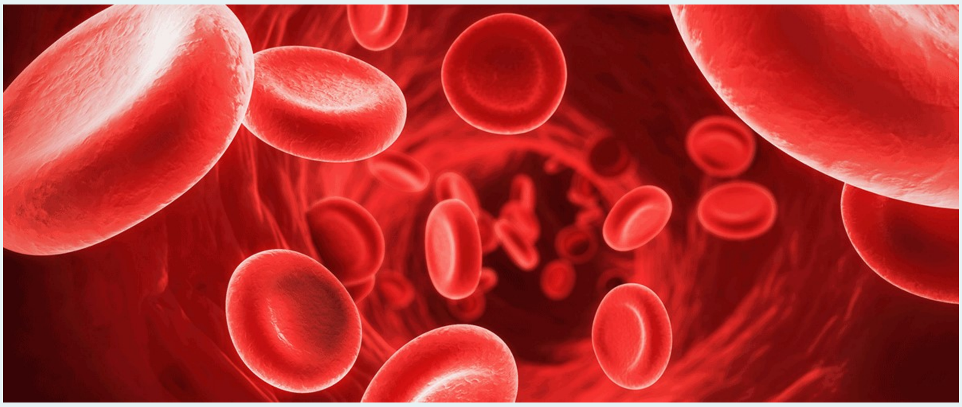
Started on	Thursday, 12 October 2023, 4:29 PM
State	Finished
Completed on	Monday, 16 October 2023, 10:47 PM
Time taken	4 days 6 hours
Marks	21.62/29.00
Grade	**7.46** out of 10.00 (**74.55**%)

Question **1**

Correct

Mark 1.00 out of 1.00



All globin proteins are closely related both in sequence and in three-dimensional structure. The primary function of the myo- and hemoglobins is to bind oxygen. They both consist of globin fold polypeptide chains with a bound heme group. The heme group is a so-called prosthetic group and is indispensable for the oxygen binding function of the protein. The heme consists of four pyrrole rings joined by methene bridges with an iron atom in the centre.

Let's explore protein sequence databases to learn much more regarding these protein families.

Question 1. The following search strategy gives no results at NCBI protein database. Why?

human[porgn] AND hemoglobin[protein name] AND refseq[filter]

Tips: Try other search options and tags to find out what is happening.

Select one:

○ a. Hemoglobin is produced by erythrocytes, and mammalian erythrocytes lack nuclei and organelles.

○ b. There are no records for human hemoglobins at the refseq.

○ c. Don't know/no answer (without penalty).

○ d. The NCBI only recognized the British word haemoglobin and not hemoglobin.

● e. The NCBI search engine requires complete names with the "[protein name]" tag and "hemoglobin" is an incomplete name for a human protein. ✓

Obviously there are some entries containing human hemoglobin sequences at refseq. The point is there are several hemoglobin variants in humans known as subunits. The right protein names are "hemoglobin subunit alpha", "hemoglobin subunit beta", etc. The search tag [protein name] at NCBI will check for the complete name of a protein without accepting incomplete names. In this case hemoglobin is an incomplete protein name in organisms with many variants. The use of the truncation wildcard "*" would help, but we must be careful as this strategy could include many false positive results.

The correct answer is: The NCBI search engine requires complete names with the "[protein name]" tag and "hemoglobin" is an incomplete name for a human protein.

Question **2**

Partially correct

Mark 2.00 out of 2.00

Question 2. Seek the highest possible number of protein sequences for human hemoglobin (including all subunits) at the NCBI refseq database. How many have you found? 10 ✓

Optionally write here your search strategy

"Homo Sapiens"[porgn] AND hemoglobin[title] AND refseq[filter] ✖ (no rate)

2.1 And for human myoglobin in the same database? 9 ✓

Tips: To do the search you should use the conclusions obtained from Question 1. You must assess whether the search tag [protein name] is convenient or not here. Remember we mention in class that There is another place where the protein name is written.

At protein database:

human[porgn] AND hemoglobin[title] AND refseq[filter]

human[porgn] AND myoglobin[title] AND refseq[filter]

If you use (human[porgn] AND hemoglobin*[title] AND refseq[filter]) the word hemoglobinization will interfere your search.

The use of hemoglobin*[protein name] would help but we must be careful as this strategy could include false positive results and miss true positives.

Question **3**

Incorrect

Mark -0.25 out of 1.00

Question 3. In Question 2 you have found two records for human hemoglobin subunit alpha with identical sequences. If refseq is considered a non-redundant database, why did it accept these identical human sequences in separate files?

Tips:

- RefSeq records contain a COMMENT section or field that includes, among other things, some explanations justifying the inclusion of sequences in the non-redundant database.
- Protein records at NCBI contain the DBSOURCE field to show the accession of the GenBank/Refseq record (nucleotide sequence) from which the protein translation was obtained.

Select one:

● a. Because they come from the same gene in the same organism, but it suffers alternative splicing producing two protein isoforms. ✖

○ b. Because they come from the same gene in different organisms, even though both genes encode the same protein sequence.

○ c. Because they come from different genes in the same organism, even though both genes encode the same protein sequence.

○ d. Because they come from different genes in different organisms, even though both genes encode the same protein sequence.

There are two refseq records for human hemoglobin subunit alpha: NP_000508.1 (from gene HBA2) and NP_000549.1 (from gene HBA1) because Humans have two different genes for this protein.

The correct answer is: Because they come from different genes in the same organism, even though both genes encode the same protein sequence.

Question 4

Correct

Mark 1.00 out of 1.00

Question 4. Note all myoglobin sequences found in Question 2.1 are also identical (except for the region not present in the shorter ones). Once again, if refseq is considered a non-redundant database, why did it accept these identical human sequences in separate files?

Tips: RefSeq records contain a COMMENT section or field that includes, among other things, some explanations justifying the inclusion of sequences in the non-redundant database.

Select one:

○ a. Don't know/no answer (without penalty).

○ b. Because refseq is a primary, non-curated database accepting all sort of row data.

● c. Because they come from different transcript variants of the same gene, even though all variants encode the same protein. ✓

○ d. Because they come from different sequencing projects and different human individuals.

○ e. Because they come from different genes in the same organism, even though all genes encode the same protein.

Read the comments associated to each record. For instance:
COMMENT .... Summary: ... Multiple transcript variants encoding distinct isoforms exist for this gene.

And take a look a the "Evidence-Data-Start"

myoglobin isoform 2 NP_001369742.1

```
            ##Evidence-Data-START##
            Transcript exon combination :: SRR5189655.183016.1,
                                           SRR7346977.669151.1 [ECO:0000332]
            ##Evidence-Data-END##
```

myoglobin isoform 2 NP_001369741.1

```
           ##Evidence-Data-START##
            Transcript exon combination :: DRR138508.22234.1, AA393926.1
                                           [ECO:0000332]
            ##Evidence-Data-END##
```

We can see that different exons have created each isoform although they end up having the same sequences:

```
Query= NP_001369742.1 myoglobin isoform 2 [Homo sapiens]

Length=99



                                                                  Score     E     Max
Sequences producing significant alignments:                      (Bits)  Value  Ident

NP_001369741.1 myoglobin isoform 2 [Homo sapiens]                  202     2e-74  100%

ALIGNMENTS
>NP_001369741.1 myoglobin isoform 2 [Homo sapiens]
 NP_001369742.1 myoglobin isoform 2 [Homo sapiens]
 AAH18001.1 MB protein [Homo sapiens]
Length=99

 Score = 202 bits (513),  Expect = 2e-74, Method: Compositional matrix adjust.
 Identities = 99/99 (100%), Positives = 99/99 (100%), Gaps = 0/99 (0%)

Query  1    MKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQV  60
            MKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQV
Sbjct  1    MKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQV  60

Query  61   LQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG  99
            LQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG
Sbjct  61   LQSKHPGDFGADAQGAMNKALELFRKDMASNYKELGFQG  99
```

The correct answer is: Because they come from different transcript variants of the same gene, even though all variants encode the same protein.

Question **5**

Correct

Mark 4.00 out of 4.00

Question 5. Check now that there is only one record for human myoglobin at Swissprot at UniProtKB database.

What is its unique accession number? P02144 ✓

Indicate the type of evidence that supports the existence of this protein:

● Experimental evidence at protein level.✓
○ Experimental evidence at transcript level.
○ Protein inferred from homology.
○ Protein predicted.
○ Don't know/no answer (without penalty).

Mark 1.00 out of 1.00

The correct answer is: Experimental evidence at protein level.

What is the subcellular localization for this protein?

○ Golgi apparatus.
○ Mitochondrion.
● Cytosol.✓
○ Cell membrane.
○ Don't know/no answer (without penalty).

Mark 1.00 out of 1.00

The correct answer is: Cytosol.

5.1. Search for the UniRef sets of proteins containing this entry. How many swiss-prot entries are included within the UniRef50 cluster for Myoglobin (including the one we've been talking about here)? 22 ✓

The accession code at UniProtKB is P02144. At NCBI it is P02144.2. (both codes are corrects)

you can find it by searching this in NCBI protein:

human[porgn] AND "swissprot"[Filter] AND myoglobin [protein]

or this:

(organism_id:9606) AND (reviewed:true) AND (protein_name:myoglobin)

in uniprot

(remember that you can use advanced search)

All information to continue answering the questions could be found at:

https://www.uniprot.org/uniprotkb/P02144/entry

or

https://www.uniprot.org/uniref/UniRef50_P02144

https://www.uniprot.org/uniref/UniRef50_P02144?facets=uniprot_member_id_type%3Auniprotkb_reviewed_swissprot

you can also make this search in UniprotKB:

(uniref_cluster_50:UniRef50_P02144) AND (reviewed:true)

where only 22 entries belong to UniProt/Swiss-prot

Question **6**

Partially correct

Mark 4.50 out of 6.00

Question 6. Localize within the reviewed entry for human myoglobin at UniProt/Swiss-prot the link to its protein structure experimentally solved. Indicate the unique ID for this structure. 3RGK ✓

Visualize the PDB file for this structure at RCSB PDB and answer the following True-False questions:

a) This protein was co-crystalized together with a residue SO4. True ⬍ ✓

b) The heme group is included in the structure. True ⬍ ✓

c) This structure was obtained with very low resolution. True ⬍ ✗

d) Two protein chains are presented in this PDB file. False ⬍ ✓

e) Histidine at position 48 participates in the binding to heme group. False ⬍ ✓

Crystal structure of human myoglobin at PDB is 3RGK. At its PDB flat text file you will find all information needed to answer the questions.

a) This protein was co-crystalized together with a residue SO4.

```
REMARK 800 SITE_IDENTIFIER: AC2
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE SO4 A 300
```

b) The heme group is included in the structure.

```
REMARK 800 SITE_IDENTIFIER: AC1
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE HEM A 154
```

c) This structure was obtained with very low resolution.

```
REMARK   2 RESOLUTION.    1.65 ANGSTROMS.
```

This is high resolution (low numbers = High resolution)

d) Two protein chains are presented in this PDB file.

```
COMPND   3 CHAIN: A;
```

