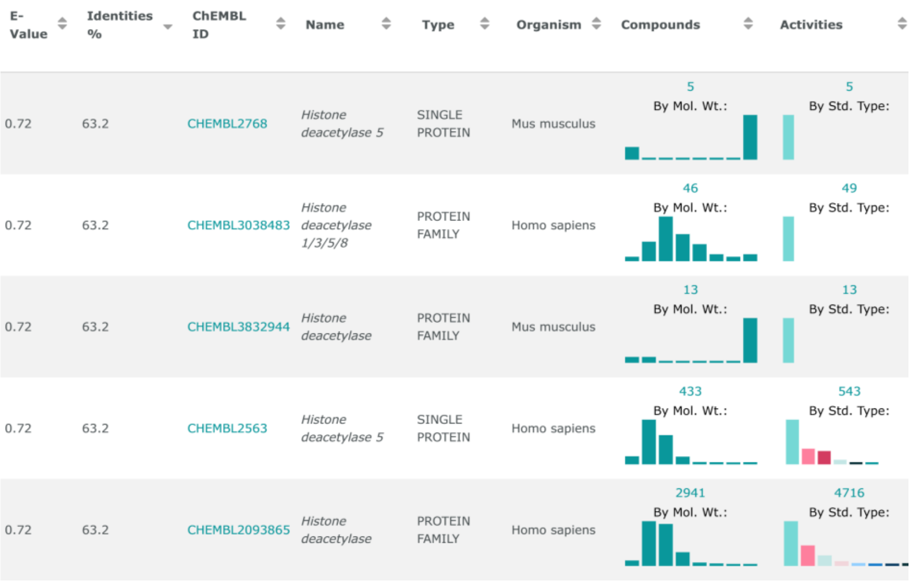


```
                                  1        10        20        30        40        50        60
Domestic_Cat_(F.catus)            MSETVICSSWATVMLYDDTNKRWLPAGTGPQAFSRVQIYHNPTANSFRVVGWKMQPDQQV
1EVH-chainA_PDBID                 MSEQSICQARAAVMVYDDANKKWVPAG.GSTGFSRVHIYHHTGNNTFRVVGRKIQ.DHQV
Chinese_Giant_Salamander_(A.davi) .SESSICQARATVMIYDDGNKKWVPAGTGPQAFSRVQIYHNPGNNAFRVVGRKMQTDQQV

                                           70        80        90       100       110       120
Domestic_Cat_(F.catus)            VINCAIVRGIKYNQATPSFHQWRDARQVWGLNFGSKEDAAQFAAGMASALEALEGGGPPP
1EVH-chainA_PDBID                 VINCAIPKGLKYNQATQTFHQWRDARQVYGLNFGSKEDANVFASAMMHALEVLN......
Chinese_Giant_Salamander_(A.davi) VMNCPIVKGLKYNQATPNFHQWRDARQVWGLNFGSKEDAALFANGMIHALEILN......

                                          130       140       150       160       170       180
Domestic_Cat_(F.catus)            PPPPAAPSTWSVQNGPAPEEVEQQKRQPPGPPEHMERRVSNAGGPPAPPAGGPPPPPGPP
1EVH-chainA_PDBID                 ............................................................
Chinese_Giant_Salamander_(A.davi) SSPDGGFSTRPVSNGPSLEELEQQKRLEQQRLEQLER.......................

                                          190       200       210       220       230       240
Domestic_Cat_(F.catus)            PPPGPPPPPGVSPSGVSAAGHGAGGGPPPAPPLPTAQGPSGGGTGAPGLAAAIAGAKLRK
1EVH-chainA_PDBID                 ............................................................
Chinese_Giant_Salamander_(A.davi) ............................................................

                                          250       260       270       280       290       300
Domestic_Cat_(F.catus)            VSKQEEASGGPSAPKADSGRSTGGGLMEEMNAMLARRRKATQVGEKPPKDESANQEEPEA
1EVH-chainA_PDBID                 ............................................................
Chinese_Giant_Salamander_(A.davi) ............................................................

                                          310       320       330       340       350       360
Domestic_Cat_(F.catus)            RVPAQSESVRRPWEKNSTTLPRMKSSSSVTTSEAHPSTPSSSDESDLERVKQELLEEVRK
1EVH-chainA_PDBID                 ............................................................
Chinese_Giant_Salamander_(A.davi) ............................................................

                                          370       380
Domestic_Cat_(F.catus)            ELQKVKEEIIEAFVQELRKRGSP
1EVH-chainA_PDBID                 .......................
Chinese_Giant_Salamander_(A.davi) .......................
```

[Q10] Perform a "Target" search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any Target Associated Assays and ligand efficiency data reported that may be useful starting points for exploring potential inhibition of your novel protein?

