# BGGN-213: FOUNDATIONS OF BIOINFORMATICS

## The find-a-gene project assignment

http://thegrantlab.org/bggn213/

~~Dr. Barry Grant~~

## Overview:

The find-a-gene project is a required assignment for BGGN-213. You should prepare a written report in PDF format that has responses to each question labeled [Q1] - [Q10] below. You may wish to consult the scoring rubric at the end of this document and the example report provided online.
The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered in class.

## Due Date:

Your responses to questions Q1-Q4 are due at the beginning of Week 5. Note that these answers can be obtained very quickly (at best within 10 or 15 minutes), so if you don't succeed at first, just keep trying.
The complete assignment, including responses to all questions, is due at the beginning of Week 10. Late responses will not be accepted under any circumstances.

## Submission instructions:

Submit your PDF document to GradeScope as directed on our class website. Please do make sure your document is in PDF format and named something like BGGN213_F20_[yourUCSDname].pdffor example, my document would be named BGGN213_F20_bjgrant.pdf

Be sure to include your UCSD email and PID number on the first page of your report.

nsasse@ucsd.edu
A59006574

Submit your preliminary report with answers to Q1-Q4 at the beginning of week 5 so we can determine if you have found a novel gene. Submit this preliminary report as one document with screen shots of the results inserted appropriately.

See the demonstration report linked to on the course website for an example of format. I will email you my decision; proceed with subsequent questions only after we are sure you have found a novel gene.

For the final report add your results for Q5-Q10 to the preliminary report and submit a final document containing the results for all questions. Please do not submit only Q5-Q10 answers as the final report.

Questions:

[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 it's 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:  corticotropin releasing hormone receptor 1 (CRHR1)

Accession: XP_038482545

Species: Canis lupus familiaris (dog)

[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).

Method: TBLASTN

Database: Expressed sequence tags (est)

Organism: yellow perch

Also include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10so 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 [].pngin your Desktop directory). It is not necessary to print out all the blast results if there are many pages.

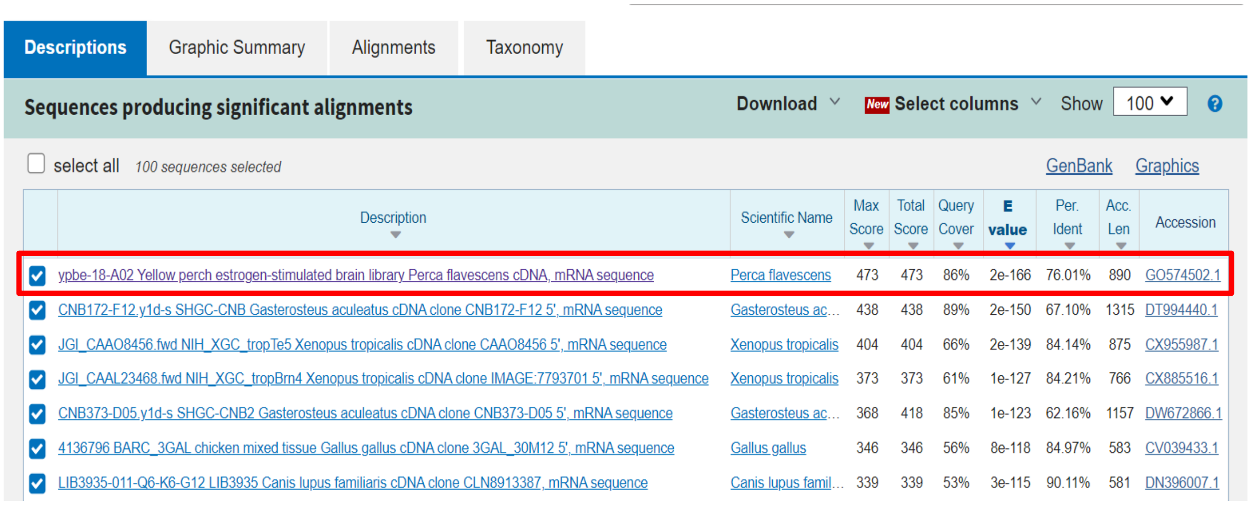
On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or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

Chosen match: Accession GO574502.1,  Yellow perch estrogen-stimulated brain library Perca flavescens cDNA, mRNA sequence.

See below for alignment details

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.

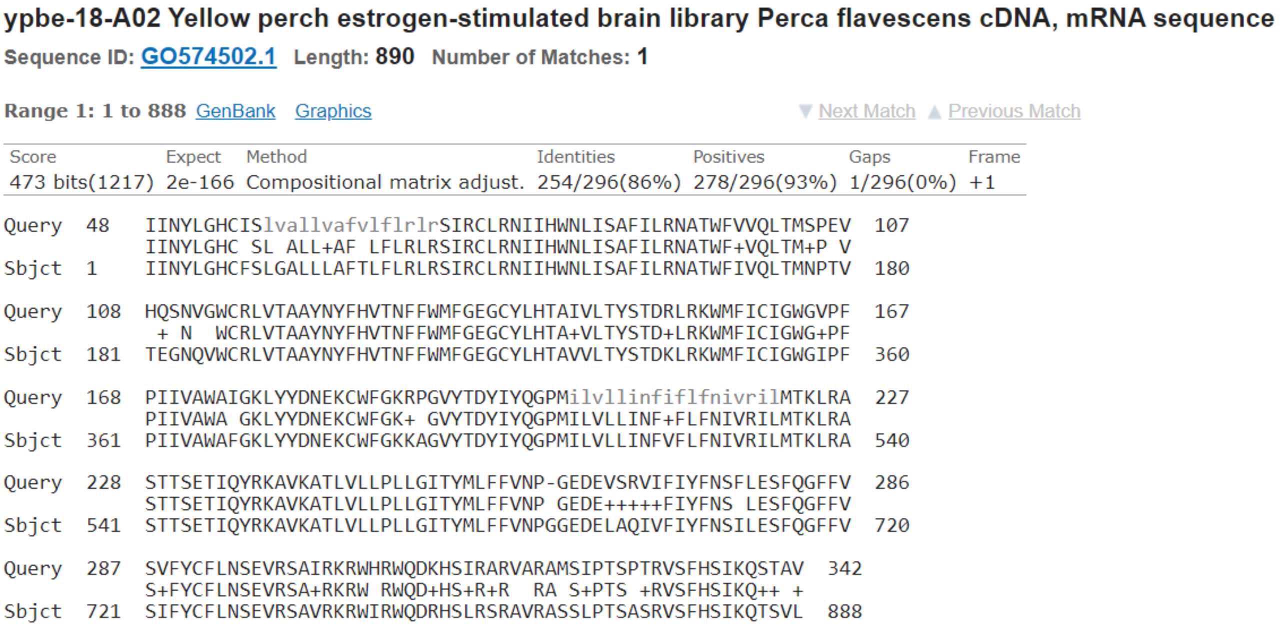
If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.



**Sequences producing significant alignments**

□ select all *100 sequences selected*

| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | ypbe-18-A02 Yellow perch estrogen-stimulated brain library Perca flavescens cDNA, mRNA sequence | Perca flavescens | 473 | 473 | 86% | 2e-166 | 76.01% | 890 | GO574502.1 |
| ☑ | CNB172-F12.y1d-s SHGC-CNB Gasterosteus aculeatus cDNA clone CNB172-F12 5', mRNA sequence | Gasterosteus ac... | 438 | 438 | 89% | 2e-150 | 67.10% | 1315 | DT994440.1 |
| ☑ | JGI_CAAO8456.fwd NIH_XGC_tropTe5 Xenopus tropicalis cDNA clone CAAO8456 5', mRNA sequence | Xenopus tropicalis | 404 | 404 | 66% | 2e-139 | 84.14% | 875 | CX955987.1 |
| ☑ | JGI_CAAL23468.fwd NIH_XGC_tropBrn4 Xenopus tropicalis cDNA clone IMAGE:7793701 5', mRNA sequence | Xenopus tropicalis | 373 | 373 | 61% | 1e-127 | 84.21% | 766 | CX885516.1 |
| ☑ | CNB373-D05.y1d-s SHGC-CNB2 Gasterosteus aculeatus cDNA clone CNB373-D05 5', mRNA sequence | Gasterosteus ac... | 368 | 418 | 85% | 1e-123 | 62.16% | 1157 | DW672866.1 |
| ☑ | 4136796 BARC_3GAL chicken mixed tissue Gallus gallus cDNA clone 3GAL_30M12 5', mRNA sequence | Gallus gallus | 346 | 346 | 56% | 8e-118 | 84.97% | 583 | CV039433.1 |
| ☑ | LIB3935-011-Q6-K6-G12 LIB3935 Canis lupus familiaris cDNA clone CLN8913387, mRNA sequence | Canis lupus famil... | 339 | 339 | 53% | 3e-115 | 90.11% | 581 | DN396007.1 |

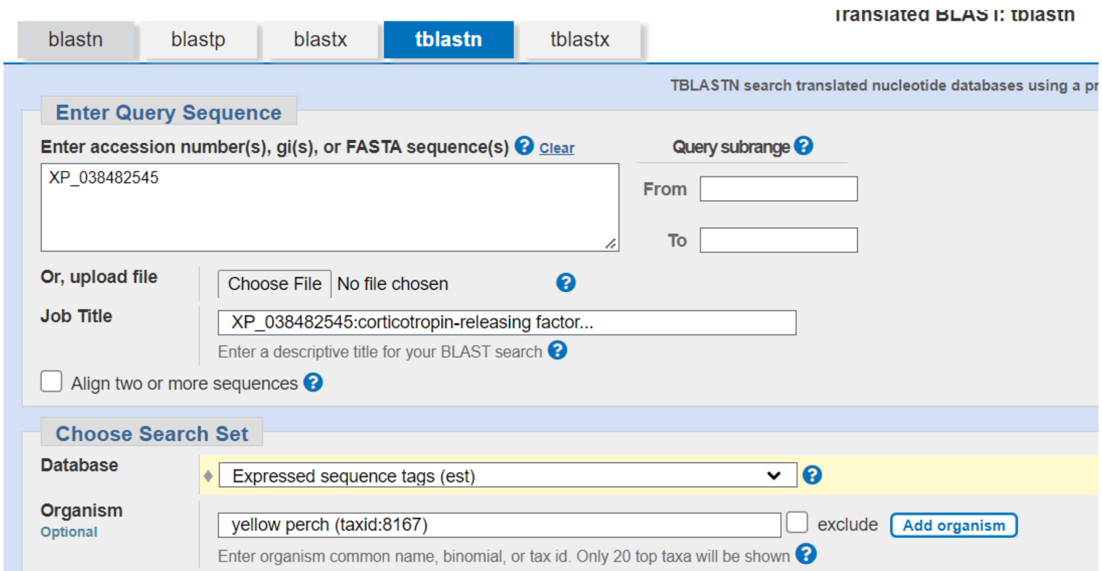


**ypbe-18-A02 Yellow perch estrogen-stimulated brain library Perca flavescens cDNA, mRNA sequence**

**Sequence ID: GO574502.1 Length: 890 Number of Matches: 1**

**Range 1: 1 to 888** GenBank Graphics

```
Score          Expect  Method                       Identities     Positives      Gaps       Frame
473 bits(1217) 2e-166  Compositional matrix adjust. 254/296(86%)   278/296(93%)   1/296(0%)  +1

Query  48    IINYLGHCISlvallvafvlflrlrSIRCLRNIIHWNLISAFILRNATWFVVQLTMSPEV  107
             IINYLGHC SL ALL+AF LFLRLRSIRCLRNIIHWNLISAFILRNATWF+VQLTM+P V
Sbjct  1     IINYLGHCFSLGALLLAFTLFLRLRSIRCLRNIIHWNLISAFILRNATWFIVQLTMNPTV  180

Query  108   HQSNVGWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAIVLTYSTDRLRKWMFICIGWGVPF  167
              + N  WCRLVTAAYNYFHVTNFFWMFGEGCYLHTA+VLTYSTD+LRKWMFICIGWG+PF
Sbjct  181   TEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLRKWMFICIGWGIPF  360

Query  168   PIIVAWAIGKLYYDNEKCWFGKRPGVYTDYIYQGPMilvllinfiflfnivrilMTKLRA  227
             PIIVAWA GKLYYDNEKCWFGK+ GVYTDYIYQGPMILVLLINF+FLFNIVRILMTKLRA
Sbjct  361   PIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMILVLLINFVFLFNIVRILMTKLRA  540

Query  228   STTSETIQYRKAVKATLVLLPLLGITYMLFFVNP-GEDEVSRVIFIYFNSFLESFQGFFV  286
             STTSETIQYRKAVKATLVLLPLLGITYMLFFVNP GEDE+++++FIYFNS LESFQGFFV
Sbjct  541   STTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIYFNSILESFQGFFV  720

Query  287   SVFYCFLNSEVRSAIRKRWHRWQDKHSIRARVARAMSIPTSPTRVSFHSIKQSTAV  342
             S+FYCFLNSEVRSA+RKRW RWQD+HS+R+R  RA S+PTS +RVSFHSIKQ++ +
Sbjct  721   SIFYCFLNSEVRSAVRKRWIRWQDRHSLRSRAVRASSLPTSASRVSFHSIKQTSVL  888
```



TBLASTN search translated nucleotide databases using a pr

**Enter Query Sequence**

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

XP_038482545

Query subrange ?

From

To

Or, upload file: Choose File No file chosen ?

Job Title: XP_038482545:corticotropin-releasing factor...

Enter a descriptive title for your BLAST search ?

□ Align two or more sequences ?

**Choose Search Set**

Database: Expressed sequence tags (est) ?

Organism (Optional): yellow perch (taxid:8167) □ exclude Add organism

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

[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.

```
> Yellow perch, GO574502.1 (from EMBOSS Transeq)

IINYLGHCFSLGALLLAFTLFLRLRSIRCLRNIIHWNLISAFILRNATWFIVQLTMNPTV
TEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLRKWMFICIGWGIPF
PIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMILVLLINFVFLFNIVRILMTKLRA
STTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIYFNSILESFQGFFV
SIFYCFLNSEVRSAVRKRWIRWQDRHSLRSRAVRASSLPTSASRVSFHSIKQTSVLX
```

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.

Name: corticotropin-releasing factor receptor 1

Species: Perca flavescens

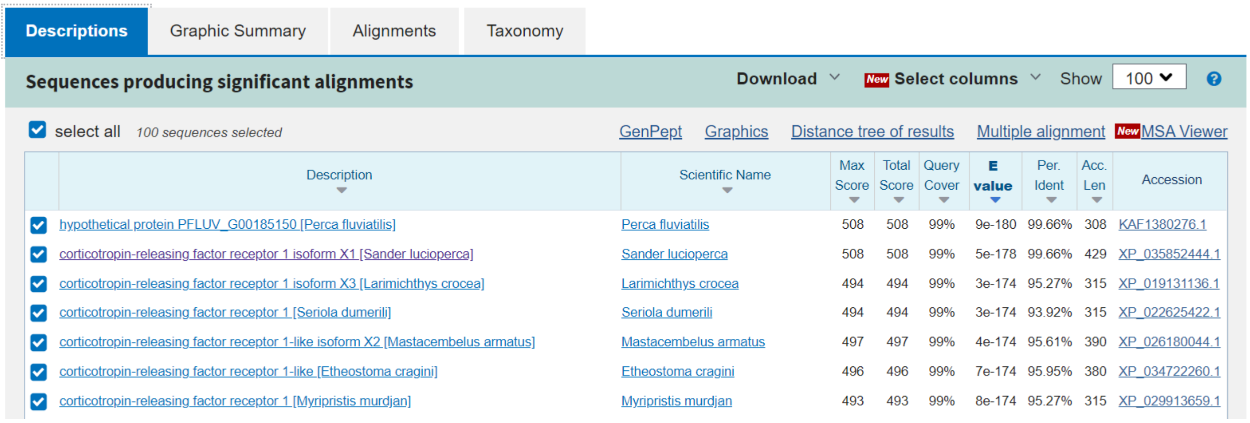
> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
> Actinopterygii; Neopterygii; Teleostei; Neoteleostei;
> Acanthomorphata; Eupercaria; Perciformes; Percoidei; Percidae;
> Percinae; Perca.

[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.

A BLASTP search against NR database (see setup in first screen-shot below) yielded a top hit result is to a protein from Sander lucioperca (pike perch). Most of the results are bony fish.

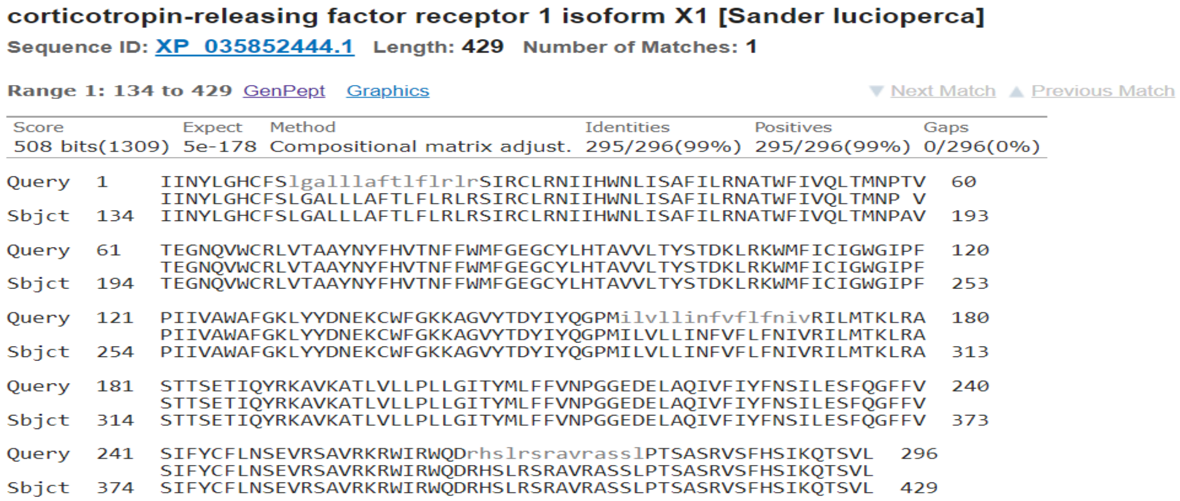
See additional screen shots below for top hits and selected alignment details:



**Sequences producing significant alignments**

☑ select all *100 sequences selected*

| | Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|---|
| ☑ | hypothetical protein PFLUV_G00185150 [Perca fluviatilis] | Perca fluviatilis | 508 | 508 | 99% | 9e-180 | 99.66% | 308 | KAF1380276.1 |
| ☑ | corticotropin-releasing factor receptor 1 isoform X1 [Sander lucioperca] | Sander lucioperca | 508 | 508 | 99% | 5e-178 | 99.66% | 429 | XP_035852444.1 |
| ☑ | corticotropin-releasing factor receptor 1 isoform X3 [Larimichthys crocea] | Larimichthys crocea | 494 | 494 | 99% | 3e-174 | 95.27% | 315 | XP_019131136.1 |
| ☑ | corticotropin-releasing factor receptor 1 [Seriola dumerili] | Seriola dumerili | 494 | 494 | 99% | 3e-174 | 93.92% | 315 | XP_022625422.1 |
| ☑ | corticotropin-releasing factor receptor 1-like isoform X2 [Mastacembelus armatus] | Mastacembelus armatus | 497 | 497 | 99% | 4e-174 | 95.61% | 390 | XP_026180044.1 |
| ☑ | corticotropin-releasing factor receptor 1-like [Etheostoma cragini] | Etheostoma cragini | 496 | 496 | 99% | 7e-174 | 95.95% | 380 | XP_034722260.1 |
| ☑ | corticotropin-releasing factor receptor 1 [Myripristis murdjan] | Myripristis murdjan | 493 | 493 | 99% | 8e-174 | 95.27% | 315 | XP_029913659.1 |

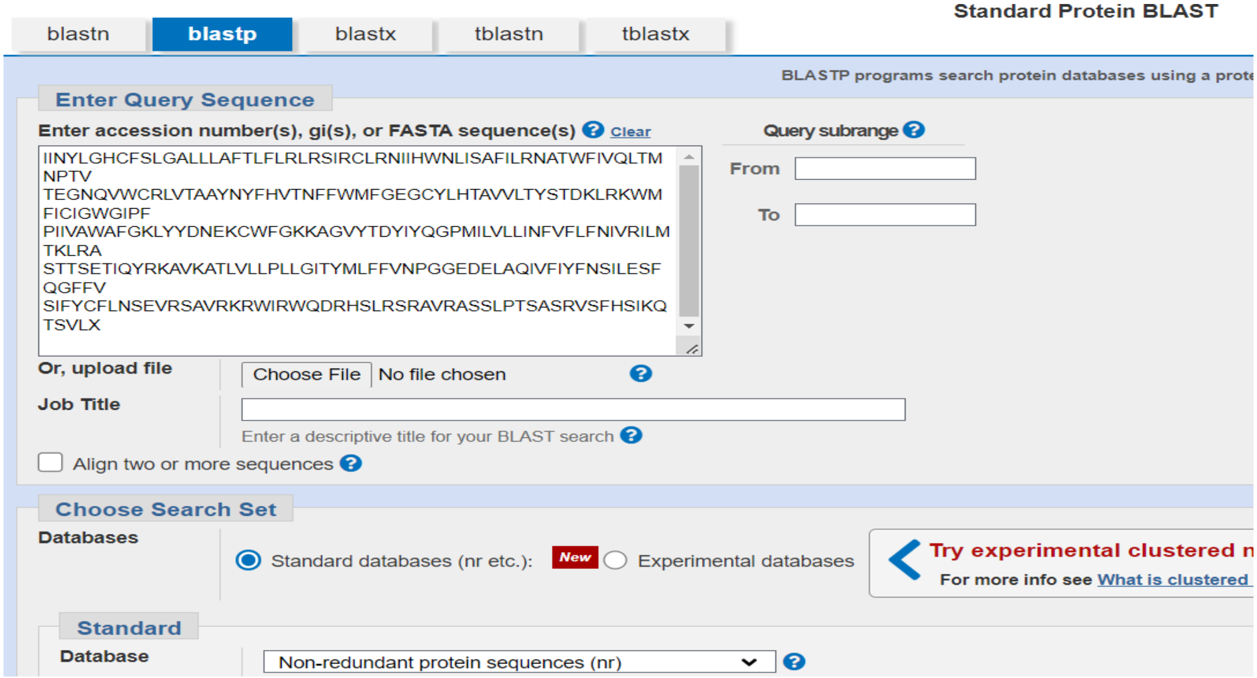


**corticotropin-releasing factor receptor 1 isoform X1 [Sander lucioperca]**

**Sequence ID: XP_035852444.1 Length: 429 Number of Matches: 1**

**Range 1: 134 to 429** GenPept Graphics

```
Score          Expect   Method                         Identities     Positives      Gaps
508 bits(1309) 5e-178   Compositional matrix adjust.   295/296(99%)   295/296(99%)   0/296(0%)

Query  1    IINYLGHCFSlgalllaftlflrlrSIRCLRNIIHWNLISAFILRNATWFIVQLTMNPTV  60
            IINYLGHCFSLGALLLAFTLFLRLRSIRCLRNIIHWNLISAFILRNATWFIVQLTMNP V
Sbjct  134  IINYLGHCFSLGALLLAFTLFLRLRSIRCLRNIIHWNLISAFILRNATWFIVQLTMNPAV  193

Query  61   TEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLRKWMFICIGWGIPF  120
            TEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLRKWMFICIGWGIPF
Sbjct  194  TEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLRKWMFICIGWGIPF  253

Query  121  PIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMilvllinfvflfnivRILMTKLRA  180
            PIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMILVLLINFVFLFNIVRILMTKLRA
Sbjct  254  PIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMILVLLINFVFLFNIVRILMTKLRA  313

Query  181  STTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIYFNSILESFQGFFV  240
            STTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIYFNSILESFQGFFV
Sbjct  314  STTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIYFNSILESFQGFFV  373

Query  241  SIFYCFLNSEVRSAVRKRWIRWQDrhslrsravrasslPTSASRVSFHSIKQTSVL  296
            SIFYCFLNSEVRSAVRKRWIRWQDRHSLRSRAVRASSLPTSASRVSFHSIKQTSVL
Sbjct  374  SIFYCFLNSEVRSAVRKRWIRWQDRHSLRSRAVRASSLPTSASRVSFHSIKQTSVL  429
```



