| | |
|---|---|
| **Started on** | Wednesday, 27 September 2023, 4:25 PM |
| **State** | Finished |
| **Completed on** | Thursday, 28 September 2023, 11:34 AM |
| **Time taken** | 19 hours 9 mins |
| **Marks** | 13.08/16.00 |
| **Grade** | **8.18** out of 10.00 (**81.77**%) |

**Question 1**

Partially correct

Mark 3.75 out of 4.00

**Question 1.** What NCBI Entrez is and what does it allow us to do? (yes-no questions)

| | |
|---|---|
| YES | It is a comprehensive website for biologists. |
| NO | It is a language used by Relational Database Management Systems. |
| YES | It is the text-based search and retrieval system used at the NCBI. |
| YES | It is an integrated database system. |
| NO | It comprises only molecular databases. |
| YES | It can be accessed through the search box at the top of the NCBI homepage. |
| YES | It facilitates constructing both simple and more sophisticated queries. |
| YES | It may be accessed programmatically for high volume searches. |

Tips
Visit Entrez Help

*Entrez* is a database system that provides integrated access to NCBI databases. It is available via the WWW.

I accepted (50%) it is a website, but IT IS NOT a website. The website to access it is ncbi.

**Question 2**

Correct

Mark 1.00 out of 1.00

**Question 2.** In a Global Search at the NCBI by the Entrez system the terms “human” and “*Homo sapiens*” in the organism field, are usually considered as one term. Why is it possible?

**A.** Taxonomy database provides controlled vocabulary for the major bio-molecular Entrez databases.
**B.** MeSH database provides controlled vocabulary for articles in PubMed.

**Tips**:
See "Controlled Vocabulary Fields and Query Mapping" here.

Select one:

- a. Both A and B are correct. ✔
- b. Both A and B are wrong.
- c. Don't know/no answer (without penalty)
- d. Only B is correct.
- e. Only A is correct.

Both are correct. However you should consider most of the source databases in Entrez Nucleotide do not use a controlled vocabulary and the way in which a submitter describes his/her sequence can vary. The only controlled vocabulary used by archival databases is that offered by the taxonomy database initiative, but it only helps to control organism names. Then in this case both A and B are correct.

The correct answer is: Both A and B are correct.

**Question 3**

Correct

Mark 1.00 out of 1.00

**Question 3**. With which of these strategies will you find in PubMed all the scientific papers containing the word "human" in the title and published in the last five years but are not review articles?

**Tips**:

See "Using search field tags" for PubMed here.

Select one:

- a. human[title] OR ("2017/09/21"[DP] : "2022/09/19"[DP]) NOT Review[PT]
- b. human[title] AND ("2023/09/21"[DP] : "2018/09/19"[DP]) AND Review[PT]
- c. human[filter] AND ("2018/09/21"[DP] : "2023/09/19"[DP]) NOT Review[PT]
- d. human[title] AND ("2018/09/21"[DP] : "2023/09/19"[DP]) NOT Review[PT] ✔
- e. Don't know/no answer (without penalty)

The correct answer is: human[title] AND ("2018/09/21"[DP] : "2023/09/19"[DP]) NOT Review[PT]

**Question 4**

Partially correct

Mark 1.33 out of 2.00

**Question 4.** What is the unique identifier (TaxID) for the house mouse at NCBI Taxonomy database. 10090

✔ (an integer, no spaces, no comas, nothing other than numerical characters)

¿What is the total number of records for this taxon at the nucleotide database to this day? 10498343 ✖ (an integer, no spaces, no comas, nothing other than numerical characters)

¿How many eukaryotic species (with formal names) are there currently in the taxonomy database 526106

✔ (an integer)

**Tips**:

- Last question requires an independent study of new field based search strategies for this database. See Taxonomy help at NCBI and Frequently Asked Questions. (I'm asking only for species, not order, genus. Please check [SubTree] , [Rank] , [prop])

The txid=10088 is for mice, genus *Mus*, but for the species "the house mouse" (*Mus musculus*) it is 10090.

A search in the nucleotide database with txid10090[porgn] finds all records for this species. [porgn] is important to select sequences really belonging to this organism. I have accepted a search txid10090[orgn] with 75%.

To see how many eukaryotic species there are, go to the taxonomy home page and click on Statistics. Or type in taxonomy database:

Eukaryota[SubTree] AND species[Rank] AND specified[prop]
Please read the first question & answer in the FAQ: https://www.ncbi.nlm.nih.gov/books/NBK54428/

That means all eukaryotic taxa in taxonomy database at the species level. "specified[prop]" helps to restrict the output of this list to species with formal Linnaean binomial names. Then, avoiding "uncultures" organisms derived from metagenomic projects or organisms named as *genus sp*.

**The option (Eukaryota[SubTree] AND species[Rank]) was rated with 75% of the final qualification and (Eukaryota[SubTree]) with 50%.**

If you type only Eukaryota[SubTree] your looking for all taxa below the superkingdom Eukaryota (species, families, genus, etc.).

**Question 5**

Correct

Mark 1.00 out of 1.00

**Question 5.** With which of these strategies will you find all the sequence from the house mouse (*Mus musculus*) stored in the nucleotide database at the NCBI.

**A**. txid10090[Primary organism]
**B**. mus musculus[Primary Organism]
**C**. mus musculus[porgn]
**D**. house mouse[porgn]

Select one:

- a. All of the search strategies are correct. ✔
- b. Only A, B and C are correct.
- c. Don't know/no answer (without penalty)
- d. Only B is correct.
- e. Only B and C are correct.

[Primary organism] = [porgn]

In taxonomy database txid10090 = "mus musculus" = "house mouse". Caution with words "mice" and "mouse"!!! There are more than one taxID using these synonyms.

The correct answers are: All of the search strategies are correct., Don't know/no answer (without penalty)

**Question 6**

Correct

Mark 1.00 out of 1.00

**Question 6.** In the following search statement in the Entrez Nucleotide database, what does the asterisk mean?

NC_0000*[Accession] AND Human[Organism]

Select one:

- a. The asterisk replaces 0 to n number of characters anywhere in the accession number
- b. The asterisk replaces just one character at the end of the accession number
- c. Don't know/no answer (without penalty)
- d. The asterisk replaces 0 to n number of characters at the end of the accession number ✔
- e. The asterisk replaces just one character anywhere in the accession number

The correct answer is: The asterisk replaces 0 to n number of characters at the end of the accession number

**Question 7**

Incorrect

Mark 0.00 out of 2.00

**Question 7**. Which of the following NCBI records cross-reference each other?

Select only two UIDs from the list

Select one or more:

- [ ] NP_000508.1
- [ ] NM_000518.5
- [ ] DQ659148.1
- [ ] AAP35454.1
- [x] BT006808.1 ✔
- [x] BC005255.1 ✖

Protein sequence AAP35454.1 was obteined from the nucleotide entry BT006808.1.

The correct answers are: BT006808.1, AAP35454.1

**Question 8**

Correct

Mark 1.00 out of 1.00

**Question 8**. What of these strategies is the easiest way (only one of them is the easiest) to download a large set of not contiguous records from the same NCBI database using a list of unique identifiers?

Select one or more:

- [x] a. Using Batch Entrez. ✔
- [ ] b. Through a File Transfer Protocol (FTP).
- [ ] c. Using the Entrez Programming Utilities: "E-utilities" or Entrez Direct: E-utilities on the Unix Command Line
- [ ] d. Using an advanced text query with the search field [Accession].

The FTP service allows users to download files containing whole subset of data. It is for large data downloading.

Entrez Programming Utilities AKA: "E-utilities" provide a way to access entrez data as a web service. It is very versatile, but it is not simpler than batch entrez. Although there is no penalty if you answered this.

The correct answer is: Using Batch Entrez.

**Question 9**

Correct

Mark 2.00 out of 2.00

**Question 9.** From the list of the NCBI accession numbers stored in the flat text file Acc_List.txt, answer the following questions:

What database belong all accession numbers? Protein ✔

How many records are there listed? 295 ✔

Open the listed records on the appropriate NCBI database and select the correct statement.

- (●) All the sequences were obtained from just one species. ✔
- ( ) All the sequences have the same length.
- ( ) The sequences belong to more than one species.
- ( ) All the sequences are related to the same protein function and biological process.
- ( ) None of the statements are correct.
- ( ) Don't know/no answer (without penalty).

Mark 1.00 out of 1.00

The correct answer is: All the sequences were obtained from just one species.

Note: If batchentrez does not work, check the last part in the last exercise in this topic. Here a hint: (WARNING: **WRITE_THE_RIGHT_DB_HERE** must be changed)

```
cat Acc_List.txt |xargs -tI% wget -O %.fasta "https://eutils.ncbi.nlm.nih.gov/entrez/eutils
/efetch.fcgi?db=WRITE_THE_RIGHT_DB_HERE&id=%&rettype=fasta&retmode=text"
```

If batchentrez did work, you can see something weird happening. Can you find it? if you do, In the forum for this topic, explain what it is, and why. If you do, you will be rewarded.

All records start for NP_ (so they are proteins), but if you don't know the NP_ meaning by memory you just chek one of them and you'll see. Or just try all of them in Batch Entrez until you get results :)

Once you get the Batch Entrez results you can check the species in the left column.

If you download all sequences in fasta format, you can answer all this questions with the following commands:

```
less -NS Acc_List.txt
grep -h \> *fasta|cut -d \[ -f 2|sort|uniq
grep -h \> *fasta|less -NS
```

**Question 10**

Correct

Mark 1.00 out of 1.00

**Question 10.** Since the beginning the NCBI have used two different unique codes (primary keys) to identify the same sequence entry: the GI and the accession number. The GI number has been used for many years by NCBI to track sequence histories in sequence databases. However, NCBI has changed the way they handle GI numbers for sequence records. That means, accession.version identifiers, rather than GI numbers, are the primary identifiers for sequence records. Presentation of GI sequence identifiers in the GenBank flatfile format was discontinued as of March 2017.

What does this E-utilities URL do?

efetch.fcgi?db=nuccore&id=663070995,568815587&rettype=acc

Note: There is something missing in this URL, finding it will be very helpful.

Select one:

- a. It returns sequences for two entries in the nucleotide database in the *acc* format.
- b. It converts two GI numbers to Accession.version numbers. ✔
- c. It returns all kown data for two entries in the nucleotide database in *acc* output format.
- d. It converts two Accession.version numbers to GI numbers.
- e. Don't know/no answer (without penalty).

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=663070995,568815587&rettype=acc

It converts two GI numbers to Accession.version numbers.

https://www.ncbi.nlm.nih.gov/nuccore/663070995

NM_001178.5

https://www.ncbi.nlm.nih.gov/nuccore/568815587

NC_000011.10

The correct answer is: It converts two GI numbers to Accession.version numbers.

| | |
|---|---|
| **Started on** | Thursday, 5 October 2023, 9:26 AM |
| **State** | Finished |
| **Completed on** | Monday, 9 October 2023, 11:08 PM |
| **Time taken** | 4 days 13 hours |
| **Marks** | 14.43/20.00 |
| **Grade** | **7.21** out of 10.00 (**72.14**%) |

**Question 1**

Correct

Mark 1.00 out of 1.00

**Question 1.** Select the description best corresponding to each NCBI accession code. (No penalty)

**Note:** Select a different commentary for each accession number.

**Tips:** The revision history format shows the various accession numbers, version numbers, and update dates for sequences that appeared in a specific GenBank record.

| | |
|---|---|
| GFSJ01000014.1 | The entry contains a sequence from a transcriptome shotgun assembly (TSA) |
| AC275307.1 | The entry contains a working draft sequence from a high-throughput genomic (HTG) approach |
| CP018631.1 | The current version of a nuccore genbank entry for a circular DNA molecule |
| M60741.1 | An old version of a nuccore genbank entry |
| LKPB02002952.1 | The entry contains a whole genome shotgun sequence (WGS) |
| NM_000517.6 | The current version of a refseq entry |

The correct answer is: GFSJ01000014.1 → The entry contains a sequence from a transcriptome shotgun assembly (TSA), AC275307.1 → The entry contains a working draft sequence from a high-throughput genomic (HTG) approach, CP018631.1 → The current version of a nuccore genbank entry for a circular DNA molecule, M60741.1 → An old version of a nuccore genbank entry, LKPB02002952.1 → The entry contains a whole genome shotgun sequence (WGS), NM_000517.6 → The current version of a refseq entry

**Question 2**

Correct

Mark 2.00 out of 2.00

**Question 2.** What sequencing technology was used to obtain sequence LKPB02002952.1?

- Illumina HiSeq2500.
- PacBio. ✔
- Ion torrent.
- Oxford nanopore.
- Don't know/no answer. (without penalty)

Mark 1.00 out of 1.00

The correct answer is: PacBio.

What was the aim of the sequencing project in which this sequence was obtained?

- To sequence only the X and Y chromosomes in a family.
- To study the expression of human genes.
- To sequence the human genome for the very first time.
- To add diverse allelic variation to the reference human genome. ✔
- Don't know/no answer (without penalty)

Mark 1.00 out of 1.00

The correct answer is: To add diverse allelic variation to the reference human genome.

GenBank files usually contains this information.

https://www.ncbi.nlm.nih.gov/nuccore/LKPB02002952.1

```
Sequencing Technology :: PacBio
```

```
Sequence from this project will be used to improve the contiguity of the human reference
sequence and add diverse allelic variation
```

You can also find the second answer in the bioproject link:

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA288807

For the 1st one you will have to go to ASSEMBLY from the "related information" link in the bioproject page:

https://www.ncbi.nlm.nih.gov/assembly?LinkName=bioproject_assembly_all&from_uid=288807
And select "NA19240_prelim_3.0" which will send you to the entry in the Assembly database:

https://www.ncbi.nlm.nih.gov/assembly/GCA_001524155.4

**Question 3**

Correct

Mark 1.00 out of 1.00

**Question 3.** Two of these search tags will help you finding sequences that are codifying at nucleotide database (CDS). Which ones?

**Tips**:

Learn about Search Field Descriptions for Sequence Database in this link.

Select one or more:

- [ ] a. protein_nucleotide[filter]
- [ ] b. All[protein]
- [x] c. nucleotide_protein[filter] ✔
- [x] d. CDS[fkey] ✔
- [ ] e. protein[fkey]
- [ ] f. protein[filter]

nucleotide_protein[filter] will deliver sequences that are in nucleotide (AKA: nuccore) and have a link to protein.

CDS[fkey]: sequences that contain one or more CDS features annotated

however you must take into account that not all annotated CDSs have a link to the protein DB

The correct answers are: nucleotide_protein[filter], CDS[fkey]

**Question 4**

Partially correct

Mark 2.00 out of 2.00

**Question 4.** Localize all of the records in the NCBI nucleotide database containing rabbit genomic sequences that are codifying. How many have you found? 4061 ✔

Optionally write here your search strategy

rabbit[Organism] AND biomol_genomic[PROP] AND CDS[fkey] ✖ (no rate)

¿How many of these sequences come from the mitochondrial compartment? 159 ✔

Here the question asked for genomic sequences, then, biomol_genomic[PROP].

rabbit[porgn] AND biomol_genomic[PROP] AND CDS[fkey]

This is not the right answer:

rabbit[PORGN] AND biomol_genomic[PROP] AND nucleotide_protein[filter]

the fact that their products don't have a link to the protein DB does not mean that they are not there.

In the left margin you can see that 159 are in the Mithocondrion

OR use this search:

rabbit[porgn] AND biomol_genomic[PROP] AND CDS[fkey] AND mitochondrion[filter]

**Question 5**

Partially correct

Mark 1.43 out of 5.00

**Question 5.** Do you think you have found in Question 4 the NCBI nucleotide reference entry for the complete rabbit mitochondrial chromosome? Yes ✔

If yes, what is its refseq access number? AJ001588.1 ✖

How many proteins does it code? 13 ✔

Between what coordinates do we find the tRNA for tryptophan? From 4971 ✔ to 5037 ✔

Click on the filter "Mithochondiron" in the question 4 results and select the only refseq result refering to the Mithocondrion complete genome "Oryctolagus cuniculus mitochondrion, complete genome". NC_001913.1 (https://www.ncbi.nlm.nih.gov/nuccore/NC_001913.1). You can recognize it is a code from RefSeq database due to its starting characters "NC_".

Or perform a search like this:

rabbit[porgn] AND biomol_genomic[prop] AND CDS[fkey] AND refseq[filter] AND mitochondrion[filter]

Use "Related information" box and click in protein: (https://www.ncbi.nlm.nih.gov/protein?LinkName=nuccore_protein&from_uid=5835526) to obtain results for the proteins encoded in this entry.

In the gbff file (https://www.ncbi.nlm.nih.gov/nuccore/NC_001913.1) search for "tRNA-Trp"

```
tRNA            4971..5037
                /product="tRNA-Trp"
```

**Question 6**

Partially correct

Mark 1.00 out of 1.00

**Question 6.** ¿How many complete mammalian mitochondrial chromosomes can be found at the NCBI nucleotide database? 77718 ✔

Optionally write here your search strategy

mammals[filter] AND "complete genome"[title] also select mitochondrion in ✖ (no rate)

**Tips**: Mitochondria has its own genome. Researchers and/or curators use to add to the title of entries containing complete genomic sequences the phrase “complete genome” for fully sequenced chromosomes or genomes of eukaryotic organelles.

mammalia[porgn] AND mitochondrion[filter] AND "complete genome"[TI]

mitochondrion[filter] and gene_in_mitochondrion[PROP] both give the same result.

txid40674 and mammalia are synonyms.

"complete genome" should be added using [title] or [word] as search tag or simply nothing. This is an approximation to the best way to get the most results. There is not a search tag to filter for the status of a genome sequencing project at NCBI nucleotide.

**Question 7**

Correct

Mark 1.00 out of 1.00

**Question 7.** One of the proteins encoded in the mitochondrial chromosomes is cytochrome b. The best strategy to search for sequences encoding for cytochrome b proteins in NCBI nucleotide database is to include in a query the expression:

Select one:

- a. AND (cytochrome b[protein name]) ✔
- b. AND “cytochrome b”[title]
- c. NOT cytochrome*
- d. AND cytochrome*[title]
- e. Don't know/no answer (without penalty)

in this case you should evaluate what would happen if you add each of the phrases in a search strategy:

AND “cytochrome b”[title]; with this phrase you are including other proteins like cytochrome b reductase (e.g. NM_001011954.1).

