MSc in Bioinformatics 2017-2018 **Module II**

Lessons: Sequence Formats and Online Databases **Contact: Daniel Yero, IBB, UAB**

Exercises **daniel.yero@uab.cat**

By completing these exercises, you will be exposed to the major databases currently used by researchers in molecular and evolutionary biology, and you will gain a better understanding of gene analysis, taxonomy, and evolution. While no advanced computer programming skills are necessary to complete the modules in this work, prior exposure to personal computers and the Internet will be assumed. The main program that you will need is an Internet browser, such as Internet Explorer or Mozilla.

I. Before starting…

1. Explore the National Center for Biotechnology information (NCBI) website and get familiar with its design and environment and its major databases.
   NCBI https://www.ncbi.nlm.nih.gov/
2. Explore other major sequence databases accessible through the Internet. The databases are available at the indicated addresses and return sequence files through an Internet browser. Many of the sites shown provide access to multiple databases.
   The European Bioinformatics Institute (EBI) https://www.ebi.ac.uk/
3. Look for a gene of your interest in the three primary nucleotide databases: compare the information given in each one of them. In each database try to find for the selected sequences all the links to other databases. Are they cross-referenced?
   GenBank (NCBI) http://www.ncbi.nlm.nih.gov/nuccore/
   ENA (EBI/EMBL) http://www.ebi.ac.uk/ena/
   ARSA (DDBJ) http://arsa.ddbj.nig.ac.jp/html/
4. Open the genbank file for a nucleotide entry at NCBI and get familiar with its format. Are the three primary nucleic acid databases using the same criteria to construct the features tables? Read the DDBJ/EMBL/GenBank feature Table definition at
   http://www.insdc.org/documents/feature_table.html
   And see the GenBank sample record at
   https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html
5. Explore major protein databases containing primary and curated data.
   The UniProt knowledge base (UniProtKB) http://www.uniprot.org/
   The Protein Data Bank (RCSB PDB) https://www.rcsb.org/pdb/home/home.do
   Protein database at NCBI https://www.ncbi.nlm.nih.gov/protein
6. Information you might need to complete the exercises:
   - Search Field Descriptions for NCBI Sequence Database at https://www.ncbi.nlm.nih.gov/books/NBK49540/
   - Supported query fields for searching specific data in UniProtKB at http://www.uniprot.org/help/query-fields

II. Exercises about the NCBI [http://www.ncbi.nlm.nih.gov/] and its primary nucleotide database.

1. How many databases are hosted by the NCBI?
    a. How many entries are at present in each NCBI database?
    Note: Try to do it through the internet with only a unique line of commands
    Tips: NCBI help manual
    https://www.ncbi.nlm.nih.gov/books/NBK3831/
    See the *E*-utilities Quick Start guide and related tutorials:
    https://www.ncbi.nlm.nih.gov/books/NBK25500
    https://www.ncbi.nlm.nih.gov/books/NBK25499/
2. Global searches at the NCBI.
    a. Perform a **Global Query** at the NCBI using the word "cancer". Analyze the result displayed on the Global Query page. How many entries (records) have been obtained for each database?
    Note: See how your query is interpreted differently in different databases (PubMed, Nucleotide, Protein, Taxonomy, etc.).
    b. How does NCBI apply **controlled vocabulary** to each search in each database?
3. Human sequences at the NCBI.
    a. How many human sequences are there today in the nucleotide database? Have they all been submitted primarily to GenBank?
    b. How many human nucleotide sequences have been submitted primarily to the GenBank database until today?
    c. How many non-redundant human nucleotide sequences are there currently at the GenBank?
    d. How many human sequences submitted primarily to the GenBank are codifying? How many are mRNA sequences?
    e. List all the human proteins at the NCBI protein database. Look at the source database filter (left panel) from which databases come all these entries. Based on this data, how many human proteins are there at the curated and non-redundant databases refseq and SwissProt? Why are these numbers different? Do you consider both sets of proteins represent the complete human proteome?
4. Taxonomy at the NCBI.
    a. How do I find all "known and validated" species with nucleotide sequences submitted to NCBI?
    Note: "known and validated" means only species with formal Linnaean binomial names and do not include candidate or uncultured taxa.
    Tips on Statistics link at https://www.ncbi.nlm.nih.gov/taxonomy and at Taxonomy FAQs https://www.ncbi.nlm.nih.gov/books/NBK54428/
    b. How do I find all known and validated species at NCBI that belong to a particular group?
    c. How many validated bacterial genera are there at NCBI?