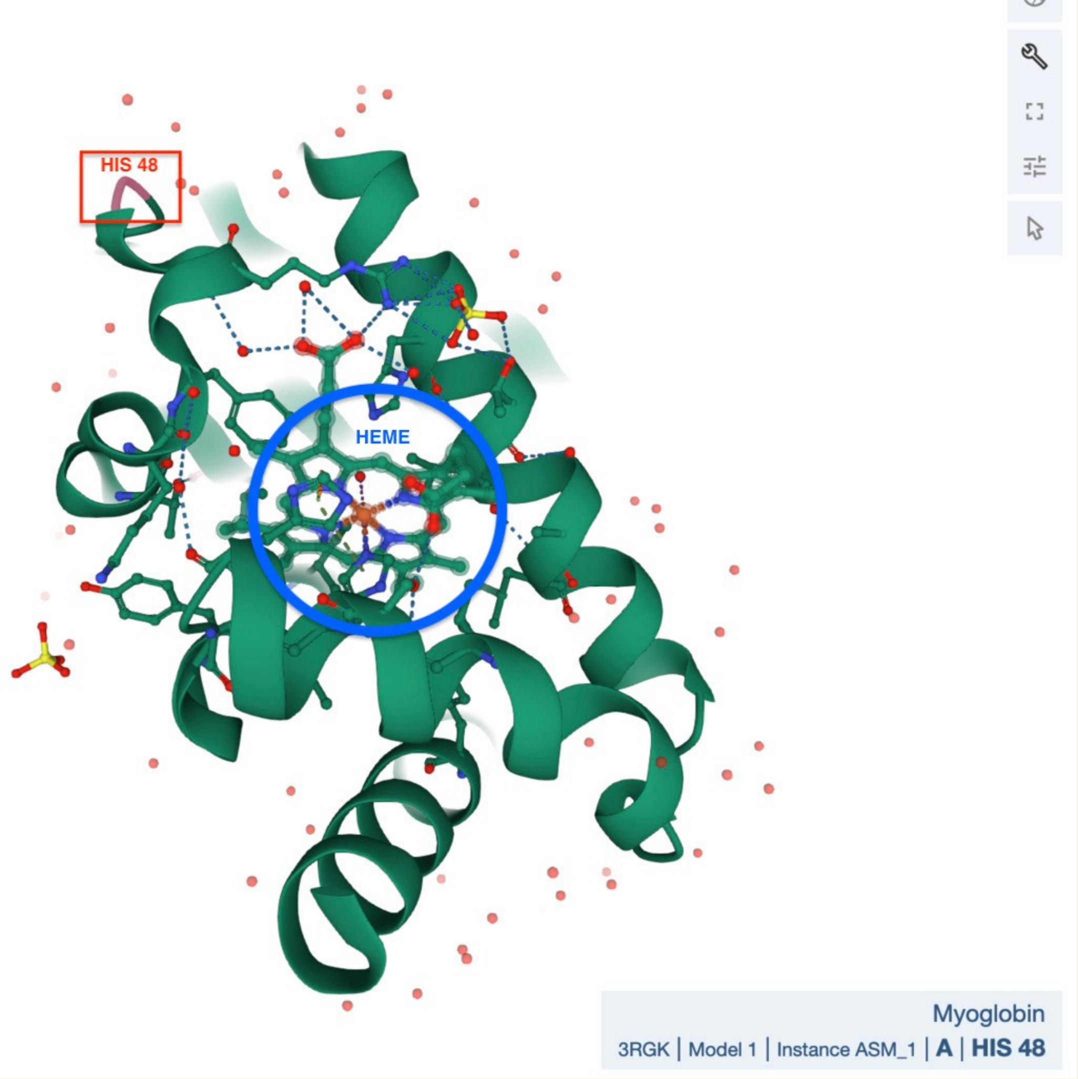
Only one chain "A"

e) Histidine at position 48 participates in the binding to heme group.

Based on the PDB site the answer is TRUE. But if you look at the 3D structure 48H is far from the pocket where the heme group accomodates. So, FALSE was also accepted!

```
REMARK 800 SITE_IDENTIFIER: AC1
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE HEM A 154
:    :      :     :      :       :
```

```
SITE     1 AC1 20 LYS A  42  PHE A  43  ARG A  45  HIS A  48
SITE     2 AC1 20 GLU A  54  ALA A  57  HIS A  64  VAL A  68
SITE     3 AC1 20 LEU A  89  SER A  92  HIS A  93  HIS A  97
SITE     4 AC1 20 ILE A  99  TYR A 103  PHE A 138  HOH A 200
SITE     5 AC1 20 HOH A 207  HOH A 211  HOH A 219  HOH A 290
```



Question 7

Partially correct

Mark 1.50 out of 3.00

Question 7. How many reviewed (swiss-prot) entries are there for human hemoglobin (including all subunits) at UniProtKB? 11 ☑

For how many of these entries are there clear experimental evidences for the existence of the protein? 11 ☑

For how many of these entries are there links to the PDB? 8 ☑

Tips: UniProt accepts incomplete protein names for the search field "protein name".

a) (protein_name:"hemoglobin") AND (organism_id:9606) AND (reviewed:true)

but, sequences Q86VB7 (Scavenger receptor cysteine-rich type 1 protein M130) and Q9NZD4 (Alpha-hemoglobin-stabilizing protein) should be excluded. So the search should be:

(protein_name:"hemoglobin subunit") AND (organism_id:9606) AND (reviewed:true)

b) (protein_name:"hemoglobin subunit") AND (organism_id:9606) AND (reviewed:true) AND (existence:1)

c) (protein_name:"hemoglobin subunit") AND (organism_id:9606) AND (reviewed:true) AND (existence:1) AND (structure_3d:true)

Question 8

Correct

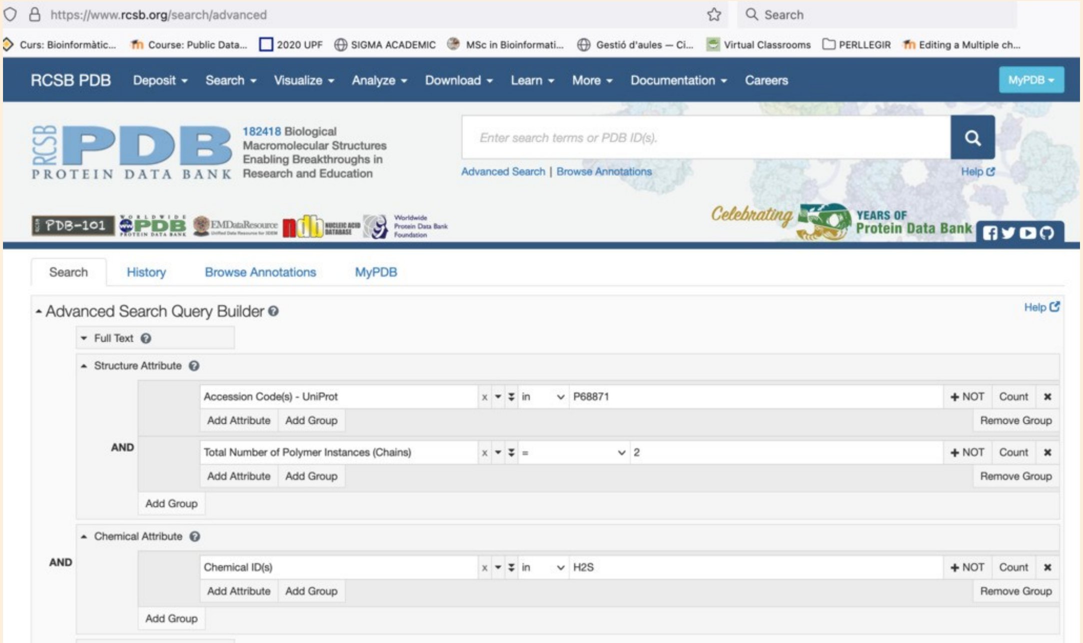
Mark 2.00 out of 2.00

Question 8. Note the vast amount of protein structures there are for human hemoglobin subunit beta by searching inside its Uniprot/Swiss-prot file with accession number P68871 ✓ .

From this list, what is the accession code at RCSB PDB for a structure containing only two chains and with hydrosulfuric acid (H_2S) as a ligand (it is a chemical) ? 5UCU ✓

Tip: Advanced Search at RCSB PDB database could help to answer this question.

