

Bioinformatics: Life Sciences on Your Computer

Practice Problems - Part 2

These problems cover the NCBI database. Answers are below the final question. Please feel free to use the discussion forum to collaborate or ask for help.

1. Go to the taxonomy database at NCBI and type in the following words (separate searches):

Pyrococcus

Danio

Clostridium

Open each record and roll your mouse over the links in the lineage without clicking!

a. To which of the three domains does each of the above organisms belong? (Hint: there are three domains and NCBI uses the term “superkingdom” to refer to a domain).

Pyrococcus: _____

Danio: _____

Clostridium: _____

Click on the link that says “root” on the lineage line.

b. In addition to the three domains, what other categories are on the “top level?”

c. What types of sequences might be listed under “other” (short answer)?

2. In the NCBI protein database, please enter AGM53857.

a. Is AGM53857 a primary key or a secondary key?

b. What is the “name” of the protein?

c. From what virus?

d. Is there a published reference (yes/no)?

e. Why would the sequence of this virus be considered possibly important for public health reasons?

3. Find the nucleotide record with the accession number NM_005368.2.

a. How is this sequence described in the definition field?

b. How many exons are annotated in the Features table.

c. What is the numerical location of the poly A signal?

d. Look at the revision history. On what date did the sequence change?

e. Compare the before and after for the sequence change. How many nucleotides were added?

4. Find the human NOD2 gene in the Gene (not Nucleotide) database at NCBI.

a. According to the summary, mutations in this gene are associated with what diseases?

b. How many gene models (green) are shown in the graphic?

c. Click the OMIM link on the right menu. Then open the Sarcoidosis record. From the Molecular Genetics section, what is the inheritance pattern of this disorder.

d. The OMIM record discusses the CARD15 gene. Do a little searching of the internet. Are CARD15 and NOD2 separate genes or the same gene?

Optional challenges for advanced users

The trick to this problem is researching how to properly format an EC number search.

5. The following paper appeared in PubMed in 2011: "Cathepsin E (EC 3.4.23.34) - a review." Find all human cathepsin E proteins in the NCBI protein database using the EC number search.
- How many records?
 - Which have identical amino acid lengths?

The trick to this problem is researching how to limit a search to the Swiss-Prot database.

6. Search for human protein records with CFTR in the protein field and limit your search to the Swiss-Prot database. You should get one record if the search is correct. If not, please ask for help in Discussions.
- What is the prevalence of cystic fibrosis in the Caucasian population (hint: CTRL-F is a Find on Page feature of most PC browsers)?
 - Click Display Settings and then Revision History. When was this record first seen at NCBI?
 - At some point, several identical sequences were merged into one record. That is version 3 in this record. When was the new version 3 first seen?

This one is very tricky and might require some collaboration.

7. Difficult combination -- this one is a real challenge for advanced users--do not get discouraged if you can't get it!

Find all human nucleotide sequences that meet the following criteria:

- Not genomic sequences (molecule)
- Not STS sequences
- Mention "histone" in the title

Should be 1014 records (as of 06/04/2014).

To answer this question, describe your search strategy.

Answers

1.a Pyrococcus: Archaea

Danio: Eukaryota

Clostridium: Bacteria

b. viroids, viruses, other, unclassified

c. artificial sequences

2.a. secondary (GI is primary)

b. hemagglutinin

c. Influenza A Virus

d. The reference is not published

e. It's the H1N1 strain of flu, or bird flu

3.a. Homo sapiens myoglobin (MB), transcript variant 1, mRNA.

b. 3

c. 1055..1060

d. March 4, 2004

e. 12

4.a. Crohn Disease, Blau Syndrome

- b. one gene model (what you see is the intron-exon pattern)
- c. autosomal dominant
- d. They are synonyms for the same gene. Terminology differences can be problematic!

5.a. five sequences

- b. two are 363 amino acids in length; two are 396 in length.

6.a. one in 2000 live births

b. April 23, 1993

c. May 18, 2007

7. The easiest way is human [orgn] AND histone [title] NOT biomol genomic [prop].

That's not obvious (eliminating genomic), a very experienced user may figure that out. See if you can find other ways to do it using the limits field and using the Advanced link. Bring it up in the discussion forum.