**Bioinformatics Competence Assessment**

**August 2022**

All code also needs to be submitted to GitHub and shared with fourie.joubert@up.acza.

1. Parsing

In the folder there is a fragment of an hmmsearch output file (hmmsearch.out). Parse the file using any method of your choice to generate a list of the top 10 hits with their descriptions and E-values, and provide the commands or code you used to do this. [10]

1. Coding
2. Write a command line programme that prompts a user for a DNA input file, and a n-mer number, and then reports the number of times each n-mer is found in the DNA sequence. Show your code and results. [10]
3. Write a command line program to prompt a user for a protein input file and an E-value cutoff, and then run a remote BLAST at the NCBI, and return the top 10 hits. Show your code and results. [10]
4. Write a program to create a scatter plot of Theoretical vs. Experimental values from a file (graphnums.tab), and draw a straight line on the plot that would represent a perfect match. Colour the dots by category listed in each line. Show your code and results.
5. Sequence analysis

Provide a Hidden Markov Model state diagram for the following pattern: [10]

A picture containing text, electronics, keyboard

Description automatically generated

1. Genomics

Perform an assembly of the bacterial raw Illumina reads provided in the MiSeq *E. coli* fastq files using software of your choice. Perform an assessment of the reads and the assembly quality and comment on this. Show your commands and results. [10]

1. Amplicon sequence analysis

Please find Illumina output for PCR amplified fragments of algae mitochondrial DNA in the file metagenomics.fastq. Perform binning of the reads (identify taxonomic units in the sample) and estimate statistical parameters of the sample: the expected species richness of the sample and an alpha-diversity index of the algaeal species (Shannon index or any other indices).  [10]

1. Metagenomics

Shotgun metagenomics has broadened our understanding of microbial communities and their evolution within a well-defined habitat, ‘the microbiome’. This includes expanding current insights regarding plant-microbe interactions, soil microbial communities, and biogeochemical cycling in the oceans. These studies require the integration and application of various bioinformatics tools. Using any benchmarked, reproducible tool used for genome resolved metagenomics, analyse the provided dataset. By generating assembled contigs, provide output files showing an overview of MAGs (metagenome assembled genomes) meeting acceptable specifications for medium to high quality drafts. [10]

1. General

If you are funded to sequence, annotate and publish interesting information on the genome of a South African endangered mammal, which would you choose? Describe which students of what levels you would involve, what your supervision process would be, what your detailed experimental and bioinformatics approaches would be, how you would distribute the different aspects between students, and how you would structure the publication (max. 2 pages). [25]