At RCSB PDB site, advanced search:



https://www.rcsb.org/search/advanced

Advanced Search Query Builder

Structure Attribute

Accession Code(s) - UniProt | in | P68871

AND

Total Number of Polymer Instances (Chains) | = | 2

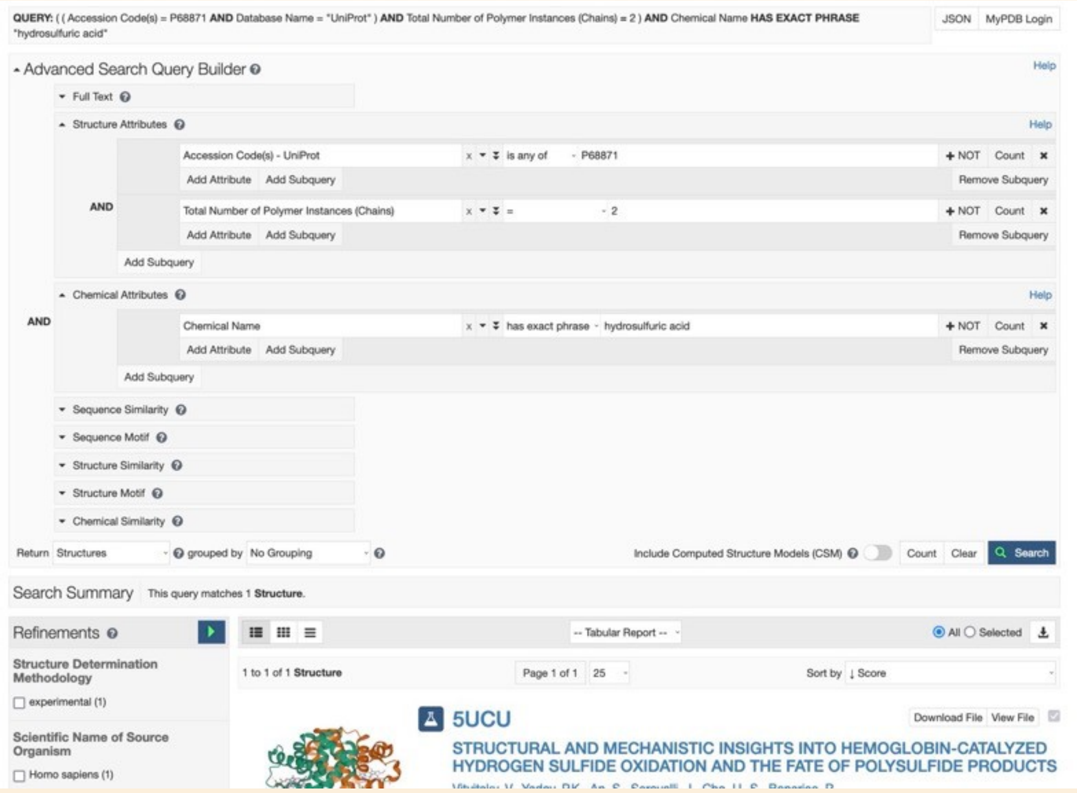
AND

Chemical Attribute

Chemical ID(s) | in | H2S

Search
Curs: Bioinformàtic... | Course: Public Data... | 2020 UPF | SIGMA ACADEMIC | MSc in Bioinformati... | Gestió d'aules — Ci... | Virtual Classrooms | PERLLEGIR | Editing a Multiple ch...
RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | Documentation | Careers | MyPDB
RCSB PDB PROTEIN DATA BANK
182418 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education
Enter search terms or PDB ID(s).
Advanced Search | Browse Annotations
Help
PDB-101 | WORLDWIDE PDB PROTEIN DATA BANK | EMDataResource | NUCLEIC ACID DATABASE | Worldwide Protein Data Bank Foundation
Celebrating 50 YEARS OF Protein Data Bank
Search | History | Browse Annotations | MyPDB
Help
Full Text
Add Attribute | Add Group | Remove Group
+ NOT | Count
Add Group

or



QUERY: ((Accession Code(s) = P68871 **AND** Database Name = "UniProt") **AND** Total Number of Polymer Instances (Chains) = 2) **AND** Chemical Name **HAS EXACT PHRASE** "hydrosulfuric acid"

Advanced Search Query Builder

Structure Attributes

Accession Code(s) - UniProt | is any of | P68871

AND

Total Number of Polymer Instances (Chains) | = | 2

AND

Chemical Attributes

Chemical Name | has exact phrase | hydrosulfuric acid

Return Structures grouped by No Grouping

Search Summary This query matches 1 **Structure**.

1 to 1 of 1 **Structure**

5UCU

STRUCTURAL AND MECHANISTIC INSIGHTS INTO HEMOGLOBIN-CATALYZED HYDROGEN SULFIDE OXIDATION AND THE FATE OF POLYSULFIDE PRODUCTS

Vitvitsky, V., Yadav, P.K., An, S., Seravalli, J., Cho, U.S., Banerjee, R.

Result: Only one record! 5UCU

Question 9

Partially correct

Mark 1.87 out of 3.00

Question 9. Given this sequence in FASTA format for human hemoglobin, to which subunit correspond?

Delta ⬍ ✓

To which of the other human subunit this sequences is more similar? Beta ⬍ ✓

And with which shares the least similarity? Gamma ⬍ ✗

Tip: you can use both NCBI blast and uniprot and restrict your search for a specific organism in any database (including swissprot)

Blastp vs. refseq or swiss-prot databases and only for human sequences. Look at table with sequences producing significant alignments the % of identity and query coverage.

a) Query sequence is identical to hemoglobin subunit delta

b) Query sequence is 93% identical to hemoglobin subunit beta

c) Query sequence is only 36% identical to hemoglobin subunit mu

Question **10**

Partially correct

Mark 2.00 out of 4.00

Question 10. Does this DNA sequence encode for hemoglobin or myoglobin?

○ Yes, it codes for hemoglobin.

● Yes, it codes for myoglobin. ✓

○ No, it codes for an unknown protein.

○ It is not a codifying sequence.

○ Don't know/no answer (without penalty).

Mark 1.00 out of 1.00

The correct answer is: Yes, it codes for myoglobin.

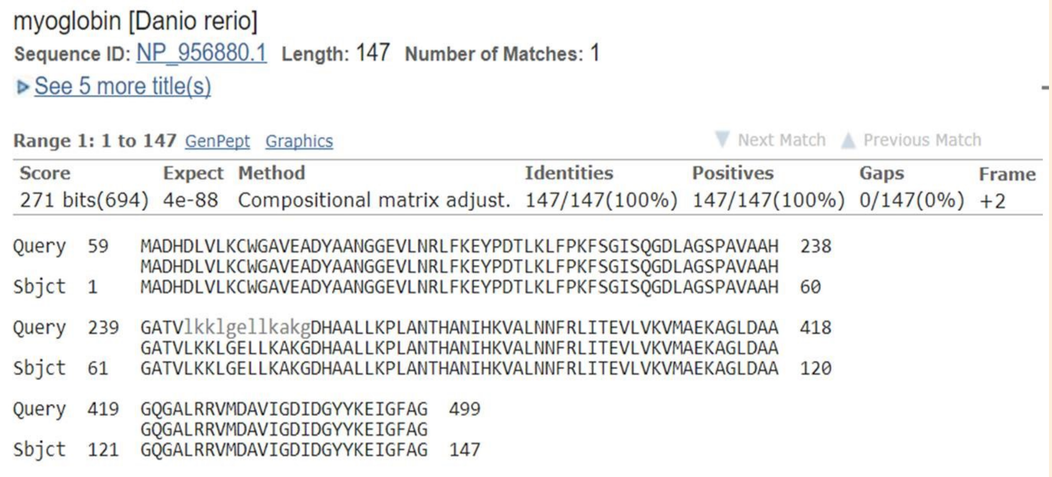
Between which coordinates (in your sequence) did you find the CDS? From 60 ✗ to 89 ✗

To which organism is more probable this sequence belong to? taxid= 7955 ✓ (an integer)

Tip: use both implementations of blast (ncbi and uniport) Which one gets your answers easily?

