BIMM-143: INTRODUCTION TO BIOINFORMATICS

The find-a-gene project assignment
https://bioboot.github.io/bimm143_S20/
Dr. Barry Grant

Overview:

The find-a-gene project is a required assignment for BIMM-143. 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 class **Tuesday May 5th** (05/05/20) at 12pm San Diego time. 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 **Friday June 5th** (06/05/20) at 12pm San Diego time.

Submission instructions:

Your report formatted as a **PDF document** should be uploaded to ***GradeScope***. Please make sure to include your UCSD email and PID number on the first page.

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

Submit your preliminary report with answers to Q1-Q4 as soon as you can 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 the final document containing your results for all questions - **Please do not send 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: Beta globin

Accession: NP_000509

Species: Homo Sapiens

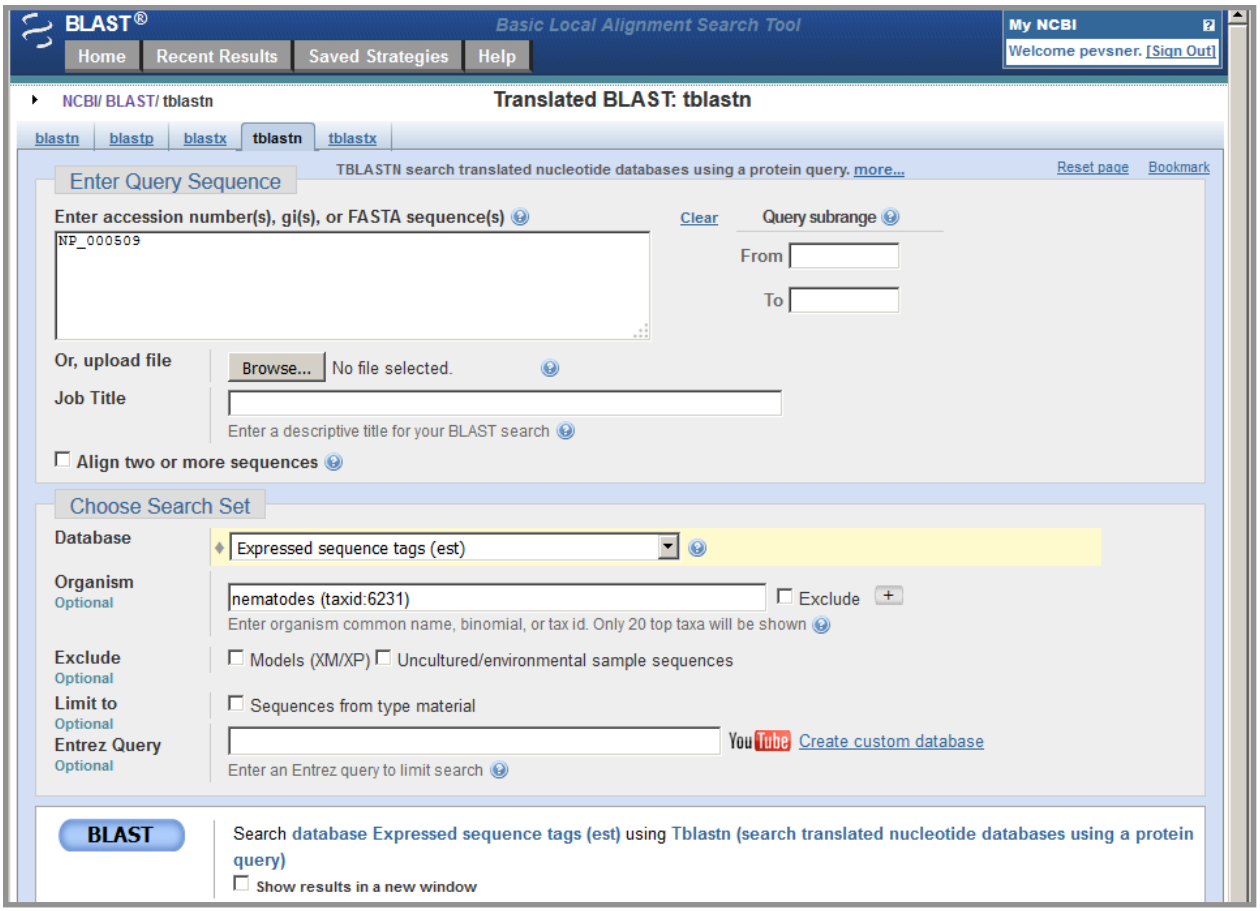
[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: TBLASTN (2.7.1) search against nematode ESTs

Database: Expressed Sequence Tags (est)