AND (cytochrome b[protein name]); this strategy will retrieve all nucleotide sequences encoding for a protein named exactly as "cytochrome b", including complete bacterial genomes and eukaryotic mitochondrial genomes.

AND cytochrome*[title]; this strategy includes other variants like cytochrome c.

NOT cytochrome*; this strategy will exclude cytochrome b from your search

The correct answer is: AND (cytochrome b[protein name])

**Question 8**

Correct

Mark 2.00 out of 2.00

**Question 8.** Find out the NCBI genomic reference sequence encoding for human (us) cytochrome b protein.

Indicate the NCBI accession numbers for both the genomic sequence NC_012920.1 ✔ and the encoded protein YP_003024038 ✔ .

Strategy

"human"[Primary Organism] AND "cytochrome b"[protein name] AND refseq[filter]

Sequence with accession number NC_012920.1 is the reference sequence in refseq for the Homo sapiens mitochondrion complete genome. This is the reference DNA sequence encoding, among other proteins, for the human cytochrome b protein (YP_003024038.1).

AC_000021 is an obsolete version of NC_012920, however, if you have answred "NC_012920, AC_000021" this is counted as correct. Please notify me if you have introduced a different variation

**Question 9**

Correct

Mark 1.00 out of 1.00

**Question 9.** By querying the NCBI nucleotide database investigate if bacteria genomes encode for cytochrome b protein.

**Tips**:

Enterobacteria and Cyanobacteria are two different groups of bacteria.

Select one:

- a. Yes, cytochrome b is exclusive of enterobacteria.
- b. No, bacteria do not have cytochrome b protein.
- c. Don't know/no answer (without penalty).
- d. Yes, but only in the group of Cyanobacteria.
- e. Yes, several bacterial groups encode for a protein named cytochrome b. ✔

bacteria[porgn] AND "cytochrome b"[protein] You will get more than 128000 entries in almost all bacterial groups.

Enterobacteria and Cyanobacteria are two groups of bacteria. If you indicate the following search strategies you will find fewer number of entries.

cyanobacteria[porgn] AND "cytochrome b"[protein]

enterobacteria[porgn] AND "cytochrome b"[protein]

The correct answer is: Yes, several bacterial groups encode for a protein named cytochrome b.

**Question 10**

Partially correct

Mark 2.00 out of 2.00

**Question 10.** Complete microbial genomes at NCBI.

**10.1**.Open entry with accession number NC_001802.1. What does this sequence represent?

- A partial sequence for a viral genome.
- A complete genomic sequence for a human gene.
- A reference sequence for a complete viral genome. ✔
- The reference sequence for the human genome.
- Don't know/no answer (without penalty)

Mark 1.00 out of 1.00

The correct answer is: A reference sequence for a complete viral genome.

**10.2**. It is possible to find complete genome sequences in the NCBI nucleotide database for any organism. Create a search strategy trying to find at the nucleotide database all the complete or nearly complete genomic sequences for Human immunodeficiency virus type 1.

How many entries have you found? 6723 ✔

Optionally write here your search strategy

txid11676[Organism] AND "complete genome"[Title] ✖ (no rate)

**Tips:**

- Use the taxid instead of the full name for the Human immunodeficiency virus type 1
- Researchers and/or curators use to add to the title of entries containing complete or nearly complete genome sequences the phrase “complete genome”

https://www.ncbi.nlm.nih.gov/nuccore/NC_001802.1

Human immunodeficiency virus 1, complete genome

Strategy at nucleotide database:

Usually you will go with:

txid11676[Primary organism] AND "complete genome"[title]

But it is not bad to check this:

txid11676[Primary organism] AND "complete genome"

you get 10 more results. Some of them are partial, but some of them are good like this one: https://www.ncbi.nlm.nih.gov/nuccore/AJ508595.1

**Question 11**

Incorrect

Mark 0.00 out of 2.00

**Question 11.** There are several genomes at the NCBI without annotations. That means none of the genetic elements like CDSs, genes or any other features annotated as “misc_feature”, have been indicated in the sequence. Search the records in the NCBI nucleotide database for HIV-1 (Human immunodeficiency virus type 1) complete genomes without any annotation.

How many entries have you found? 4491 ✖

Indicate here the accesion.version number of only one of them. AB253704.1 ✖

Optionally indicate here your search strategy

((txid11676[Organism]) AND "complete genome"[Title]) NOT "misc_featur ✖ (no rate)

Option 1.

txid11676[porgn] AND "complete genome"[title] NOT CDS[fkey] NOT Gene[fkey] NOT misc_feature[fkey]

Only 5 entries fulfill these criteria. JN571034.1, AY781127.1, AY781125.1, AY781128.1 and AY781126.1.

UPDATE: now it is 10!

JN571034.1, AY781127.1, AY781125.1, AY781128.1, AY781126.1, ON245430.1, ON245428.1, ON245429.1, ON245427.1, ON245431.1

so both answers are considered right.

However, there is a second option:

txid11676[Primary organism] AND complete genome NOT CDS[fkey] NOT Gene[fkey] NOT misc_feature[fkey]

Here we are not forcing the words "complete" and "genome" neither to the title nor to be together so we can view more results

We use to get 21 good entries (2LDL_A is not a complete genome) with accession numbers:

```
JX503079.1
M93259.1
M93258.1
JX503077.1
JX503078.1
JX503080.1
JX503073.1
JX503074.1
JX503082.1
JX503071.1
JN571034.1
JX503072.1
JX503075.1
JX503081.1
JX503076.1
JX503083.1
AY781127.1
AY781125.1
AY781128.1
AY781126.1
MK457954.1
```

UPDATE: now is 26 (again 2LDL_A is not a complete genome)

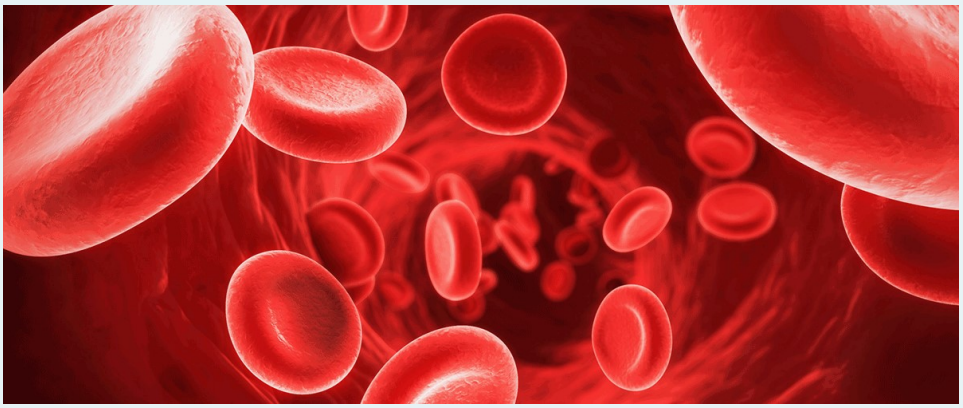
```
JX503079.1
M93259.1
M93258.1
JX503077.1
JX503078.1
JX503080.1
JX503073.1
JX503074.1
JX503082.1
JX503071.1
JN571034.1
JX503072.1
JX503075.1
JX503081.1
JX503076.1
JX503083.1
AY781127.1
AY781125.1
AY781128.1
AY781126.1
ON245430.1
ON245428.1
MK457954.1
ON245429.1
ON245427.1
ON245431.1
```

| | |
|---|---|
| **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;
```

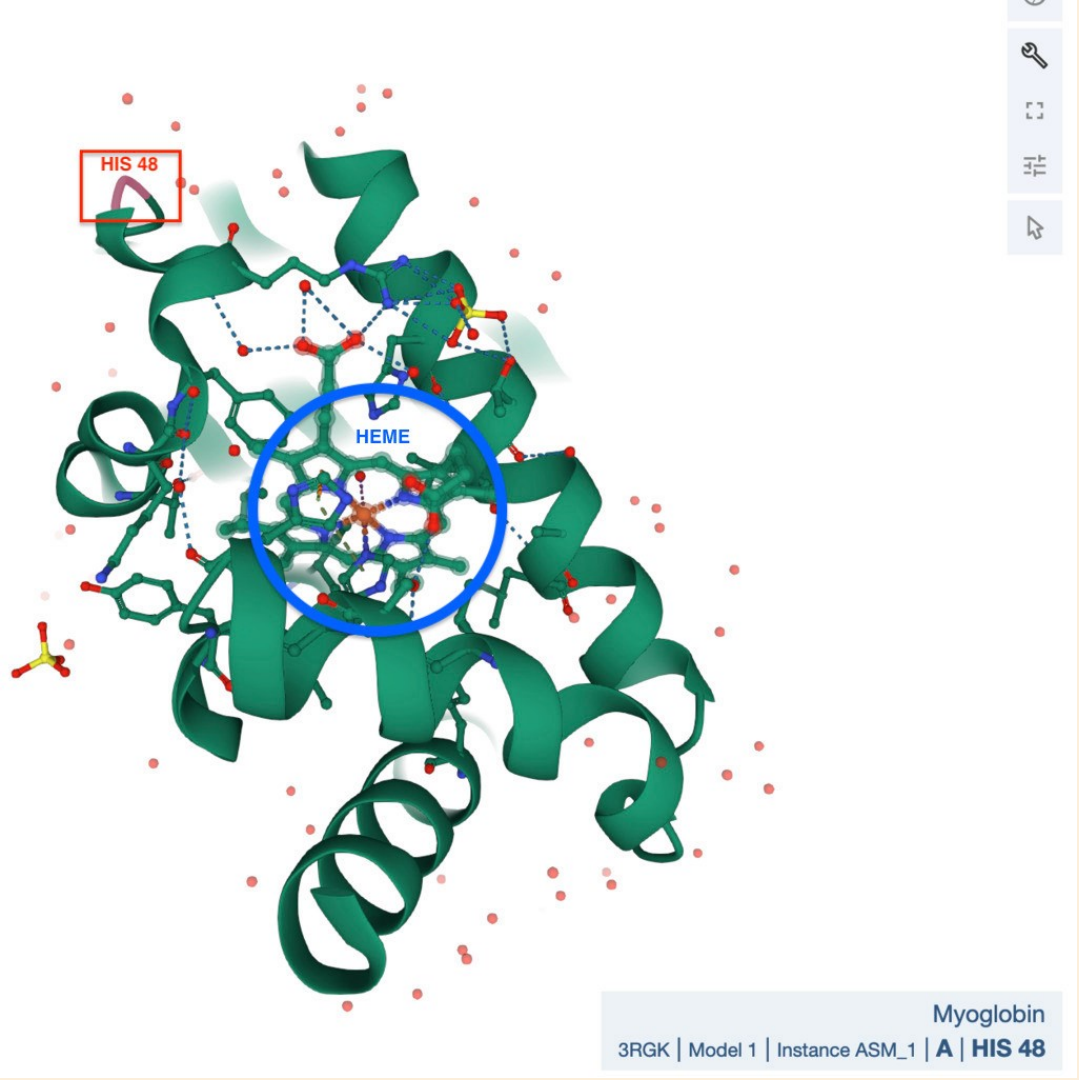
Only one chain "A"

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

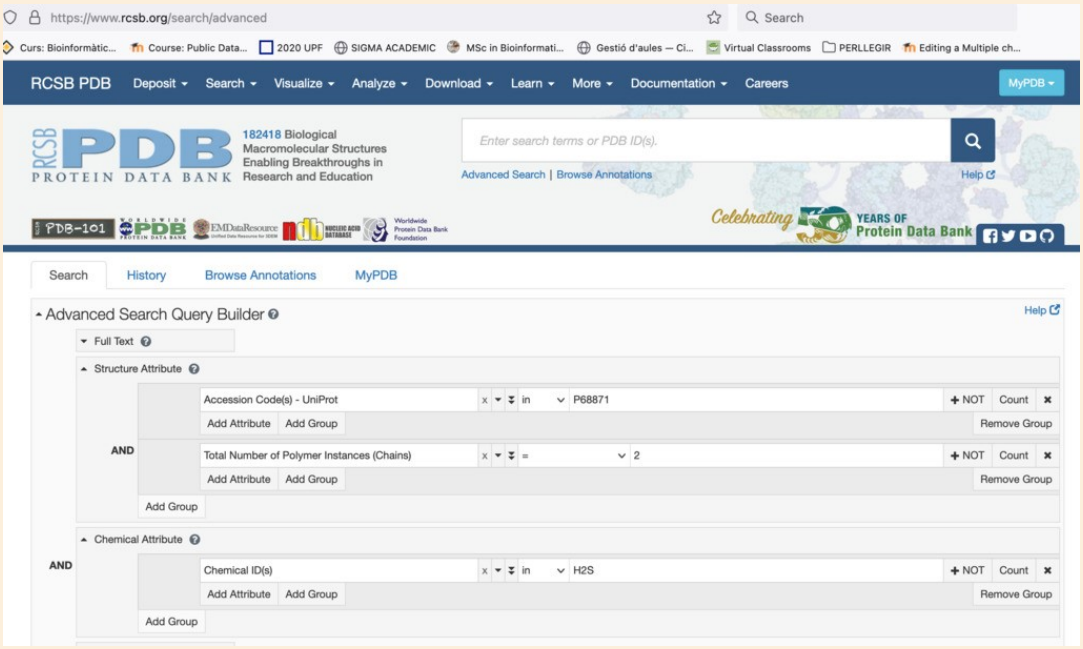
Mark 2.00 out of 2.00

**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 ✔ .

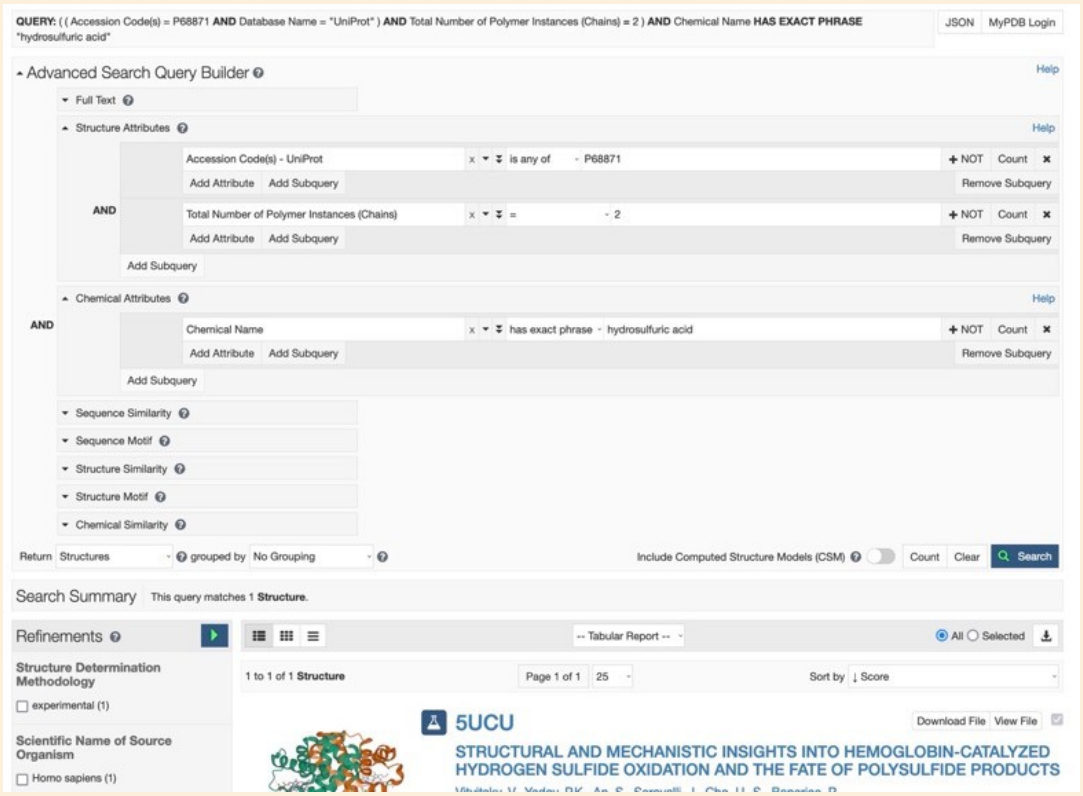
From this list, what is the accession code at RCSB PDB for a structure containing only two chains and with hydrosulfuric acid (H2S) as a ligand (it is a chemical) ? 5UCU ✔

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

At RCSB PDB site, advanced search:




or




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.

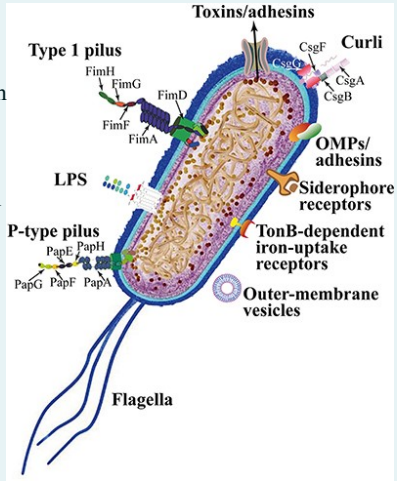
| | |
|---|---|
| **Started on** | Saturday, 21 October 2023, 8:31 AM |
| **State** | Finished |
| **Completed on** | Monday, 23 October 2023, 11:59 PM |
| **Time taken** | 2 days 15 hours |
| **Marks** | 9.50/12.00 |
| **Grade** | **7.92** out of 10.00 (**79.17**%) |

**Question 1**

Correct

Mark 1.00 out of 1.00

*Escherichia coli* is a significant cause of diarrheal illness and mortality worldwide each year, especially among children in developing countries. Pathogenic *E. coli* strains produce diverse virulence factors that allow them to overcome or subvert host defenses and to colonize, injure, and invade host cells or tissues. Today the existence of thousands of *E. coli* genomes in public databases has opened new perspectives for identification of novel virulence factors that could act as targets for treatment of infectious diseases. However, around 25% of the open reading frames (ORFs) in *E. coli* genomes remain as hypothetical genes without known function because they codify for uncharacterized proteins.




One of those ORF is the one that codifies for the uncharacterized protein YeeJ (UniProt P76347) in the *E. coli* K12 genome. Here, to better annotate function for this protein, we will combine similarity searches using protein databases with searches at pattern, domain and family databases.

**Question 1.** What can you say about the type of evidence that supports the existence of the protein P76347?

Select one:

- a. The existence of this protein is probable because clear orthologs exist in closely related species. ✔
- b. There are clear experimental evidences for the existence of this protein.
- c. The existence of this protein is unsure.
- d. Don't know/no answer (without penalty).
- e. Expression data indicate the existence of a transcript for this protein.

https://www.uniprot.org/uniprotkb/P76347/entry

https://www.uniprot.org/help/protein_existence

The value **'Protein inferred by homology'** indicates that the existence of a protein is probable because clear orthologs exist in closely related species.

The correct answer is: The existence of this protein is probable because clear orthologs exist in closely related species.

**Question 2**

Correct

Mark 1.00 out of 1.00

**Question 2.** In what biological process does this protein seem to participate?

Select one:

- a. Don't know/no answer (without penalty).
- b. Amino acid transporter
- c. Protein kinase signal transduction
- d. Cell adhesion ✔
- e. DNA replication

https://www.uniprot.org/uniprotkb/P76347/entry#function

https://www.ebi.ac.uk/QuickGO/annotations?geneProductId=P76347

The correct answer is: Cell adhesion

**Question 3**

Correct

Mark 2.00 out of 2.00

**Question 3.** With which nucleotide entry at the NCBI RefSeq database has this UniProt protein cross-reference? (In case of multiple possibilities, use the oldest one) NC_000913 ✔ (Indicate the refseq accession number)

What does this refseq entry represent? An annotated complete bacterial genome. ✔

https://www.uniprot.org/uniprotkb/P76347/entry#sequences

NC_000913.3 and NZ_LN832404.1 are Refseq nuccore entries linked by this Uniprot accession.

NC_000913.3 is the oldest (2006 vs 2014, you can see it in the reference paper)

NC_000913.3 is nucleotide NP_416485.4 is protein.

NC_000913.3 is complete since it only contains one contig

it is annotated: Customize view -> customize-> all features

**Question 4**

Correct

Mark 1.00 out of 1.00

Question 4. In other *Escherichia coli* strains different to K12, the function of this protein has been annotated as:

Select one:

- a. VacA-like protein
- b. Inverse autotransporter adhesin YeeJ / Bacterial Ig-like domain family protein ✔
- c. Agglomeration and penetration protein
- d. AIDA-1 autotransporter adhesion protein
- e. Don't know/no answer (without penalty)

We can find it in similar proteins

https://www.uniprot.org/uniprotkb/P76347/entry#similar_proteins

The correct answer is: Inverse autotransporter adhesin YeeJ / Bacterial Ig-like domain family protein

**Question 5**

Partially correct

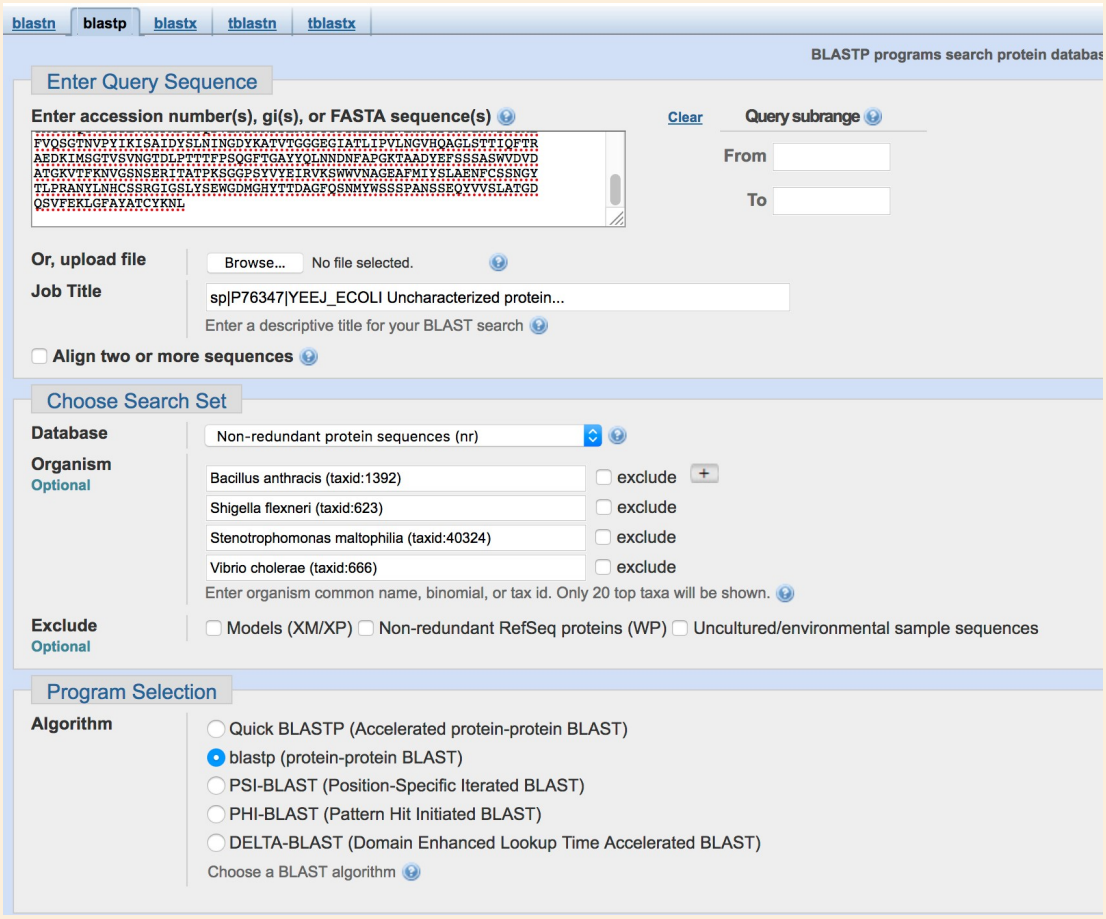
Mark 0.75 out of 2.00

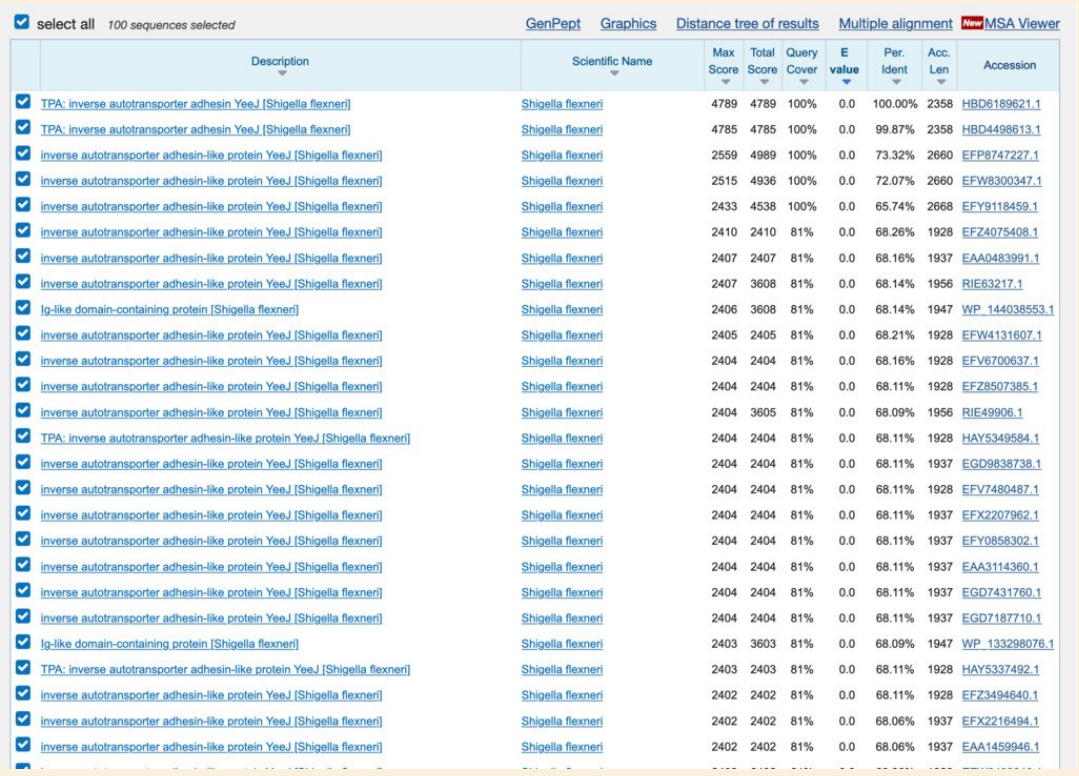
**Question 5.** Based exclusively on our protein sequence (UniProt P76347), which of these bacterial pathogens seems to be the closest to *E. coli* from an evolutionary point of view? Shigella flexneri ✔

What function has been assigned to the homologous protein in the selected closest species? Bacterial motility ✖

In NCBI:

Blast vs nr only for those organisms.



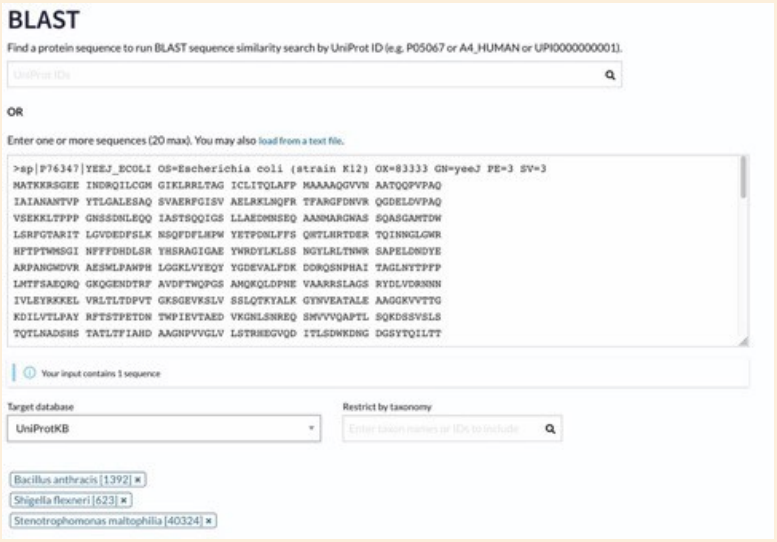



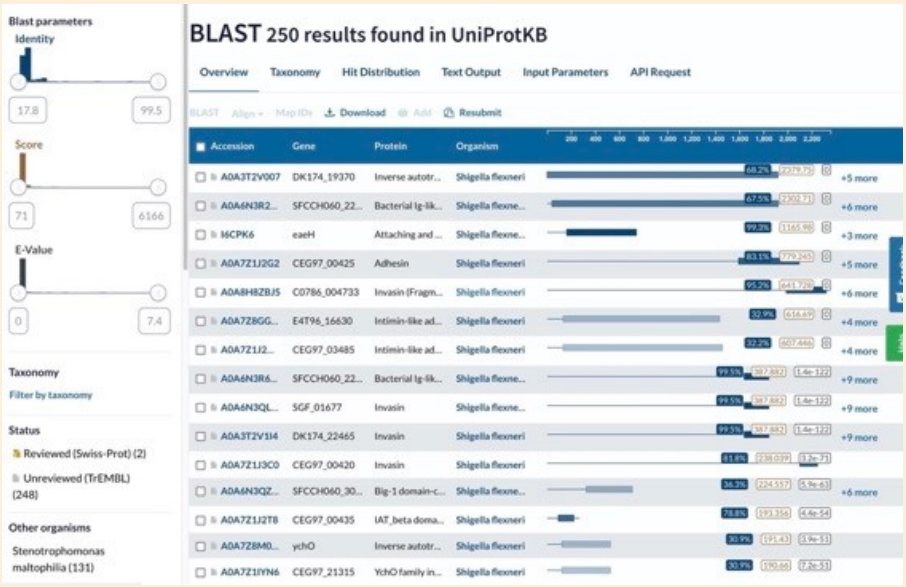
select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

| Description | Scientific Name | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession |
|---|---|---|---|---|---|---|---|---|
| TPA: inverse autotransporter adhesin YeeJ [Shigella flexneri] | Shigella flexneri | 4789 | 4789 | 100% | 0.0 | 100.00% | 2358 | HBD6189621.1 |
