Microbiome Assembly

Patrycja Adamowicz, Natalia Radzikowska, Kaja Wróblewska

*Zadanie 1.*

1. Który znak ASCII odpowiada najmniejszej mierze jakości Phred (*ang*. PHRED score) dla platformy Illumina 1.8+?

Najmniejszej jakości Phred dla platformy Ilumina 1.8 odpowiada znak ASCI „!”.

1. Jaka jest miara jakości Phred dla trzeciego nukleotydu sekwencji pierwszego readu?

Nukleotydowi temu przypisany jest znak „I”, odpowiada on wartości Phred 40.

1. Jaka jest dokładność odczytu (ang. accuracy) trzeciego nukleotydu?

Accuracy wyliczono z poniższego wzoru:

Dla wartości Q=40 i uzyskano wynik P = 0.0001.

1. Jakiej platformy do sekwencjonowania użyto do wygenerowania tych danych?

Sanger/Ilummina 1.9

1. Który z dwóch plików charakteryzuje się lepszą jakością? Dlaczego? Czy jest to znacząca różnica?

Lepszą jakością charakteryzują się odczyty short\_reads\_1. Dla nich mediany klasyfikują się na poziomie 40 punktów, a żółte wykresy boxplot nie spadają (poza jednym) na pomarańczowy rejon. Natomiast dla odczytu drugiego (short\_reads\_2) jedna mediana spada nawet na pomarańczowy obszar (wartość Phred, około 24). Dodatkowo boxploty dla końcowych nukleotydów również spadają do pomarańczowego obszaru.

Obraz zawierający tekst, zrzut ekranu, diagram, linia

Opis wygenerowany automatycznie

Rysunek 1. Wykres oceny jakości Phred dla sekwencji short\_reads\_1.

Obraz zawierający tekst, zrzut ekranu, diagram, Równolegle

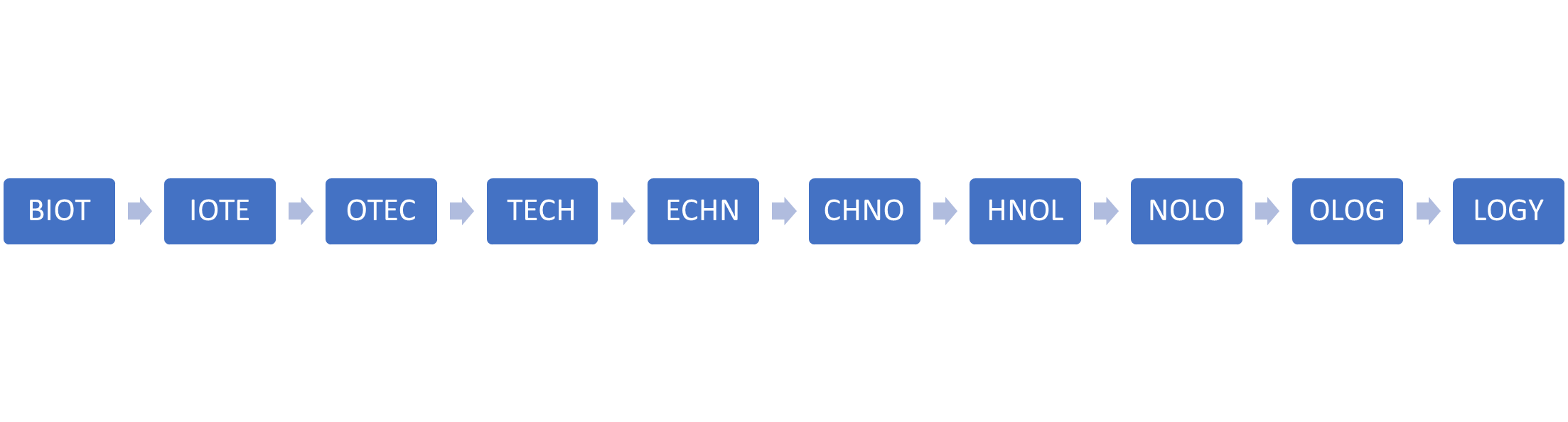
Opis wygenerowany automatycznie

Rysunek 2. Wykres oceny jakości Phred dla sekwencji short\_reads\_2.

