BIMM 143
Find-A-Gene Project

[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.

I am interested in studying the retinol binding protein 4 (RBP4), specifically isoform b, which is a transporter protein that is involved in binding and transporting vitamin A (i.e. retinol) from the liver to other tissues. This protein is found in humans (Homo sapiens) and its accession number from RefSeq is NP_001310447.1.

Name: RBP4
Accession: NP_001310447.1
Species: Homo sapiens

Link: https://www.ncbi.nlm.nih.gov/protein/NP_001310447.1

retinol-binding protein 4 isoform b [Homo sapiens]

NCBI Reference Sequence: NP_001310447.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       NP_001310447             199 aa            linear   PRI 19-FEB-2023
DEFINITION  retinol-binding protein 4 isoform b [Homo sapiens].
ACCESSION   NP_001310447
VERSION     NP_001310447.1
DBSOURCE    REFSEQ: accession NM_001323518.2
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 199)
  AUTHORS   Hu R, Yang X, He X and Song G.
  TITLE     The relationship between NAFLD and retinol-binding protein 4 - an
            updated systematic review and meta-analysis
  JOURNAL   Lipids Health Dis 22 (1), 8 (2023)
   PUBMED   36670387
  REMARK    GeneRIF: The relationship between NAFLD and retinol-binding protein
            4 - an updated systematic review and meta-analysis.
            Review article
            Publication Status: Online-Only
REFERENCE   2  (residues 1 to 199)
  AUTHORS   Deng X, Ren J, Bi Z and Fu Z.
  TITLE     Positive Expression of Retinol-Binding Protein 4 Is Related to the
            Malignant Clinical Features Leading to Poor Prognosis of
            Glioblastoma
  JOURNAL   Genet Res (Camb) 2022, 5435523 (2022)
   PUBMED   36632438
  REMARK    GeneRIF: Positive Expression of Retinol-Binding Protein 4 Is
            Related to the Malignant Clinical Features Leading to Poor
            Prognosis of Glioblastoma.
```

[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). 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. 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. 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.

I used NCBI BLAST, specifically tblastn, to find genes relating to the RBP 4 [Homo sapiens] (NP_001310447.1) with the organism set to abalones (taxid:6451). The database I searched was the expressed sequence tags (est) database. The top hit was GOF163 Haliotis discus cDNA library (GOF) Haliotis discus cDNA, mRNA sequence, which corresponds to Haliotis discus (disk abalone), has a query cover of 83%, an E-value of 5e-08, a percent identity of 26.74%, and the accession number is CX726648.1.

Method: TBLASTN search
Database: Expressed Sequence Tags (est)
Organism: abalones (taxid:6451)

Screenshots are attached with results from the BLAST search:

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

Your search is limited to records that include: abalones (taxid:6451)

Job Title	NP_001310447:retinol-binding protein 4 isoform...
RID	7UUP5HWG013 *Search expires on 06-06 12:55 pm* Download All
Program	TBLASTN Citation
Database	est See details
Query ID	NP_001310447.1
Description	retinol-binding protein 4 isoform b [Homo sapiens]
Molecule type	amino acid
Query Length	199
Other reports	

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity [ ] to [ ] E value [ ] to [ ] Query Coverage [ ] to [ ]

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *5 sequences selected* GenBank Graphics

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
GOF163 Haliotis discus cDNA library (GOF) Haliotis discus cDNA, mRNA sequence	Haliotis discus	50.1	50.1	83%	5e-08	26.74%	825	CX726648.1
zsbca0_003935.z1.scf Normalized Bacteria Challenged Haliotis diversicolor Reeve whole body library Haliotis dive...	Haliotis diversicolor	46.6	46.6	72%	5e-07	27.45%	555	GT869549.1
zsbca0_008908.z1.scf Normalized Bacteria Challenged Haliotis diversicolor Reeve whole body library Haliotis dive...	Haliotis diversicolor	43.5	43.5	72%	6e-06	25.68%	598	GT867546.1
zsbca0_009685.z1.scf Normalized Bacteria Challenged Haliotis diversicolor Reeve whole body library Haliotis dive...	Haliotis diversicolor	42.4	42.4	56%	1e-05	28.32%	483	GT871442.1
TH00191 Haliotis discus hannai RwRh cDNA library Haliotis discus hannai cDNA clone TH004G09 5', mRNA sequ...	Haliotis discus ha...	38.9	38.9	51%	3e-04	27.78%	679	EE675606.1

Descriptions | Graphic Summary | Alignments | Taxonomy

Alignment view Pairwise Restore defaults Download

5 sequences selected

Download GenBank Graphics ▼Next ▲Previous ◀◀Descriptions

GOF163 Haliotis discus cDNA library (GOF) Haliotis discus cDNA, mRNA sequence

Sequence ID: CX726648.1 Length: 825 Number of Matches: 1

Range 1: 62 to 514 GenBank Graphics ▼Next Match ▲Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
50.1 bits(118)	5e-08	Compositional matrix adjust.	46/172(27%)	77/172(44%)	27/172(15%)	+2

```
Query  6    IPAQVDLRRQTERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDE  65
            + A V ++ Q     R     V+ NF+   + GTWY     + E  F    I AE +
Sbjct  62   LAALVFVQGQVPALWRCPDVPVQRNFNLTSYLGTWY-----EDERFFAVYEIGAECA---  217

Query  66   TGQMSATAKGRVRLLNNW---DVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWI  122
               S  A G VR+LN+     +  D       D  +PAK  +K+      F  +   ++W+
Sbjct  218  QANYSMKANGHVRVLNSGTGRSIEGDAYAP--DPSEPAKLLVKF-----FWWQRPGNYWV  376

Query  123  VDTDYDTYAVQYSC---RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQR  171
            V+TDY  Y+V YSC     ++LD       +++ +R  +G+P + +  +  R
Sbjct  377  VETDYSNYSVVYSCIDLYFMHLDT------AWLLTRTRDGVPDDMKNDIYGR  514
```

[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. 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.

>CX726648.1 Haliotis discus protein

LAALVFVQGQVPALWRCPDVPVQRNFNLTSYLGTWY-----EDERFFAVYEIGAECA---QANYSMKA NGHVRVLNSGTGRSIEGDAYAP--DPSEPAKLLVKF-----FWWQRPGNYWVVETDYSNYSVVYSCID LYFMHLDT------AWLLTRTRDGVPDDMKNDIYGR

The protein I found is from disk abalone (Haliotis discus) and is called GOF163 Haliotis discus cDNA library (GOF) Haliotis discus cDNA, mRNA sequence. The accession ID is CX726648.1.

GOF163 Haliotis discus cDNA library (GOF) Haliotis discus cDNA, mRNA sequence

GenBank: CX726648.1

FASTA Graphics

Go to:

```
LOCUS       CX726648                825 bp    mRNA    linear   EST 01-DEC-2005
DEFINITION  GOF163 Haliotis discus cDNA library (GOF) Haliotis discus cDNA,
            mRNA sequence.
ACCESSION   CX726648
VERSION     CX726648.1
DBLINK      BioSample: SAMN00154170
KEYWORDS    EST.
SOURCE      Haliotis discus
  ORGANISM  Haliotis discus
            Eukaryota; Metazoa; Spiralia; Lophotrochozoa; Mollusca; Gastropoda;
            Vetigastropoda; Lepetellida; Haliotoidea; Haliotidae; Haliotis.
REFERENCE   1  (bases 1 to 825)
  AUTHORS   Kim,Y.O., Nam,B.H., Lee,J.H., Kim,K.K., Moon,E.J., Jee,Y.J.,
            Lee,S.J., Nam,Y.K. and Kim,D.S.
  TITLE     Haliotis discus cDNA
  JOURNAL   Unpublished
COMMENT     Contact: Young-Ok Kim
            Biotechnology Research Center
            National Fisheries Research & Development Institute
            #408-1, Sirang-ri, Gijang-eup, Gijang-gun, Busan, Republic of
            Korea, 619-902
            Tel: 051 720 2452
            Fax: 051 720 2456
            Email: yokim@momaf.go.kr.
FEATURES             Location/Qualifiers
     source          1..825
                     /organism="Haliotis discus"
                     /mol_type="mRNA"
                     /db_xref="taxon:36094"
                     /clone_lib="SAMN00154170 Haliotis discus cDNA library
                     (GOF)"
                     /dev_stage="Adult"
                     /lab_host="E.coli SOLR"
                     /note="Organ: Gonad of female; Vector: pBluescript SK(-);
                     Site_1: EcoR I; Site_2: Xho I"
ORIGIN
        1 ggctgcagga attcggcacg agggtcctgt gaagatggaa gctgtgttct gtgtggttct
```

Customize view

Analyze this sequence
Run BLAST
Pick Primers
Find in this Sequence

Related information
BioSample
Taxonomy

Recent activity
Turn Off Clear
GOF163 Haliotis discus cDNA library (GOF) Haliotis discus cDNA, mRNA sequer Nucleotide
PREDICTED: Spea bombifrons retinol binding protein 4 (RBP4), mRNA Nucleotide
Huntingtin [Camelus dromedarius] Protein
huntingtin [Callithrix jacchus] Protein
(huntingtin) AND "Aspergillus fumigatus" [porgn] (264) Protein
See more...

Haliotis discus

Taxonomy ID: 36094 (for references in articles please use NCBI:txid36094)

current name

Haliotis discus Reeve, 1846

homotypic synonym: ***Nordotis discus***

NCBI BLAST name: **gastropods**
Rank: **species**
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 5 (Invertebrate Mitochondrial)

Lineage*(full)*

cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Spiralia; Lophotrochozoa; Mollusca; Gastropoda; Vetigastropoda; Lepetellida; Haliotoidea; Haliotidae; Haliotis

[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.

Since my top hit, apolipoprotein D-like [Haliotis rufescens], which comes from California red abalone, has a percent identity of 98.01%, it is novel. Therefore, we have succeeded at identifying a novel protein.

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

Job Title	H. discus protein
RID	7UUUJ6YH016 *Search expires on 06-06 12:57 pm* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_145643
Description	H. discus protein
Molecule type	amino acid
Query Length	151
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity [ ] to [ ] E value [ ] to [ ] Query Coverage [ ] to [ ]

Filter Reset

Compare these results against the new Clustered nr database BLAST

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Select columns — Show 100

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

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
apolipoprotein D-like [Haliotis rufescens]	Haliotis rufescens	313	313	100%	5e-107	98.01%	179	XP_046353923.2
apolipoprotein D-like [Haliotis rubra]	Haliotis rubra	278	278	100%	4e-93	85.43%	179	XP_046571640.1
hypothetical protein KUTeg_007313 [Tegillarca granosa]	Tegillarca granosa	159	159	96%	7e-46	50.99%	198	KAJ8315163.1
apolipoprotein D [Lingula anatina]	Lingula anatina	154	154	100%	3e-44	46.25%	192	XP_013385924.1
APOD [Mytilus coruscus]	Mytilus coruscus	144	144	92%	4e-40	49.31%	192	CAC5384473.1
apolipoprotein D-like isoform X1 [Mytilus californianus]	Mytilus californianus	144	144	92%	5e-40	49.31%	192	XP_052084767.1

Download GenPept Graphics ▼Next ▲Previous ◀Descriptions

apolipoprotein D-like [Haliotis rufescens]

Sequence ID: XP_046353923.2 **Length:** 179 **Number of Matches:** 1

Range 1: 10 to 160 GenPept Graphics ▼Next Match ▲Previous Match

Score	Expect	Method	Identities	Positives	Gaps
313 bits(801)	5e-107	Compositional matrix adjust.	148/151(98%)	149/151(98%)	0/151(0%)

```
Query  1    LAALVFVQGQVPALWRCPDVPVQRNFNLTSYLGTWYEDERFFAVYEIGAECAQANYSMKA  60
            LAALV VQGQVPALWRCPDVPVQ NF+LTSYLGTWYEDERFFAVYEIGAECAQANYSMKA
Sbjct  10   LAALVIVQGQVPALWRCPDVPVQGNFSLTSYLGTWYEDERFFAVYEIGAECAQANYSMKA  69

Query  61   NGHVRVLNSGTGRSIEGDAYAPDPSEPAKLLVKFFWWQRPGNYWVVETDYSNYSVVYSCI  120
            NGHVRVLNSGTGRSIEGDAYAPDPSEPAKLLVKFFWWQRPGNYWVVETDYSNYSVVYSCI
Sbjct  70   NGHVRVLNSGTGRSIEGDAYAPDPSEPAKLLVKFFWWQRPGNYWVVETDYSNYSVVYSCI  129

Query  121  DLYFMHLDTAWLLTRTRDGVPDDMKNDIYGR  151
            DLYFMHLDTAWLLTRTRDGVPDDMKNDIYGR
Sbjct  130  DLYFMHLDTAWLLTRTRDGVPDDMKNDIYGR  160
```

Related Information

Genome Data Viewer - aligned genomic context

[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:

>Disk_abalone | Haliotis discus protein | CX726648.1
LAALVFVQGQVPALWRCPDVPVQRNFNLTSYLGTWY-----EDERFFAVYEIGAECA---QANYSMKA
NGHVRVLNSGTGRSIEGDAYAP--DPSEPAKLLVKF-----FWWQRPGNYWVVETDYSNYSVVYSCI
DLYFMHLDT------AWLLTRTRDGVPDDMKNDIYGR

>Red_abalone | XP_046353923.2:10-160 apolipoprotein D-like [Haliotis rufescens]
LAALVIVQGQVPALWRCPDVPVQGNFSLTSYLGTWYEDERFFAVYEIGAECAQANYSMKANGHV
RVLNSGTGRSIEGDAY
APDPSEPAKLLVKFFWWQRPGNYWVVETDYSNYSVVYSCIDLYFMHLDTAWLLTRTRDGVPDD
MKNDIYGR

>Blacklip_abalone | XP_046571640.1:10-160 apolipoprotein D-like [Haliotis rubra]
LTALVTIQGQVPALWRCPGVPVQDNFTLSEYLGTWYEYERFFTIYEIGAECAQANYSMKADGHV
KVLNSGRGRSIEGDAY
APDPSVPAKLLVKFFWWQRPGNYWVVETDYSNYSVVYSCTDLYFMHFDTAWLLTRTRDGIPDDL
KYPIYSR

>Korean_mussel | CAC5384473.1:21-164 APOD [Mytilus coruscus]
QVPRWGRCPDIPVKQNFDADRYLGKWYENEKFFFLAELGLKCISANYSLNNDGSIKVLNAGVNT
RTGRPSEAVGRATLGE
TEPAKLSVKFSRFQPAGAYWVLDTDYDTYSLVWSCSDFYFARTEIAWILTRERDGVDEATTNRL

>California_mussel | XP_052084767.1:21-164 apolipoprotein D-like isoform X1 [Mytilus californianus]
QVPRWGRCPDIPVKQNFDADRYLGKWYENEKFFFLAELGLKCISANYSLNNDGSIKVLNAGVNT
RTGRPSEAVGRATLGE
TEPAKLSVKFSRFQPAGAYWVLDTDYDTYSLVWSCSDFYFARTEIAWILTRDRDGVDEATTNRL

>Blue_mussel | CAG2223514.1:21-164 APOD [Mytilus edulis]

QVPRWGSCPDIPVKQNFDADRYLGKWYENENFFFIAEIGLKCISANYSLNSDGSIKVLNAGINART
GRPSESVGRATIGE
KEPAKLSVKFSRFQPAGAYWVLDTDYDTYSLVWSCSDFYFARAEIAWILTRERDGVDETTTNRL

>Blood_clam | KAJ8315163.1:26-176 hypothetical protein KUTeg_007313 [Tegillarca granosa]
VAGQVLKWGRCPNIPVQPNFDVNSYMGEWYEYQRFFAIFELGVKCAKADYSIIDESTVRVVNTG
FNKWTGKTSIANGTAT
ISDPAVPAKLKVKFSRFQPAGNYWVLDTDYNTYSLVYSCQDLYIAHSEITWILTRDRAGISDDTTN
RLYSR

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