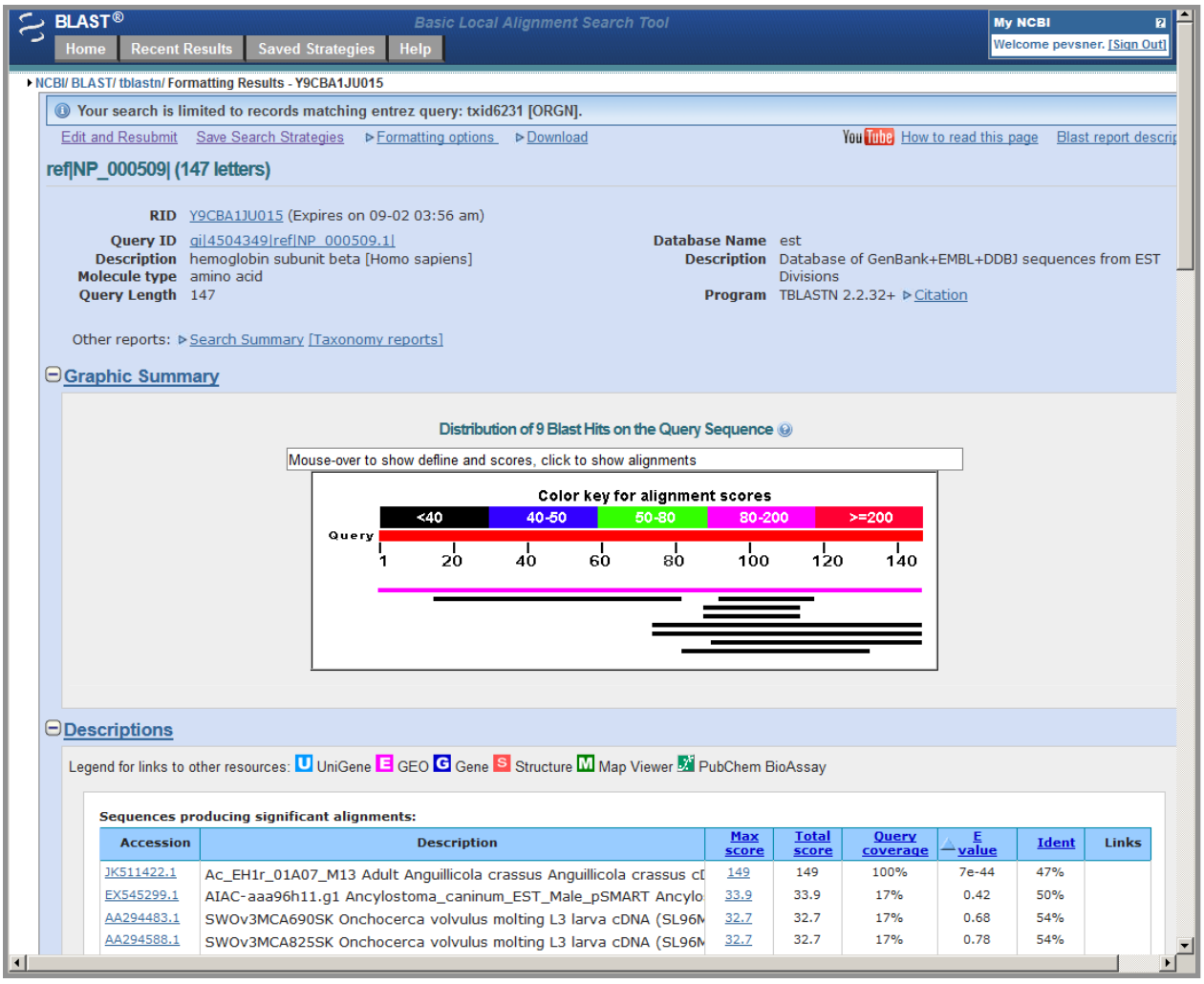
Organism: Nematodes (Taxid: 6231)

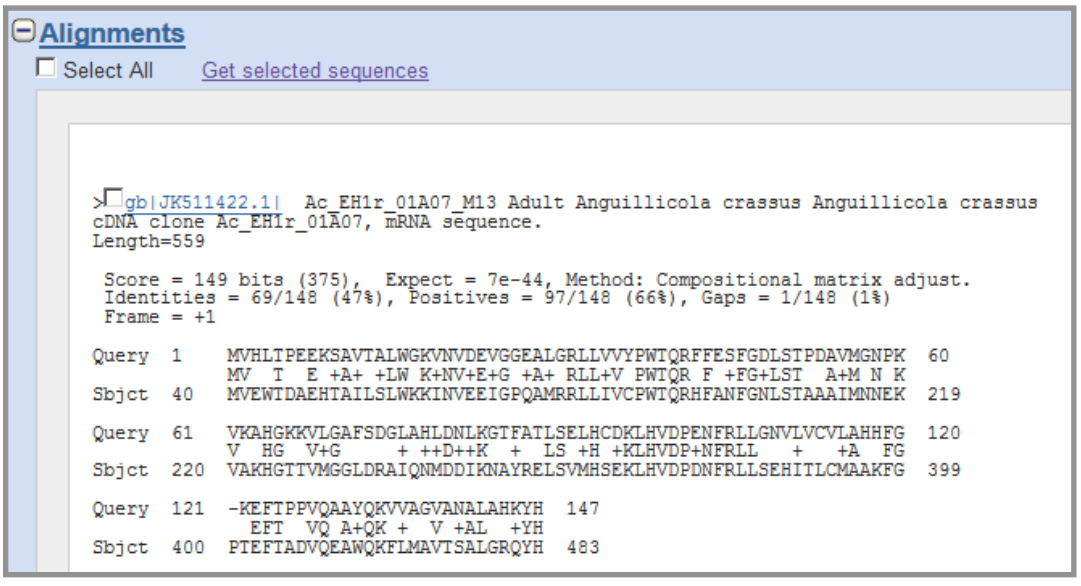
Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select 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.

Chosen match: Accession JK511422.1, a 559 base pair clone from *Anguillicola crassus*. See below for alignment details.


BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
Welcome pevsner. [Sign Out]
NCBI/ BLAST/ tblastn/ Formatting Results - Y9CBA1JU015
Your search is limited to records matching entrez query: txid6231 [ORGN].
Edit and Resubmit
Save Search Strategies
Formatting options
Download
How to read this page
Blast report descrip
ref|NP_000509| (147 letters)
RID Y9CBA1JU015 (Expires on 09-02 03:56 am)
Query ID gi|4504349|ref|NP_000509.1|
Description hemoglobin subunit beta [Homo sapiens]
Molecule type amino acid
Query Length 147
Database Name est
Description Database of GenBank+EMBL+DDBJ sequences from EST Divisions
Program TBLASTN 2.2.32+ Citation
Other reports: Search Summary [Taxonomy reports]
Graphic Summary
Distribution of 9 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
20
40
60
80
100
120
140
Descriptions
Legend for links to other resources: UniGene GEO Gene Structure Map Viewer PubChem BioAssay
Sequences producing significant alignments:
Accession
Description
Max score
Total score
Query coverage
E value
Ident
Links
JK511422.1 Ac_EH1r_01A07_M13 Adult Anguillicola crassus Anguillicola crassus cD 149 149 100% 7e-44 47%
EX545299.1 AIAC-aaa96h11.g1 Ancylostoma_caninum_EST_Male_pSMART Ancylo 33.9 33.9 17% 0.42 50%
AA294483.1 SWOv3MCA690SK Onchocerca volvulus molting L3 larva cDNA (SL96M 32.7 32.7 17% 0.68 54%
AA294588.1 SWOv3MCA825SK Onchocerca volvulus molting L3 larva cDNA (SL96M 32.7 32.7 17% 0.78 54%


Alignments
Select All
Get selected sequences
>gb|JK511422.1| Ac_EH1r_01A07_M13 Adult Anguillicola crassus Anguillicola crassus cDNA clone Ac_EH1r_01A07, mRNA sequence.
Length=559
Score = 149 bits (375), Expect = 7e-44, Method: Compositional matrix adjust.
Identities = 69/148 (47%), Positives = 97/148 (66%), Gaps = 1/148 (1%)
Frame = +1
Query 1 MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60
MV T E +A+ +LW K+NV+E+G +A+ RLL+V PWTQR F +FG+LST A+M N K
Sbjct 40 MVEWTDAEHTAILSLWKKINVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNEK 219
Query 61 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG 120
V HG V+G + ++D++K + LS +H +KLHVDP+NFRLL + +A FG
Sbjct 220 VAKHGTTVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFG 399
Query 121 -KEFTPPVQAAYQKVVAGVANALAHKYH 147
EFT VQ A+QK + V +AL +YH
Sbjct 400 PTEFTADVQEAWQKFLMAVTSALGRQYH 483

Alignment details:

```
>gb|JK511422.1|  Ac_EH1r_01A07_M13 Adult Anguillicola crassus Anguillicola
crassus
cDNA clone Ac_EH1r_01A07, mRNA sequence.
Length=559

 Score = 149 bits (375),  Expect = 7e-44, Method: Compositional matrix
adjust.
 Identities = 69/148 (47%), Positives = 97/148 (66%), Gaps = 1/148 (1%)
 Frame = +1

Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MV  T  E +A+ +LW K+NV+E+G +A+ RLL+V PWTQR F +FG+LST  A+M N K
Sbjct  40   MVEWTDAEHTAILSLWKKINVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNEK  219

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
            V  HG  V+G     + ++D++K  +  LS +H +KLHVDP+NFRLL   +   +A  FG
Sbjct  220  VAKHGTTVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFG  399

Query  121  -KEFTPPVQAAYQKVVAGVANALAHKYH  147
              EFT  VQ A+QK +  V +AL  +YH
Sbjct  400  PTEFTADVQEAWQKFLMAVTSALGRQYH  483
```

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. crassus protein (sequence taken from BLAST result)
MVEWTDAEHTAILSLWKKINVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNEKVAKH
GTTVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFGPTEFTADV
QEAWQKFLMAVTSALGRQYH
```

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: *Anguillicola* globin
Species: Anguillicola crassus

Eukaryota; Metazoa; Ecdysozoa; Nematoda; Chromadorea; Spirurida; Dracunculoidea; Anguillicolidae; Anguillicola.

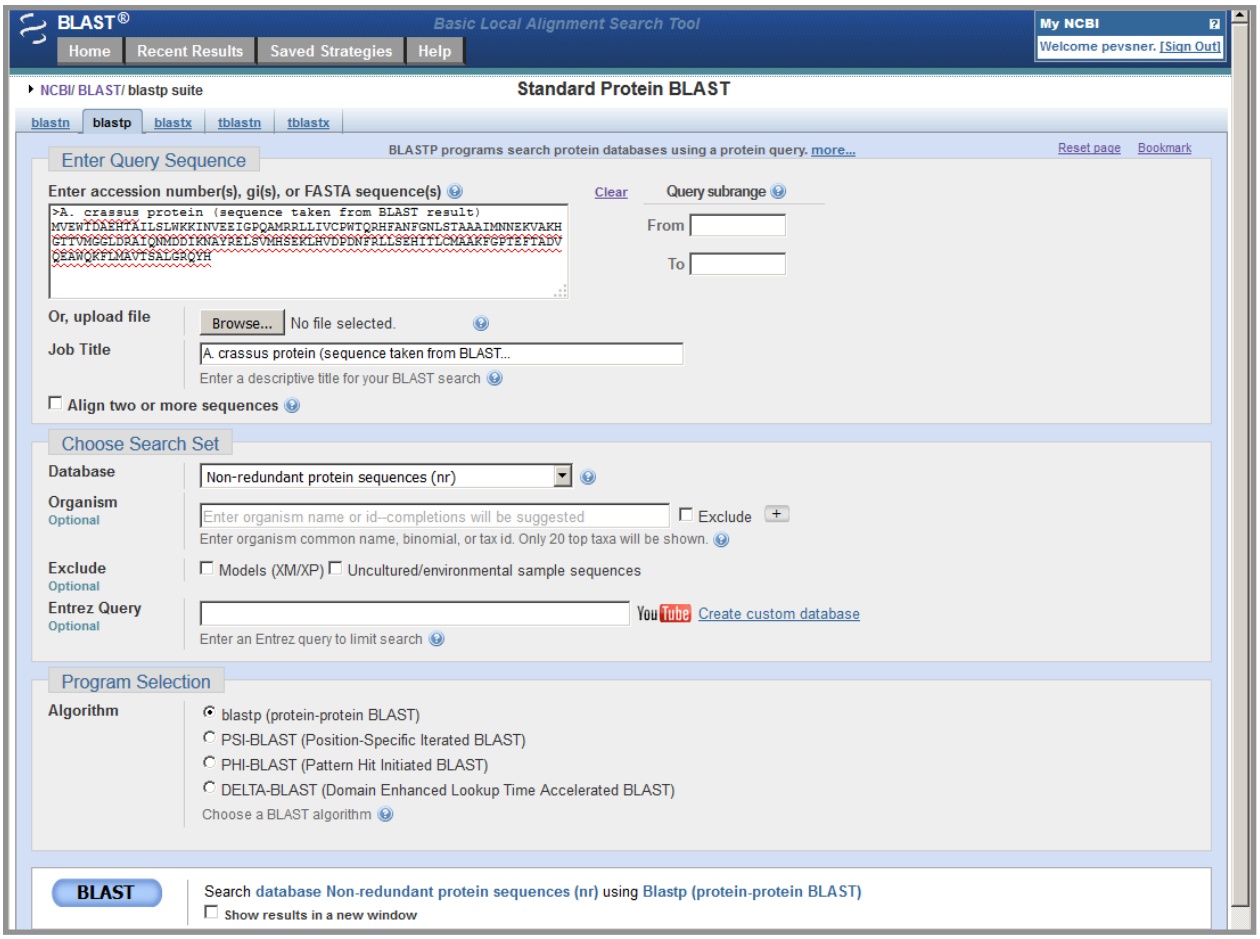
[**Q4**] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, "novel" is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

- 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.

Details:

A BLASTP search against NR database (see setup in first screen-shot below) yielded a top hit result is to a protein from *Anguilla anguilla* (European eel).

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



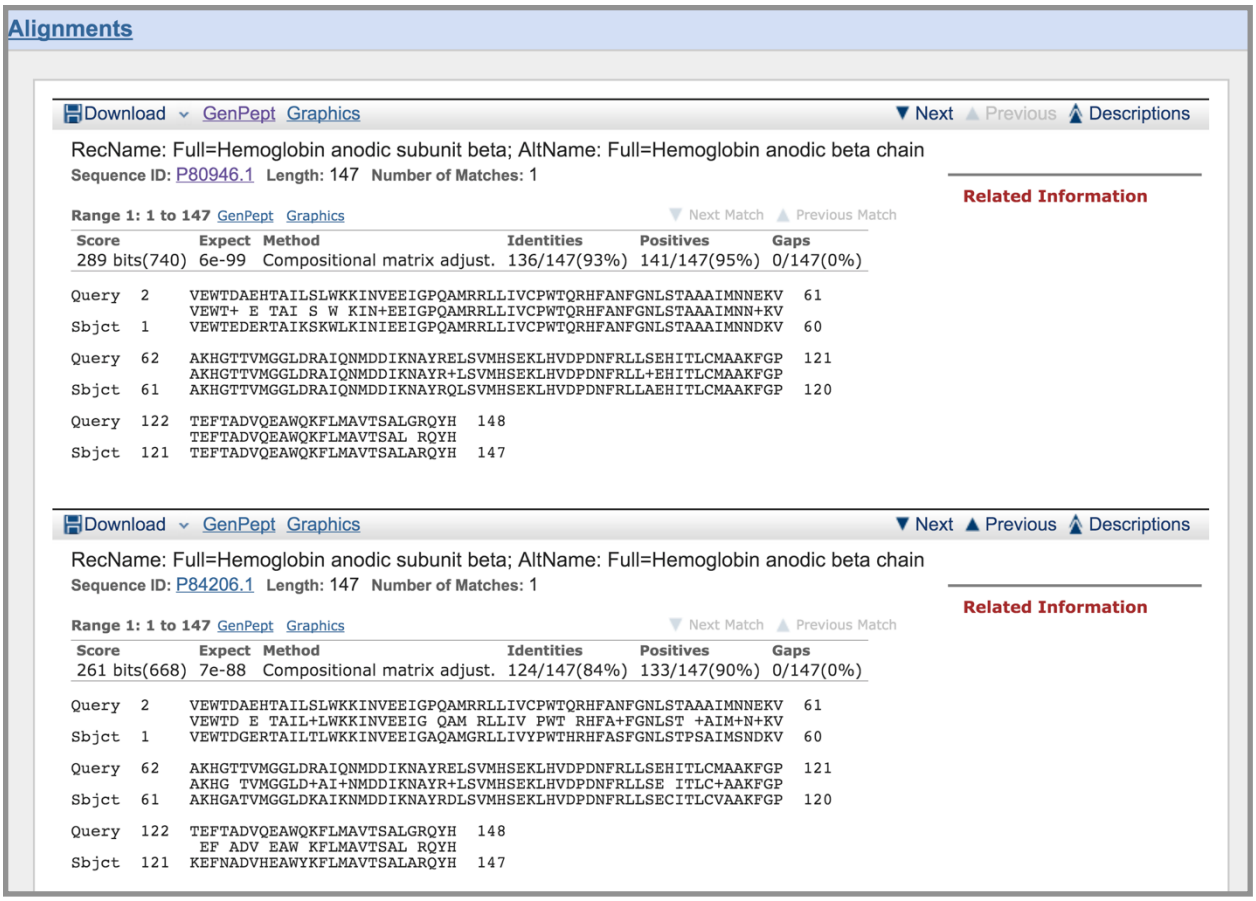
The top result is to a protein from *Anguilla anguilla* (European eel), see second screen shot below for alignment details:

Sequences producing significant alignments:

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
RecName: Full=Hemoglobin anodic subunit beta; AltName: Full=Hemoglobin anodic beta chai	289	289	99%	6e-99	93%	P80946.1
RecName: Full=Hemoglobin anodic subunit beta; AltName: Full=Hemoglobin anodic beta chai	261	261	99%	7e-88	84%	P84206.1
RecName: Full=Hemoglobin subunit beta-3; AltName: Full=Beta-3-globin; AltName: Full=Hem	253	253	99%	2e-84	80%	Q7LZC1.1
PREDICTED: hemoglobin subunit beta-A [Lates calcarifer]	245	245	100%	3e-81	76%	XP_018560551.1
PREDICTED: hemoglobin subunit beta-A-like [Paralichthys olivaceus]	244	244	100%	5e-81	75%	XP_019950077.1
PREDICTED: hemoglobin subunit beta-A-like [Lates calcarifer]	243	243	100%	1e-80	76%	XP_018560550.1



Alignments

Download GenPept Graphics ▼ Next ▲ Previous Descriptions

RecName: Full=Hemoglobin anodic subunit beta; AltName: Full=Hemoglobin anodic beta chain

Sequence ID: P80946.1 Length: 147 Number of Matches: 1

Related Information

Range 1: 1 to 147 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
289 bits(740)	6e-99	Compositional matrix adjust.	136/147(93%)	141/147(95%)	0/147(0%)

```
Query  2    VEWTDAEHTAILSLWKKINVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNEKV  61
            VEWT+ E TAI S W KIN+EEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNN+KV
Sbjct  1    VEWTEDERTAIKSKWLKINIEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNDKV  60

Query  62   AKHGTTVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFGP  121
            AKHGTTVMGGLDRAIQNMDDIKNAYR+LSVMHSEKLHVDPDNFRLL+EHITLCMAAKFGP
Sbjct  61   AKHGTTVMGGLDRAIQNMDDIKNAYRQLSVMHSEKLHVDPDNFRLLAEHITLCMAAKFGP  120

Query  122  TEFTADVQEAWQKFLMAVTSALGRQYH  148
            TEFTADVQEAWQKFLMAVTSAL RQYH
Sbjct  121  TEFTADVQEAWQKFLMAVTSALARQYH  147
```

Download GenPept Graphics ▼ Next ▲ Previous Descriptions

RecName: Full=Hemoglobin anodic subunit beta; AltName: Full=Hemoglobin anodic beta chain

Sequence ID: P84206.1 Length: 147 Number of Matches: 1

Related Information

Range 1: 1 to 147 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
261 bits(668)	7e-88	Compositional matrix adjust.	124/147(84%)	133/147(90%)	0/147(0%)

```
Query  2    VEWTDAEHTAILSLWKKINVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNEKV  61
            VEWTD E TAIL+LWKKINVEEIG QAM RLLIV PWT RHFA+FGNLST +AIM+N+KV
Sbjct  1    VEWTDGERTAILTLWKKINVEEIGAQAMGRLLIVYPWTHRHFASFGNLSTPSAIMSNDKV  60

Query  62   AKHGTTVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFGP  121
            AKHG TVMGGLD+AI+NMDDIKNAYR+LSVMHSEKLHVDPDNFRLLSE ITLC+AAKFGP
Sbjct  61   AKHGATVMGGLDKAIKNMDDIKNAYRDLSVMHSEKLHVDPDNFRLLSECITLCVAAKFGP  120

Query  122  TEFTADVQEAWQKFLMAVTSALGRQYH  148
             EF ADV EAW KFLMAVTSAL RQYH
Sbjct  121  KEFNADVHEAWYKFLMAVTSALARQYH  147
```

[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:

```
>Human_HBB gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLA
HLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH

>Anguillicola_globin (sequence taken from BLAST result)
```

```
MVEWTDAEHTAILSLWKKINVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNEKVAKHGTTVMGGLDRAIQ
NMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFGPTEFTADVQEAWQKFLMAVTSALGRQYH

>Eel gi|2494788|sp|P80946.1|HBBA_ANGAN RecName: Full=Hemoglobin anodic
subunit beta; AltName: Full=Hemoglobin anodic beta chain
VEWTEDERTAIKSKWLKINIEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNDKVAKHGTTVMGGLDRAIQN
MDDIKNAYRQLSVMHSEKLHVDPDNFRLLAEHITLCMAAKFGPTEFTADVQEAWQKFLMAVTSALARQYH


>Atlantic_salmon gi|469832338|gb|AGH92521.1| hemoglobin subunit beta [Salmo
salar]
MVDWTDAERSAIVGLWGKISVDEIGPQALARLLIVSPWTQRHFSTFGNLSTPAAIMGNPAVAKHGKTVMHGLDRAVQ
NLDDIKNAYTALSVMHSEKLHVDPDNFRLLADCITVCVAAKLGPTVFSADIQEAFQKFLAVVVSALGRQYH

>Rainbow_trout gi|1183021|dbj|BAA11632.1| beta-globin [Oncorhynchus mykiss]
MVDWTDPERSAIVGLWGKISVDEIGPQALARLLIVSPWTQRHFSTFGNLSTPAAIMGNPAVAKHGKTVMHGLDRAVQ
NLDDIKNTYTALSVMHSEKLHVDPDNFRLLADCITVCVAAKLGPAVFSADTQEAFQKFLAVVVSALGRQYH

>Zebrafish gi|18858329|ref|NP_571095.1| hemoglobin subunit beta-1 [Danio
rerio]
MVEWTDAERTAILGLWGKLNIDEIGPQALSRCLIVYPWTQRYFATFGNLSSPAAIMGNPKVAAHGRTVMGGLERAIK
NMDNVKNTYAALSVMHSEKLHVDPDNFRLLADCITVCAAMKFGQAGFNADVQEAWQKFLAVVVSALCRQYH

>Channel_catfish gi|318171215|ref|NP_001187115.1| hemoglobin-beta [Ictalurus
punctatus] >gi|38606322|gb|AAR25199.1| hemoglobin-beta [Ictalurus punctatus]
MVHWTDAERHIIADLWGKINHDEIGGQALARLLIVYPWTQRYFSSFGNLSNAAAIIGNPKVAAHGKVVLGGLTKAVQ
NLDNIKGIYTQLSTLHSEKLHVDPSNFTLLGDTFTVTLAANFGPSVFTPEVHETWQKFLNVVVAALGKQYH
```

Alignment:

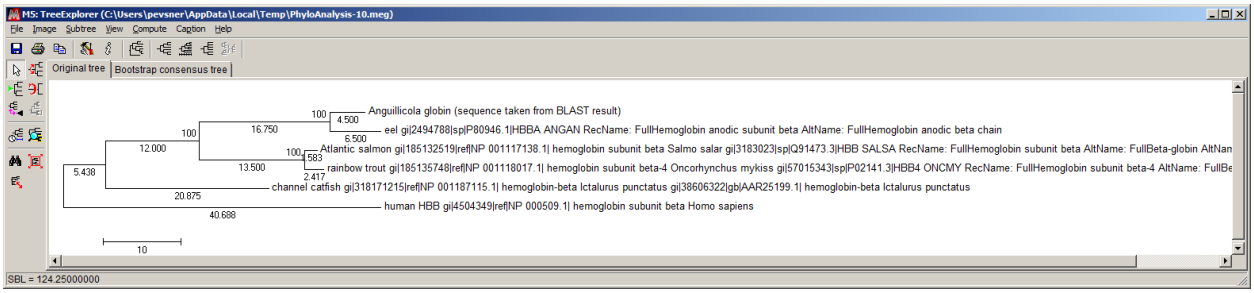
Obtained using MUSCLE (version 3.8) at EBI:

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


Human_HBB         MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Channel_catfish   MVHWTDAERHIIADLWGKINHDEIGGQALARLLIVYPWTQRYFSSFGNLSNAAAIIGNPK
Anguillicola      MVEWTDAEHTAILSLWKKINVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNEK
Eel               -VEWTEDERTAIKSKWLKINIEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNDK
Zebrafish         MVEWTDAERTAILGLWGKLNIDEIGPQALSRCLIVYPWTQRYFATFGNLSSPAAIMGNPK
Atlantic_salmon   MVDWTDAERSAIVGLWGKISVDEIGPQALARLLIVSPWTQRHFSTFGNLSTPAAIMGNPA
Rainbow_trout     MVDWTDPERSAIVGLWGKISVDEIGPQALARLLIVSPWTQRHFSTFGNLSTPAAIMGNPA
                   *  *  *.  :   * *:. :*:* :*: * *:* *****.* .**:**.. *::.*

Human_HBB         VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Channel_catfish   VAAHGKVVLGGLTKAVQNLDNIKGIYTQLSTLHSEKLHVDPSNFTLLGDTFTVTLAANFG
Anguillicola      VAKHGTTVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFG
Eel               VAKHGTTVMGGLDRAIQNMDDIKNAYRQLSVMHSEKLHVDPDNFRLLAEHITLCMAAKFG
Zebrafish         VAAHGRTVMGGLERAIKNMDNVKNTYAALSVMHSEKLHVDPDNFRLLADCITVCAAMKFG
Atlantic_salmon   VAKHGKTVMHGLDRAVQNLDDIKNAYTALSVMHSEKLHVDPDNFRLLADCITVCVAAKLG
Rainbow_trout     VAKHGKTVMHGLDRAVQNLDDIKNTYTALSVMHSEKLHVDPDNFRLLADCITVCVAAKLG
                  *  **  *: .:  .: ::*::*. :  ** :*.:******.** **.: :.   * ::*
```

```
Human_HBB          KE-FTPPVQAAYQKVVAGVANALAHKYH
Channel_catfish    PSVFTPEVHETWQKFLNVVVAALGKQYH
Anguillicola       PTEFTADVQEAWQKFLMAVTSALGRQYH
Eel                PTEFTADVQEAWQKFLMAVTSALARQYH
Zebrafish          QAGFNADVQEAWQKFLAVVVSALCRQYH
Atlantic_salmon    PTVFSADIQEAFQKFLAVVVSALGRQYH
Rainbow_trout      PAVFSADTQEAFQKFLAVVVSALGRQYH
                      *..  : ::**.:  *. ** .:**
```

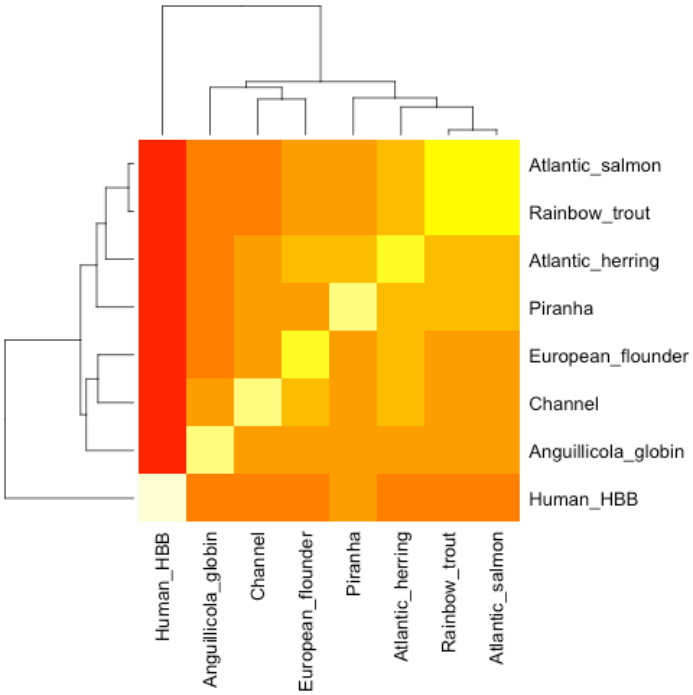
[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

Import the sequences into MEGA, align with MUSCLE, and create a neighbor-joining tree:



[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.