*Zadanie 2.*

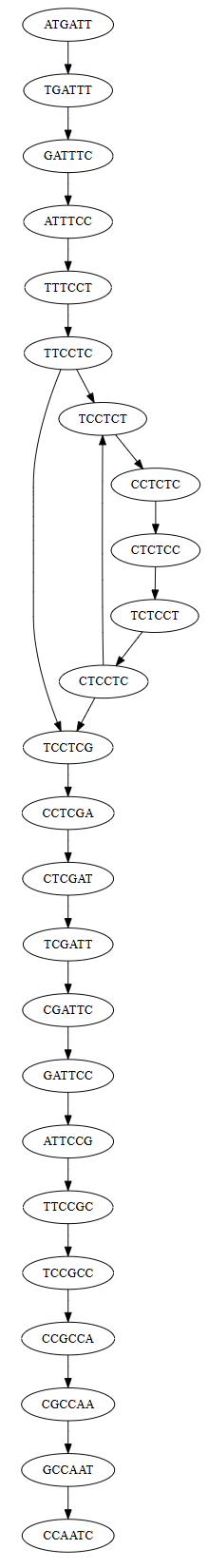
Tabela 1: k-mery o długości 4 dla słowa "BIOTECHNOLOGY"

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | BIOT | IOTE | OTEC | TECH | ECHN | CHNO | HNOL | NOLO | OLOG | LOGY |
| BIOT | - | + | - | - | - | - | - | - | - | - |
| IOTE | - | - | + | - | - | - | - | - | - | - |
| OTEC | - | - | - | + | - | - | - | - | - | - |
| TECH | - | - | - | - | + | - | - | - | - | - |
| ECHN | - | - | - | - | - | + | - | - | - | - |
| CHNO | - | - | - | - | - | - | + | - | - | - |
| HNOL | - | - | - | - | - | - | - | + | - | - |
| NOLO | - | - | - | - | - | - | - | - | + | - |
| OLOG | - | - | - | - | - | - | - | - | - | + |
| LOGY | - | - | - | - | - | - | - | - | - | - |



Rysunek 3: Graf de Bruijina dla słowa „BIOTECHNOLOGY” i długości k-merów równej 4

*Zadanie 3.*

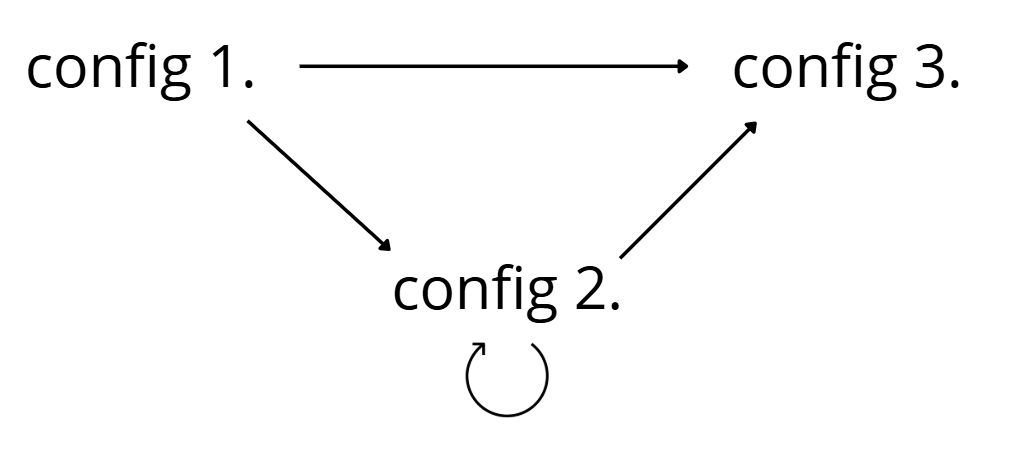


Rysunek 4. Graf de Bruijna dla sekwencji "ATGATTTCCTCTCCTCGATTCCGCCAATC" o długości k-merów równej 6.

Config 1. ATGATTTCCTC

Config 2. TCCTCTCCTC

Config 3. TCCTCGATTCCGCCAATC



Rysunek 5. Assembly graph

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

Z powodu występujących pętli nie można jednoznacznie odczytać grafu.

