Final project description and hints

Introduction

Recently, a genome of an archaea from microbial communities from various locations in Russia was sequenced. This is an unclassified Thermoplasmata. The archaea is not characterized yet. To each of you, a fasta file with a fragment (~20000 nt long) of genomic DNA of this archaea is assigned. Your aim is to annotate the given region and write a structured report on your results and findings.

What do we expect from you?

We want you to annotate the given region and give us as much biologically relevant information as possible. You can use any computational tools that you want. Practical instructions are listed at the end of the document.

A more detailed plan is provided below.

1) Annotate all coding and non-coding genes;

2) For coding genes (especially hypothetical ones) try to find their functions;

3) Describe the structure of found operons (if any);

4) For the long operons (longer than 4 genes) try to find whether they have some known regulatory regions;

5) Try to find genes that were obtained by this archea through horizontal gene transfer (HGT);

6) Try to find genes associated with secondary metabolites.

Report

We expect you to write a report about your findings. Think of this report as a scientific paper, which means that the results you present have to be reproducible, data that you've got have to be clearly presented and discussed. Include plots or any other visual materials, if necessary. The methods section has to include all the information important to reproduce your results.

- **Please keep in mind that it is equivalent to an exam where you demonstrate how well you have mastered the course material.**
- **As with all other courses, remember that plagiarism is unacceptable. If you have consulted somebody, it is OK, but mention it explicitly. And don't forget about generous citing.**

Reports have to be in PDF. You have to submit your report on Canvas.

The deadline for your submission is Dec, 23, 23:59 (Moscow time).

You'll be penalized for submitting the report after the deadline (you could get only 50% of the grade after the deadline).

Report structure

The report should include several sections:

1. Introduction

It should include a few general statements about the subject to provide a background to your project.

2. Methods

Describe step-by-step which tools/programs with the parameters and thresholds and for what purposes you used while conducting the project. There should be enough information to allow another researcher to repeat your data procedures and reproduce your results.

3. Results

Summarize your main findings in the text. Overall, you are required to describe the results obtained by completing the tasks from the "What do we expect from you?" section above. You may provide plots, tables, figures, etc. if required. Additional data you've got could be presented in the Supplements section. Tables and figures should be numbered and include a brief description, with the necessary information in a legend; graphs should have axis labels. You can divide Results into several subsections. If you didn't find something, this is also a result worth mentioning. Do not provide raw screenshots, unless they are essential; you may put them to Supplements.

4. Discussion

Interpret your results. Highlight the most significant results in your opinion, but do not just repeat what you've written in the Results section. You may provide an explanation of how the results differ from what you expected.

You can combine Results and Discussion into one section.

5. Conclusions

Summarise in a few sentences the outcome of your work - what are the most interesting (including negative) findings.

6. References (optional)
7. Supplements (optional)

Technical instructions

Getting the data

The file with fasta assignments is on Canvas (Files_assignment_Final_project.pdf).

Fasta files are uploaded to the server (10.30.194.110) and are located in this folder:

/home/Shared/data/Final_Project/student_regions

Annotation

For preliminary annotation, you can use any of the pipelines that we've discussed during the seminars.

After you'll get ORF positions, you need to find anything you can about the functions of the genes.

You can start with BLAST and domain predictions.

Pay special attention to hypothetical proteins, try to figure out what they do.

Don't forget about the ncRNAs (check Rfam database).

Let's define operon as a set of co-linear genes with intergenic regions shorter than 150 bp.

Horizontal gene transfer

Usually, significant BLAST hits from phyla different from the phylum of a studied archea may be a sign of HGT. We suggest you analyze several BLAST hits with a very low e-value, use them to build a phylogenetic tree, inspect it and suggest whether HGT could happen and what could be the source of it.

Key grading criteria

- Basic annotation (whether you have run the pipeline(s) correctly and provided the results in an understandable form: as a table or a figure);
- Manual curation of functional prediction;
- Variety of used methods (for curation);
- Description of the operons;
- Discussion and conclusion section;
- Additional findings (HGT, regulatory sequences, RNA secondary structure, etc.);
- Report structure;
- Extra points (for exceptional cases), maximum 1 point.