

VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

Lab 1

Atliko: 4 kurso 1 grupės studentas
Ignas Šileika

2024 m. spalio 14 d.

Turinys

| | |
|---------------------|---|
| 1. ĮVADAS | 3 |
| 2. METODAI | 3 |
| 3. REZULTATAI | 4 |
| 3.1. Codonų | 5 |
| 3.2. Dicodonų | 5 |
| 4. IŠVADOS | 6 |

1. Įvadas

Šio darbo tikslas buvo palyginti kodonų ir dikodonų dažnius virusuose, kurie užkrečia žinduolius ir bakterijas. Išsiaiškinti, ar šie virusai sudaro atskiras grupes pagal kodonų ir dikodonų dažnius.

Darbas buvo atliktas naudojant python programavimo kalbą.

2. Metodai

Kodas: <https://github.com/ignasvoveriukas9/BioinformatikaLab1>

Pateiktos virusų sekos buvo FASTA formatu. Iš kiekvienos sekos:

Randoma 6 variantus kaip galima skaityti seką (nuo pirmo, antro ir trečio simbolio + reverse complimentas)

Tada šias sekas vertėme į aminorūgščių sekas pagal kodonų lentelę.

Suradome pradžios (M) ir pabaigos kodonus (*). Tarp šių kodonų ištraukėme baltymus koduojančias sekas, jei jos buvo ilgesnės nei 100 bazinių porų.

Skaiciavome kiekvieno viruso ir bakterijų kodonų ir dikodonų dažnius.

Panaudojome Euklido atstumą, kad apskaičiuotume skirtumus tarp virusų bei bakterijų pagal kodonų ir dikodonų dažnius.

$$d(A, B) = \sqrt{\sum_{i=1}^n (f_{A_i} - f_{B_i})^2}$$

Gauname atstumo matricas ir panaudojome jas sukurti medžiams, naudodamiesi neighbour joining metodu.

3. Rezultatai

Dažniausi kodonai ir dikodonai:

```
Lactococcus_phage
codon: L 0.10297267197806904
dicodon: LL 0.010560027698433307
KM389305.1
codon: L 0.09202417634510547
dicodon: LL 0.008812421317666807
NC_028697.1
codon: L 0.10135006655257653
dicodon: LL 0.010501376704872894
KC821626.1
codon: L 0.09952559794425776
dicodon: LK 0.0102477365436275
coronavirus
codon: L 0.11321952595936795
dicodon: LL 0.015211828262723912
adenovirus
codon: L 0.09906172839506173
dicodon: SS 0.01086739780658026
U18337.1
codon: I 0.10437680115273776
dicodon: LL 0.011562272396212672
herpesvirus
codon: R 0.09375433726578765
dicodon: RR 0.015918653576437587
```

