Lab Report- Microbial Metagenomics 12/06/2024

Script for creating inversions (Inversions generator)

This script was developed as part of a project during the Microbial Metagenomics course at the University of Padova for the academic year 2023-2024. The project was supervised by Prof. Stefano Campanaro and Dr. Sofia Frauini. This project was made possible by the contribution of Atefe Rostami, Lisia Peqini, Mohsen Rastgoo Shahrestani and Jaydeep Sarjerao Desai.

• Introduction

The objective of the project is to write a script in Python that creates inversion mutations to the full genome of *Methanoculleus marisnigri* JR1. *Methanoculleus marisnigri* JR1 is a well-studied methanogenic archaeon with significant environmental, ecological, and biotechnological importance. Its complete genome and one single scaffold provide a robust framework for research into methanogenesis, microbial ecology, and potential industrial applications. This genome was subjected to inversion mutation on a specific genome length. An inversion mutation is a type of chromosomal rearrangement in which a segment of a chromosome is reversed end to end. This means that a section of DNA is cut out, flipped 180 degrees, and then reinserted back into the chromosome. Inversion mutations can occur within a single chromosome and can involve small or large segments of the DNA sequence. Keeping in mind that when creating an inversion, it is also needed to generate the reverse-complement of the DNA sequence (not only the reverse one). After generating the modified version of the genome, changing the mutation length and the mutation percentage, ANI (Average Nucleotide Identity) was calculated. ANI measures the average percentage of nucleotide sequences that are identical between two genomes. Upon the calculation of ANI, a correlation analysis was performed.

•Materials and Methods

1) Writing the script

To write the script we use Python and import Standard Library Packages like ‘random’, ‘argparse’, ‘subprocess’ and Third-Party Packages like ‘Pandas’ and ‘Numphy’ for tabular data manipulation, ‘scipy.stats’ for statistical functions, ‘Bio.SeqIO’ part of BioPython used for reading and writing sequence file formats, ‘Bio.Seq’ which contains the Seq and MutableSeq classes for representing and manipulating biological sequences, ‘Bio.SeqFeature’ part of BioPython used for representing sequence features. And packages for statistical data visualization and statistical graphics like ‘seaborn’, ‘matplotlib.pyplot’, ‘plotly.express’, etc. By integrating these packages, the script can effectively handle genomic data, perform random mutations, run external analysis tools, process the results and visualize the outcomes.

2) The input and output files

The input file of the reference genome is FASTA format. Before running the script, the user is also required to input the mutation length and the mutation percentage as integer. According to those specified parameters, the genome file is parsed and randomic regions are inverted. The modified regions are stored as an output file in a .txt format. From each modification there are two .txt files obtained. The first one shows the number of mutations applied and all the modified regions listed. While the second one shows the whole modified genome.

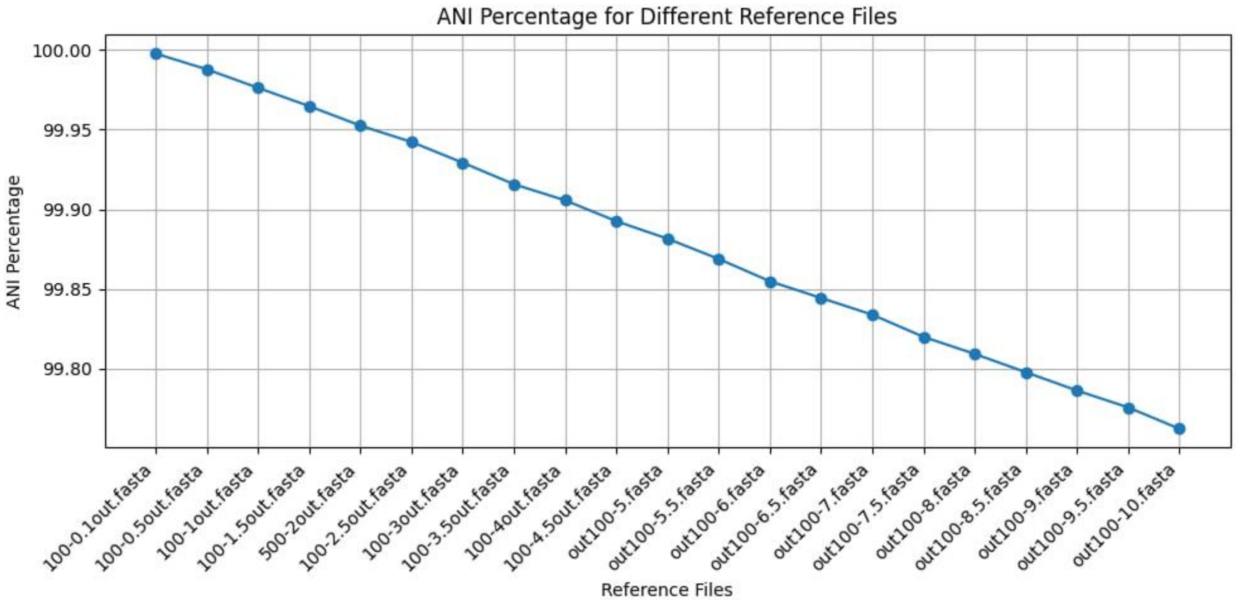
3) Run FastANI

After running the script multiple times, keeping the mutation length at 100 and changing the mutation percentage from 0.1% to 10%, a .txt file is created containing the list of paths to the modified fasta files. Then FastANI is performed on the original file and on the modified versions of the genome to compute ANI.

The file generated is a .txt file which is composed of five columns. The first column shows the filenames of the modified genome sequences. The second column shows the filename of the reference genome against which all the modified sequences are being compared. The third column represents the ANI percentage. The fourth column indicates the number of fragments or alignment windows used in the comparison that matched with the reference. The fifth column is the total number of fragments or alignment windows in the reference sequence that were used for comparison.

4) Correlation plot

At the end a correlation plot is generated, showing the relationship between the mutation percentage of the files on the X-axis and the ANI percentage in the Y-axis. As shown in the figure, there is a linear negative correlation. This means that the increase of mutation percentage is associated with a decrease in the ANI percentage. This was expected, because the more mutations introduced, the less similar the two genomes will be.



- Results and Discussions

The script successfully introduced inversions in the *Methanoculleus marisnigri JR1* genome at various mutation percentages. The FastANI analysis revealed that as the mutation percentage increased, the average nucleotide identity (ANI) between the original and modified genomes decreased, indicating lower similarity with higher mutation rates. The correlation plot confirmed a clear negative relationship between mutation percentage and ANI.