



Vilniaus universitetas

Matematikos ir informatikos fakultetas

Informatikos katedra

Programų sistemų studijų programa

Bioinformatika

Pirmojo laboratorinio darbo ataskaita

Ataskaitą tikrino: Prof. Dr. Gediminas Alzbutas

Ataskaitą parengė: Simonas Nausėda

Vilnius

Išvadas

Laboratorinio darbo tikslas: Įvertinti kodonų ir dikodonų dažnio skirtumus žinduolių ir bakterijų virusuose.

Laboratorinio darbo uždutis:

1. Pateiktoje sekoje fasta formatu surasti visas start ir stop kodonų poras, tarp kurių nebūtų stop kodono (ir tiesioginei sekai ir jos reverse komplementui).
2. Kiekvienam stop kodonui parinkti toliausiai nuo jo esantį start kodoną (su sąlyga, kad tarp jų nėra kito stop kodono)
3. Atfiltruoti visus fragmentus, kurie trumpesni nei 100 fragmentų.
4. Parašyti funkcijas, kurios įvertintų kodonų ir dikodonų dažnius.
5. Palyginti kodonų bei dikodonų dažnius tarp visų sekų.
6. Įvertinti, ar bakteriniai ir žinduolių virusai sudaro atskirus klasterius vertinant kodonų/dikodonų dažnių aspektu.

Atstumo funkcijos skaičiavimo aprašymas

Atstumų matricai formuoti buvo pasirinktas toks algoritmas: kiekvienos sekos kiekvieno kodono ar dikodono dažnis buvo atimamas iš kitos sekos kodono ar dikodono dažnio ir pasiimtas šio rezultato modulis ir pridamas prie bendro taškų skaičiaus.

Pavyzdys: tarkime turime tokius kodonų dažnius sekose:

1 sekoje: AAA – 1.344, TTT – 0.324...

2 sekoje: AAA – 0.344, TT – 2.324...

Imtume pirmos sekos kodoną AAA ir atliktume tokį veiksmą:

total = 0

$|AAA(1 \text{ sekos}) - AAA(2 \text{ sekos})|$, t.y. $|1.344 - 0.344| = 1$, total += 1

$|TTT(1 \text{ sekos}) - TTT(2 \text{ sekos})|$, t.y. $|0.324 - 2.324| = 2$, total += 2

atlikus šiuos veiksmus gautume, kad bendra suma yra 3, o perėjus per visus kodonus esančius sekoje ir sudėjus rezultatą gautume galutinį kodonų dažnio rezultatą.

Vadinasi sekos, kurios yra identiškos kodonų ar dikodonų dažniu turės rezultatą 0. Sekos, kurios turi kuo didesnę rezultatą yra vis mažiau panašios viena į kitą.

Gauta atstumų matrica:

1. Kodonams

8

Lactococcus_phage	0.00	44.26	22.47	38.98	27.52	51.69	35.90	71.86
KM389305.1	44.26	0.00	34.05	69.75	39.22	27.35	46.43	39.58
NC_028697.1	22.47	34.05	0.00	46.43	21.73	41.48	32.76	61.79
KC821626.1	38.98	69.75	46.43	0.00	49.99	76.73	47.92	92.65
coronavirus	27.52	39.22	21.73	49.99	0.00	42.92	37.00	65.34
adenovirus	51.69	27.35	41.48	76.73	42.92	0.00	53.78	25.31
U18337.1	35.90	46.43	32.76	47.92	37.00	53.78	0.00	70.78
herpesvirus	71.86	39.58	61.79	92.65	65.34	25.31	70.78	0.00