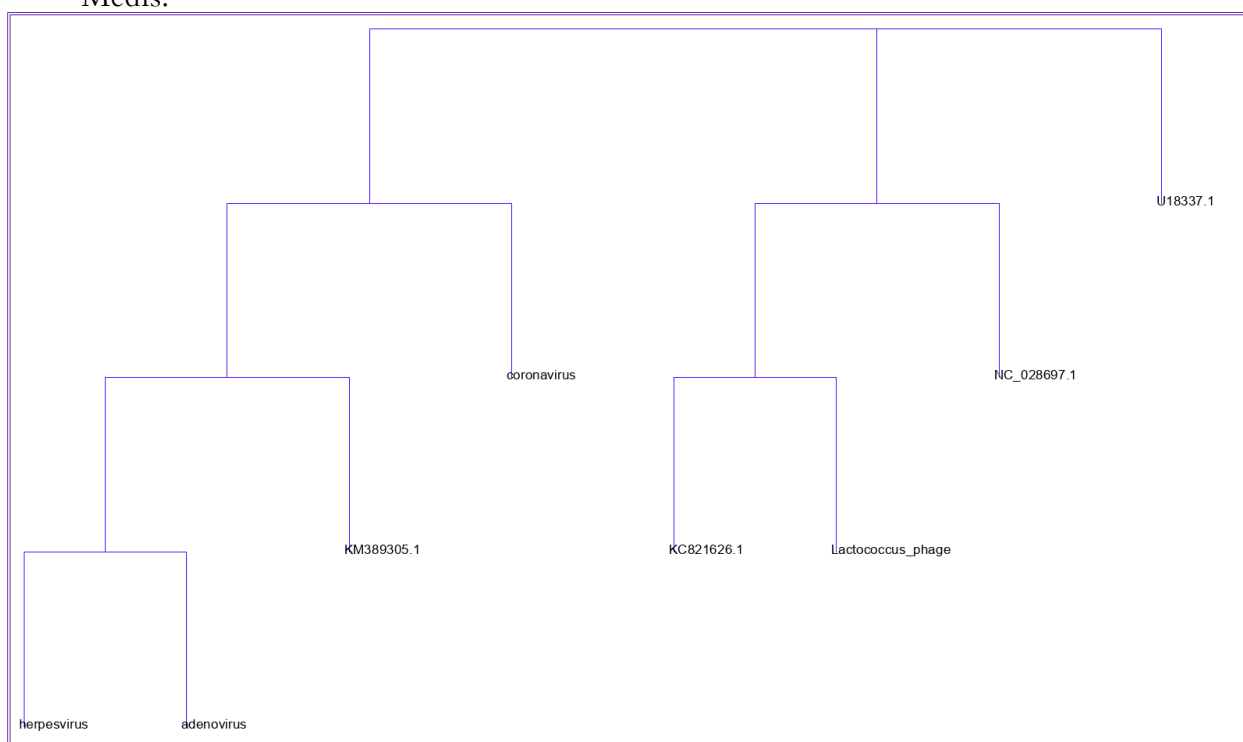
3.1. Codonų

Atstumų matrica:

| | | | | | | | | |
|-------------------|--------|--------|--------|--------|--------|--------|--------|--------|
| Lactococcus_phage | 0.0 | 0.0569 | 0.0293 | 0.0296 | 0.0570 | 0.0871 | 0.0654 | 0.1158 |
| KM389305.1 | 0.0569 | 0.0 | 0.0404 | 0.0585 | 0.0500 | 0.0438 | 0.0686 | 0.0698 |
| NC_028697.1 | 0.0293 | 0.0404 | 0.0 | 0.0316 | 0.0416 | 0.0668 | 0.0686 | 0.0950 |
| KC821626.1 | 0.0296 | 0.0585 | 0.0316 | 0.0 | 0.0627 | 0.0889 | 0.0688 | 0.1152 |
| coronavirus | 0.0570 | 0.0500 | 0.0416 | 0.0627 | 0.0 | 0.0604 | 0.0749 | 0.0856 |
| adenovirus | 0.0871 | 0.0438 | 0.0668 | 0.0889 | 0.0604 | 0.0 | 0.1052 | 0.0371 |
| U18337.1 | 0.0654 | 0.0686 | 0.0686 | 0.0688 | 0.0749 | 0.1052 | 0.0 | 0.1275 |
| herpesvirus | 0.1158 | 0.0698 | 0.0950 | 0.1152 | 0.0856 | 0.0371 | 0.1275 | 0.0 |

1 lentelė. Codon Distance Matrix

Medis:



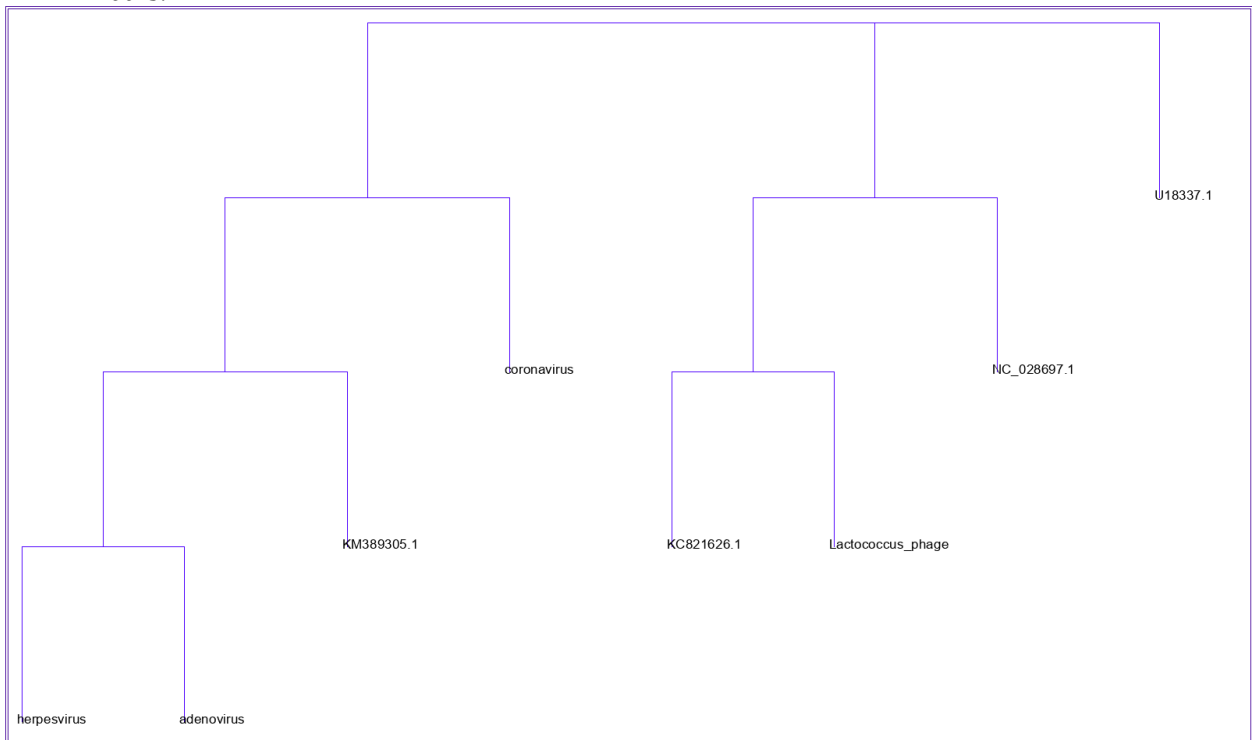
3.2. Dicodonų

Atstumų matrica:

| | | | | | | | | |
|-------------------|--------|--------|--------|--------|--------|--------|--------|--------|
| Lactococcus_phage | 0.0 | 0.0239 | 0.0166 | 0.0197 | 0.0248 | 0.0333 | 0.0271 | 0.0441 |
| KM389305.1 | 0.0239 | 0.0 | 0.0188 | 0.0252 | 0.0215 | 0.0196 | 0.0277 | 0.0301 |
| NC_028697.1 | 0.0166 | 0.0188 | 0.0 | 0.0197 | 0.0201 | 0.0270 | 0.0284 | 0.0377 |
| KC821626.1 | 0.0197 | 0.0252 | 0.0197 | 0.0 | 0.0268 | 0.0339 | 0.0293 | 0.0439 |
| coronavirus | 0.0248 | 0.0215 | 0.0201 | 0.0268 | 0.0 | 0.0244 | 0.0291 | 0.0345 |
| adenovirus | 0.0333 | 0.0196 | 0.0270 | 0.0339 | 0.0244 | 0.0 | 0.0385 | 0.0186 |
| U18337.1 | 0.0271 | 0.0277 | 0.0284 | 0.0293 | 0.0291 | 0.0385 | 0.0 | 0.0471 |
| herpesvirus | 0.0441 | 0.0301 | 0.0377 | 0.0439 | 0.0345 | 0.0186 | 0.0471 | 0.0 |

2 lentelė. Dicodon Distance Matrix

Medis:



4. Išvados

Dažniausias Kodonas visuose bakterijose buvo L, dikodonas LL išskyrus KC821626.1 kur dažniausias LK dicodonas.

Virusuose daugiau skirtumu. Corona ir adeno virusų dažniausias taip pat L kodonas, tačiau U18337.1 dažniausias I, o herpesvirus R. Idomu tai kad adenovirus ir U18337.1 dažniausiuose dikodonuose nėra jų atitinkamų dažniausių kodonų, kai visuose kituose ir bakterijose ir virusuose dažniausias kodonas yra ir dažniausiuose kodonų porose.

Pagal atstumus tarp virusų ir bakterijų iš bakterijų labiausiai išsiskyrė KM389305.1 kuris buvo didesniu atstumu nuo kitų bakterijų (net arčiau adenoviruso nei Lactococcus ir KC821626.1). Iš virusų coronavirus ir U18337.1 buvo arčiau bakterijų nei kitų virusų, ir U18337.1 turėjo didžiausią atstumą nuo kitos artimiausios sekos.