```
Blacklip_abalone     LTALVTIQGQVPALWRCPGVPVQDNFTLSEYLGTWYEYERFFTIYEIGAECAQANYSMKA
Disk_abalone         LAALVFVQGQVPALWRCPDVPVQRNFNLTSYLGTWYEDERFFAVYEIGAECAQANYSMKA
Red_abalone          LAALVIVQGQVPALWRCPDVPVQGNFSLTSYLGTWYEDERFFAVYEIGAECAQANYSMKA
Blood_clam           ------VAGQVLKWGRCPNIPVQPNFDVNSYMGEWYEYQRFFAIFELGVKCAKADYSIID
Korean_mussel        ---------QVPRWGRCPDIPVKQNFDADRYLGKWYENEKFFFLAELGLKCISANYSLNN
California_mussel    ---------QVPRWGRCPDIPVKQNFDADRYLGKWYENEKFFFLAELGLKCISANYSLNN
Blue_mussel          ---------QVPRWGSCPDIPVKQNFDADRYLGKWYENENFFFIAEIGLKCISANYSLNS
                     **    **.:**: **   *:* *** :.** : *:* :* .*:**:

Blacklip_abalone     DGHVKVLNSG----RGRSIE--GDAYAPDPSVPAKLLVKFFWWQRPGNYWVVETDYSNYS
Disk_abalone         NGHVRVLNSG----TGRSIE--GDAYAPDPSEPAKLLVKFFWWQRPGNYWVVETDYSNYS
Red_abalone          NGHVRVLNSG----TGRSIE--GDAYAPDPSEPAKLLVKFFWWQRPGNYWVVETDYSNYS
Blood_clam           ESTVRVVNTGFNKWTGKTSIANGTATISDPAVPAKLKVKFSRFQPAGNYWVLDTDYNTYS
Korean_mussel        DGSIKVLNAGVNTRTGRPSEAVGRATLGE-TEPAKLSVKFSRFQPAGAYWVLDTDYDTYS
California_mussel    DGSIKVLNAGVNTRTGRPSEAVGRATLGE-TEPAKLSVKFSRFQPAGAYWVLDTDYDTYS
Blue_mussel          DGSIKVLNAGINARTGRPSESVGRATIGE-KEPAKLSVKFSRFQPAGAYWVLDTDYDTYS
                     :. :.*:*:*    *..   * *  :  **** *** .:* .* ***:.***..**

Blacklip_abalone     VVYSCTDLYFMHFDTAWLLTRTRDGIPDDLKYPIYSR
Disk_abalone         VVYSCIDLYFMHLDTAWLLTRTRDGVPDDMKNDIYGR
Red_abalone          VVYSCIDLYFMHLDTAWLLTRTRDGVPDDMKNDIYGR
Blood_clam           LVYSCQDLYIAHSEITWILTRDRAGISDDTTNRLYSR
Korean_mussel        LVWSCSDFYFARTEIAWILTRERDGVDEATTNRL---
California_mussel    LVWSCSDFYFARTEIAWILTRDRDGVDEATTNRL---
Blue_mussel          LVWSCSDFYFARAEIAWILTRERDGVDETTTNRL---
                     :*:** *:*: . : :*:*** * *: :  .  :
```

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Blacklip_abalone       LTALVTIQGQVPALWRCPGVPVQDNFTLSEYLGTWYEYERFFTIYEIGAECAQANYSMKA
Disk_abalone           LAALVFVQGQVPALWRCPDVPVQRNFNLTSYLGTWYEDERFFAVYEIGAECAQANYSMKA
Red_abalone            LAALVIVQGQVPALWRCPDVPVQGNFSLTSYLGTWYEDERFFAVYEIGAECAQANYSMKA
Blood_clam             ------VAGQVLKWGRCPNIPVQPNFDVNSYMGEWYEYQRFFAIFELGVKCAKADYSIID
Korean_muscle          ---------QVPRWGRCPDIPVKQNFDADRYLGKWYENEKFFFLAELGLKCISANYSLNN
California_mussel      ---------QVPRWGRCPDIPVKQNFDADRYLGKWYENEKFFFLAELGLKCISANYSLNN
Blue_mussel            ---------QVPRWGSCPDIPVKQNFDADRYLGKWYENENFFFIAEIGLKCISANYSLNS
                                 **     **.:**: **    *:* *** :.** : *:* :* .*:**:

Blacklip_abalone       DGHVKVLNSG----RGRSIE--GDAYAPDPSVPAKLLVKFFWWQRPGNYWVVETDYSNYS
Disk_abalone           NGHVRVLNSG----TGRSIE--GDAYAPDPSEPAKLLVKFFWWQRPGNYWVVETDYSNYS
Red_abalone            NGHVRVLNSG----TGRSIE--GDAYAPDPSEPAKLLVKFFWWQRPGNYWVVETDYSNYS
Blood_clam             ESTVRVVNTGFNKWTGKTSIANGTATISDPAVPAKLKVKFSRFQPAGNYWVLDTDYNTYS
Korean_muscle          DGSIKVLNAGVNTRTGRPSEAVGRATLGE-TEPAKLSVKFSRFQPAGAYWVLDTDYDTYS
California_mussel      DGSIKVLNAGVNTRTGRPSEAVGRATLGE-TEPAKLSVKFSRFQPAGAYWVLDTDYDTYS
Blue_mussel            DGSIKVLNAGINARTGRPSESVGRATIGE-KEPAKLSVKFSRFQPAGAYWVLDTDYDTYS
                       :. :.*:*:*     *..    * *   :   **** *** .:* .* ***::***..**

Blacklip_abalone       VVYSCTDLYFMHFDTAWLLTRTRDGIPDDLKYPIYSR
Disk_abalone           VVYSCIDLYFMHLDTAWLLTRTRDGVPDDMKNDIYGR
Red_abalone            VVYSCIDLYFMHLDTAWLLTRTRDGVPDDMKNDIYGR
Blood_clam             LVYSCQDLYIAHSEITWILTRDRAGISDDTTNRLYSR
Korean_muscle          LVWSCSDFYFARTEIAWILTRERDGVDEATTNRL---
California_mussel      LVWSCSDFYFARTEIAWILTRDRDGVDEATTNRL---
Blue_mussel            LVWSCSDFYFARAEIAWILTRERDGVDETTTNRL---
                       :*:** *:*: . : :*:*** * *: :  .  :
```

[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.

Creating a phylogenetic tree using EBI simple phylogeny set to neighbor-joining:

[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.

```{r}
library("bio3d")
library(pheatmap)
alignment <- read.fasta("q5_align.clw")
identity_matrix <- seqidentity(alignment)
p <- pheatmap(identity_matrix)
```

[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 choose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

```r
library(bio3d)

# Define PDB IDs
pdb_ids <- c("2HZQ", "7YX1", "7O2Y")

# Initialize list to store the information for each PDB ID
info_list <- list()

# For each PDB ID
for (i in 1:length(pdb_ids)) {
  # Get the annotation
  pdb_annotation <- pdb.annotate(pdb_ids[i])

  # Check the structure of the annotation
  print(str(pdb_annotation))

  # Save the information
  info_list[[i]] <- list(
    pdb_id = pdb_ids[i],
    experimentalTechnique = pdb_annotation$experimentalTechnique[1],
    resolution = as.numeric(pdb_annotation$resolution[1]),
    source = pdb_annotation$source[1]
  )
}

# Convert the list to a data frame
info_df <- do.call(rbind.data.frame, info_list)

# Add column names
names(info_df) <- c("PDB ID", "Experimental Technique", "Resolution", "Source Organism")

# Print the data frame
print(info_df)
```

