Ex\_7

Podobnie jak w poprzednich zadaniach, załączam kod i output tutaj:

from Bio import SeqIO

fasta\_file = "genome\_data.fasta"

for record in SeqIO.parse(fasta\_file, "fasta"):

print(f"ID: {record.id}, length: {len(record.seq)}")

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

import pandas as pd

vcf\_file = "data.vcf"

vcf\_data = pd.read\_csv(vcf\_file, sep="\t", comment="#"

print("Podgląd danych VCF:")

print(vcf\_data.head())

print("\nStatystyki opisowe:")

print(vcf\_data.describe())

Obraz zawierający tekst, zrzut ekranu, wyświetlacz, oprogramowanie

Zawartość wygenerowana przez AI może być niepoprawna.

# filtrujemy tylko SNP

import pandas as pd

colnames = ["CHROM","POS","ID","REF","ALT","QUAL","FILTER","INFO"]

vcf\_data = pd.read\_csv("data.vcf", sep="\s+", comment="#", names=colnames)

print(vcf\_data)

mutations = vcf\_data[vcf\_data["INFO"] == "SNP"]

print("\nMutacje typu SNP:")

print(mutations)

Obraz zawierający tekst, zrzut ekranu, wyświetlacz, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

I ostatnia wizualizacja:

Obraz zawierający tekst, zrzut ekranu, Czcionka

Zawartość wygenerowana przez AI może być niepoprawna.

Teraz kolejna część, wpierw uruchomiłem coś takiego:

Obraz zawierający tekst, zrzut ekranu, oprogramowanie

Zawartość wygenerowana przez AI może być niepoprawna.

Teraz z 4 partycjami:

Obraz zawierający tekst, zrzut ekranu

Zawartość wygenerowana przez AI może być niepoprawna.

Generalnie spark pozawala szybko policzyć mutacje SNP w pliku VCF, równolegle na wielu partycjach. Przy dużych plikach ma to kolosalne znaczenie, że nie pracujemy na jednej instancji pamięci RAM.

Stosując Big Data możemy:

Identyfikować warianty genetyczne w dużych kohortach

Wykrywać mutacje chorobotwórcze

Prowadzić takie projekty jak 1000 GENOMES

Filtrować potężne datasety

Wyzwaniami jest głównie konfiguracja środowiska, dlatego pewnie lepiej korzystać z gotowych images. Łatwo jest też przekroczyć założone koszta bez odpowiedniej kontroli. Natomiast pracuje się po prostu szybciej i o wiele łatwiej można skalować.