Started on	Tuesday, 5 December 2023, 3:02 PM
State	Finished
Completed on	Tuesday, 5 December 2023, 5:02 PM
Time taken	1 hour 59 mins
Grade	**6.87** out of 10.00 (**68.65**%)

Question 1

Correct

Mark 0.40 out of 0.40

Q 01/01. What NCBI Entrez is and what does it allow us to do? (yes-no questions)

YES ✔ It gives access to all NCBI databases.

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. It can be accessed as a web service through the e-utilities

Question 2

Partially correct

Mark 0.36 out of 0.60

Q 01/02. Which of the following NCBI entries have cross-references and can be linked (with others in this list)?

i) J01643.1

j)AAA24067.1

k)WP_000646078.1

m)WP_181201725.1

Select one:

- All are linked
- None are related
- i) and j) are related (GenBank nucleotide, GenBank CDS) k) and m) are unrelated to them ✔
- I don't answer the question
- i) j) and K) are related (GenBank nucleotide, GenBank CDS, identical Refseq protein) m) is unrelated to them

Nucleotide sequence J01643.1 encodes Protein entry AAA24067.1 which is identical (identical protein groups) to WP_000646078.1.

https://www.ncbi.nlm.nih.gov/ipg/AAA24067.1

WP_181201725.1 is also a LexA protein from an E. coli organism but its sequence is different and has no relation with the other three.

The correct answer is: i) j) and K) are related (GenBank nucleotide, GenBank CDS, identical Refseq protein) m) is unrelated to them

Question 3

Partially correct

Mark 0.33 out of 0.50

Q 02/01. 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.

U00096.1	An old version of a nuccore genbank entry ✔
ERX3647991	The entry contains a short reads obtained with an Illumina sequencer ✔
AM743169.1	The current version of a refseq entry ✖
NC_010943.1	The current version of a nuccore genbank entry for a circular DNA molecule ✖
AAGX00000000.1	The entry contains a whole genome shotgun sequence (WGS) ✔
AC275307.1	The entry contains a working draft sequence from a high-throughput genomic (HTG) approach ✔

Remember that the number after the "." is the version, which indicates a change in the sequence.

For example U00096.1 is an old version. U00096.3 is the current one.

The correct answer is: U00096.1 → An old version of a nuccore genbank entry, ERX3647991 → The entry contains a short reads obtained with an Illumina sequencer, AM743169.1 → The current version of a nuccore genbank entry for a circular DNA molecule, NC_010943.1 → The current version of a refseq entry, AAGX00000000.1 → The entry contains a whole genome shotgun sequence (WGS), AC275307.1 → The entry contains a working draft sequence from a high-throughput genomic (HTG) approach

Question 4

Incorrect

Mark 0.00 out of 0.50

Q 02/02. Localize all of the records in the NCBI nucleotide database containing rabbit (no subspecies) genomic sequences that are codifying for a protein and do not have a link to protein database. How many have you found? 4064 ✖

Write here your search strategy

rabbit[Organism] AND biomol_genomic[PROP] AND CDS[fkey] ✖

¿How many of these sequences come from RefSeq? 1409 ✖

Rabbit organism: (rabbit[PORGN], txid9986[porgn])

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

It have to codify for protein, so it must have a CDS :CDS[fkey]

It must not have a link to protein database so NOT nucleotide_protein[filter]

So several possiblities:

rabbit[porgn] AND biomol_genomic[PROP] AND CDS[fkey] NOT nucleotide_protein[filter]

rabbit[porgn] AND biomol_genomic[PROP] NOT nucleotide_protein[filter] AND CDS[fkey]

txid9986[porgn] AND biomol_genomic[PROP] AND CDS[fkey] NOT nucleotide_protein[filter]

txid9986[porgn] AND biomol_genomic[prop] AND CDS[fkey] NOT nucleotide_protein[filter]

"Oryctolagus cuniculus"[Primary Organism] AND biomol_genomic[Properties] AND CDS[Feature key] NOT nucleotide_protein[Filter]

txid9986[porgn] AND CDS[fkey] NOT nucleotide_protein[Filter] AND biomol_genomic[PROP]

rabbit[porgn] AND biomol_genomic[PROP] AND CDS[fkey] NOT nuccore_protein[filter]

rabbit[porgn] AND biomol_genomic[PROP] AND CDS[fkey] NOT nucleotide_protein[filter]

etc...

in the left margin you can see that 9 are in the RefSeq

Question 5

Correct

Mark 1.00 out of 1.00

Q 02/03

Find the NCBI entry containing the complete genome of Xylella fastidiosa Temecula1 *and the gene rpff*

AE009442.1 ✔

Wite the coordinates of this gene (from low number to high number) START: 499318 ✔ .. END: 500190 ✔ .

