Prokaryotic gene annotation

You will search for genes present in the first few thousands nucleotides of a contig of *Methanococcus maripaludis*'s genome. Here is the sequence:

```
>M.maripaludis
CAGGGTTTAGGATATATTTCTAAAAAGAAAGCAAAACAGAAATTAAAAGAAGTAATTCAA
GAAGTTATTTAATTTTTAAGTTTTTTTATTTTTTATTTGTTAAGTAATTAAATTAATTCA
ACGAAGTTTTCAATCATTTTTAAACCGTTTTTACCACTTTTTTCAGGATGGAACTGGGTA
GCATAAACGTTTTTTTTATTCAAAATGCACGGGAATTCGTAACCGTAATTAGTAGTTCCT
GAAATTACGTCTTTTTCTGAAGGATTCACGTGGTATGAATGTACAAAGTAGAAATATTCA
TTATTTGCTATTCCTTCAAAAAGGGGAATATCCTGAACCTGATTTACAGTATTCCAACCC
ATATGGGGATTTTTTTCAGAATGTTTGAATTTTATAACGTCCCCTTTTATCACGCCAAGA
CCTGGAGTTTCTGGACATTCTTCACTTTTTTCAAGTAACAACTGCATACCTAAACAAATT
CCTAAAAATGGAACCTTTTGAACGCATTTATTAATTATTTCATTTAAAGAGCAGTCTCCT
GTTTTTTGGGAAATATTTTTCATTGAATCTCCAAAATTTCCAACACCCGGGAGGACTAGC
TTGTCAGCACTTAAAATAGTTTCAGGGTCACTTGTAACAACGATGTTTTTTGTGTATAAT
TCAAGTGCCTTTTCGATACTCCTCAAGTTGCCTGCATTATAATCAATTATTGCAATCACG
ATTATCCCGTTTAATATAATTCTTCTTCAAGTTCTTTTAATGTTTTAGATATTTTATCAA
ATGCTATGTATGCATCTTCGATTACTTTTAATCCGGTAATTACAACTTTTCCACTACCAA
ATATTAATACAACAACTTTAGGTTCACTCAATCTGTAAACTAATCCAGGGAACTGTTCTG
GTTCGTATTCTGTACATTCTAATGTGGATATGTCATCTAAGTTAGGTTCCATTCCAAGTT
CGGTTGTAGCAACCATATTTTGTACTTTTACTTCAGGATT
```

Study of Open Reading Frames

To begin, we will study open reading frames thanks to the Orf Finder software. Choose the genetic code n°11 'Bacterial code'.

[ORFfinder results]

> **Question 1**
>
> What are the predicted open reading frames (position, size, frame) ?

When you click on an ORF, the latter is selected and you have the ability to launch a BlastP (or SmartBlast). You can also launch a Blast directly on all ORFs simultaneously. To do so, click on "Mark subset", then in the unfolding menu, choose "All ORFs". Finally, click on the "Blast" button (or "SmartBlast") under "Marked set". In this way, launch the Blast by choosing the "NR" bank, and answer the following questions **for each ORF** :

[37 aa ORF result], [59 aa ORF result], [30 aa ORF result], [81 aa ORF result], [151 aa ORF result]

> **Question 2**
>
> Does the protein encoded by the ORF look like known proteins (present in the banks) ?
> If yes, does the request sequence align on the **totality** of one or several protein sequences from the bank ?
> From it, can you deduce precisely the start and stop positions of the CDS present within the studied sequence ?
> Save, in FASTA format, the sequence of the bank protein which looks like the ORF the most.

Finer determination of the start and stop positions of coding sequences.

BlastP's goal is to select proteins from the bank which look the most like a request sequence. In our case, we are dealing with the translation of an ORF. Yet, the ORF could be incomplete because of sequencing errors, or because of alternative initiation codons. Furthermore, BlastP is not dedicated to the identification of a gene's structure.
The WISE software is dedicated to the alignment of a protein sequence with a genomic sequence. It tries to find the zones from the DNA sequence which encode for the protein. Compare the entire DNA sequence to the proteins which you previously saved.

Beware : The 81 and 151 AA ORFs are predicted on the "-" DNA sequence strain. Yet, this Wise version does not allow for comparison on the 2 strands of the DNA sequence. It is therefore necessary to first turn in into a reverse-complement thanks to this tool, for example, before submitting it to Wise.

[WP011170201.1 81 aa ORF Wise result], [WP_104837782.1 151 aa ORF Wise result]

> **Question 3**
>
> Does Wise find different boundaries than those deduced from the results from BlastP ?
> Where do those differencies come from?
> Are the genes found in their entirety on the sequence ?
> If no, why ?

Statistical prediction

We will use the **hmm** version of GeneMark (cf the "Gene Prediction in Bacteria, Archaea and Metagenomes" section of the software's welcome page). You will select :

- in *Select species*, a model close to *Methanococcus Maripaludis*,
- in *Output options*, the PDF output.

> **Question 4**
>
> How many genes are predicted by GeneMark ?
> Check the calculations made by GeneMark graph("PDF"), do all the genes shown by the software have a curve greater than 0.5 ?
> What can you deduce from it regarding the credibility of the predicted genes ?

You can also launch the **hmm with heuritic models** version as a complement...

Conclusion

We will now reflect on the results obtained with the help of the 2 possible methods: comparison to pre-existing proteins (Orf Finder + BlastP + Wise) or *ab initio* prediction (GeneMark).

> **Question 5**

According to you, how many CDS are present on the sequence, and what are their positions?
Which method seems the most reliable for this study ?
Compute the the protein sequences encoded by the CDS thanks to a translation software.