After blastx vs nr. protein or vs swissprot or vs refseq this DNA sequence encodes for a protein 100% identical to myoglobin in zebrafish (Danio rerio; taxid=7955) between coordinates 59 and 499 in the frame +2.

3' coordinate in 502 has also been accepted if we include stop codon.



myoglobin [Danio rerio]

Sequence ID: NP_956880.1 Length: 147 Number of Matches: 1

▶ See 5 more title(s)

Range 1: 1 to 147 GenPept Graphics

▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
271 bits(694)	4e-88	Compositional matrix adjust.	147/147(100%)	147/147(100%)	0/147(0%)	+2

```
Query  59   MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH  238
            MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH
Sbjct  1    MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH  60

Query  239  GATVlkklgellkakgDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKAGLDAA  418
            GATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKAGLDAA
Sbjct  61   GATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKAGLDAA  120

Query  419  GQGALRRVMDAVIGDIDGYYKEIGFAG  499
            GQGALRRVMDAVIGDIDGYYKEIGFAG
Sbjct  121  GQGALRRVMDAVIGDIDGYYKEIGFAG  147
```

Or a blastn provides this best match

https://www.ncbi.nlm.nih.gov/nucleotide/BC056727.1

Were wee can see that the CDS (codifying region) starts at 59 and ends at 502 (it includes the stop)

Question 11

Correct

Mark 2.00 out of 2.00

Question 11 Let’s investigate if organisms lacking red blood cells produce a protein similar to hemoglobin.

Do molluscs produce a protein similar to human alpha hemoglobin?

○ Yes, there is an almost identical protein with > 90% identity.

○ Yes, there is a similar protein with > 60% similarity (positives) along the entire sequence.

● Yes, there is a similar protein with > 45% similarity (positives) in at least half of the sequence.✓

○ No, there is no protein similar to human alpha hemoglobin among mollusc.

○ Don't know/no answer (without penalty).

Mark 1.00 out of 1.00

The correct answer is: Yes, there is a similar protein with > 45% similarity (positives) in at least half of the sequence.

Do flowering plants produce a protein similar to human alpha hemoglobin?

○ Yes, there is an almost identical protein with > 90% identity.

○ Yes, there is a similar protein with > 60% similarity (positives) along the entire sequence.

○ Yes, there is a similar protein with > 45% similarity (positives) in at least half of the sequence.

● No, there is no protein similar to human alpha hemoglobin among flowering plants.✓

○ Don't know/no answer (without penalty).

Mark 1.00 out of 1.00

The correct answer is: No, there is no protein similar to human alpha hemoglobin among flowering plants.

Tip: positives (AKA similarity) is not provided by the blast GUI implemented in uniprot, use the NCBI one.

NCBI BLAST!

Human alpha hemoglobin at refseq is NP_000549.1 or NP_000508.1.Both identical. Either one can be used for the search. At UniProt/Swiss-prot is P69905.

NCBI: human[PORGN] AND hemoglobin subunit alpha[TI] AND refseq[filter]

Then we search for mollusc and flowering plants in the ncbi taxonomy database to make sure we are using the right name and the right taxid

a) blastp vs. nr. protein database for molluscs (taxid:6447)

There are several hit sequences with >45% similarity and >50% query coverage. For instance, sequence ACN90947.1 (69% query coverage and 48% similarity) annotated as an hemoglobin in *Anadara kagoshimensis*, a kind of mollusk.

b) blastp vs. nr. protein database for flowering plants (taxid:3398)

There is no protein similar to human alpha hemoglobin among flowering plants.