15.01.2018

LAB 2

PUBMED (http://www.ncbi.nlm.nih.gov/pubmed)

1. Find all articles authored by James Dewey Watson:
   (A) What query did you use to search the database and how was it interpreted by the search engine (*see Search Details*)?
   (B) How many articles did you find?

2. Find all articles concerning tropical rain forest. Using filters in the side bar limit the search to papers published in 2009 and 2010, in Spanish.
   (A) How many records did you find?
   (B) How was the filtering interpreted (*Search details*)?
   (C) Write the citation details of the article which is available as a free full text (authors, title, year, journal, volume, pages).

3. Find all articles about dementia. Limit the search to clinical trials about male patients over 65.
   (A) How many records did you find?
   (B) Sort records by Best Match. Write the citation details of the first article. Check its similar articles. How many are there?

4. Using 'Advanced Search' find the article published in October 2007 in the journal "Genes and Development" that starts on page 2539. Write the full citation details of this paper. How many PubMed Central articles have cited this paper?

BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

1. Open the BLAST page and select the 'nucleotide blast'. Using the sequence in the file **bt.fa** search the 'nucleotide collection (nr/nt)' database limiting the taxonomy range to vertebrates.
   (A) What sequence is it and from which organism does it come from?
   (B) Write accession number of sequence that show the highest similarity to the query sequence.
   (C) What is 'Max_score', 'Ident' and 'Query_cover'?
   (D) What is the 'E-value'?
   (E) Using the 'Taxonomy reports' tab give accession number of human sequence that show the highest similarity to the query sequence.

2. Using BLAST check if there is a protein whose sequence contains a motif "GANDALF". Give the example Sequence ID. What is the lowest E-value in the search? Is it statistically relevant?

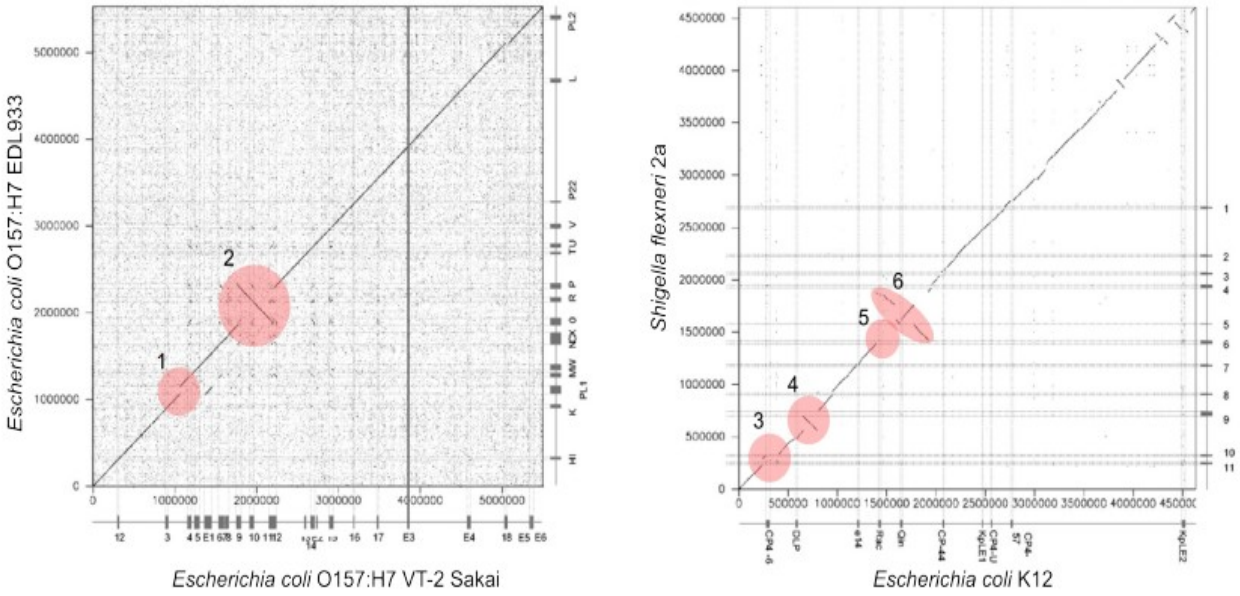
3. In the Protein database (NCBI) find the record NP_000508. What is the name of the protein?
   Using the side bar 'Analyze this sequence' go to the BLAST page (*Run BLAST*). Search the

UniProtKB/Swiss-Prot database excluding human (taxid: 9606) records from the search. What is the lowest E-value in the search? Give the accession number of this record.

DOT PLOT

1. The file **dotplot.fa** contains nucleotide sequences. Using the dotmatcher program from the EMBOSS package (http://emboss.bioinformatics.nl/cgi-bin/emboss/dotmatcher) perform dot-plot analysis for each of the sequence pairs given below. Use Window size 10 and Threshold 35. Save and interpret all graphic files (include in the report).
   (a) s1:s1 (b) s1:s10 (c) s2:s2 (d) s4:s5 (e) s7:s8 (f) s4:s4

2. Figures below show dot plot analysis of whole genomes from two *E. Coli* strains (left) and from *E. Coli* and *Shigella flexneri* (right). What kind of genomic rearrangements are highlighted?



3. Search the Protein database (NCBI) with query: “ERK kinase” AND “MAP kinase”. Using filters limit results to *Arabidopsis thaliana* sequences from the RefSeq database that are 350-360 aa long. Sort records by the accession number and open the last one.
   (A) What is the accession number of this sequence?
   (B) Go to the record of this protein in the Gene database. What is the Gene ID of the gene coding the protein?
   (C) Go to the NCBI Reference Sequences (RefSeq) section of the record and save the genomic sequence and the first mRNA sequence in FASTA format. Using the dotmatcher program from the EMBOSS package (http://emboss.bioinformatics.nl/cgi-bin/emboss/dotmatcher) compare these sequences using different values of Window size and Threshold parameters:
   (a) 10:20 (b) 10:30 (c) 10:40 (d) 30:10 (e) 30:40 (f) 30:70
   Save all graphic files (include in the report). Why does the increase of window size and threshold reduce the level of the plot background? What can you say about the structure of this gene?