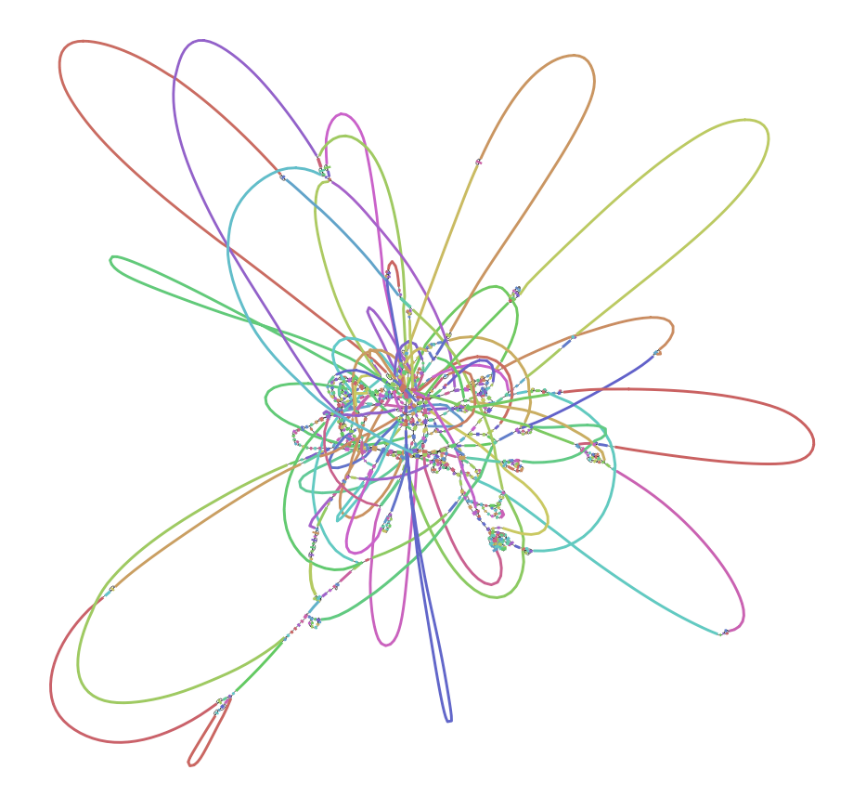
*Zadanie 4.*

1. czym różnią się grafy tworzone dla różnych długości k?

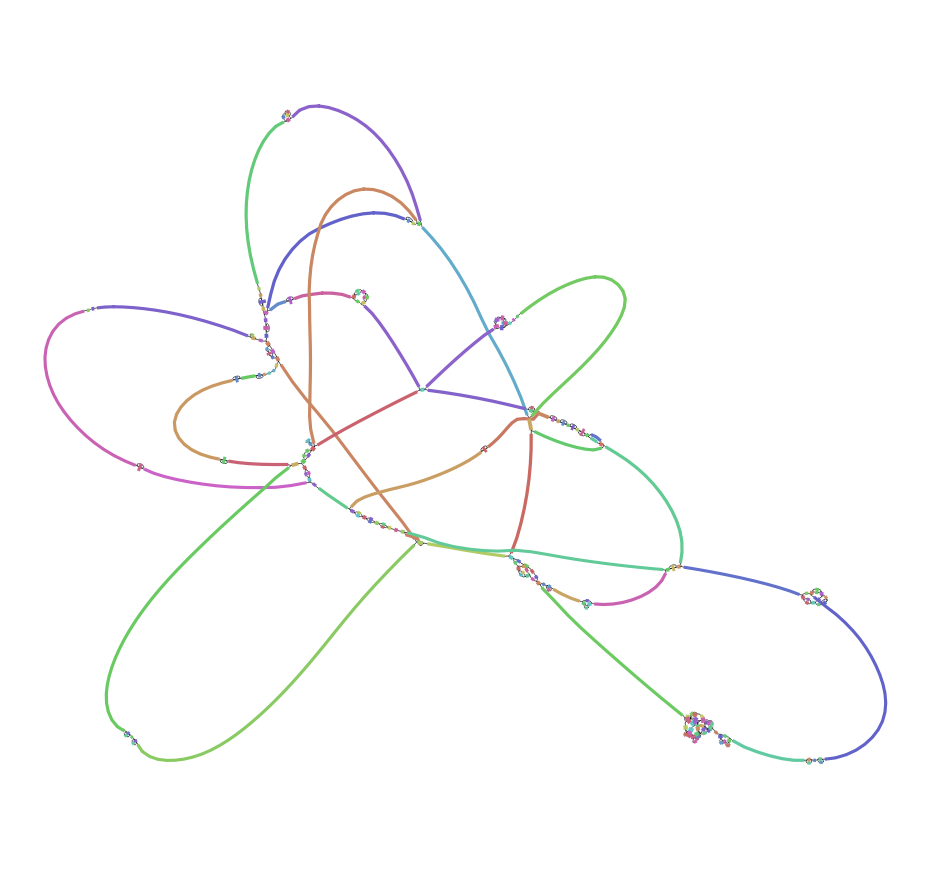
Grafy utworzone dla różnej długości k-merów różnią się ilością węzłów na uzyskanym grafie. Graf dla k-merów o długości 19 ma najwięcej węzłów, czyli najwięcej połączeń dla sekwencji oraz wizualnie przypomina węzeł Gordyjski. Dla k-merów o długości 31 nadal obserwujemy dużą ilość węzłów, natomiast jest on bardziej czytelny. Natomiast dla najdłuższych k-merów, długość 89, obserwujemy nawet sekwencje, które nie są w żaden sposób połączone z resztą grafu, co sprawi, że nie jesteśmy w stanie odtworzyć sekwencji.

