Homework 3
2-AIN-501/1-BIN-301: Methods in bioinformatics
Deadline: Tuesday December 13, 2022 22:00

This homework is assigned to both computer scientists and biologists, and its goal is to introduce to you several bioinformatics tools and databases. The questions are written in English to practice English terminology. You can write your answers in English or Slovak, as you wish.

Download a short DNA sequence from `http://compbio.fmph.uniba.sk/vyuka/mbi-data/a3.fasta`. This sequence comes from an unknown bacteria contaminating a sequencing sample (real story).

a) Use BLAST to compare this sequence with known genomes to find out from what species it probably comes from. Go to `http://blast.ncbi.nlm.nih.gov`, select *Nucleotide BLAST*, enter the provided sequence as a query, and as the database, select *Refseq Genome Database*. In the box *Organism* enter *Gammaproteobacteria (taxid: 1236)*, so the program will search only one group of bacteria (in real use, we would not have this information, and thus we would have to look for it for example in all bacteria). Leave the other settings at default values and start the search.

   List details of the alignment with the highest score found by the program (% identity, E-value, source organism). What can you infer about this sequence based on the results?

b) We would like to see if this DNA sequence encodes some proteins and what their function might be. Instead of an HMM for gene finding, we will simply consider open reading frames (ORF). An ORF is a sequence of codons that does not contain a stop codon and could therefore potentially encode some protein. We need to consider ORFs on both strands and also in all three reading frames, which are three possible shifts where codons could start within the sequence. In bacteria, genes usually do not have introns, and so the search for ORFs is a relatively successful and simple procedure for finding candidate genes.

   Find the ORFs in our sequence using the tool `http://www.bioinformatics.org/sms2/orf_find.html`. Look for ORFs with at least 180 codons. Check all three reading frames and both strands. Allow any start codon and use the standard genetic code. List protein sequences of the found ORFs; we will use these in the subsequent subtasks. Most programs require protein sequences in the FASTA format, in which each sequence is preceded by a line starting with > sign followed by the name of the sequence, e.g. `>orf1`. Write the protein sequences in this format.

   You should obtain 4 ORFs, which we will name as orf1 to orf4. The starts of the corresponding protein sequences should be as follows: orf1 QPKA, orf2 PEKD, orf3 TRVA, orf4 QVRN.

c) For orf1, run BLAST on the website `http://www.uniprot.org/blast` against *"UniProtKB SwissProt"* database (change the *Target database* parameter). The SwissProt database contains only manually curated proteins for which more information is usually available. Among the BLAST results, select the *P06959* protein and two more proteins (e.g. the two with the highest score, but you can also choose others). Build a multiple alignment of orf1 and the selected three proteins (directly on the BLAST results page in the Alignments section select checkboxes of the desired proteins, press Align and select that you want to include the query as well). Copy the resulting alignment to your homework.

   Display the description of the *P06959* protein and find the so called *active sites* participating in the enzymatic reaction catalyzed by this protein. Highlight these positions in the multiple alignment (you can do it manually or using Select annotation dopdown menu on the UniProt website). Are these active sites conserved in evolution?

d) Enter orf2, orf3, and orf4 into the tool for finding domains using HMM profiles in the Pfam database at `https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan`. What domains did you find, what was the E-value? Based on the results, which of these three ORFs has the greatest chance of being an evolutionary conserved protein?

   Press arrow > next to the match with the lowest E-value to display the full alignment and copy it to your homework. In the Model line, you see for each position in the HMM profile, the most common amino acid with capital letters highlighting the most strongly conserved positions. Does our protein match the HMM row in the columns, where the HMM has capital letters?

e) In this subtask, we determine the structure of orf1 and orf3 proteins using AlphaFold2. If you wish, you can run AlphaFold2 yourself, using Google Colab `https://bit.ly/alphafoldcolab`. However the computation may take quite long, so you can find produced images on the course website.

   In addition to the image of the structure, AlphaFold2 provides two plots representing the confidence in the structure, abbreviated pLDDT and PAE. Read about these indicators at `https://alphafold.ebi.ac.uk/faq`, sections "How confident should I be in a prediction?" and "How should I interpret the relative positions of domains?".

Based on this, describe what you can infer about reliability of the structure predictions for the two proteins. Are some parts more reliable than others? Which of the proteins has an overall more reliable prediction? Are any domains visible on the PAE plots? Can you make any other observations?

f) Compare the original DNA from part a) with the profiles of the families of functional RNAs from the Rfam database at `https://rfam.org/search`, using *Sequence search* tab. To speed up the search, enter only part of the sequence from the line that begins with TCGTTT to the line that begins with TAGCAG. With which RNA family have you found a match?

g) **Bonus (max. 3 points)** What else can you find out about the ORFs or the DNA sequence? Describe how you arrived at your findings.

One possibility is to examine the relationship of the RNA from part f) and orf2 to manganese and their possible interaction. What is their mutual position in our sequence?