Wydział Matematyki
i Nauk Informacyjnych
POLITECHNIKA WARSZAWSKA

Bioinformatyka

Projekt 2

Analiza filogenetyczna

Michał Rdzany

Spis treści

Wprowadzenie

Opis zadania

Głównym zadaniem w projekcie jest przeprowadzenie analizy filogenetycznej sekwencji białka S wirusa SARS-CoV-2.

Cel badań

Głównym celem projektu jest zbudowanie prostego drzewa filogenetycznego szczepów wirusa oraz oszacowanie tempa ich ewolucji.

Opis danych

Wykorzystywane dane zostały pobrane ze strony NCBI. W sumie analizowanych jest 26622 sekwencji w formacie FASTA, a także powiązane z nimi lokalizacje i daty (kolumny *Geo Location* i *Collection Date*).

Opis wykonanych zadań

Przygotowanie danych

Sekwencje zawierające braki usunąłem korzystając z wyrażeń regularnych w Notepad++ (*Find: ">.*\n[^>]*X[^>]*(?=>)", Replace: ""*). Do dalszych przekształceń używałem pythona (kod dostępny w załączonym pliku .ipynb, a także w wersji wyrenderowanej do .html). Połączyłem sekwencje z ich lokalizacjami i datami, a następnie usunąłem zduplikowane sekwencje, zachowując najstarsze.

Drzewo przewodnie

Tworzenie drzewa

Sekwencje z przygotowanych danych zapisałem w formacie FASTA. Do utworzenia drzewa przewodniego wykorzystałem program dostępny na stronie EMBL-EBI. Po wykonaniu się algorytmu pobrałem drzewo przewodnie w formacie Newick. Drzewo to zwizualizowałem korzystając z biblioteki ETE Toolkit. Dalsze operacje na drzewach również wykonywałem przy pomocy tej biblioteki.

Podział na grupy

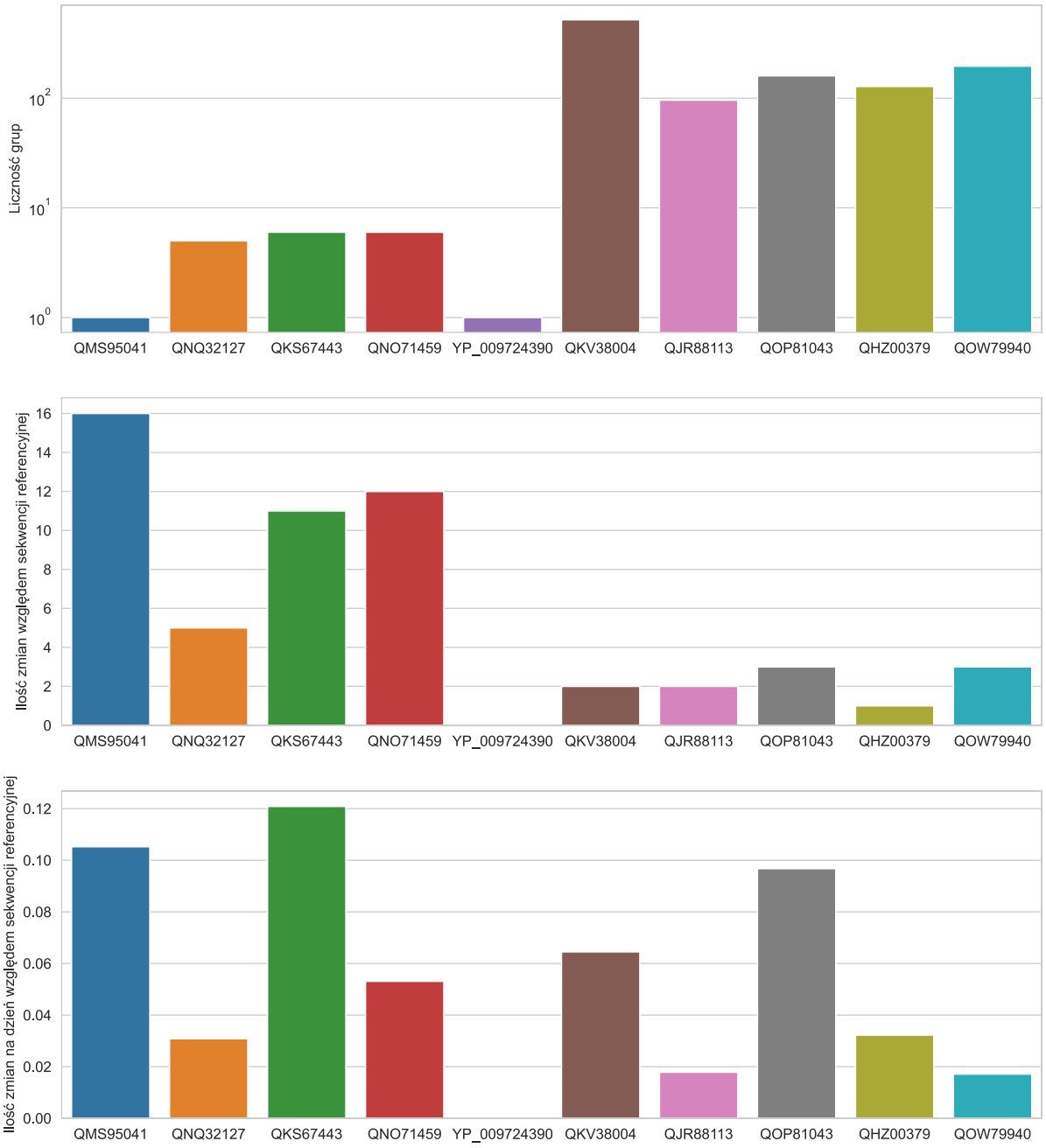
Grupy wydzieliłem poprzez iteracyjny podział drzewa. Z listy drzew (początkowo zawierającej tylko drzewo przewodnie) wybierane jest drzewo, dla którego następniki korzenia są najdalej oddalone od siebie i jest ono dzielone na dwa poddrzewa, które wcześniej były połączone z korzeniem. Zdecydowałem się na utworzenie 10 grup.