1. z czego wynikają te różnice?

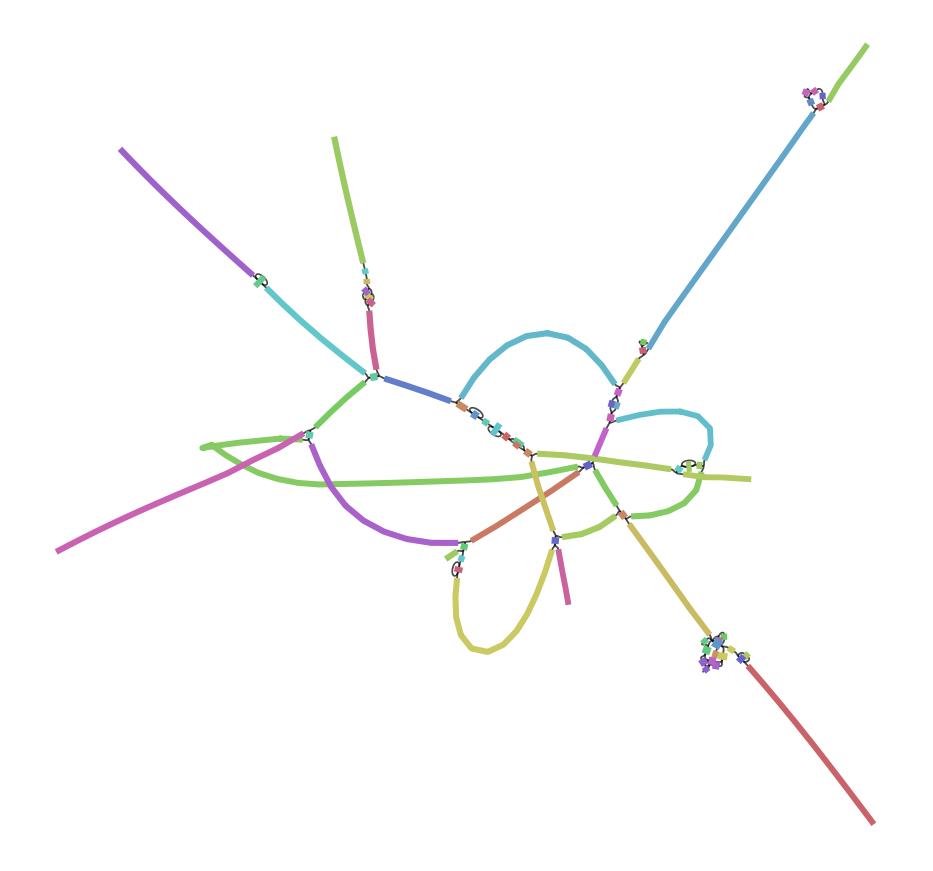
Różnice wynikają z tego, że jeśli weźmiemy dłuższe k-mery, to mamy mniejsze prawdopodobieństwo, że wszystkie znaki w danym kamerze będą identyczne z innym k-merem.