5. The dynamic of the NCBI entries.
    a. It could be possible to track the history of an entry at the GenBank? (use the accession number NM_000770.3 as an example)
    b. What are the oldest entries at the NCBI? Protein or nucleotide?
6. The goal of this exercise is to retrieve genetic sequence data from the NCBI database that identifies the 16S rRNA gene sequence. The nucleotide sequence of these genes is used today to identify the taxonomic position in the Tree of Life of the organism harboring this gene sequence. These genes are therefore referred to as “phylogenetic genes”.
    a. Create a strategy to retrieve unambiguous bacterial 16S rRNA gene sequences from the NCBI.
    b. ¿Are there sequences from metagenomic projects in your search?
    c. ¿Are there sequences from other high throughput projects?
7. The WGS database.
    a. Localize all of the rabbit DNA records in the NCBI database that are whole genome shotgun (WGS).
    b. Identify the master record for the project that gathers all of the contigs.
    c. Justify all the entries annotated as whole genome shotgun sequences.
8. We'd like to retrieve all nucleotide sequences associated with “cancer”.
    a. We’d like to focus only on human sequences associated with “cancer” not included in a partial or complete genome record.
    b. We are now interested in patented DNA sequences of women associated with “cancer” from all molecule types.
9. Gene Database at NCBI.
    a. The completion of the human genome sequencing project resulted in the availability of sequence data for many human diseases and other traits. Some users, however, might be interested in identifying loci that are associated with a known phenotype but not yet associated with sequence data, because these loci might be of interest as possible research topics. Find human loci that are associated with a phenotype but do not currently have the corresponding sequence data.
10. Viral genomes at NCBI. We are interested to study the evolution and spread of HIV worldwide. But first we need to answer some questions.
    a. How many different isolates of human immunodeficiency virus type 1 are represented in the NCBI sequence databases? How many of these have a complete genome sequence record in *Entrez*?
    b. We are now interested in those sequences that have been already annotated.
11. Bacterial genomes at NCBI. We want to study how many bacterial genome sequencing projects have been completed and which are in progress by searching the different NCBI databases. Design and discuss different approaches to answer the following questions.
    a. How many known species of bacteria have genome sequencing projects at NCBI?
    b. How many bacterial genomes have been submitted to the NCBI regardless of their status? How many are completely sequenced?
    c. How many in “*contigs*” status?

III. Advanced searches at UNIPROT [http://www.uniprot.org/].

1. From the non-reduntant set of human proteins at the UniProt/Swiss-prot database answer these questions:
    a. Does this number of proteins correspond with the number of proteins annotated in the **human proteome** at UniProt? Why?
    b. Try to answer these questions first.
        i. What is the complete human proteome definition of UniProt?
        ii. How to retrieve sets of UniProt protein sequences?
        iii. What are complete proteomes in UNIPROT?
        iv. What is the curation status of UniProt complete proteomes?

    Tips at http://www.uniprot.org/help/human_proteome
    c. For how many of these human sequences there are clear experimental evidences of their existence?
    d. How many human proteins have a 3D structure experimentally solved?
    e. Investigate how many proteins do not have an assigned function either experimentally or by homology. Which is the same, how many are annotated as hypothetical or uncharacterized proteins?
2. Localize the shortest sequences in UNIPROT
    a. Are all full/real protein?
    b. Does the shortest complete protein sequence have a cross-reference at the GenBank?
3. In the UniProtKB/Swiss-Prot how many human proteins inferred from homology are there greater than 10 KDa?
4. List all UniProtKB entries with 5 or more transmembrane regions.
5. NCBI and UNIPROT cross-reference, ID mapping and redundancy.
    a. The protein sequences in the NCBI Protein database come from several different sources, such as UniProt. List with a simple search all the proteins at the NCBI coming from the Swiss-prot.
    b. Look for cross-reference in UNIPROT for these protein sequences with NCBI accession numbers: CAA24952, CAB06032 and NP_003520. How redundant are sequences in UniProtKB?
    c. Why does UNIPROT protein P42284 cross-references with so many entries at the NCBI, including proteins NP_724945.1, NP_788320.1 and NP_995807.1 with different sizes? How redundant are sequences in UniProtKB?
    d. Look for record WP_003094337.1 at the protein NCBI database. What kind of entry is it? What species it belongs? From what nucleotide entry this protein sequence was obtained? How redundant are sequences at the NCBI? What is the refseq?

To continue studying in more depth . . . . .

1. Visit the NCBI educational resources. NCBI Education provides educational tutorials, software, and mini-courses. All NCBI educational materials are available for anyone to re-use and distribute.
   https://www.ncbi.nlm.nih.gov/home/tutorials/
   https://www.ncbi.nlm.nih.gov/home/coursesandwebinars/
   https://www.ncbi.nlm.nih.gov/home/documentation/
2. Free online courses at EMBL-EBI Train Online.
   https://www.ebi.ac.uk/training/online/
3. To further investigate the intricacies of the world of sequence formatting; here is an extensive resource for keeping your formats straight: http://emboss.sourceforge.net/docs/themes/SequenceFormats.html. Additionally, go to these web sites http://bioperl.org/formats/sequence_formats/ to check for all the currently used formats.
4. A great deal of all the information about GenBank format and databases updates can be found in the GenBank release notes, *gbrel.txt*, on the GenBank web site at ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt