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

☑ BT006808.1 ✓

☑ 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:

☑ 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.