LAB 4: Gene finding, Blast and sequence alignment

**LAB PREPARATION**

**The** **scenario**

There is an outbreak of mysterious infectious diseases in your town. Doctors do not know what is causing it, but they do know it is spreading fast. Patients come into the hospital presenting very different symptoms and doctors suspect that there might be more than one pathogen behind it. There are three variations of the mysterious disease:

Disease 1 mostly affects children and their families. It is a severe form of diarrhoea that can lead to death if the patient isn't carefully rehydrated. Many, but not all, patients also present vomiting.

Disease 2 seems to be sexually transmitted. It causes the skin or mucosa of the affected areas to form blisters, similar to burning. These open wounds in the anal-genital area often lead to other infections. Many patients have to take intravenous serum instead of normal food, and some must be kept under total isolation in the intensive care unit.

Disease 3 has been spreading the fastest. Patients initially have flu-like symptoms, but in a matter of days or weeks it evolves to a life-threatening pneumonia, associated to chest pain and bloody coughs.

Doctors have been able to isolate bacteria from the blood of a few patients from each of these conditions (healthy individuals usually do not have any bacteria in the blood stream). They have extracted DNA from the bacteria and had it sequenced. DNA sequencing produces very small reads that have to be assembled (put together) into longer fragments of contiguous bases, called contigs. A bioinformatician at the sequencing centre has already done the initial work. It is now up to you and your colleagues to find out as much as possible about this pathogen. The patients are counting on you!

**Preparation questions**

We will start each lab discussing a few preparatory questions. You will not gain or lose points from them, but you might be called to discuss them in front of your classmates, so be prepared!

**i.** What is genome annotation? What is its goal?

**ii.** What is an Open Reading Frame? How does this relate to genome annotation?

**iii.** What does it mean to “align two sequences”? What is the goal?

**iv.** What is a p-value? How is that different from an e-value (expectation value)?

**LAB INSTRUCTIONS**

Log into Canvas and go to the BB2441 page. Now go to Modules → Computer exercises → Input files. Select one of the files Bacteria1.fasta, Bacteria2.fasta or Bacteria3.fasta, which correspond to diseases 1, 2 and 3, respectively. Make a folder called bioinformatics in your local computer account, download one of the fasta files and save it there.

**Q1** Which of the 3 unknown bacteria have you chosen to work with?

Open the fasta file in a text editor such as gedit. Alternatively, view it through the command line.

**Q2** How is a fasta file organised? What information can be found in it? Is this a practical format? Why/why not?

To understand the metabolism and life-cycle of an unknown species based on its DNAcontent, we have to study the functions of its genes. The first step for doing so is finding the gene sequences within the genome. Fortunately, there are tools that can find the genes inside a genome, based on certain sequence characteristics. Check the Bioinformatics Tools booklet and look for online tools for gene finding.

**Q3**  Which tools did you find?

Take a look at their websites. Feel free to explore them for a few minutes. Then, pick one tool to use in this assignment. It’s important to have the nucleotide sequences as output.

**Q4** Which tool did you choose? Why? Did you change any parameters from the default settings? Which, how, and why?

Make a fasta file of all the candidate genes you’ve found. Make sure to erase all comments or other lines that don’t fit the fasta format.

**Q5** How many genes did this tool find? Is this a good estimate for the number of genes in this organism? Why/why not?

Now download the total set of predicted genes. If this is not possible, select everything with the mouse, paste it to a plain text document and save it.

Now that we have identified the genes (at least the most likely genes according to the gene finder you employed), we can start studying their functions. One approach for doing this is to compare these new sequences with sequences from better known organisms. A very popular tool for doing this is Blast. Check the Bioinformatics Tools booklet on instructions in how to use online Blast for nucleotides. Blast the first 5 genes you have found.

**Q6** Which Blast variant have you chosen? Why? Did you change any parameters from the default settings? Which, how, and why?

**Q7** How many hits did you find for each gene? What do they correspond to? Does this make sense?

*Note:* if there are too many hits, just describe the top ones!

**Q8** Considering how many genes you have found, is it practical to examine the function of each corresponding protein by online Blast?

Blast can also be run locally, through the command line interface. This allows whole genomes or collections of genomes to be scanned very fast. However, it takes a bit more bioinformatics expertise to go through the very large files that are produced, so we'll do this in a slightly simplified way this time. Let's look for RNA-polymerases, that is, the enzymes that transcribe DNA into RNA. They have already been downloaded from NCBI as described here<https://www.youtube.com/watch?v=OC74-DpkWjE>, using “Bacteria” and “RNA polymerase” as keywords. You can retrieve this file directly from Canvas as Modules → Computer exercises → Input files → polymerases.fasta

Now look into this fasta file. A lot of the sequences are described as “CDS”.

**Q9** What does that mean? What is the difference between CDS, EST and ORF?

**Q10** How many sequences are there in the fasta file? This can be quickly counted through the command line.

The first step to running Blast through the command line is to prepare a database. You have the necessary fasta files to compare, ie, your unknown bacterium and the collection of bacterial RNA-polymerases. One of these files is going to be your database, and the other one contains all of your queries (the sequences to be identified).

**Q11** Which of these files should be the database, and which one should be the query? Why? What would happen if you did it the other way around?

Look into the Bioinformatics Tools booklet to see how to prepare a Blast database.

**Q12** Which command did you run? Describe what each part of it does.

Now that the database is ready, it's time to run nucleotide Blast.

**Q13** Which command did you run? Describe what each part of it does.

**Q14** From all the results you got, can you pick RNA polymerases within your bacterial genome? Give its position within the genome, together with its e-value, length and bit-score.

*Hint:* You might get several hits for the same region, but take into account that a good hit should have more or less the same length as the sequence it is matching to!

Now download from Canvas the file fewer\_polymerases.fasta. This file contains only 50 of the sequences from the polymerase database. Format a Blast database from this file, too, and run Blast using it.

**Q15** Which commands did you run? Are there any differences compared to what you did before? Did you get the same hits as you had before?

Now look at the e-values and compare them to what you had before.

**Q16** Is there any change between the e-values you had before and what you got now? How do you explain that?

Not all polymerases are the same. Let's compare a few of the ones in the database. You can find all of them in the smaller dataset. For this, you can go back to using online Blast. For the following sequences, justify your answer with your own words, but also include a *dot plot* and at least part of the *sequence alignment.*

Compare the first two sequences in the file, gi|328835344 and gi|328835342

**Q17** Which type of Blast did you run? Why?

**Q18** How similar are these sequences? Looking at the sequence headers, is this expected?

Now compare the first sequence, gi|328835344, with the one that has ID gi|619834969.

**Q19** How similar are these sequences? Looking at the sequence headers, is this expected?

For the final test, compare the first sequence to the one with ID gi|627755269.

**Q20** Considering what you learned in the biostatistics part, but also your biological knowledge, is this match relevant? Why/why not?