Get the fasta DNA sequence for this gene (only this gene, not the full genome) and paste here the header as it is generated by ncbi (do not include the "enter" at the end of the line) Remember: "fasta format" has some requirements in the header:

>AE009442.1:499318-500190 Xylella fastidiosa Temecula1, complete genome ✔

Taxid for Xylella fastidiosa Temecula1

https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=183190&lvl=3&lin=f&keep=1&srchmode=1&unlock

search:

txid183190[ORGN] AND rpff[Gene]

this is the result:

https://www.ncbi.nlm.nih.gov/nuccore/AE009442.1

NOTE: NC_004556.1 is also accepted (RefSeq version of AE009442.1) although the rpff gene is not explicitly annotated in here so you will end up needing AE009442.1

Ctrl-F and search for rpff to find the gene and its coordinates

click on the word "gene"

click on the word FASTA (display: FASTA) bottom right corner

https://www.ncbi.nlm.nih.gov/nuccore/AE009442.1?from=499318&to=500190&report=fasta&strand=2

you could also select de region of the gene:

https://www.ncbi.nlm.nih.gov/nuccore/AE009442.1?from=499318&to=500190

click on change the format to "Fasta" (upper left corner where it says "FASTA")

https://www.ncbi.nlm.nih.gov/nuccore/AE009442.1?report=fasta&from=499318&to=500190

NOTE: START 500190 and END 499318 is wrong! it is clearly stated in the question that is from low number to high number.

The following fasta headers could be obtained depending on how you reach the sequence, all are considered correct:

>AE009442.1:499318-500190 Xylella fastidiosa Temecula1, complete genome

>NC_004556.1:499318-500190 Xylella fastidiosa Temecula1, complete sequence

>AE009442.1:c500190-499318 Xylella fastidiosa Temecula1, complete genome

>NC_004556.1:c500190-499318 Xylella fastidiosa Temecula1, complete sequence

>AE009442.1:499318-500190 Xylella fastidiosa Temecula1, complete genome

I also accept headers like this:

>lcl|AE009442.1_gene_407 [gene=rpfF] [locus_tag=PD_0407] [location=complement(499318..500190)] [gbkey=Gene]

coming from the coding sequences file

If ">" is not included in your answer, it is not correct, since fasta format must start with it. (I told you in the question: *Remember: "fasta format" has some requirements in the header*)

Including the sequence is also incorrect (we even commented that the spanish translation for header is "cabecera")

Question 6

Partially correct

Mark 1.67 out of 2.00

Q 03_04/01

WHEN WRITING INTEGERS USE ONLY NUMERICAL CHARACTERS (NOT "," or " ")

Perform a search at UniProtKB main page for proteins in the mouse reference proteome (proteome id UP000000589) with the following criteria:

i) Belonging to the Swiss-Prot database
ii) With clear experimental evidences for its existence at the protein level.
iii) With cross-reference to PDB (available structure)
iv) With subcellular location in the Mitochondrion
v) Containing a binding site for a group termed "heme".

How many proteins in the mouse proteome id meet criteria i) and ii) at the same time? 13102 ✔

Indicate the UniProt AC (Accession) for the shortest protein in the mouse proteome id that meets all criteria (i to v) at the same time. P00015 ✔

For the protein previously selected. What is the accession code for the database containing the 3D structure experimentally solved for this protein? 2AIU ✔ What method was used to solve this structure? X-ray Crystallography ✔

Verify in this 3D structure that the Heme group (HEM) has been detected as a ligand for this protein. Which of the following amino acids are thought to participate in the binding of this ligand?

Histidine (HIS) at position 19 ✔

Note: For information related to 3D structures I recommend the link to the PDB through RCSB PDB And remember that information can also be found in the pdb file.

Which one is the most frequent amino acid in the 12th position of this protein pfam family profile (three letter code, all capitals) - ✖

The mouse (*Mus musculus*) reference proteome at UniProt had 55311 proteins on Dec 7th 2022.

https://www.uniprot.org/proteomes/UP000000589

(proteome:UP000000589) AND (reviewed:true) AND (existence:1)

13102

(proteome:UP000000589) AND (reviewed:true) AND (existence:1) AND (structure_3d:true) AND (cc_scl_term:SL-0173) AND (ft_binding:"CHEBI:30413")

P00015

https://www.rcsb.org/structure/2AIU

In the PDB file: (See Question #6 Quiz Topic 3)

https://files.rcsb.org/view/2AIU.pdb

```
::::::::::::::::::::::::::::::::::::::::::::::::::
REMARK 800 SITE_IDENTIFIER: AC2
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE PO4 A 202
REMARK 800 SITE_IDENTIFIER: AC3
REMARK 800 EVIDENCE_CODE: SOFTWARE
REMARK 800 SITE_DESCRIPTION: BINDING SITE FOR RESIDUE HEM A 200
DBREF  2AIU A    2   105  UNP    P00015   CYC2_MOUSE       1    104
:::::::::::::::::::::::::::::::::
HET    PO4  A 201       5
HET    PO4  A 202       5
HET    HEM  A 200      43
HETNAM     PO4 PHOSPHATE ION
HETNAM     HEM PROTOPORPHYRIN IX CONTAINING FE
HETSYN     HEM HEME
FORMUL   2  PO4    2(O4 P 3-)
FORMUL   4  HEM    C34 H32 FE N4 O4
FORMUL   5  HOH   *224(H2 O)
::::::::::::::::::::::::::::::::::::
SITE     1 AC1  8 LYS A  14  SER A  48  LYS A  74  HOH A 304
SITE     2 AC1  8 HOH A 361  HOH A 369  HOH A 477  HOH A 483
SITE     1 AC2 10 LYS A  54  ASN A  55  GLY A  57  HOH A 339
SITE     2 AC2 10 HOH A 360  HOH A 416  HOH A 446  HOH A 466
SITE     3 AC2 10 HOH A 472  HOH A 506
SITE     1 AC3 22 LYS A  14  CYS A  15  GLN A  17  CYS A  18
SITE     2 AC3 22 HIS A  19  THR A  29  PRO A  31  THR A  41
SITE     3 AC3 22 GLY A  42  TYR A  49  THR A  50  ASN A  53
SITE     4 AC3 22 LYS A  56  TRP A  60  TYR A  68  THR A  79
SITE     5 AC3 22 LYS A  80  MET A  81  ILE A  82  PHE A  83
SITE     6 AC3 22 LEU A  99  HOH A 313
::::::::::::::::::::::::::::::::::::::::::::::::
```

P00015 pfam entry:
http://pfam.xfam.org/protein/P00015
I has the domain Cytochrom_C (PF00034)
In the logo for this family profile:
https://www.ebi.ac.uk/interpro/entry/pfam/PF00034/logo/
we can see a big "C" in the 12th position, so the most frequent AA in this position is cysteine, in three letter code is CYS)

Question 7

Correct

Mark 1.50 out of 1.50

Marfan syndrome is an inherited disorder that affects connective tissue, the fibers that support and hold organs and other structures in the body. Marfan syndrome most commonly affects the heart, eyes, blood vessels, and skeleton. According to the Open Targets Platform:

1. How many targets are associated to this disease? 307 ✔ **(0.3 pts)** *(write a number)*
2. How many targets associated to this disease are supported by genetic associations? 18 ✔ **(0.3 pts)** *(write a number)*
3. How many targets associated to this disease have the highest overall association score of >= 0.5? 3 ✔ **(0.3 pts)** *(write a number)*
4. What is the symbol of the top, most-associated target? FBN1 ✔ **(0.3 pts)** *(write a gene symbol)*
5. Regarding the top, most-associated target: Which evidences support the association of this target with Marfan syndrome? Select all that apply. **(0.04286 / -0.04286 pts each)**
   - Yes ✔ Genetic associations
   - No ✔ Somatic mutation
   - No ✔ Drugs
   - No ✔ Pathways and Systems biology
   - Yes ✔ Text mining
   - No ✔ RNA expression
   - Yes ✔ Animal models

Question 8

Partially correct

Mark 1.36 out of 2.00

According to Ensembl, and regarding the **top, most-associated target** you found in the previous question:

1. In which chromosome is the gene located? 15 ✔ **(0.25 pts)** *(Enter either a number, X or Y only)*
2. How many different transcripts does the gene encode? 13 ✔ transcripts **(0.125 pts)**, of which 3 ✔ are classified as "nonsense mediated decay" **(0.125 pts)**
3. How many different proteins does the gene encode? 2 ✔ **(0.25 pts)** *(excluding nonsense mediated decay)*
4. What are the closest protein-coding genes? CEP152 ✔ upstream **(0.125 pts)**, and DUT ✔ downstream **(0.125 pts)** *(enter the symbols of the corresponding genes)*
5. How many paralogues does the gene have in humans? 2 ✔ paralogues **(0.125 pts)**. Are they located on the same chromosome? **(0.125 / -0.125 pts)** No ✔
6. In which biological processes is the gene involved? **(0.25 / -0.0625 pts)**
   - ○ DNA repair
   - ○ Lipid metabolism
   - ◉ Skeletal system development ✔
   - ○ Protein transport
   - ○ Blank answer

   Mark 2.00 out of 2.00

   The correct answer is: Skeletal system development

According to Expression Atlas:

1. In which tissues is this gene mostly expressed according to GTEx? **(0.25 / -0.0625 pts)**
   - ○ Liver
   - ◉ Lung ✖
   - ○ Aorta
   - ○ Blood
   - ○ Blank answer

   Mark -0.50 out of 2.00

   The correct answer is: Aorta

2. Is this gene differentially expressed in patients with Klinefelter's Syndrome (fibroblasts) compared to controls (embryonic stem cells)? **(0.25 / -0.0833 pts)**
   - ○ Yes, the gene is upregulated
   - ◉ Yes, the gene is downregulated ✖
   - ○ No, this phenotype does not associate with the gene
   - ○ Blank answer

   Mark -0.66 out of 2.00

   The correct answer is: Yes, the gene is upregulated

Question 9

Partially correct

Mark 0.25 out of 0.75

Use the **HumanMine** to answer the following questions:

1. How many human genes are associated with diseases containing the keyword *MARFAN SYNDROME* according to OMIM, ClinVar and Human Phenotype Ontology? *(Note: use the Template "Disease --> Gene(s)"; List name: MarfanSyndrome_genes)* 1 ✔ **(0.25 pts)**
    - How many human genes are associated with diseases containing the keyword *ECTOPIA LENTIS* according to OMIM, ClinVar and Human Phenotype Ontology? *(Note: use the Template "Disease --> Gene(s)"; List name: EctopiaLentis_genes)*
    - How many human genes are associated with diseases containing the keyword *WEILL-MARCHESANI* according to OMIM, ClinVar and Human Phenotype Ontology? *(Note: use the Template "Disease --> Gene(s)"; List name: WeillMarchesani_genes)*
2. How many genes do the previous three lists have in total if combined? *(List name: Combined_genes)* 1 ✖ **(0.25 pts)**
3. How many of them share nodes in the Gene Interaction Network? (Interaction Type: Both) 0 ✖ **(0.25 pts)**

Question 10

Not answered

Marked out of 0.75

Open **Cytoscape**. *Import* a *Network from Public Databases* starting from the gene symbols found in the *Combined_genes* list from the previous exercise: enter the gene symbols and import all protein-protein interactions from *IMEx*.

1. The network diagram that represents these interactions is: **(0.15 / -0.05 pts)**
   - ○ Blank answer
   - ○ Weighted
   - ○ Directed
   - ○ Undirected

   Mark 0.00 out of 1.00

   The correct answer is: Undirected

2. How many interactions were imported? ✖ **(0.15 pts)**

Now perform the following operations and answer all subsequent questions on the simplified network:

- remove duplicated edges
- remove self-loops
- select nodes corresponding to human genes (*Taxonomy ID=9606*) and create a new network with them (all edges)

Find where the protein products of genes *ADAMTSL4* and *FYN* are located within the simplified network and focus on all the nodes in the shortest paths in between them. Select them and create a new network with them, all paths. Ignore duplicated edges between any two nodes.

1. Calculate the shortest path between *ADAMTSL4* and *FYN (Uniprotkb_accession IDs: Q6UY14 and P06241)*. ✖ **(0.15 pts)**
2. Analyze betweenness centrality for all the nodes in the shortest paths between *ADAMTSL4* and *FYN* and identify the most central node/s accordingly. Which of the following is a most central node? **(0.15 pts)**
   - ○ FBN1
   - ○ FBLN5
   - ○ LTBP2
   - ○ All are correct
   - ○ Blank answer

   Mark 0.00 out of 1.00

   The correct answer is: All are correct

3. Which of the following is the node with the highest degree? **(0.15 pts)**
   - ○ FBN1
   - ○ FBLN5
   - ○ LTBP2
   - ○ All are correct
   - ○ Blank answer

   Mark 0.00 out of 1.00

   The correct answer is: FBN1