Universitat Autònoma de Barcelona
MSc in Bioinformatics Module 2

Sequence Formats and Online Databases

Homework #1 - NCBI

Online Sequence Databases

Edgar Garriga

DATE: *3rd November 2016*
DUE : *3rd November 2016*

1 PROBLEM DEFINITION

Exercises about advanced searches at the NCBI:

1-TAXONOMY How do I find all the bacterial genera with nucleotide sequences at the NCBI?

```
"txid2"[Subtree] AND ("genus"[Rank] AND
"taxonomy nucleotide"[filter])
```

2- THE DYNAMIC OF THE SEQUENCE ENTRIES It is possible to track the history of an entry at the GenBank. When sequence with accession number NM_000770.3 was first seen at NCBI?

```
Yes, using the nucleotides database, we can select the format
to see the result. By default we will see the GenBank format,
but changing the view by "Revision history" we will have all
the modifications for the entry we are looking for. The first
time for the NM_0007700 in the NCBI was on:
03/24/1995 05:10 PM
```

3- METAGENOMIC DATA Why this strategy seen in class gave no results? ("bacteria"[Organism] AND txid408169[primary organism])

```
This strategy doesn't work because the metagenomes at NCBI are
indexed without using the environment metagenomic study as a
filter.
```

4- THE WGS DATABASE Localize all of the rabbit DNA records in the NCBI database that are whole genome shotgun (WGS).

```
rabbit[Organism] AND "wgs"[filter]
```

5- CROSS-REFERENCE The protein sequences in the NCBI Protein database come from several different sources, such as UniProt. How to list with a simple search all the proteins at the NCBI coming from the Swiss-prot?

```
srcdb_swiss prot[prop]
```

6- COMPLEX SEARCHES We'd like to retrieve all nucleotide sequences associated with "cancer". We'd like to focus only on human sequences associated with "cancer" not included in a partial or complete genome record.

```
cancer[Disease/Phenotype] AND human[Organism]
NOT "genome"[Assembly Name]
```

7- REDUNDANCIES Look for record WP_003094337.1 at the protein NCBI database. What kind of entry is it? What species itt belongs? From what nucleotide entry this protein sequence was obtained? Why the NCBI has created this sort of records?

```
The WP_003094337 is a non redundant entry. This kind of entry
is used to solve the problem with the increasing redundant
in the prokaryotic proteins
```

8- BACTERIAL GENOMES Discuss how to retrieve all the complete proteomes from all sequenced strains for given bacterial specie.