Genomika mikrobów – ćwiczenia (2024)

Przed rozpoczęciem

Proszę pobrać
1) dane do ćwiczeń dostępne są pod następującym linkiem.
2) program fastQC (dostępny tutaj)
3) program Bandage (dostępny tutaj)

Zadanie 1

Część 1.
Proszę otworzyć plik w pobranych danych

```
dane/czesc1-qc/short_reads_1.fastq
```

używając edytora do tekstu (albo opcji "head -4" + nazwa_pliku w terminalu). Następnie proszę o odpowiedź na następujące pytania:
a) Który znak ASCII odpowiada najmniejszej mierze jakości
b) Phred (*ang*. PHRED score) dla platformy Illumina 1.8+?
c) Jaka jest miara jakości Phred dla trzeciego nukleotydu sekwencji pierwszego readu?
d) Jaka jest dokładność odczytu (ang. accuracy) trzeciego nukleotydu?

Przeliczniki z ASCII na Phred oraz dokładność odczytu znajdziecie Państwo na https://en.wikipedia.org/wiki/FASTQ_format#Encoding

Część 2.
Proszę otworzyć pliki

```
dane/czesc1-qc/short_reads_1.fastq
```

oraz

```
dane/czesc1-qc/short_reads_2.fastq
```

za pomocą programu fastQC (link powyżej). Proszę przeanalizować dostępne wykresy jakości i odpowiedzieć na następujące pytania:
a) Jakiej platformy do sekwencjonowania użyto do wygenerowania tych danych?
b) Który z dwóch plików charakteryzuje się lepszą jakością? Dlaczego? Czy jest to znacząca różnica?

Zadanie 2

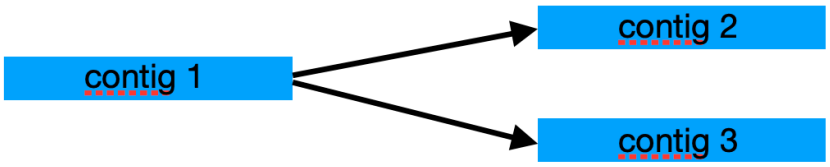
Proszę skonstruować graf de Bruijna dla słowa
`BIOTECHNOLOGY`
używając długości kmera k=4. W odpowiedzi proszę podać zarówno tabelkę jak i sam graf skierowany.

Zadanie 3

Dany jest "genom" o następującej sekwencji:
`ATGATTTCCTCTCCTCGATTCCGCCAATC`

1) Proszę skonstruować z niego graf de Bruina dla k=6 (można również napisać w tym celu krótki program w dowolnym języku programowania, np Python).

2) Następnie proszę z tego grafu skonstruować *assembly graph*, który powinien mieć następującą formę contigów połączonych krawędziami skierowanymi (przykład poniżej). Proszę również podać sekwencje wszystkich contigów.



contig 1: ACTGACGATG
contig 2: GCTATCGGCAAA
contig 3: GTTAAACTGATT

3) Czy z otrzymanego assembly graphu można jednoznacznie rozczytać genom? Odpowiedź uzasadnij.

Zadanie 4

Proszę otworzyć pliki

`dane/czesc2-assembly/H_pylori-k*.gfa`

za pomocą programu Bandage (Load graph -> Draw graph), które pokazują assembly grafy genomu bakterii *Helicobacteri pylori* dla wartości k = {19, 31, 63, 89}. Proszę porównać ze sobą grafy i odpowiedzieć na następujące pytania:

a) czym różnią się grafy tworzone dla różnych długości k?
b) z czego wynikają te różnice?

Zadanie 5

Dla wybranego izolatu bakterii *Klebsiella pneumoniae* wykonano sekwencjonowanie technologiami Illumina (krótkie odczyty) i ONT (długie odczyty). Następnie zrobiono QC zebranych odczytów i wykonano hybrydowe assembly. W pliku

`dane/czesc3-dziwny_contig/congit5.fasta`

zawarto jeden z contigów otrzymanych w wyniku tego assembly.

Dla wspomnianego izolatu istnieje już assembly genomu w bazie NCBI Nucleotide, jednak zostało ono otrzymane z innego zestawu odczytów. Do jego uzyskania użyto tej samej metody hybrydowego assembly.

1) Za pomocą programu Nucleotide BLAST (np. blast.ncbi.nlm.nih.gov) ustal skąd pochodzi sekwencja – podaj rodzaj i gatunek organizmu, nazwę, identyfikator NCBI oraz replikon, z którego pochodzi sekwencja referencyjna (znaleziona).
2) Pobierz sekwencje referencyjną i porównaj ją programem BLAST z analizowanym contigiem (opcja: *Align two or more sequences*). Przeanalizuj wyniki za pomocą opcji DotPlot. Jakie zjawisko obserwujesz? Podaj dwie potencjalne biologiczne przyczyny występowania tego zjawiska.