FH Krems	ATBI6ILV – Algorithms and Tools in Bioinformatics Pt 1: Data, Tools and Technologies in Bioinformatics	SS 2024

Name: _____ Effort [h]: _____

Points: _____

1. Next-generation Sequencing (25 pt)

SARS-CoV2 data in FASTQ file format is provided in the *ATBI_Excercise* folder (*ATBI_Exercise1_R1.fastq* and *ATBI_Exercise1_R2.fastq* – **please note** the paired-end reads). Please analyze the provided data and generate the following outputs:

- **a)** Perform quality analysis and provide a quality score boxplot (using e.g., sequana-fastqc in Python or FastQC or FastQC in https://usegalaxy.org/) – describe what you see
- **b)** Generate a SAM/BAM file by aligning all reads to the provided *reference.fasta* file (using e.g., bowtie2 or bwa – also available on https://usegalaxy.org/)
- **c)** Build a consensus sequence (using e.g., "ivar consensus" on https://usegalaxy.org/)
- **d)** Perform multiple sequence alignment with other SARS-CoV2 variant data (*SARS_CoV2_data.fasta*) (using e.g., https://www.ebi.ac.uk/jdispatcher/msa/clustalo) – please add the phylogenetic tree and describe what you see (by which SARS-CoV2 variant is our patient most likely affected?)
- **e)** Use your consensus sequence and sequence of the Alpha variant and
  (1) translate the sequences to amino acid sequences (starting from ATGGTACCACATATATCACG until CTCTCTACTACCCTTCTGCT) and
  (2) perform a sequence alignment – describe what you see

2. PDB (25 pt)

Two amino acid sequences are provided:

```
> ATBI_Exercise2_A
QDNSRYTHFLTQHYDAKPQGRDDRYCESIMRRRGLTSPCKDINTFIHGNKRSIKAICENKNGNPHRENLRISKSS
FQVTTCKLHGGSPWPPCQYRATAGFRNVVVACENGLPVHLDQSIFRRP
> ATBI_Exercise2_B
KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQTNCYQS
YSTMSITDCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHFDASV
```

- **a)** Use the "Basic Local Alignment Search Tool" (BLAST; https://blast.ncbi.nlm.nih.gov/Blast.cgi) and identify the species (!) – report the top species and the description (protein names)
- **b)** Search for these two proteins on PDB (https://www.rcsb.org/) and report the PDB IDs you are going to use for further analysis
  Hint: do not hesitate to use the "Advanced Search Query Builder" and use the amino acid sequence
- **c)** Choose one of the two structures (do not forget to report which one you chose) and use AlphaFold2 to predict the structure of the protein based on the amino acid sequence
- **d)** Load all three files with a molecular visualization software (e.g., PyMOL) and align the proteins – report the RMSD value and describe the results (or use the python/colab script we used in the lecture)

3. Gene Expression **(50 pt)**

Go to the gene expression omnibus (GEO) platform (https://www.ncbi.nlm.nih.gov/geo/browse/) or ArrayExpress (https://www.ebi.ac.uk/arrayexpress/browse.html) and search for a dataset you are interested in.

a) Describe the dataset (incl. descriptive statistics – e.g., age distribution, distribution by gender, overall survival etc.)

b) Implement a data acquisition function (using e.g. GEOparse) – report the code snipped/screenshot

c) Analyze the following gene expressions:

- IL-1 beta
- PERK (EIF2AK3)
- RRAS, NRAS and KRAS
- + 1 other gene (incl. reference to a KEGG pathway: https://www.genome.jp/kegg/)

d) Report the results (boxplots, heatmaps, correlations (if any – e.g. LM-plots))

e) Extract (parts of) the dataset and perform a classification, regression or clustering using machine learning algorithms (e.g., RF, NN, LR etc.) and report the results. Target variables could be for example healthy vs. diseased, gene expression prediction (regression), disease type A vs. disease type B etc. Describe the results (also use e.g., http://biogps.org if you receive feature importances)

Remarks: Hand in your elaborations electronically,

on MS Teams or per mail to julia.vetter@fh-hagenberg.at

Accepted format: One (!) .pdf file