**Standard Protein BLAST**

BLASTP programs search protein databases using a prote

**Enter Query Sequence**

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

```
IINYLGHCFSLGALLLAFTLFLRLRSIRCLRNIIHWNLISAFILRNATWFIVQLTM
NPTV
TEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLRKWM
FICIGWGIPF
PIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMILVLLINFVFLFNIVRILM
TKLRA
STTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIYFNSILESF
QGFFV
SIFYCFLNSEVRSAVRKRWIRWQDRHSLRSRAVRASSLPTSASRVSFHSIKQ
TSVLX
```

**Query subrange**

From

To

**Or, upload file** Choose File No file chosen

**Job Title**

Enter a descriptive title for your BLAST search

□ Align two or more sequences

**Choose Search Set**

**Databases**

● Standard databases (nr etc.): New ○ Experimental databases

**Try experimental clustered n**
For more info see What is clustered

**Standard**

**Database** Non-redundant protein sequences (nr)

[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.

## Re-labeled sequences for alignment:

### Original query protein:

```
>Yellow Perch: 16-433 corticotropin-releasing factor receptor 1-like isoform X3
[Perca flavescens]
MMCVCLFLSGRVSPTQLTCETLMLLSTNLTARMLVFLNQTFGIRNSSGVFCDLSVDGIGTCWPLSAAGQLISRPCPEQFN
```

### Novel protein:

```
>Pike Perch: 1-429 corticotropin-releasing factor receptor 1 isoform X1 [Sander
lucioperca]
MEKLLSQMMCVCLFLSGRVSPTQLTCETLILLSTNLTARTLVFLNQTFGVRNSSGVFCDLSVDGIGTCWPLSAAGQLISR
```

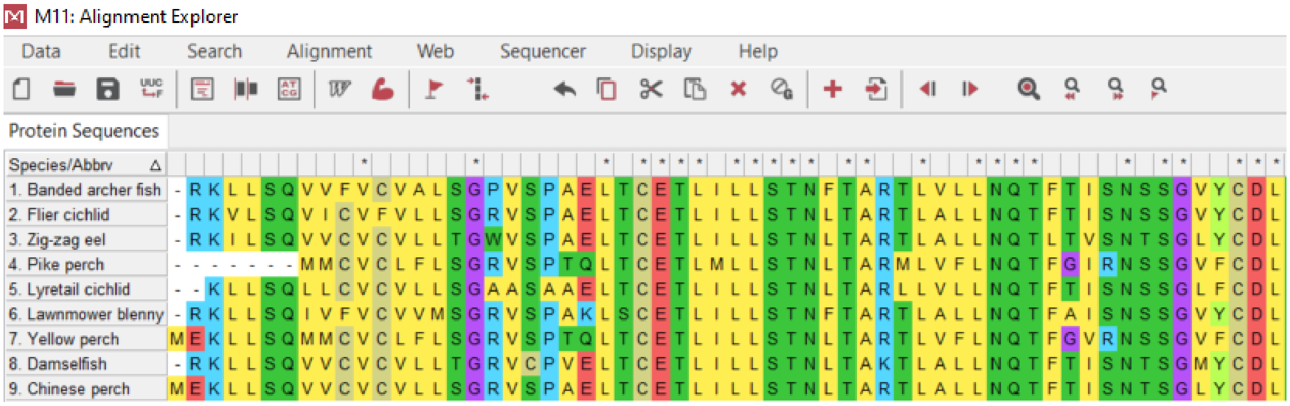
### Other sequences for alignment:

```
>Chinese Perch: 4-432 corticotropin-releasing factor receptor 1 isoform X1
[Siniperca chuatsi]
MEKLLSQVVCVCVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNTSGLYCDLSVDGIGTCWPRSAAGELISR

>Damselfish: 3-430 corticotropin-releasing factor receptor 1-like isoform X2
[Acanthochromis polyacanthus]
RKVLSQVICVFVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNSSGVYCDLSVDGIGTCWPRSAAGELISRP

>Flier cichlid: 5-432 corticotropin-releasing factor receptor 1-like
[Archocentrus centrarchus]
RKLLSQIVFVCVVMSGRVSPAKLSCETLILLSTNFTARTLALLNQTFAISNSSGVYCDLSVDGIGTCWPRSAAGELVSRP

>Zig-zag eel: 3-430 corticotropin-releasing factor receptor 1-like isoform X1
[Mastacembelus armatus]
RKILSQVVCVCVLLTGWVSPAELTCETLILLSTNLTARTLALLNQTLTVSNTSGLYCDLSVDGIGTCWPRSAAGELISRP

>Lawnmower blenny: 4-430 corticotropin-releasing factor receptor 1-like
[Salarias fasciatus]
KLLSQLLCVCVLLSGAASAAELTCETLILLSTNLTARLLVLLNQTFTISNSSGLFCDLSVDGIGTCWPRSAAGELVSRPC

>Banded archerfish:6-433 corticotropin-releasing factor receptor 1 isoform X1
[Toxotes jaculatrix]
RKLLSQVVCVCVLLTGRVCPVELTCETLILLSTNLTAKTLALLNQTFTISNTSGMYCDLSVDGIGTCWPRSAAGELISRP

>Lyretail cichlid: 3-430 corticotropin-releasing factor receptor 1
[Neolamprologus brichardi]
RKLLSQVVFVCVALSGPVSPAELTCETLILLSTNFTARTLVLLNQTFTISNSSGVYCDLSVDGIGTCWPRSAAGELVSRP
```