2. Dikodonams

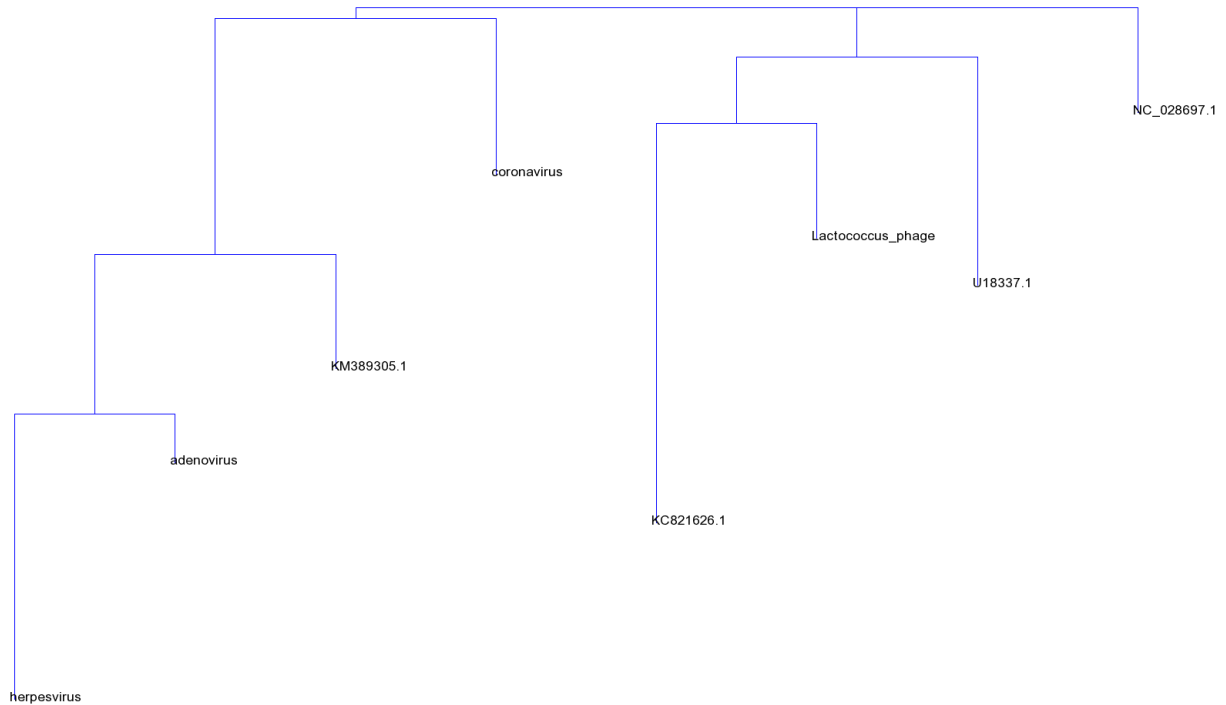
8

Lactococcus_phage	0.00	99.32	82.32	99.19	88.73	106.42	93.14	127.48
KM389305.1	99.32	0.00	89.73	122.26	95.07	80.84	99.12	94.25
NC_028697.1	82.32	89.73	0.00	102.78	81.53	93.35	92.59	113.25
KC821626.1	99.19	122.26	102.78	0.00	103.72	125.79	103.87	142.72
coronavirus	88.73	95.07	81.53	103.72	0.00	93.26	93.28	114.08
adenovirus	106.42	80.84	93.35	125.79	93.26	0.00	105.21	82.85
U18337.1	93.14	99.12	92.59	103.87	93.28	105.21	0.00	123.65
herpesvirus	127.48	94.25	113.25	142.72	114.08	82.85	123.65	0.00

