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: Mantas Jakaitis

Vilnius

Įvadas

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

Laboratorinio darbo užduoties formuluotė:

1. Pateiktoje sekoje fasta formatu surastu visas start ir stop kodonų poras, tarp kurių nebūtu stop kodono (ir tiesioginei sekai ir jos reverse komplementui).
2. Kiekvienam stop kodonui parinkti toliausiai nuo jo esanti start kodoną (su sąlyga, kad tarp jų nėra kito stop kodono)
3. Atfiltruokite visus fragmentus ("tai butu baltymų koduojančios sekos"), kurie trumpesni nei 100 fragmentų.
4. Parašykite funkcijas, kurios įvertintu kodonu ir dikodonu dažnius (visi įmanomi kodonai/dikodonai ir jų atitinkamas daznis - gali būti nemažai nuliu, jei jų sekoje nerasite).
5. Palyginkite kodonų bei dikodonų dažnius tarp visu seku (atstumu matrica - kokia formule naudosite/kaip apskaičiuosite - parašykite ataskaitoje).
6. Įvertinkite, ar bakteriniai ir žinduolių virusai sudaro atskirus klasterius vertinant kodonu/dikodonu dažniu aspektu.

Laboratorinio darbo eigos aprašymas

Kadangi pagrindinis laboratorinio darbo tikslas yra išgauti matricas, kurios lygintų kodonų bei dikodonų dažnius skirtingose DNR sekose, pirmiausia reikėjo susirasti kiekvienos sekos visas įmanomas kombinacijas, t.y. tiesioginei sekai ir jos reverse komplementui (tokių sekų iš viso buvo 6). Vėliau reikėjo pereiti per kiekvieną kombinaciją ir iš jos ištraukti atitinkamus ORFus, tai buvo padaryta paprasčiausiai surandant ORFo pradžia, kuri visada bus ATG ir ieškant bet kurio iš trijų stop kodonų. Vėliau buvo galima pasiimti ORFus ir jų viduje surasti kodonų bei dikodonų dažnius. Kai turime visų DNR sekų kodonų bei dikodonų dažnius galima buvo pradėti formuoti atstumų matricą. Atstumų matricai formuoti buvo pasirinktas paprastas algoritmas: vienos sekos kiekvieno kodono ar dikodono dažnis buvo atimamas iš kitos sekos kodono ar dikodono dažnio ir pasiimtas šio skaičiavimo rezultato modulis bei buvo pridedamas prie bendro taškų skaičiaus.

Pavyzdys: tarkime turime tokius kodonų dažnius sekose:
1 sekoje: GGG – 1.244, AAA – 1.324…

2 sekoje: GGG – 0.889, AAA – 4.213…

Imtume pirmos sekos kodoną GGG ir atliktume tokį veiksmą:

total = 0

|GGG(1 sekos) – GGG(2 sekos)|, t.y. |1.244 – 0.889| = 0.355, total += 0.355

|AAA(1 sekos) – AAA(2sekos)|, t.y. |1.324 – 4.213| = 2.889, total += 2.889

atlikus šiuos veiksmus gautume, kad bendra suma yra 3.244, 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 didesnį rezultatą yra vis mažiau panašios viena į kitą.

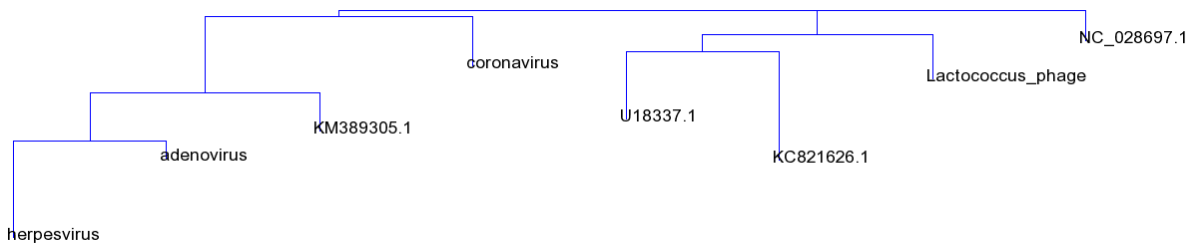
Gauta atstumų matrica:

1. Kodonams
   8
   Lactococcus_phage 0.00 39.63 19.45 31.65 24.70 45.81 31.08 65.17
   KM389305.1 39.63 0.00 30.74 59.45 34.78 24.49 46.54 35.98
   NC_028697.1 19.45 30.74 0.00 37.23 19.00 36.69 31.40 55.39
   KC821626.1 31.65 59.45 37.23 0.00 42.82 64.72 36.97 80.38
   coronavirus 24.70 34.78 19.00 42.82 0.00 38.30 35.17 58.44
   adenovirus 45.81 24.49 36.69 64.72 38.30 0.00 52.33 23.82
   U18337.1 31.08 46.54 31.40 36.97 35.17 52.33 0.00 67.14
   herpesvirus 65.17 35.98 55.39 80.38 58.44 23.82 67.14 0.00
2. Dikodonams
   8
   Lactococcus_phage 0.00 86.92 71.36 86.06 76.91 92.96 79.76 112.65
   KM389305.1 86.92 0.00 78.16 106.78 82.74 71.31 86.84 82.85
   NC_028697.1 71.36 78.16 0.00 88.64 69.85 80.87 78.61 99.80
   KC821626.1 86.06 106.78 88.64 0.00 90.47 109.12 85.68 125.15
   coronavirus 76.91 82.74 69.85 90.47 0.00 80.46 79.23 101.12
   adenovirus 92.96 71.31 80.87 109.12 80.46 0.00 93.23 73.77
   U18337.1 79.76 86.84 78.61 85.68 79.23 93.23 0.00 108.94
   herpesvirus 112.65 82.85 99.80 125.15 101.12 73.77 108.94 0.00