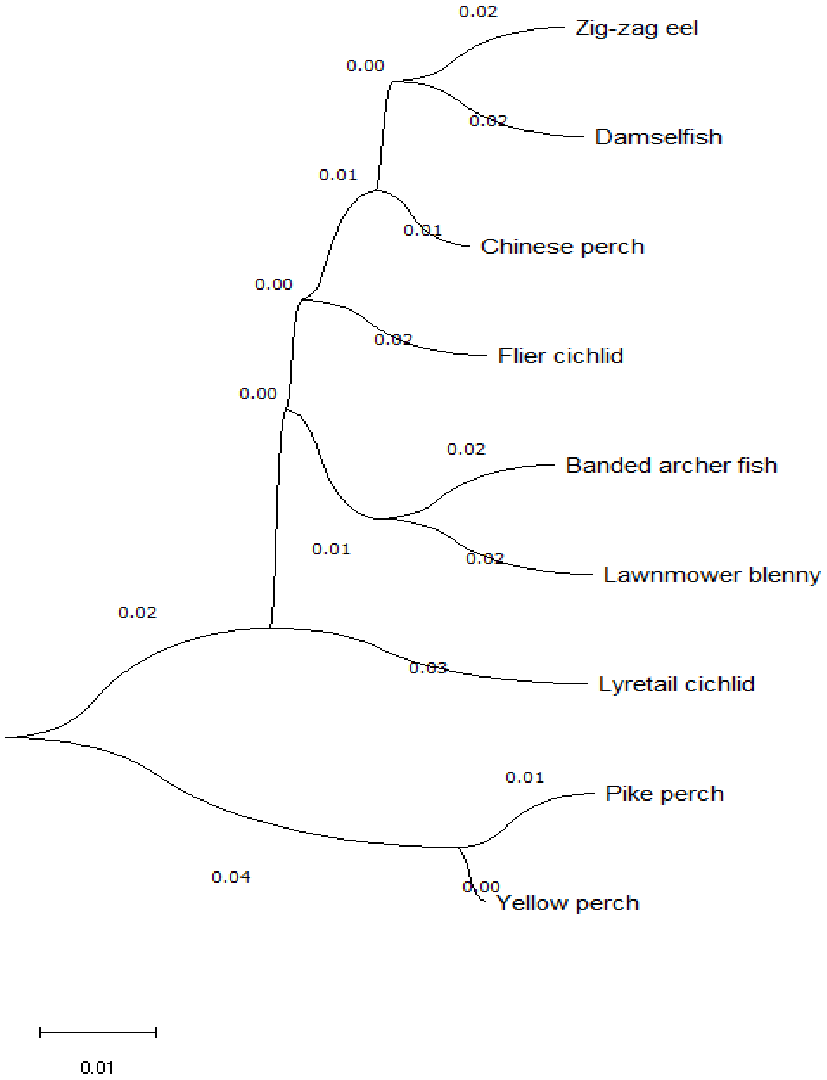
```
CLUSTAL format alignment by MAFFT (v7.503)

Yellow perch:   MEKLLSQMMCVCLFLSGRVSPTQLTCETLILLSTNLTARTLVFLNQTFGVRNSSGVFCDL
Pike perch:     -------MMCVCLFLSGRVSPTQLTCETLMLLSTNLTARMLVFLNQTFGIRNSSGVFCDL
Chinese perch:  MEKLLSQVVCVCVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNTSGLYCDL
Damselfish:     -RKLLSQVVCVCVLLTGRVCPVELTCETLILLSTNLTAKTLALLNQTFTISNTSGMYCDL
Flier cichlid:  -RKVLSQVICVFVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNSSGVYCDL
Zig-zag eel:    -RKILSQVVCVCVLLTGWVSPAELTCETLILLSTNLTARTLALLNQTLTVSNTSGLYCDL
Lawnmower:      -RKLLSQIVFVCVVMSGRVSPAKLSCETLILLSTNFTARTLALLNQTFAISNSSGVYCDL
Banded archer:  -RKLLSQVVFVCVALSGPVSPAELTCETLILLSTNFTARTLVLLNQTFTISNSSGVYCDL
Lyretail:       --KLLSQLLCVCVLLSGAASAAELTCETLILLSTNLTARLLVLLNQTFTISNSSGLFCDL
                       :: * : ::* ....:*:****:*****:**: *.:****: : *:**::***

XP_035852444.1: SVDGIGTCWPLSAAGQLISRPCPEQFNGIHYNTSNRVFRECQTNGSWAPRGNYSQCTEII
XP_028454502.1: SVDGIGTCWPLSAAGQLISRPCPEQFNGIHYNTSNRVFRECQTNGSWAPRGNYSQCTEII
XP_044037885.1: SVDGIGTCWPRSAAGELISRPCPEQFNGIHYNTTNRVYRECQSNGSWAPRGNYSQCTEII
XP_040917920.1: SVDGIGTCWPRSAAGELISRPCPEQFNGIHYNTTNRVYRECQSNGSWALRGNYSQCTEII
XP_022067791.1: SVDGIGTCWPRSAAGELISRPCPEQFNGIHYNTTNRVFRECLSNGSWAPRGNYSQCTEII
XP_026180043.1: SVDGIGTCWPRSAAGELISRPCPEQFNGIHYNTSNRVYRECQFNGSWAPRGNYSQCTEII
XP_030591515.1: SVDGIGTCWPRSAAGELVSRPCPEQFNGIHYNTTNRVYRECQVNGSWAPRGNYSQCTEII
XP_006805303.1: SVDGIGTCWPRSAAGELVSRPCPEQFNGIHYNTTNRVYRECQVNGSWAPRGNYSQCTEII
XP_029944739.1: SVDGIGTCWPRSAAGELVSRPCPEQFNGIHYNTTNRVYRDCQSNGSWAPRGNYSQCTEII
                ********** ****:*:***************:***:*:*  ***** ***********

XP_035852444.1: VMRKSKLHYQVAVIINYLGHCFSLGALLLAFTLFLRLRSIRCLRNIIHWNLISAFILRNA
XP_028454502.1: IMRKTKLHYQVAVIINYLGHCFSLGALLLAFTLFLRLRSIRCLRNIIHWNLISAFILRNA
XP_044037885.1: VLRKSKVHYQVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_040917920.1: VLRKSKVHYQVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_022067791.1: ILRKSKVHYHVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_026180043.1: VLRKSKVHYQVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_030591515.1: ILRKSKVHYQVAVIINYMGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_006805303.1: VLRKSKVHYQVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
XP_029944739.1: VMRKSKVHYHVAVIINYLGHCISLGALLLAFTLFMRLRSIRCLRNIIHWNLISAFILRNA
                ::**:*:**:*******:***:************:*************************

XP_035852444.1: TWFIVQLTMNPAVTEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_028454502.1: TWFIVQLTMNPTVTEGNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_044037885.1: TWFIVQLTMTSAVTESNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_040917920.1: TWFIVQLTMTPAVTESNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_022067791.1: TWFIVQLTMNPAVTESNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAIVLTYSTDKLR
XP_026180043.1: TWFIVQLTMNPAVTESNQVWCRLVTAAYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_030591515.1: TWFIVQLTMNPAVTESNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_006805303.1: TWFIVQLTMNPAVTERNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
XP_029944739.1: TWFIVQLTMNPAVTESNQVWCRLVTAGYNYFHVTNFFWMFGEGCYLHTAVVLTYSTDKLR
                *********..:*** **********.**********************:**********

XP_035852444.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMILVLLINFVFL
XP_028454502.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKKAGVYTDYIYQGPMILVLLINFVFL
XP_044037885.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_040917920.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_022067791.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_026180043.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_030591515.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
```

```
XP_006805303.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
XP_029944739.1: KWMFICIGWGIPFPIIVAWAFGKLYYDNEKCWFGKRAGVYTDYIYQGPMILVLLINFVFL
                ***********************************:************************

XP_035852444.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIY
XP_028454502.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDELAQIVFIY
XP_044037885.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_040917920.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_022067791.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_026180043.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_030591515.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVAQIVFIY
XP_006805303.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVARIVFIY
XP_029944739.1: FNIVRILMTKLRASTTSETIQYRKAVKATLVLLPLLGITYMLFFVNPGGEDEVSQIVFIY
                ****************************************************:::*****

XP_035852444.1: FNSILESFQGFFVSIFYCFL-NSEVRSAVRKRWIRWQDRHSLRSRAVRASSLPTSASRVS
XP_028454502.1: FNSILESFQGFFV-----FLNNSEVRSAVRKRWIRWQDRHSLRSRAVRASSLPTSASRVS
XP_044037885.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSFRSRAVRATSLPTSPSRVS
XP_040917920.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRTVRATSLPTSPSRVS
XP_022067791.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRAVRATSLPTSPSRVS
XP_026180043.1: FNSILESFQGFFVSVFYCFL-NSEVRSAARKRWIRWQDRHSIRSRAVRATSLPTSPSRVS
XP_030591515.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRAVRATSLPTSPSRVS
XP_006805303.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRVIRATSLPTSPSRVS
XP_029944739.1: FNSILESFQGFFVSVFYCFL-NSEVRSAVRKRWIRWQDRHSIRSRTVRATSLPTSPSRVS
                *************     ** *******.************:***.:**:*****.****
```

[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.



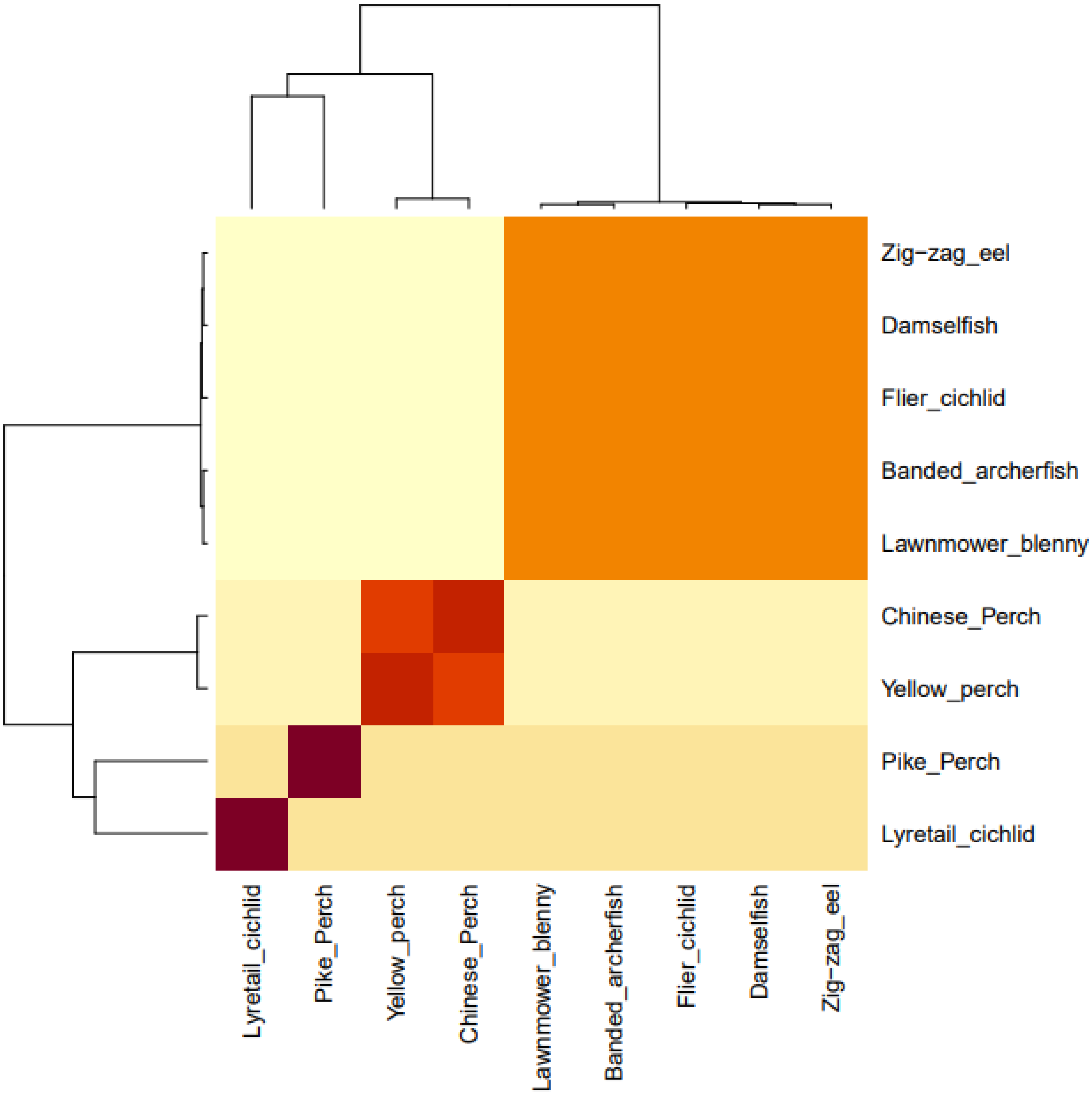
```
Species/Abbrv                    *     *      * **** ***** **  *  ****    * **  ***
1. Banded archer fish  -RKLLSQVVFVCVALSGPVSPAELTCETLILLSTNFTARTLVLLNQTFTISNSSGVYCDL
2. Flier cichlid       -RKVLSQVICVFVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNSSGVYCDL
3. Zig-zag eel         -RKILSQVVCVCVLLTGWVSPAELTCETLILLSTNLTARTLALLNQTLTVSNTSGLYCDL
4. Pike perch          -------MMCVCLFLSGRVSPTQLTCETLMLLSTNLTARMLVFLNQTFGIRNSSGVFCDL
5. Lyretail cichlid    --KLLSQLLCVCVLLSGAASAAELTCETLILLSTNLTARLLVLLNQTFTISNSSGLFCDL
6. Lawnmower blenny    -RKLLSQIVFVCVVMSGRVSPAKLSCETLILLSTNFTARTLALLNQTFAISNSSGVYCDL
7. Yellow perch        MEKLLSQMMCVCLFLSGRVSPTQLTCETLILLSTNLTARTLVFLNQTFGVRNSSGVFCDL
8. Damselfish          -RKLLSQVVCVCVLLTGRVCPVELTCETLILLSTNLTAKTLALLNQTFTISNTSGMYCDL
9. Chinese perch       MEKLLSQVVCVCVLLSGRVSPAELTCETLILLSTNLTARTLALLNQTFTISNTSGLYCDL
```



0.02
Zig-zag eel
0.00
0.02
Damselfish
0.01
0.01
Chinese perch
0.00
0.02
Flier cichlid
0.00
0.02
Banded archer fish
0.01
0.02
Lawnmower blenny
0.02
0.03
Lyretail cichlid
0.01
Pike perch
0.04
0.00
Yellow perch
0.01

[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.



[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.

| ID | Technique | Resolution | Source | E-value | Identity |
|---|---|---|---|---|---|
| 6P9X | Electron microscopy | 2.910 | Homo sapiens | 4.9123 e-205 | 80% |
| 6PB0 | Electron microscopy | 3.000 | Homo sapiens | 1.302 e-198 | 81% |
| 4Z9G | X-ray diffraction | 3.183 | Entero-bacteria | 7.816 e-121 | 53% |

[**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?

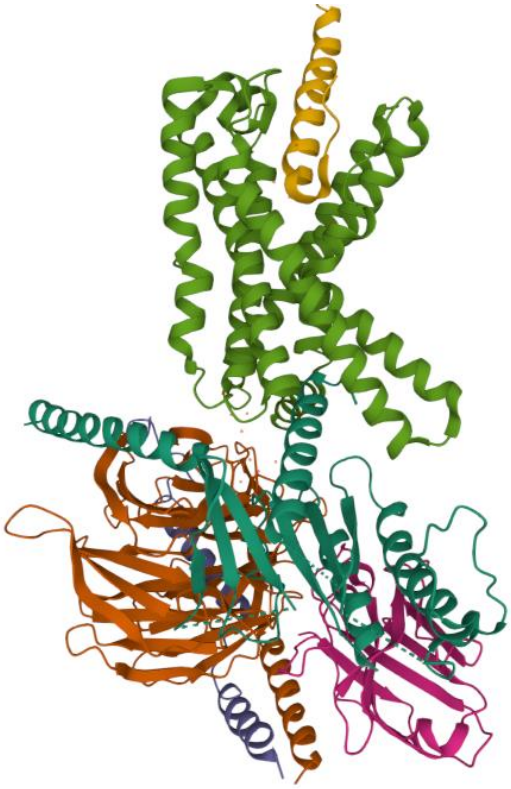
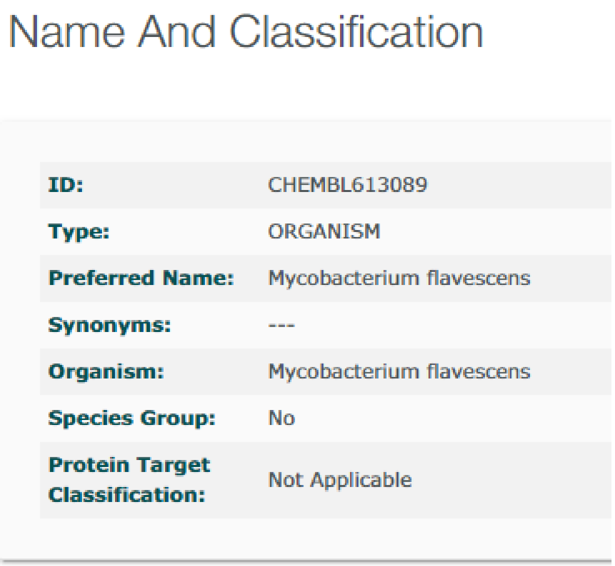


*Figure 1: 6P9X*

[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?

CHEMBL details 6 Functional Assays for CHEMBL613089; No ligand efficiency data.

https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL613089/



## Name And Classification

| | |
|---|---|
| **ID:** | CHEMBL613089 |
| **Type:** | ORGANISM |
| **Preferred Name:** | Mycobacterium flavescens |
| **Synonyms:** | --- |
| **Organism:** | Mycobacterium flavescens |
| **Species Group:** | No |
| **Protein Target Classification:** | Not Applicable |