| TPA: inverse autotransporter adhesin YeeJ [Shigella flexneri] | Shigella flexneri | 4785 | 4785 | 100% | 0.0 | 99.87% | 2358 | HBD4498613.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2559 | 4989 | 100% | 0.0 | 73.32% | 2660 | EFP8747227.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2515 | 4936 | 100% | 0.0 | 72.07% | 2660 | EFW8300347.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2433 | 4538 | 100% | 0.0 | 65.74% | 2668 | EFY9118459.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2410 | 2410 | 81% | 0.0 | 68.26% | 1928 | EFZ4075408.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2407 | 2407 | 81% | 0.0 | 68.16% | 1937 | EAA0483991.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2407 | 3608 | 81% | 0.0 | 68.14% | 1956 | RIE63217.1 |
| Ig-like domain-containing protein [Shigella flexneri] | Shigella flexneri | 2406 | 3608 | 81% | 0.0 | 68.14% | 1947 | WP_144038553.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2405 | 2405 | 81% | 0.0 | 68.21% | 1928 | EFW4131607.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.16% | 1928 | EFV6700637.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.11% | 1928 | EFZ8507385.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 3605 | 81% | 0.0 | 68.09% | 1956 | RIE49906.1 |
| TPA: inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.11% | 1928 | HAY5349584.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.11% | 1937 | EGD9838738.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.11% | 1928 | EFV7480487.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.11% | 1937 | EFX2207962.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.11% | 1937 | EFY0858302.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.11% | 1937 | EAA3114360.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.11% | 1937 | EGD7431760.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2404 | 2404 | 81% | 0.0 | 68.11% | 1937 | EGD7187710.1 |
| Ig-like domain-containing protein [Shigella flexneri] | Shigella flexneri | 2403 | 3603 | 81% | 0.0 | 68.09% | 1947 | WP_133298076.1 |
| TPA: inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2403 | 2403 | 81% | 0.0 | 68.11% | 1928 | HAY5337492.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2402 | 2402 | 81% | 0.0 | 68.11% | 1928 | EFZ3494640.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2402 | 2402 | 81% | 0.0 | 68.06% | 1937 | EFX2216494.1 |
| inverse autotransporter adhesin-like protein YeeJ [Shigella flexneri] | Shigella flexneri | 2402 | 2402 | 81% | 0.0 | 68.06% | 1937 | EAA1459946.1 |

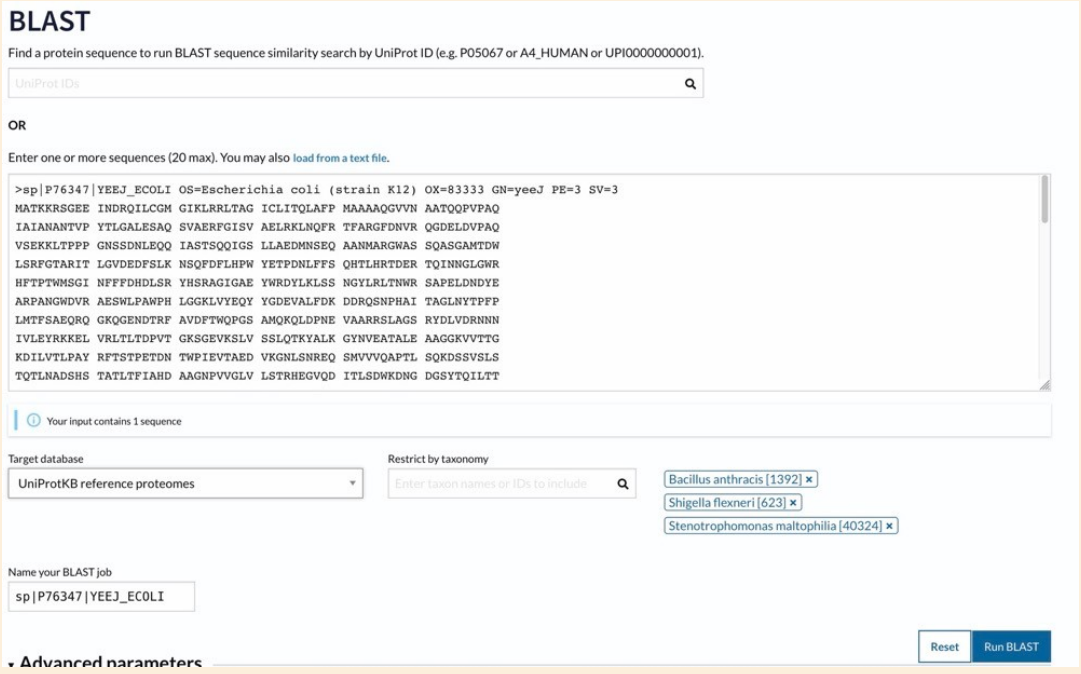
In Uniprot (Slower)

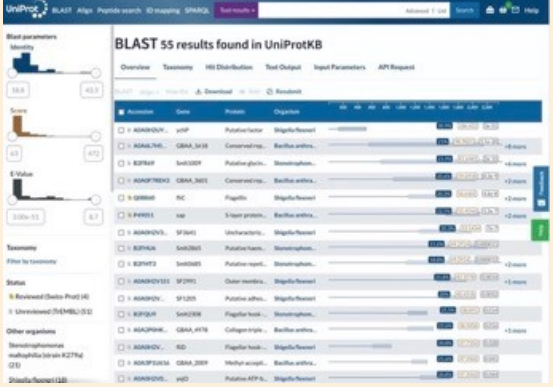
Blast vs UniprotKB, only for those organisms.




Don't use the "reference proteomes" because if your protein is not in one of them, you will end with the wrong conclusions:






If you use uniprot blast vs UniprotKB we just click on the UP accession code of the first result (https://www.uniprot.org/uniprotkb/A0A3T2V007/entry ) and we see in in Function->GO Annotations->Biological process: Cell adhesion

if you use ncbi blast

We get the best 40 NCBI IDs from the last column in the the ncbi blast results (Download -> csv, open with excel, copy the accession column) and map them to uniprot:
https://www.uniprot.org/id-mapping

From EMBL/GeneBank/DDBJ CDS to UniprotKB

and we get :

https://www.uniprot.org/uniprotkb/A0A3T2V007/entry mapped from EAA0483991.1 (the one previously mentioned with the uniprotKB blast search)

in the function section we see a GO code for "cell adhesion"

**Question 6**

Correct

Mark 1.00 out of 1.00

**Question 6.** What can we learn from our protein by searching at the PROSITE database?

Select one:

- a. It contains one sequence fragment (domain) seems to play a role in a recombinational process of DNA repair. And several big1 domains
- b. It contains several sequence fragments (domains) seem to play a role in a recombinational process of DNA repair.
- c. It contains one sequence fragment (domain) that recognizes polysaccharides (carbohydrates ) in the bacterial cell wall. And several big1 domains ✔
- d. It contains several sequence fragments (domains) that recognize polysaccharides in the bacterial cell wall.
- e. Don't know/no answer (without penalty).

https://prosite.expasy.org/

and we paste P76347 sequence or this Uniprot accession and we pres scan

The correct answer is: It contains one sequence fragment (domain) that recognizes polysaccharides (carbohydrates ) in the bacterial cell wall. And several big1 domains

**Question 7**

Correct

Mark 1.00 out of 1.00

**Question 7.** Which of these Pfam domain architectures does our protein exhibit?

Select one:

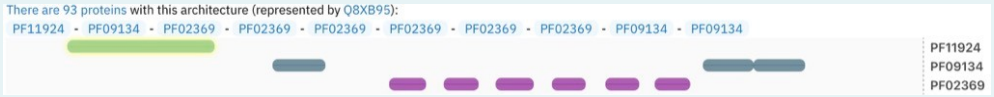
a.


b.


c. ✔


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

e.




From Uniprot jumpt to pfam.

you will be redirected to interpro

https://www.ebi.ac.uk/interpro/protein/UniProt/P76347/

pay atantion only to the pfam domains (PF...)

you will see one PF11924 folleowed by one PF09134 followed by thirteen PF02369

The correct answer is:


**Question 8**

Correct

Mark 1.00 out of 1.00

**Question 8.** Analyzing our protein at InterPro it seems that it contains much more domains or important protein signatures when compared with the result of Pfam. Why?

Select one:

a. InterPro is a primary database and contains information for all known protein structures.

b. InterPro consists of a collection of annotated multiple sequence alignment models available exclusively as position-specific score matrices (PSSMs).