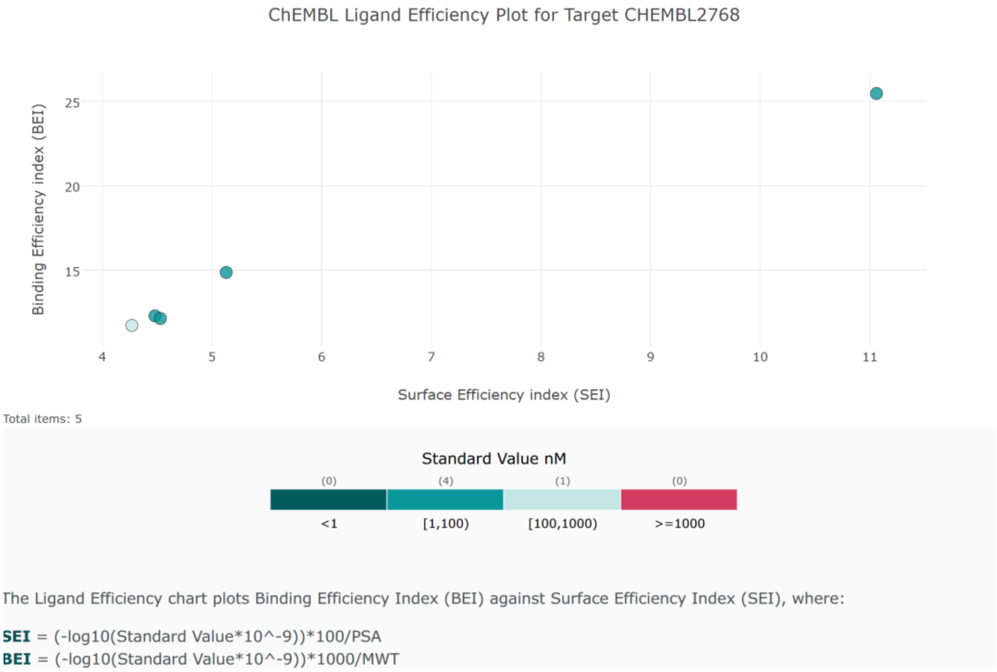
I searched my novel sequence on ChEMBL and got 14 hits. Here is a screenshot of the top six. I selected the first target hit, which is a different class of proteins from my original VASP search.

CHEMBL details for CHEMBL2768: Histone deacetylase 5 from Mus musculus
https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL2768/
One Binding Assay (ligand efficiencies shown below)
Zero Functional Assays



E-Value	Identities %	ChEMBL ID	Name	Type	Organism	Compounds	Activities
0.72	63.2	CHEMBL2768	*Histone deacetylase 5*	SINGLE PROTEIN	Mus musculus	5 By Mol. Wt.:	5 By Std. Type:
0.72	63.2	CHEMBL3038483	*Histone deacetylase 1/3/5/8*	PROTEIN FAMILY	Homo sapiens	46 By Mol. Wt.:	49 By Std. Type:
0.72	63.2	CHEMBL3832944	*Histone deacetylase*	PROTEIN FAMILY	Mus musculus	13 By Mol. Wt.:	13 By Std. Type:
0.72	63.2	CHEMBL2563	*Histone deacetylase 5*	SINGLE PROTEIN	Homo sapiens	433 By Mol. Wt.:	543 By Std. Type:
0.72	63.2	CHEMBL2093865	*Histone deacetylase*	PROTEIN FAMILY	Homo sapiens	2941 By Mol. Wt.:	4716 By Std. Type:

Ligand efficiency data is pasted below:



Total items: 5

The Ligand Efficiency chart plots Binding Efficiency Index (BEI) against Surface Efficiency Index (SEI), where:

SEI = (-log10(Standard Value*10^-9))*100/PSA
BEI = (-log10(Standard Value*10^-9))*1000/MWT