R Console

data.frame
3 x 4

PDB ID <chr>	Experimental Technique <chr>	Resolution <dbl>	Source Organism <chr>
2HZQ	X-ray	1.80	Homo sapiens
7YX1	X-ray	2.65	Sander
7O2Y	X-ray	2.50	Sander vitreus

ID	Technique	Resolution	Source	E-value	% Identity
2HZQ	**X-ray diffraction**	**1.80**	**Homo sapiens**	**3e-31**	**44.96%**
7YX1	**X-ray diffraction**	**2.65**	**Sander**	**6e-22**	**40.48%**
7O2Y	**X-ray diffraction**	**2.50**	**Sander vitreus**	**2e-21**	**40.48%**

[Q9] Generate a molecular figure of one of your identified PDB structures using the NGL viewer online (or VMD/PyMol). 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 of 2HZQ:

Based on sequence similarity, the structure of my top hit (PDB ID: 2HZQ) is likely a good representation of the structure of my "novel" protein, since they share a 44.96% sequence identity. Though this may not be considered a high sequence identity, the structure of proteins is generally more conserved than that of protein sequences, so the structure is probably similar.

[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?

Link: https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL3305/

Associated Assays

ChEMBL / Explore Assays

Assays

34 items

Tools

RELATED ACTIVITIES | SHARE | CSV | TSV

Filters

Custom

Custom Filtering

Per Property

Type Label

HISTOGRAM | FIND_TERM

B - Binding	26
A - ADME	7
F - Functional	1

Items per page 20

Showing 1-20 out of 34 records

1 2

ChEMBL ID	Assay Type	Description	Organism	Compounds	Document ChEMBL ID	BAO Format	Source
CHEMBL801821	B	Relative binding affinity against sex steroid binding protein (SBP)	Homo sapiens	12	CHEMBL1126705	single protein format	Scientific Literature
CHEMBL1025921	B	Displacement of [3H]DHT from human serum SHBG	Homo sapiens	4	CHEMBL1159196	single protein format	Scientific Literature
CHEMBL1039997	A	Binding affinity to human sex hormone binding globulin at 10 to 20 uCi up to 6 hrs by size-exclusion HPLC analysis	Homo sapiens	2	CHEMBL1155242	single protein format	Scientific Literature
CHEMBL1025920	B	Binding affinity to SHBG at 4 degC	No Data	1	CHEMBL1159196	single protein format	Scientific Literature

Ligand Efficiencies

See all bioactivities for target CHEMBL3305 used in this visualisation

Total items: 100

Standard Value nM

(15)	(44)	(23)	(19)
<1	[1,100)	[100,1000)	>=1000

Browse Activities

Edit Querystring

Show Full Query

Table

101 Activities

0 Selected - Select All

Browse Compounds

CSV TSV

Filters

Standard Type

Kd 91

IC50 10

Target Type

SINGLE PROTEIN 101

Organism Taxonomy L1

Eukaryotes 101

Organism Taxonomy L2

Mammalia 101

Records per page: 20

Show/Hide Columns

Showing 1-20 out of 101 records

1 2 3 4 5 ...

Molecule ChEMBL ID	Compound Key	Standard Type	Standard Relation	Standard Value	Standard Units	pChEMBL Value	Comment	Assay ChEMBL ID	Assay Description
CHEMBL2107797	19	Kd	=	1.23	nM	8.91	No Data	CHEMBL922711	Displacement of [3H]5alpha dihydrotestoster from human sex hormone binding globulin
	77	Kd	=	1202.26	nM	5.92	No Data	CHEMBL922711	Displacement of [3H]5alpha dihydrotestoster from human sex hormone binding

CHEMBL801821 may be a useful target-associated assay for exploring the potential inhibition of my novel protein.