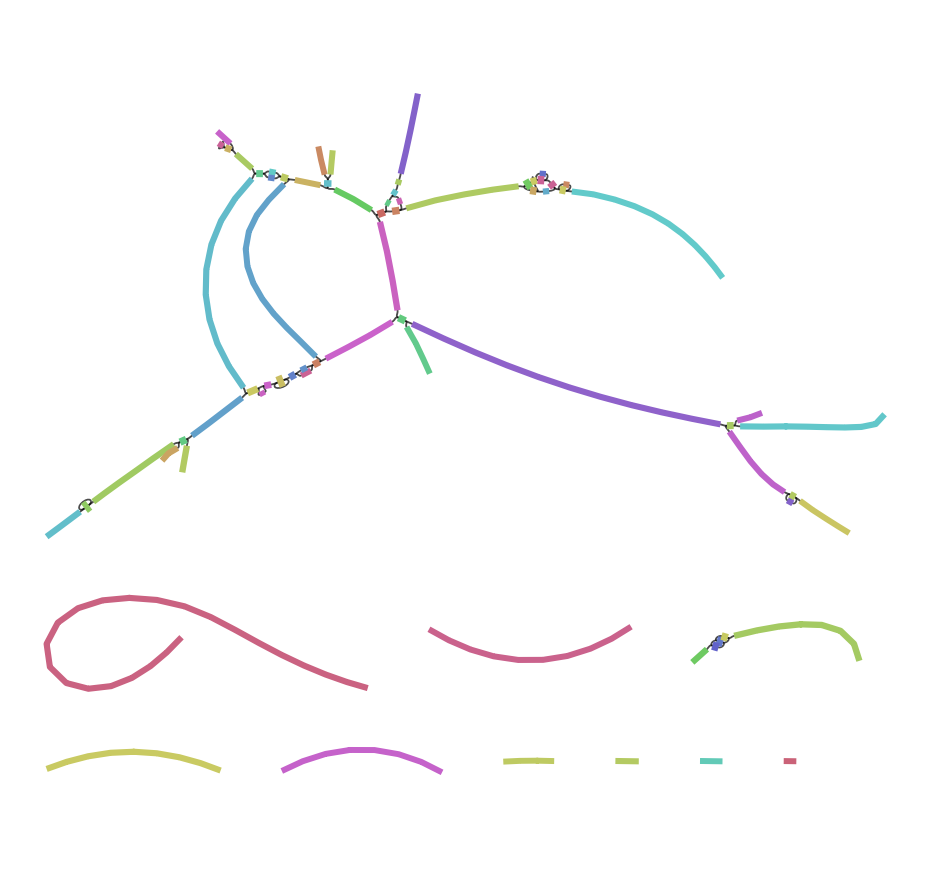
Rysunek 6. Graf dla k-merów o długości k=19



Rysunek 7. Graf dla k-merów o długości k=31



Rysunek 8. Graf dla k-merów o długości k=63.



Rysunek 9. Graf dla k-merów o długości k=89

*Zadanie 5.*

1. Rodzaj: Klebsiella

Gatunek: Klebsiella oxytoca

Nazwa: Klebsiella oxytoca strain IR5392 plasmid unnamed1, complete sequence

Identyfikator NCBI: CP064109.1

Replikon, z którego pochodzi sekwencja referencyjna: strain IR5392

1. Wyniki zaprezentowane za pomocą wykresu Dotplot

Obraz zawierający linia, Wykres, diagram, Równolegle

Opis wygenerowany automatycznie

Rysunek 10: Wykres Dotplot (macierz kropkowa) przyrównania sekwencji referencyjnej i badanej za pomocą blastn

Można zaobserwować przesunięcia i przerwy w macierzy kropkowej, które mogą świadczyć o występowaniu rearanżacji w genomie, takich jak delecje czy insercje.

Potencjalne biologiczne przyczyny występowania tego zjawiska to:

1. Zmiany w plazmidach takich jak inwersje fragmentów DNA lub wstawienie dodatkowych elementów np. transpozonów czy kaset genowych
2. Różnice wynikające z mechanizmów adaptacyjnych – bakteria, której fragment badamy znana jest z dużej zmienności genomowej związanej z adaptacja do różnych środowisk lub presji selekcyjnej. Mogły wystąpić mutacje prowadzącej do różnic w sekwencji, takie jak punktowe zmiany nukleotydów, delecje lub duplikacje regionów genomu