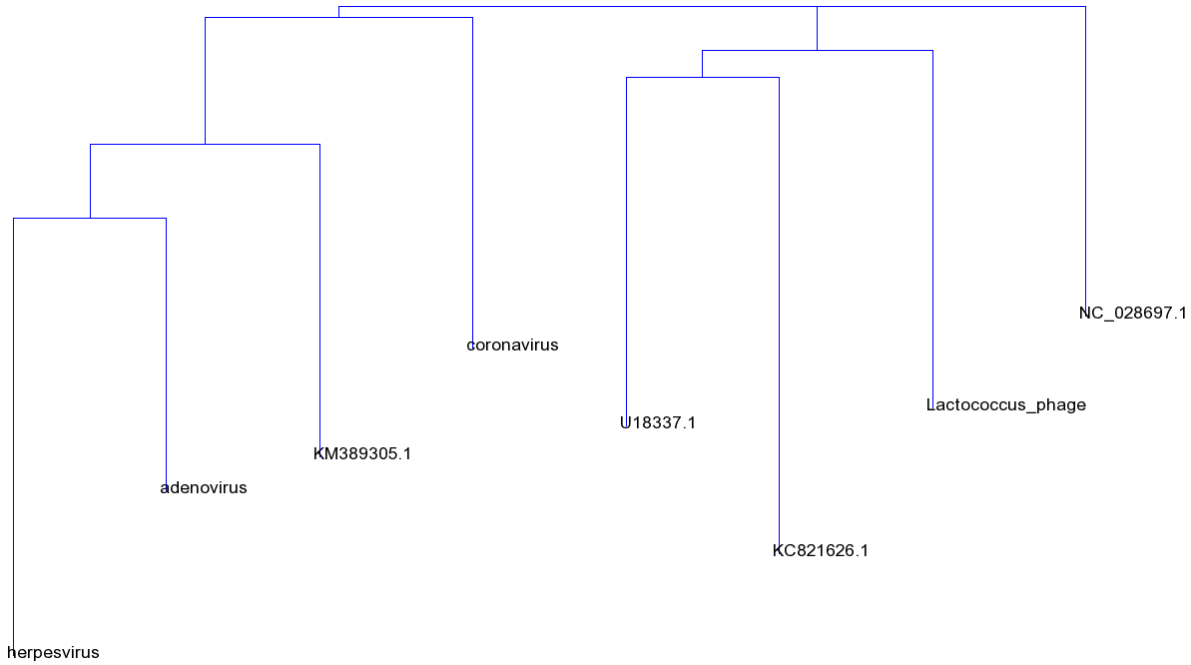
Atstumų matricos pateiktos Phylip formatu, kad vėliau būtų galima gauti atitinkamus medžius, kurie vaizduos atitinkamą klasterizavimą neighbour joining metodu.

Gauti medžiai:

1. Medis vaizduojantis kodonų klasterizavimą:



2. Medis vaizduojantis dikodonų klasterizavimą:



Iš gautų rezultatų man labiausiai išsiskiria trys virusai: herpevirus, coronavirus, NC_028697.1, U18337.1 ir KM389305.1.

Herpevirusas labiausiai išsiskiria tuo, kad jis nepanašus tiek į žinduolių, tiek į bakterijų virusus. O Coronavirusas ir NC_028697.1 yra panašūs tiek į bakterijų, tiek į žinduolių virusus. Taip pat patebime, kad U18337.1 virusas yra panašesnis į bakterinius virususir grupuojamas šalia tokių, nors pats yra žinduolinis, o KM389305.1 atvirkščiai – grupuojamas prie žinduolinių ir panašesnis į tokius, nors pats yra bakterinis.

Likusių virusų kodonų bei dikodonų dažnių klasterizavimą galima buvo nuspėt, t.y. bakteriniai virusai Lactococcus_phage, KC821626.1, NC_028697.1 buvo grupuojami labiau prie bakterinių virusų ir dažnių panašumai būtent ir rodė, kad jie panašesni į bakterinius nei į žinduolinius, o herpesvirus, adenovirus ir coronavirus buvo panašiausi į žinduolinius ir būtent taip ir grupuojami.

Herpevirusas pasiekė didžiausia kodonų dažnio balų skaičių, kai buvo lyginamas su KC821626.1 bakteriniu virusu, o žemiausią[1] kodonų dažnio balų skaičių pavyko pasiekti NC_028697.1 bakteriniui virusui, kai šis buvo lyginamas su coronavirusu.

Žiūrint į dikodonų dažnių lentelę matome, jog didžiausią dažnio balų skaičių pasiekė herpeviruso ir NC_028697.1 palyginimai, o žemiausią - NC_028697.1 ir coronavirusas.

[1] Žemiausias kodonų dažnio balų skaičius laikomas tarp dviejų skirtingų virusų, nes nebūtų prasmės aprašyti lyginimą tarp dviejų vienodų, kadangi toks visada bus 0.