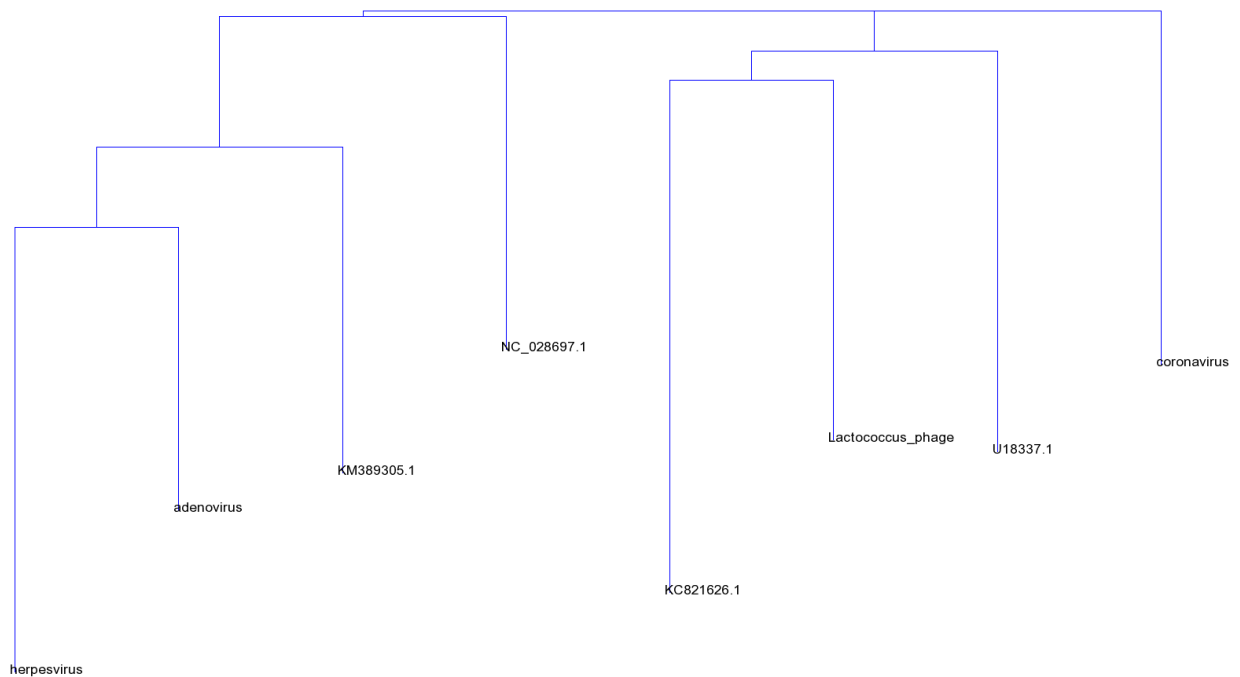
Atstumų matricos pateiktos Phylip formatu, kad vėliau būtų galima gauti atitinkamus medžius, rodančius atitinkamą klasterizavimą neighbour joining metodu.

Gauti medžiai:

1. Medis vaizduojantis kodonų klasterizavimą:



2. Medis vaizduojantis dikodonų klasterizavimą:



Iš gautų rezultatų labiausiai išsiskiria penki virusai: herpevirus, coronavirus ir NC_028697.1, U18337.1 ir KC821626.1.

Herpesvirusas labiausiai išsiskiria tuo, kad jis visiškai nepanašus tiek į žinduolių, tiek į bakterijų virusus. O Coronavirusas ir NC_028697.1 abu yra panašūs į visus lygintus tiek bakterijų, tiek žinduolių virusus.

Pirmame medyje, matomas įdomus dalykas, kad žinduolių virusas U18337.1 yra grupuojamas labiau prie bakterinių virusų, o bakterinis virusas KM389305.1 yra labiau grupuojamas prie žinduolinių.

Likusių virusų kodonų bei dikodonų dažniai nenustebino. Bakteriniai virusai Lactococcus_phage ir KC821626.1 labiau panašūs su kitais bakteriniais virusais, o taip ir buvo grupuojami, taip pat matome tokią pačią situaciją tarp žinduolinių virusų – jie labiau panašūs su kitais žinduolių virusais, nors ir čia yra tam tikrų išimčių, pavyzdžiui KM389305.1.

Herpevirusas pasiekė didžiausią kodonų dažnio balų skaičių, kai buvo lyginamas su KC821626.1 bakteriniu virusu, o žemiausią kodonų dažnio balų skaičių pavyko pasiekti bakteriniam virusui NC_028697.1, kai buvo lyginamas su coronavirusu.

Dikodonų dažnių lentelėje didžiausią dažnio balų skaičių pasiekė herpeviruso ir KC821626.1 palyginimai, o žemiausią - adenovirus ir KM389305.1.