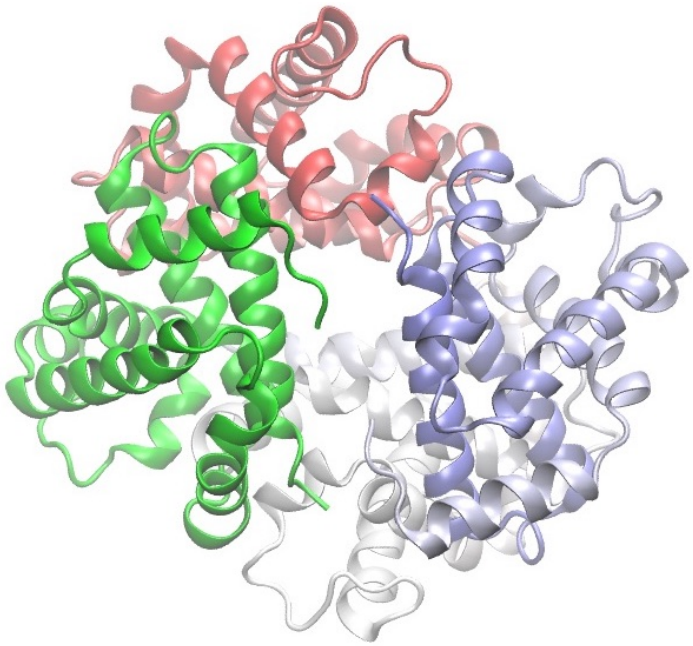
Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

ID	Technique	Resolution	Source	Evalue	Identity
3BOM	X-RAY DIFFRACTION	1.35	Oncorhynchus mykiss	6.59E-63	81.4
1SPG	X-RAY DIFFRACTION	1.95	Leiostomus xanthurus	3.16E-58	75.9
3BCQ	X-RAY DIFFRACTION	2.4	Brycon cephalus	5.11E-57	77.2

[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?

Very likely to be similar in structure to *Anguillicola* globin given the high sequence similarity (>80%). In the figure below the beta globin chain B is colored green and corresponds to the *Anguillicola* globin subject of this report.



[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 1 Binding Assay (CHEMBL695842) and 5 Functional Assays; No ligand efficiency data.

https://www.ebi.ac.uk/chembl/target/inspect/11544

Binding assay linked manuscript tested a set of 2-(N,N-dimethylamino)ethyl isothiocyanate compounds with results suggesting “substantial promise as lead compounds for development of therapeutic agents for sickle cell disease”.

Park S, Hayes BL, Marankan F, Mulhearn DC, Wanna L, Mesecar AD, Santarsiero BD, Johnson ME, Venton DL. Regioselective covalent modification of hemoglobin in search of antisickling agents. J Med Chem. 2003 Mar;46(6) 936-953. doi:10.1021/jm020361k. PMID: 12620071.

http://europepmc.org/abstract/MED/12620071

Scoring Rubric:

[45 total points available]

Q1 (4 points)	
Protein name	1
Species	1
Accession number	1
Function known	1
Q2 (6 points)	
Blast method	1
Database searched	1
Limits applied	1
Search output list (top hits)	1
Alignment of choice	1
Evalue and other alignment stats	1
Q3 (3 points)	
Protein sequence of choice matches Subject above	1
Name in header	1
Species	1
Q4 (3 point)	
Blastp output list with identities & Evalue	1
Top alignment shown with alignment statistics	1
Results indicates a “novel” gene found	1
Q5 (3 points)	
MSA labeled with useful names	1
MSA trimmed appropriately (i.e. no gap overhangs)	1
Pasted MSA fits report page width (i.e. font, format)	1
Q6 (1 point)	
Figure illustrates sequence clustering pattern	1

Q7 (10 points)

Heatmap figure included in report	5
Heatmap is legible (i.e. no labels obscured)	5

Q8 (10 points)

PDB identifiers from multiple species reported	5
Annotation of PDB source, resolution and technique	4
Annotation of Evalue and Sequence Identity	1

Q9 (4 points)

Structure figure provided	2
Uses white background for molecular figure	1
Figure of high resolution (i.e. not just snapshot)	1

Q10 (1 point)

Evidence of ChEMBEL searches	1