Analiza wybranych grup

Liczność grup jest bardzo zróżnicowana. Najmniejsza grupa zawiera tylko jedną sekwencję, a największa 519. Co ciekawe sekwencja referencyjna (YP_009724390) utworzyła własną grupę (o liczności 1).

Część grup zawiera sekwencje pochodzące ze wspólnych obszarów geograficznych (w jednej z grup 158 ze 160 sekwencji pochodzi z *Australia: Victoria*), jednak istnieją także grupy z dużą różnorodnością lokalizacji.

Porównanie z sekwencją referencyjną zostało przeprowadzone tylko dla reprezentantów grup z uwagi na długi czas liczenia uliniowień. Okazało się, że reprezentanci mało licznych grup charakteryzują się większą liczbą różnic z sekwencją referencyjną. Patrząc na tempo mutacji (dzieląc liczbę zmian przez czas pomiędzy uzyskaniem danej sekwencji, a sekwencji referencyjnej) podobna zależność nie jest już widoczna.


Liczność grup
10^2
10^1
10^0
QMS95041
QNQ32127
QKS67443
QNO71459
YP_009724390
QKV38004
QJR88113
QOP81043
QHZ00379
QOW79940
Ilość zmian względem sekwencji referencyjnej
0
2
4
6
8
10
12
14
16
QMS95041
QNQ32127
QKS67443
QNO71459
YP_009724390
QKV38004
QJR88113
QOP81043
QHZ00379
QOW79940
Ilość zmian na dzień względem sekwencji referencyjnej
0.00
0.02
0.04
0.06
0.08
0.10
0.12
QMS95041
QNQ32127
QKS67443
QNO71459
YP_009724390
QKV38004
QJR88113
QOP81043
QHZ00379
QOW79940

Wybór reprezentantów

W każdej grupie konieczne było wybranie reprezentanta dobrze charakteryzującego sekwencje w grupie. Zauważyłem, że przechodząc z wierzchołka do następnika suma odległości do wszystkich liści zmienia się w określony sposób. Wykonując przejście do następnika x wzdłuż krawędzi o długości $x.dist$, zmniejszamy o $x.dist$ odległość do liści z poddrzewa x i zwiększamy o $x.dist$ odległość do pozostałych liści. Wzór na zmianę sumy odległości można zatem zapisać jako $\Delta = x.dist \cdot (|T| - |x|) - x.dist \cdot |x| = x.dist(|T| - 2 \cdot |x|)$, gdzie $|T|, |x|$ oznaczają ilość liści w całym drzewie T i w poddrzewie x. Wybór reprezentanta jest wykonywany poprzez iteracyjny wybór następnika minimalizującego tę zmianę.

Drzewo filogenetyczne

Z wybranych reprezentantów zostało utworzone drzewo filogenetyczne. Do utworzenia drzewa filogenetycznego wykorzystałem samodzielnie zaimplementowany algorytm UPGMA.

Dystans ewolucyjny pomiędzy sekwencjami wyznaczyłem jako minimalną ilością zmian jakie należy wprowadzić w jednej sekwencji, aby przeprowadzić ją w drugą sekwencję. Do tych obliczeń wykorzystałem własną implementację algorytmu Needlemana-Wunsha. Z uwagi na duże podobieństwo sekwencji, wykorzystałem macierz *blosum80*. Skorzystałem z macierzy dostępnej w bibliotece *Biopython* (*Bio.SubsMat.MatrixInfo*). Jako karę za przerwę ustaliłem -6 wzorując się na macierzy dostępnej na stronie NCBI. Uliniowienia liczyły się dość długo, bo pętle w pythonie są wolne, ale nie znalazłem dostępnej, szybszej implementacji. Słownik z wyliczonymi dystansami zapisałem do pliku, aby nie powtarzać już tej procedury.

Utworzone drzewo filogenetyczne wyrenderowałem do pliku .png. Ma ono bardzo regularną strukturę, kolejne sekwencje są coraz bardziej oddalone od sekwencji referencyjnej. Taka struktura ma sens, jeśli popatrzymy na metodę budowania drzewa. Sekwencja referencyjna jest najbardziej podobna do innych i jest łączona jako pierwsza. Skutkuje to jednak tym, że sekwencja referencyjna jest w najgłębszym liściu, podczas gry teoretycznie powinna być „w korzeniu", przez co drzewo jest niejako „odwrócone". Być może lepszym podejściem byłaby analiza tylko „nowych" sekwencji, które mogłyby wykazywać większe różnice względem siebie i w stosunku do sekwencji referencyjnej.