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

d. InterPro is a protein annotation resource for protein domain classifications based exclusively on hidden Markov models (HMMs).

e. InterPro is an integrated database that combines protein signatures from a number of member databases. ✔

https://www.ebi.ac.uk/interpro/about/

The correct answer is: InterPro is an integrated database that combines protein signatures from a number of member databases.

**Question 9**

Correct

Mark 1.00 out of 1.00

**Question 9.** Based on information taken from derived protein databases (Interpro or Pfam) , what is the most likely function of the repetitive domains in our protein?

Select one:

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

b. This repetitive unit is a phospho-relay system usually found in regulator proteins. This signal transduction system enables bacterium to sense, respond, and adapt to a wide range of environments, stressors, and growth conditions.

c. These repetitive units have the capacity to interact with hyaluronic acid, which promotes the stability of the extra-cellular matrix.

d. Big-1 proteins are surface-expressed proteins that mediate mammalian host cell invasion or attachment. The tandem of Ig-like domains appears to form a rod to link the bacterial outer membrane anchor to the C-terminal lectin-like domain to interact with their receptors in the host cell membrane ✔

e. This multi-domain adopts a classic alpha/beta fold and contains an unusual metal ion coordination site at its surface. It has been suggested that this site represents a general metal ion-dependent adhesion site for binding protein ligands.

The repetitive domains are the "Big-1"

Big-1 domain description in interpro

https://www.ebi.ac.uk/interpro/entry/InterPro/IPR003344/

Take a look also at the description in the DBs that originated this interpro entry:

http://smart.embl-heidelberg.de/smart/do_annotation.pl?DOMAIN=SM00634

https://prosite.expasy.org/PDOC51127

Se that it is almost the same.

The correct answer is: Big-1 proteins are surface-expressed proteins that mediate mammalian host cell invasion or attachment. The tandem of Ig-like domains appears to form a rod to link the bacterial outer membrane anchor to the C-terminal lectin-like domain to interact with their receptors in the host cell membrane

**Question 10**

Incorrect

Mark -0.25 out of 1.00

**Question 10.** We finally want to know if our protein, or a protein similar to this, has a 3D structure experimentally solved.

Select one:

- a. Don't know/no answer (without penalty).
- b. Yes, protein P76347 has its 3D structure already solved. ✖
- c. There is no 3D structure for this protein. However the structures of all its known domains are known in other proteins or organisms.
- d. Yes, there is a 3D structure for this protein but only for the whole region that contains the repetitive domain Big-1.
- e. No, there is no 3D structure for this protein or for any of its domains in other proteins or close organisms.

We can see in interpro that this protein has this 3 pfam domains, and here you can see that these domains' structures have been solved in other proteins:

https://www.ebi.ac.uk/interpro/entry/pfam/PF02369/structure/PDB/#table

https://www.ebi.ac.uk/interpro/entry/pfam/PF09134/structure/PDB/#table

https://www.ebi.ac.uk/interpro/entry/pfam/PF11924/structure/PDB/#table

We can also see that it contains 4 CATH structural domains (Gene3D). Since CATH is a DB created from the the structure of proteins experimentally solved we can see other proteins with similar domains

http://www.cathdb.info/version/latest/superfamily/2.60.40.10

http://www.cathdb.info/version/latest/superfamily/3.10.100.10

http://www.cathdb.info/version/latest/superfamily/2.40.160.160

http://www.cathdb.info/version/latest/superfamily/2.60.40.1080

The correct answer is: There is no 3D structure for this protein. However the structures of all its known domains are known in other proteins or organisms.

| | |
|---|---|
| **Started on** | Saturday, 28 October 2023, 8:04 PM |
| **State** | Finished |
| **Completed on** | Sunday, 29 October 2023, 9:14 PM |
| **Time taken** | 1 day 2 hours |
| **Marks** | 9.00/9.00 |
| **Grade** | **10.00** out of 10.00 (**100**%) |

**Question 1**

Correct

Mark 9.00 out of 9.00

1 What is the **Entrez Gene ID** of *BRCA1* in the human genome?

- (•) 672 ✔
- ( ) ENSG00000012048
- ( ) NC_000017.11
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: 672

2 Where is this gene located? Chromosome [17] ✔ (answer with either a number, X, or Y, only), on the [Reverse] ✔ strand.

3 How many transcripts does *BRCA1* encode according to Ensembl?

- ( ) 1 transcript
- ( ) <5 transcripts
- (•) >5 transcripts ✔
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: >5 transcripts

4 Which are the closest **protein-coding genes** to *BRCA1* according to Ensembl? [RND2] ✔ downstream (closest to its 3' end) and [NBR1] ✔ upstream (closest to its 5' end) (answer with the corresponding gene symbols).

5 Is *BRCA1* involved in any known phenotype?

- ( ) Skin color
- ( ) Parkinson
- (•) Various types of cancer ✔
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: Various types of cancer

6 Browse the Variant table and find SNP rs1800704. Which is the most abundant allele?

- (•) C ✔
- ( ) T
- ( ) It depends on the population
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: C

7 What is the main objective of the 1000 Genomes Project?

- ( ) To characterize human genes
- (•) To study human genetic diversity in diverse populations ✔
- ( ) To compare dozens of sequenced mammalian genomes
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: To study human genetic diversity in diverse populations

| | |
|---|---|
| **Started on** | Friday, 3 November 2023, 12:44 PM |
| **State** | Finished |
| **Completed on** | Sunday, 5 November 2023, 6:21 PM |
| **Time taken** | 2 days 5 hours |
| **Grade** | **10.00** out of 10.00 (**100**%) |

**Question 1**

Correct

Mark 10.00 out of 10.00

1 The aim of functional genomics is to explore the function of the whole organism's genes and gene products.

True ✔

2 Functional genomic experiments generate… large amounts of data ✔

3 Choose the appropriate description for the EMBL-EBI ArrayExpress resource:

- A database providing information about gene expression in different biological conditions
- A database of functional genomics experiments storing information about experiments and associated data ✔
- Blank answer

Mark 1.00 out of 1.00

The correct answer is: A database of functional genomics experiments storing information about experiments and associated data

4 Choose the appropriate description for the EMBL-EBI Expression Atlas resource:

- A database providing information about gene expression in different biological conditions ✔
- A database of functional genomics experiments storing information about experiments and associated data
- Blank answer

Mark 1.00 out of 1.00

The correct answer is: A database providing information about gene expression in different biological conditions

5 You should use the ArrayExpress…

- to submit histology experiments
- to submit microarray experiments
- to submit high-throughput sequencing experiments
- (b) and (c) are correct ✔
- Blank answer

Mark 1.00 out of 1.00

The correct answer is: (b) and (c) are correct

6 You should use Expression Atlas…

- to retrieve sample annotation and raw data files for an experiment accession number you came across in a publication
- to search for gene expression in healthy, untreated human tissues
- to search for gene-condition information
- (b) and (c) are correct ✔
- Blank answer

Mark 1.00 out of 1.00

The correct answer is: (b) and (c) are correct

7 Search for the gene *BRCA1* in humans in the Expression Atlas. What effect on the *BRCA1* gene expression has the treatment of A375 melanoma cells with Nutlin 5 micromolar compared to DMSO 0.065 percent, in GL2?

- The gene *BRCA1* is upregulated
- The gene *BRCA1* is downregulated ✔
- The expression of the gene *BRCA1* is not altered
- Blank answer

Mark 1.00 out of 1.00

The correct answer is: The gene *BRCA1* is downregulated

8 Which GO term/s is/are enriched in the previous treatment?

- GO:0016363 – nuclear matrix
- GO:0006396 – RNA processing

- GO:0006310 – DNA recombination
- None of the above✔
- Blank answer

Mark 1.00 out of 1.00

The correct answer is: None of the above

9 Go back to the Expression Atlas main page and search for the biological condition "*drought environment*" in the organism "*Arabidopsis thaliana*". What is the effect of drought on the *LEA7* gene expression?

- The gene *LEA7* is upregulated✔
- The gene *LEA7* is downregulated
- The expression of the gene *LEA7* is not altered
- Blank answer

Mark 1.00 out of 1.00

The correct answer is: The gene *LEA7* is upregulated

10 Lets assume you submit the complete list of genes from the previous result to the Gene Ontology Enrichment Analysis tool and find "*response to stress*" as one of the most significantly enriched "*biological process*" term. Choose the correct interpretation of this result:

- All the genes in the list are involved in response to stress in *Arabidopsis thaliana*
- Response to stress is statistically overrepresented as biological process term associated to the genes in the list ✔
- The genes in the list are only expressed under stress conditions in *Arabidopsis thaliana*
- Blank answer

Mark 1.00 out of 1.00

The correct answer is: Response to stress is statistically overrepresented as biological process term associated to the genes in the list

| | |
|---|---|
| **Started on** | Wednesday, 8 November 2023, 6:05 PM |
| **State** | Finished |
| **Completed on** | Sunday, 12 November 2023, 3:36 PM |
| **Time taken** | 3 days 21 hours |
| **Marks** | 50.00/50.00 |
| **Grade** | **10.00** out of 10.00 (**100**%) |

**Question 1**

Correct

Mark 50.00 out of 50.00

1 Protein interaction networks must be of directional nature. False ✔

2 Networks that only have undirected edges can be represented with symmetrical adjacency matrices. True ✔

3 The main topological characteristics of protein interaction networks as we currently understand them are (select all that apply):

- No ✔ All nodes have a roughly similar number of interactions.
- No ✔ The network structure is vulnerable to random node mutations.
- Yes ✔ They have a small diameter.
- Yes ✔ There are regions within the network that more connected internally than to the rest of the network.
- Yes ✔ Edges represent physical relationships between proteins.

4 Select the tools that can be used for topological network analysis (select all that apply):

- No ✔ g:Profiler
- Yes ✔ Gephi
- Yes ✔ iGraph
- Yes ✔ Cytoscape
- No ✔ BiNGO

5 In a protein interaction network, an internally inter-connected module …

- … can only be found in networks with directed edges.
- … can represent a protein complex or a group of proteins with a defined biological function. ✔
- … is also known as a hub.
- … only has nodes with a low number of interactions.
- … changes its intrinsic functional properties when placed in a different context.
- Blank answer

Mark 5.00 out of 5.00

The correct answer is: … can represent a protein complex or a group of proteins with a defined biological function.

6 The degree of the nodes in a network can be considered as a local measure of centrality. True ✔

7 Annotation enrichment analysis is often limited in its usefulness due to the complexity and detail of annotation associated with large gene/protein sets. True ✔

8 The degree of a node can be defined as:

- The number of edges that connect to/from a node. ✔
- The smallest number of steps between two nodes.
- The number of edges in a network.
- The number of nodes in a network.
- Blank answer

Mark 5.00 out of 5.00

The correct answer is: The number of edges that connect to/from a node.

9 Transitivity...

- ... measures the number of hubs in a network.
- ... measures the shortest distance between two nodes.
- ... gives an estimation on how important that node/edge is for the connectivity of the network.
- ... measures the tendency of the nodes to cluster together. ✔

Blank answer

Mark 5.00 out of 5.00

The correct answer is: ... measures the tendency of the nodes to cluster together.

10 Annotation enrichment analysis uses gene/protein annotations provided by knowledge-bases such as Gene Ontology (GO) or Reactome to infer which annotations are over-represented in the network. True ✔

| | |
|---|---|
| **Started on** | Saturday, 18 November 2023, 7:20 PM |
| **State** | Finished |
| **Completed on** | Friday, 24 November 2023, 8:45 PM |
| **Time taken** | 6 days 1 hour |
| **Marks** | 63.00/63.00 |
| **Grade** | **10.00** out of 10.00 (**100**%) |

**Question 1**

Correct

Mark 63.00 out of 63.00

1 Which phase of the drug discovery timeline can Open Targets optimize and accelerate?

- (●) 1. Discovery ✔
- ( ) 2. Development
- ( ) 3. Delivery
- ( ) Blank answer

Mark 7.00 out of 7.00

The correct answer is: 1. Discovery

2 What data sources can provide useful information for drug discovery?

- ( ) Gene, protein and metabolite expression
- ( ) SNPs and GWAS
- ( ) Reactions, interactions and pathways
- (●) All of the above ✔
- ( ) Blank answer

Mark 7.00 out of 7.00

The correct answer is: All of the above

3 Which Evidence sources are integrated into Open Targets? (tick all that apply)

- Yes ✔ Genetic associations
- Yes ✔ RNA expression
- Yes ✔ Somatic mutations
- Yes ✔ Pathways/Systems Biology
- No ✔ Association of fly models with disease
- Yes ✔ Known drugs
- Yes ✔ Published articles (text mining)

4 How many targets are associated with asthma according to Open Targets (See Associated Targets tab)?

4562 ✔

5 Explore the association of *PDE4D* with asthma further. What type of evidence exists for such an association? (tick all that apply)

- No ✔ Affected pathways
- No ✔ Animal models
- Yes ✔ Drugs
- Yes ✔ Genetic associations
- No ✔ RNA expression
- No ✔ Somatic mutations
- Yes ✔ Text mining

6 How many different drugs are in phase IV in the association of *PDE4D* with asthma? (Phase: IV) 2 ✔

7 How many diseases are associated with *PDE4D*? 338 ✔

8 How many "Respiratory or thoracic diseases" are associated with *PDE4D*? 23 ✔

9 After asthma, which is the disease showing the highest overall association score with PDE4D

acrodysostosis ✔

10 How many genetic associations does this 2nd ranked disease have with PDE4D, according to ClinVar?

24 ✔

| | |
|---|---|
| **Started on** | Saturday, 25 November 2023, 4:42 PM |
| **State** | Finished |
| **Completed on** | Saturday, 25 November 2023, 11:55 PM |
| **Time taken** | 7 hours 12 mins |
| **Marks** | 8.00/8.00 |
| **Grade** | **10.00** out of 10.00 (**100**%) |

**Question 1**

Correct

Mark 8.00 out of 8.00

1 Controlled vocabularies or ontologies

- (•) Provide a way to organize knowledge ✔
- ( ) Allow for unambiguous identification of molecules
- ( ) Are annotations that describe your experimental data; e.g., patient/environmental details, cell lines, instrumentation, etc.
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: Provide a way to organize knowledge

2 What is the name of a set of minimum data that is needed to be stored in a study in order to follow the quality standards?

- ( ) Controlled vocabulary
- (•) Information guideline ✔
- ( ) Schema
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: Information guideline

3 Which is the Gene Ontology (GO) term for “microtubule binding”? (please use the syntax GO:#######)

GO:0008017 ✔

4 Which of the following are GO terms in the “microtubule binding” hierarchy?

- (•) GO:0003674, molecular function ✔
- ( ) GO:0051959, dynein light intermediate chain binding
- ( ) GO:0009987, cellular process
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: GO:0003674, molecular function

5 Which of the following is FALSE regarding the use of identifiers and mapping tables?

- ( ) Identifiers allow for unambiguous identifications of molecules or conceptual entities
- (•) Different identifiers most of the times have a 1:1 relationship ✔
- ( ) Identifiers might disappear or get updated
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: Different identifiers most of the times have a 1:1 relationship

6 Translate the "GeneID" identifiers ("Genome Annotation Databases" category) from the file human_GeneIDs.txt to UniProtKB accessions using the UniProt ID mapping tool. How many GeneID identifiers were successfully mapped? 13 ✔ How many UniProtKB identifiers were retrieved? 34 ✔

7 Which of the following is NOT a good practice?

- ( ) Use open-source software
- ( ) Use ontology terms to annotate metadata
- (•) Define new data formats that fit your specific needs ✔
- ( ) Blank answer

Mark 1.00 out of 1.00

The correct answer is: Define new data formats that fit your specific needs