**Student name:**

**Registration number:**

**Practical course in bioinformatics (FOR 271)**

**Gene annotation. 20.05.2025.**

**Deadline 10th June, submit to** [**zilan.wen@helsinki.fi**](mailto:zilan.wen@helsinki.fi) **and rename the file as ‘Artemis\_yourname.PDF’.**

1. **Annotation of an unknown nucleotide fragment using Artemis**

In this exercise, we will use Artemis to annotate a small nucleotide fragment. To make it easier and less time-consuming, the provided fragment contains no introns, because it comes from a genome where introns are very rare.

Following a simple annotation pipeline, you should be able to detect potential open reading frames (ORFs), to identify putative start and stop codons, to discard ORFs without start codons, and to predict function of the identified ORFs. For the last step, we will rely on BLAST searches against non-redundant (nr) database.

Open the provided file (unknown sequence.fas) in Artemis. The provided small fragment has a length of 2.52 kbp, and contains several protein-coding genes. To identify potential protein-coding open reading frames (ORF), go to the main menu – ‘Create’ -> ‘Mark open reading frames’ – >‘100’ -> ‘OK’. This operation will highlight all ORFs with the length of >100 codons, irrespectively whether they contain start codons or not. How many ORFs were identified by the program?

A: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Next step will allow us to align the start of each ORF with suitable start codons and to discard ORF without start codons. First, select all identified ORFs (‘Select’ -> ‘All’ or Ctrl+A), then trim them to the next available start codon (‘Edit’ -> ‘Trim Selected Features’ -> ‘To Met’). The program will automatically shorten the ORFs to the first available start codon. You can check that all of them now start with the codon ATG. However, if predicted ORFs have no appropriate start codon, the program will give a warning message. You can delete all ORFs without start codons, as they are unlikely to encode any proteins. How many predicted ORFs have no start codon?

A: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Take a closer look at the remaining ORFs. You will see that some of them are overlapping, i.e. they are located within the same region, but differ in their reading frame and, in some cases, in their orientation. How many of the predicted ORFs are located on the positive DNA strand? And how many are on the negative (complementary) DNA strand?

A:

ORFs in eukaryotes (and the provided sequence is of eukaryotic origin) normally do not overlap. It means that some of the predicted ORFs are ‘false-positives’, i.e. they do not encode functional proteins. At the next step, we will try to predict a function of each of the identified ORFs, and to discard the ‘false-positives’.

Follow the next steps for each of the identified ORFs. Select an ORF by on the corresponding light-blue rectangle in the upper panel of Artemis window, or on its description in the lower panel. Then generate the FASTA sequence of the putative protein encoded by the selected ORF by going to the main menu -> ‘View’ -> ‘Amino acids’ -> ‘Amino acids of selection as FASTA’. It will open a new window with the amino acids sequence of the putative protein. Submit this sequence to NCBI BLAST against non-redundant (nr) database. Summarize information on the each of the predicted ORFs by filling this form:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ORF number | Start position | End position | DNA strand (+/-) | Description of the best BLASTP hit | Is the length of the predicted protein the same as the length of the best BLASTP hit (yes / no)? |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| 3 |  |  |  |  |  |
| 4 |  |  |  |  |  |
| 5 |  |  |  |  |  |
| 6 |  |  |  |  |  |
| 7 |  |  |  |  |  |
| 8 |  |  |  |  |  |
| 9 |  |  |  |  |  |
| 10 |  |  |  |  |  |

Based on the results of the BLAST searches, can you predict, which of the identified ORFs are ‘false-positive’ and should be discarded? Pay a special attention to overlapping ORFs occupying the same nucleotide region. How many ‘good’ ORFs were identified on the provided sequence? List the numbers of the ‘good’ ORFs according to their placement in the table above.

A:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ORF4: