Shanzida Jahan Siddique
SRavichandran
BIFX-550
5th May 2018

Find -a -Gene project

[1] Tell me the name of a protein you are interested in. Include the species and the accession number. If you do not have a favorite protein, select a protein that is associated with a disease.

Ans.
Name of the protein: Interferon-induced transmembrane protein, IFITM3
Species: *Homo sapiens*
Accession number: NP_066362
Uniprot ID: Q01628
Function of the protein: The protein encoded by this gene is an interferon-induced membrane protein that helps confer immunity to influenza A H1N1 virus, West Nile virus, and dengue virus.

[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.

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.

In general, step [2] 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 step [4]), and a non-homologous result.

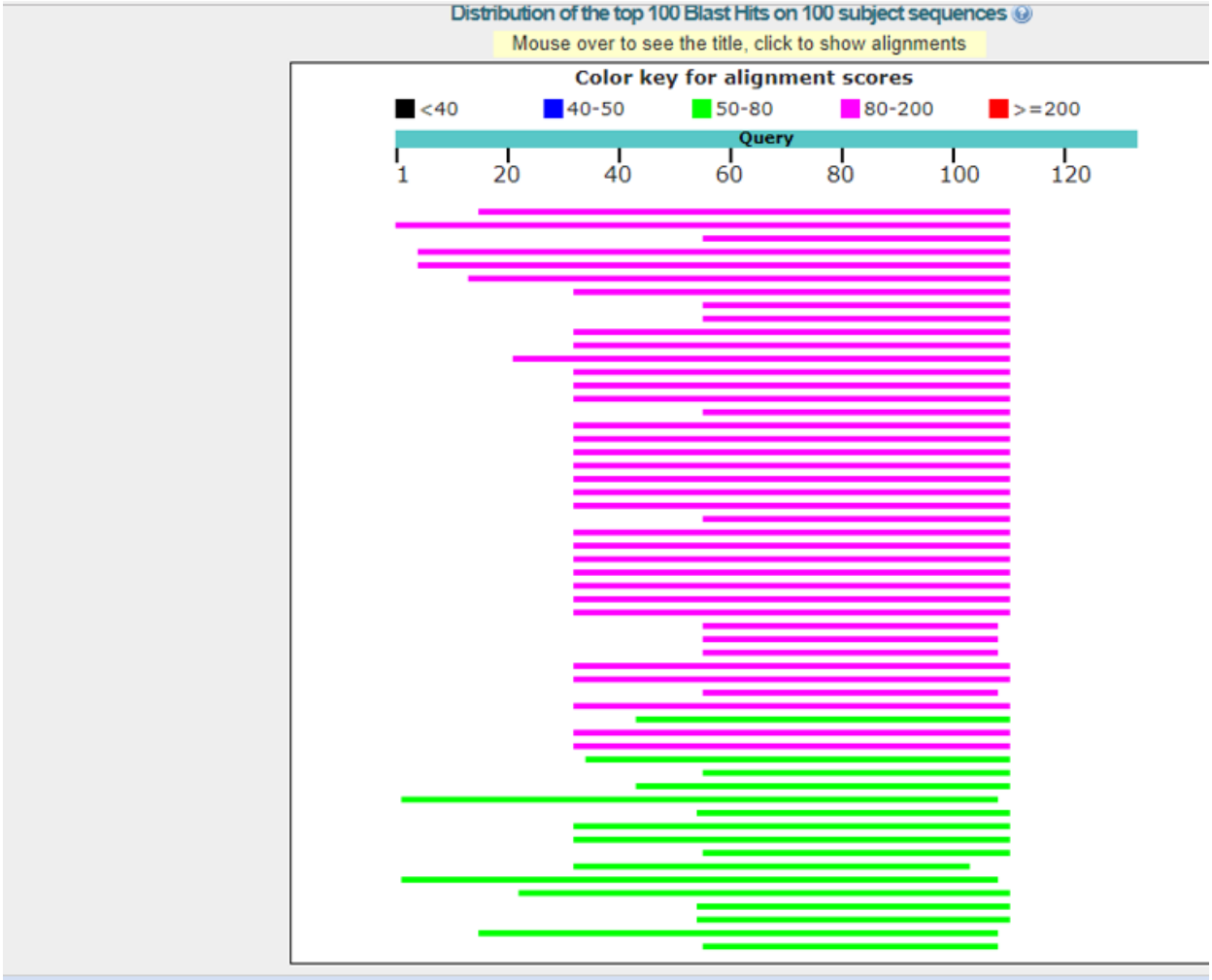
Ans.

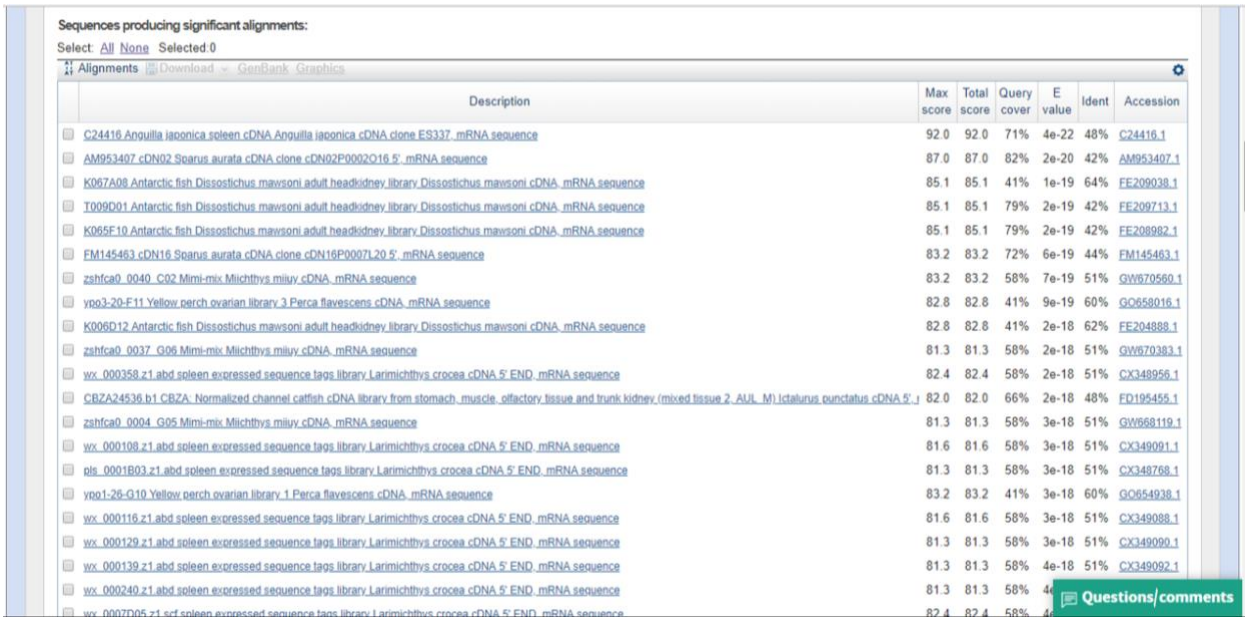
First step:

Program: TBLASTN
Database: EST
Searched against: Fish (Taxid 7868)

Algorithm parameter:

Max target sequence:100
Expect threshold:1000
Word size:2
Matrix: BLOSUM62
Filter: Low complexity region
Blast output:





Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
C24416 Anguilla japonica spleen cDNA Anguilla japonica cDNA clone ES337, mRNA sequence	92.0	92.0	71%	4e-22	48%	C24416.1
AM953407 cDN02 Sparus aurata cDNA clone cDN02P0002O16 5', mRNA sequence	87.0	87.0	82%	2e-20	42%	AM953407.1
K067A08 Antarctic fish Dissostichus mawsoni adult headkidney library Dissostichus mawsoni cDNA, mRNA sequence	85.1	85.1	41%	1e-19	64%	FE209038.1
T009D01 Antarctic fish Dissostichus mawsoni adult headkidney library Dissostichus mawsoni cDNA, mRNA sequence	85.1	85.1	79%	2e-19	42%	FE209713.1
K065F10 Antarctic fish Dissostichus mawsoni adult headkidney library Dissostichus mawsoni cDNA, mRNA sequence	85.1	85.1	79%	2e-19	42%	FE208982.1
FM145463 cDN16 Sparus aurata cDNA clone cDN16P0007L20 5', mRNA sequence	83.2	83.2	72%	6e-19	44%	FM145463.1
zshfca0_0040_C02 Mimi-mix Miichthys miiuy cDNA, mRNA sequence	83.2	83.2	58%	7e-19	51%	GW670560.1
ypo3-20-F11 Yellow perch ovarian library 3 Perca flavescens cDNA, mRNA sequence	82.8	82.8	41%	9e-19	60%	GO658016.1
K006D12 Antarctic fish Dissostichus mawsoni adult headkidney library Dissostichus mawsoni cDNA, mRNA sequence	82.8	82.8	41%	2e-18	62%	FE204888.1
zshfca0_0037_G06 Mimi-mix Miichthys miiuy cDNA, mRNA sequence	81.3	81.3	58%	2e-18	51%	GW670383.1
wx_000358.z1.abd spleen expressed sequence tags library Larimichthys crocea cDNA 5' END, mRNA sequence	82.4	82.4	58%	2e-18	51%	CX348956.1
CBZA24536.b1 CBZA: Normalized channel catfish cDNA library from stomach, muscle, olfactory tissue and trunk kidney (mixed tissue 2, AUL_M) Ictalurus punctatus cDNA 5'	82.0	82.0	66%	2e-18	48%	FD195455.1
zshfca0_0004_G05 Mimi-mix Miichthys miiuy cDNA, mRNA sequence	81.3	81.3	58%	3e-18	51%	GW668119.1
wx_000108.z1.abd spleen expressed sequence tags library Larimichthys crocea cDNA 5' END, mRNA sequence	81.6	81.6	58%	3e-18	51%	CX349091.1
pls_0001B03.z1.abd spleen expressed sequence tags library Larimichthys crocea cDNA 5' END, mRNA sequence	81.3	81.3	58%	3e-18	51%	CX348768.1
ypo1-26-G10 Yellow perch ovarian library 1 Perca flavescens cDNA, mRNA sequence	83.2	83.2	41%	3e-18	60%	GO654938.1
wx_000116.z1.abd spleen expressed sequence tags library Larimichthys crocea cDNA 5' END, mRNA sequence	81.6	81.6	58%	3e-18	51%	CX349088.1
wx_000129.z1.abd spleen expressed sequence tags library Larimichthys crocea cDNA 5' END, mRNA sequence	81.3	81.3	58%	3e-18	51%	CX349090.1
wx_000139.z1.abd spleen expressed sequence tags library Larimichthys crocea cDNA 5' END, mRNA sequence	81.3	81.3	58%	4e-18	51%	CX349092.1
wx_000240.z1.abd spleen expressed sequence tags library Larimichthys crocea cDNA 5' END, mRNA sequence	81.3	81.3	58%	4e		
wx_0007D05.z1.scf spleen expressed sequence tags library Larimichthys crocea cDNA 5' END, mRNA sequence	82.4	82.4	58%	4e		

Questions/comments

Sequence taken (The topmost hit)

C24416 Anguilla japonica spleen cDNA Anguilla japonica cDNA clone ES337, mRNA sequence.

Sequence ID: C24416.1Length: 601Number of Matches: 1

Related Information

Range 1: 245 to 508GenBankGraphicsNext MatchPrevious Match

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
92.0 bits(227)	4e-22	Compositional matrix adjust.	46/95(48%)	63/95(66%)	7/95(7%)

```
Query  16   QPPNYEMLKEEHEVAVLGAPHNPAPPTSTVIHIRSETSVPDHVVWSLFNTLFMNPCCLGF  75
            QP  YE LK+  + A++   +N +P     I +   +   DH+VWSLFN  +MN  CLGF
Sbjct  508  QPNTYERLKDPRDAAMI---NNQSPR----IMVAPISPPRDHIVWSLFNFFYMNAFCLGF  350

Query  76   IAFAYSVKSRDRKMVGDVTGAQAYASTAKCLNIWA  110
            +A  +S+KSRDRK+VGD+ GA+ Y STA+CLNI A
Sbjct  349  VALYFSIKSRDRKVVGDLEGAREYGSTARCLNIVA  245
```

Second step:

Collection of mRNA sequence:

C24416 Anguilla japonica spleen cDNA Anguilla japonica cDNA clone ES337, mRNA sequence

GenBank: C24416.1

```
>C24416.1 C24416 Anguilla japonica spleen cDNA Anguilla japonica cDNA clone
ES337, mRNA sequence
TTTTTTTTAATGTAACCAAGATAAGTGCATACGAAAAATAGACATTCATATCAGGGATATAATGGTTTTA
AATCAGAATCCACAGAGGACGTTTCAGAAAAAAGATGCAAGCAAGGAAAGAACCACGTGATCGCGTTGGC
TAATGCCTCCCCTCCATCATCTGTGAGATGTTTTGCTGGAGGGCGATAGCGCCCATGGCCATCAGGACGA
TGATGACGATTATGAACAGCAGGGTGAGACAGAGCGCCACAATGTTGAGGCAGCGCGCGGTGGAGCCGTA
TTCTCGCGCCCCTTCCAAGTCGCCCACCACCTTCCGGTCTCTGGACTTAATTGAGAAATAAAGAGCCACA
AAGCCGAGGCAGAACGCGTTCATGTAGAAAAAGTTGAACAGGGACCACACAATGTGGTCCCGAGGAGGCG
AAATCGGTGCCACCATTATCCGCGGCGACTGATTGTTGATCATCGCCGCATCCCGGGGATCTTTCAGTCG
CTCGTAGGTGTTGGGCTGCATTGGCACGCACTCCGCTGGGTACTGCATGGTCTCCGTCTTAATGTGATCA
GCGCTGAGCTGGGCGCAAGTTGTCAGGAGTCCTGTTCGTGA
```

Program: blastx

Database: Non-redundant protein sequence

Algorithm parameter:

- Max target sequence:100
- Expect threshold:100
- Word size:6
- Matrix: BLOSUM62
- Filter: Low complexity region

Blast output: According to blastx result there are no protein which is 100% identical.

[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.

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 primates or protozoa.

Ans.
Name of the novel protein: C24416 Anguilla japonica-IFITM3
Family: CD225
Predicted Sequence of the novel protein:
MQYPAECVPMQPNTYERLKDPRDAAMINNQSPRIMVAPISPPRDHIVWSLFNFFYMNAFCLGFV ALYFSIKSRDRKVVGDLEGAREYGSTARCLNIVALCLTLLFIIVIIVLMAMGAIALQQNISQMM EGRH

Species from which it was derived: *Anguilla japonica* (Japanese eel)

[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.

--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 there is a match with 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 succeeded in finding a novel gene.

--If there are no database matches to the original query from step [1], 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.

Ans.

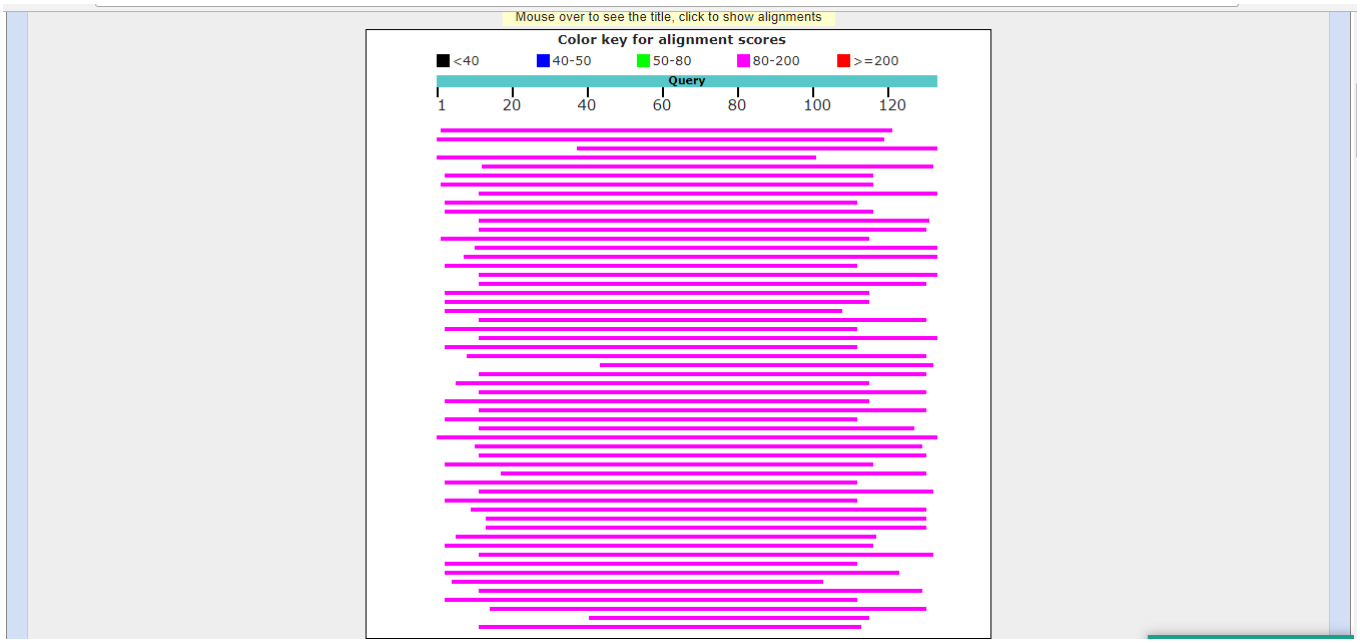
Third step:

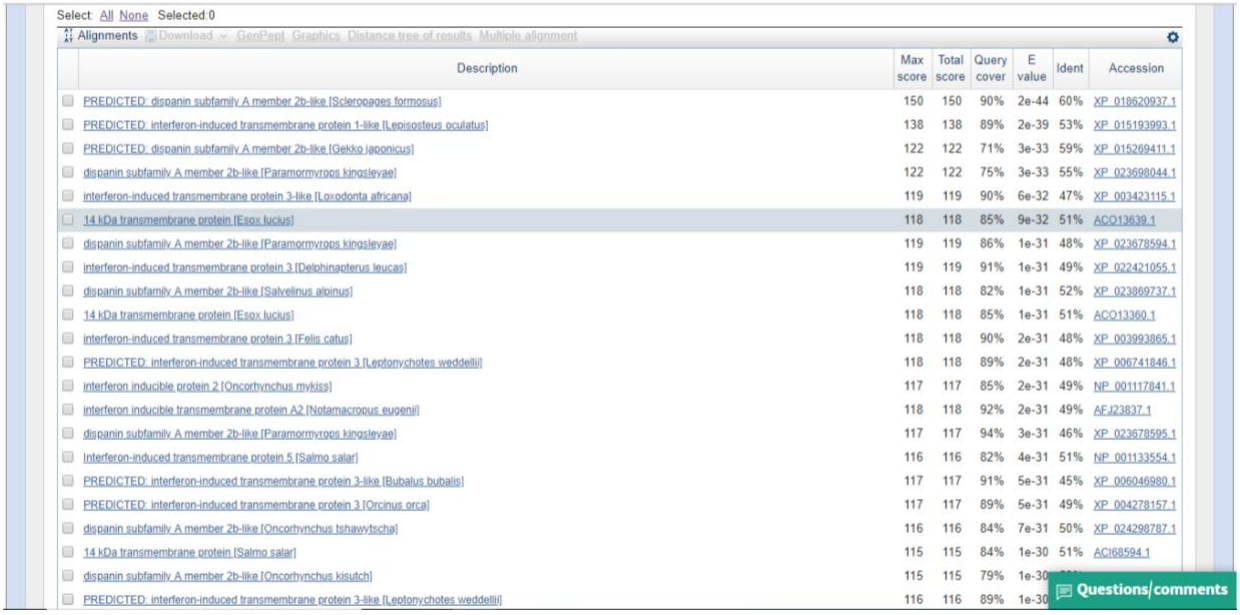
Novel gene's mRNA sequence has translated by ExPASy and taken the longest open reading frame sequence.
Program: blastp
Database: Non-redundant protein sequence
Algorithm parameter:
Max target sequence:100
Expect threshold:100
Word size:2
Matrix: BLOSUM62
Compositional adjustment: no adjustment
Blast output: According to blastp result there are no protein which is 100% identical as shown in the graphics below.

In addition, the identity/similarity of my novel protein with the original query (human IFITM3) is as follows:

```
Identity/similarity between IFITM3 vs Novel Protein:
  Alignment length: 144
Identical residues: 54
  Similar residues: 25
  Percent identity: 37.50
Percent similarity: 54.86
```

This information suggests that my discovered novel protein might play role in IFITM3-like anti-viral resistance function.





Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: dispanin subfamily A member 2b-like [Scleropages formosus]	150	150	90%	2e-44	60%	XP_018620937.1
PREDICTED: interferon-induced transmembrane protein 1-like [Lepisosteus oculatus]	138	138	89%	2e-39	53%	XP_015193993.1
PREDICTED: dispanin subfamily A member 2b-like [Gekko japonicus]	122	122	71%	3e-33	59%	XP_015269411.1
dispanin subfamily A member 2b-like [Paramormyrops kingsleyae]	122	122	75%	3e-33	55%	XP_023698044.1
interferon-induced transmembrane protein 3-like [Loxodonta africana]	119	119	90%	6e-32	47%	XP_003423115.1
14 kDa transmembrane protein [Esox lucius]	118	118	85%	9e-32	51%	ACO13639.1
dispanin subfamily A member 2b-like [Paramormyrops kingsleyae]	119	119	86%	1e-31	48%	XP_023678594.1
interferon-induced transmembrane protein 3 [Delphinapterus leucas]	119	119	91%	1e-31	49%	XP_022421055.1
dispanin subfamily A member 2b-like [Salvelinus alpinus]	118	118	82%	1e-31	52%	XP_023869737.1
14 kDa transmembrane protein [Esox lucius]	118	118	85%	1e-31	51%	ACO13360.1
interferon-induced transmembrane protein 3 [Felis catus]	118	118	90%	2e-31	48%	XP_003993865.1
PREDICTED: interferon-induced transmembrane protein 3 [Leptonychotes weddellii]	118	118	89%	2e-31	48%	XP_006741846.1
interferon inducible protein 2 [Oncorhynchus mykiss]	117	117	85%	2e-31	49%	NP_001117841.1
interferon inducible transmembrane protein A2 [Notamacropus eugenii]	118	118	92%	2e-31	49%	AFJ23837.1
dispanin subfamily A member 2b-like [Paramormyrops kingsleyae]	117	117	94%	3e-31	46%	XP_023678595.1
Interferon-induced transmembrane protein 5 [Salmo salar]	116	116	82%	4e-31	51%	NP_001133554.1
PREDICTED: interferon-induced transmembrane protein 3-like [Bubalus bubalis]	117	117	91%	5e-31	45%	XP_006046980.1
PREDICTED: interferon-induced transmembrane protein 3 [Orcinus orca]	117	117	89%	5e-31	49%	XP_004278157.1
dispanin subfamily A member 2b-like [Oncorhynchus tshawytscha]	116	116	84%	7e-31	50%	XP_024298787.1
14 kDa transmembrane protein [Salmo salar]	115	115	84%	1e-30	51%	ACI68594.1
dispanin subfamily A member 2b-like [Oncorhynchus kisutch]	115	115	79%	1e-30		
PREDICTED: interferon-induced transmembrane protein 3-like [Leptonychotes weddellii]	116	116	89%	1e-30		

Questions/comments

Pfam Domain Analyses:

Significant Pfam-A Matches

Show or hide all alignments.

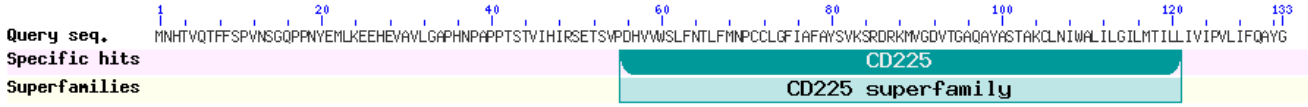
Family	Description	Entry type	Clan	Envelope Start	Envelope End	Alignment Start	Alignment End	HMM From	HMM To	HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
CD225	Interferon-induced transmembrane protein	Family	n/a	41	109	41	109	1	68	68	98.1	2.1e-28	n/a	Hide

```
#HMM      vpkdylvwsifntlf...cclplGivAivySvksrdrkaVGdlegAqsasstAkilniialvlgilliili
#MATCH    +p+d++vws+fn+++   +cl  G+vA+++S+ksrdrk vGdlegA++++stA++lni+al+l++l+ii+i
#PP       699***********99999..*****************************************986
#SEQ      PPRDHIVWSLFNFFYmnaFCL--GFVALYFSIKSRDRKVVGDLEGAREYGSTARCLNIVALCLTLLFIIVI
```

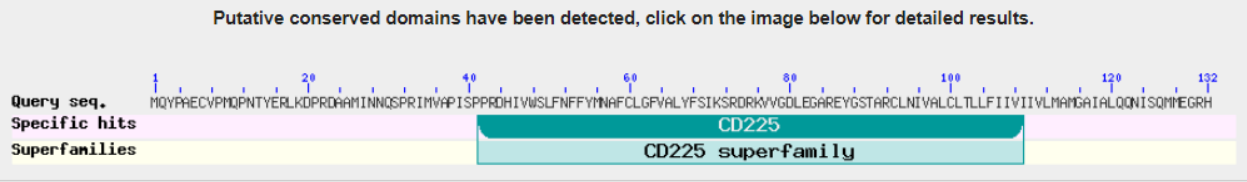
One domain has found after analyzing the protein sequence by Pfam and that is CD225.

Protein Domain Conservation Between IFITM3 and C24416:

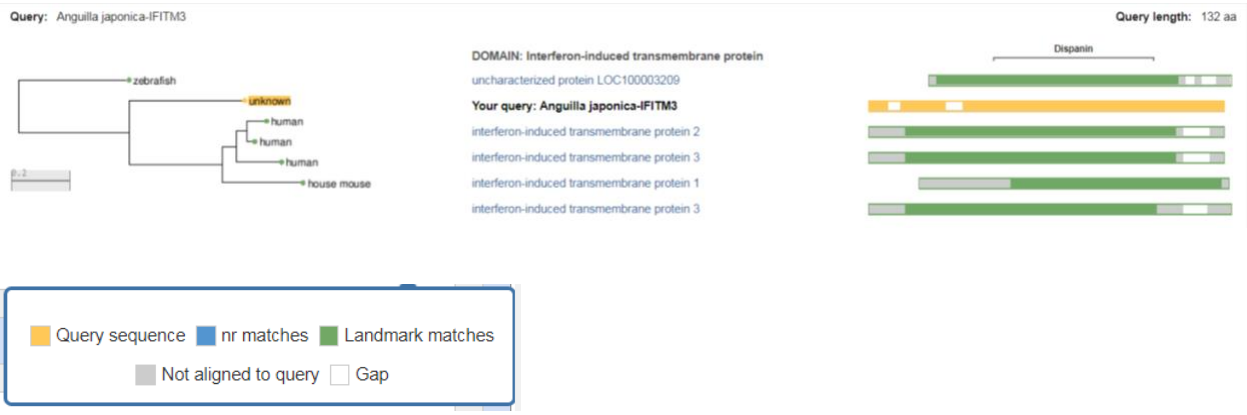
IFITM3:



C24416:



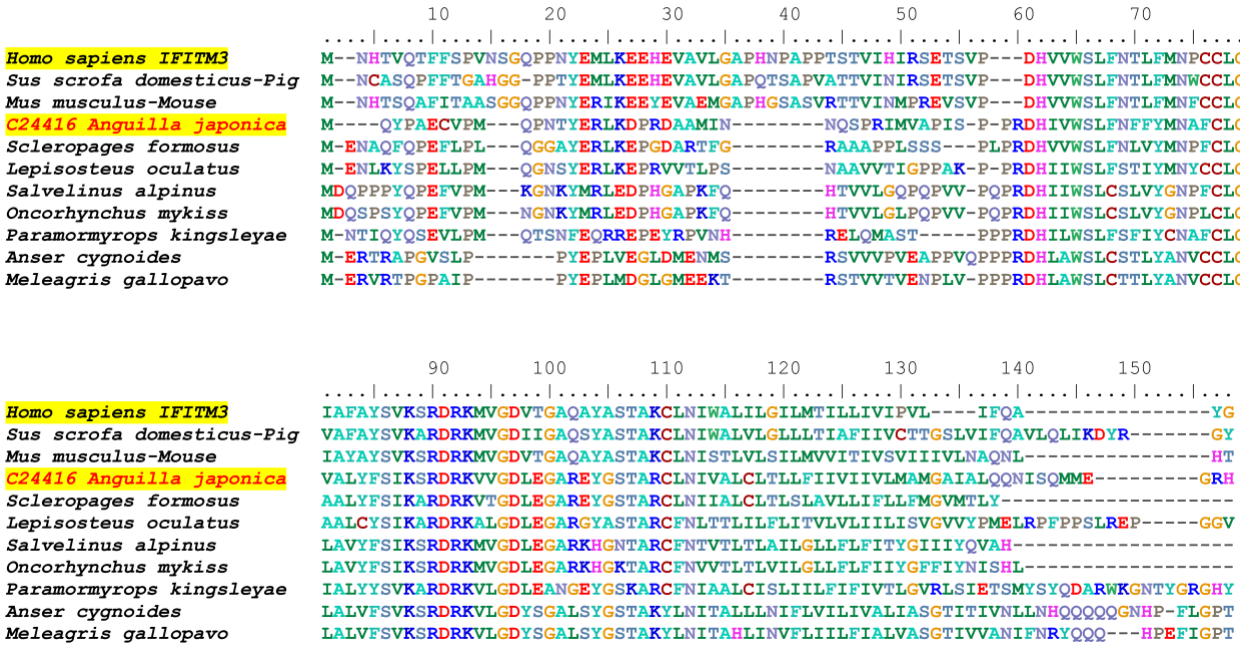
SMART-BLAST: Five Best Protein Matches from Well-Studied Reference Species



[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.

Ans: To generate multiple sequence alignment, I have used C24416 predicted protein sequence to search blastp and took top 5 hits. In addition, to include diversity I have taken the characterized IFITM3 proteins from human, domestic pig, mouse, duck or turkey, and then

aligned by MAFFT software, manually edited by Bioedit software. The alignment is given below.



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

Ans: To generate the phylogenetic tree, I have used C24416 predicted protein sequence to search blastp and taken top 5 hits. In addition, to include diversity I have taken the characterized IFITM3 proteins from human, domestic pig, mouse, duck or turkey, and then aligned by Muscle software and made the phylogenetic tree by MEGA7 using parsimony approach. The tree is given below as well as the description of making the tree from MEGA

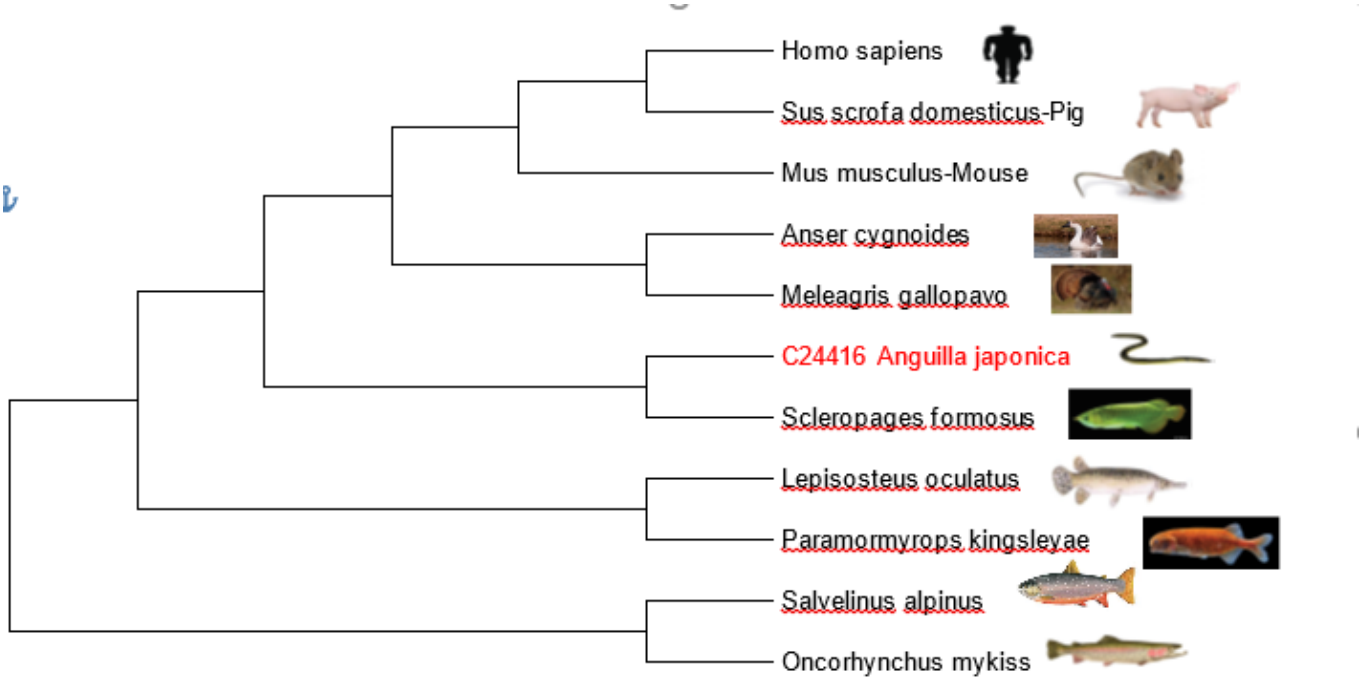


Figure. Maximum Parsimony analysis of taxa

The evolutionary history was inferred using the Maximum Parsimony method. Tree #1 out of 2 most parsimonious trees (length = 329) is shown. The consistency index is 0.820669 (0.807190), the retention index is 0.700508 (0.700508), and the composite index is 0.574885 (0.565442) for all sites and parsimony-informative sites (in parentheses). The MP tree was obtained using the Subtree-Pruning-Regrafting (SPR) algorithm (pg. 126 in ref. [1]) with search level 0 in which the initial trees were obtained by the random addition of sequences (10 replicates). The analysis involved 11 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 107 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [2].

[7] Compare the predicted structure of your protein to that of a known structure. You can use the online protein prediction servers for modeling.

Ans: No structure has been solved yet for Human IFITM3 or any of its homologues from other organisms. Therefore, prediction of C24416 structure by Phyre2 couldn't detect IFITM3 structure for modeling. Rather, Phyre2 modeled 49 of the residues by *ab initio* and rest of them was modeled based on three known structures as shown below.

Figure left: Predicted structure of C24416 by Phyre2

Known structures used for modeling are as follows.

Multi-template and *ab initio* information

3 templates were selected to model your protein based on heuristics to maximise confidence, percentage identity and alignment coverage. Below is a table indicating where your sequence was covered by each template, colour-coded by the confidence of the match to that template overall.

49 residues were modelled by *ab initio*. Please note: *ab initio* modelling is **highly** unreliable.

Template	Confidence	1
c2o7aA	58%	
c2bg9A	51%	
c3ipdB	77%	

Template	Confidence	101
c2o7aA	58%	
c2bg9A	51%	
c3ipdB	77%	

C207aA is the structure of lysozyme from hydrolase family, c3ipdB is the structure of Syntaxin-1 that is involved in exocytosis and the c2bg9A is the structure of acetylcholine receptor protein from ion channel family. It does make sense that the protein was modeled against other membrane proteins as the protein is also predicted to reside in the membrane. Interestingly, a prediction of the human IFITM3 by Phyr2 also resulted in C24416-like structure (not shown here) and was also modelled against the above structures used for C24416 modeling suggesting that both IFIMT3 and C24416 might be structurally similar.

[8] Optional: show whether this gene is under positive or negative evolutionary selection.

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

Ans: From my in-depth bioinformatics analysis, I could confidently say that my discovered gene, C24416 from *A. japonicum,* is a novel gene with a unique sequence, which was not discovered before. However, extreme conservation of its CD225 domain with the characterized IFITM3 proteins suggest that C22416 might play IFITM3-like anti-viral and other cellular roles. In addition, the phylogenetic position of CD24416 with other fishes and multiple sequence alignment suggest that IFITM3 proteins might evolved based on the needs of each species but kept the most important domain (CD225 domain) highly intact for functional conservation. However, further studies are needed to confirm if C24416 protein have similar role as IFITM3 and also it would be interesting to know if the gene could be induced by interferons in fish.