ATASKAITA LAB1

Repozitorija – https://github.com/jokubassilgalys/bio-info-lab1

Programos paleidimas: py find_orfs.py viruses/data/

Paleidžiama per komandinę eilutę paduodant vieną fasta formato failą arba path direktorijos su keliais fasta failais. Vieno failo atveju, duotame genome randamos start/stop poros, jos filtuojamos ir sekos ilgesnės nei 100bp konvertuojamos į baltymus, rezultatai rodomi konsolėje. Kelių failų atveju, procesas vykdytas vienam failui yra padaromas visiem aplanke rastiem fasta failams ir papildomai apskaičiuojami baltymų dažniai (kodonams ir dikodonams atskirai). Pagal tai sudaromos dvi dažnių matricos ir išsaugomos codon_distance.phy ir dicodon_distance.phy failuose (išsaugomi ten pat kur randasi find_orfs.py failas).

Start/stop porų radimas

Seka iš fasta failo yra skaitoma 6 kartus, 3 kartai tiesioginei sekai ir 3 jos reverse komplementui. Kiekvieną kartą skaitant seka ieškoma pabaigos kodono (TAA, TAG, TGA), jį radus judama atgal ieškant tolimiausio pradžios (ATG) kodono.
Programoje taip pat yra alternatyvi start/stop porų ieškojimo funkcija (--method alternative programos paleidimo argumentuose), kuri vieną kartą perskaičius seką įsimena kiekvieną start/stop kodoną ir jo pozicija ir tada iš šio sąrašo surenka ilgiausias poras, tokiu būdu panaikindamas važinėjimą seka atgal. Abi funkcijos gražina vienodus rezultatus, bet alternatyvi funkcija šiek tiek greitesnė.

Atstumo funkcijos apskaičiavimas

Kiekvienam virusui sudaromi du dažnių vektoriai: kodonų (21 aminorūgšties, įskaitant stop „*“) ir dikodonų (visų galimų aminorūgščių porų, 21x21).
Atstumui tarp baltymų skaičiuoti naudojama Euklidinė atstumo funkcija, taikoma normalizuotiems kodonų ir dikodonų dažnių vektoriams.

Atstumo formulė:

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

$d(A,\ B)$ – atstumas tarp A ir B virusų
n – dažnių vektorių ilgis. Kodonų - 21, dikodonų - 441.
f_A ir f_B – dažnių vektoriai virusui A ir B

Implementacija kode:

Normalizacija:

```
freq_arrays = [np.array([f[k] for k in keys]) / sum(f.values()) for f in frequency_list]
```

Atliekama atstumo matricos apskaičiavimo funkcijoje. Užtikrina, kad ilgesni baltymai neturėtų didesnės įtakos rezultatui, taip pat vienodą amino rūgščių išrykiavimą kiekviename dažnio vektoriuje.

Matricos sudarymas:

```
n = len(freq_arrays)
matrix = np.zeros((n, n))
for i in range(n):
    for j in range(n):
        if metric == 'euclidean':
            matrix[i, j] = np.linalg.norm(freq_arrays[i] - freq_arrays[j])
```

Kiekvienai i ir j virusų porai (atitinka A ir B porą iš aukščiau esančios formulės) apskaičiuojamas atstumas pagal Euklidinę atstumo formulę. Šia funkciją galima būtų paoptimizuoti simetrišku matricos užpildymu.

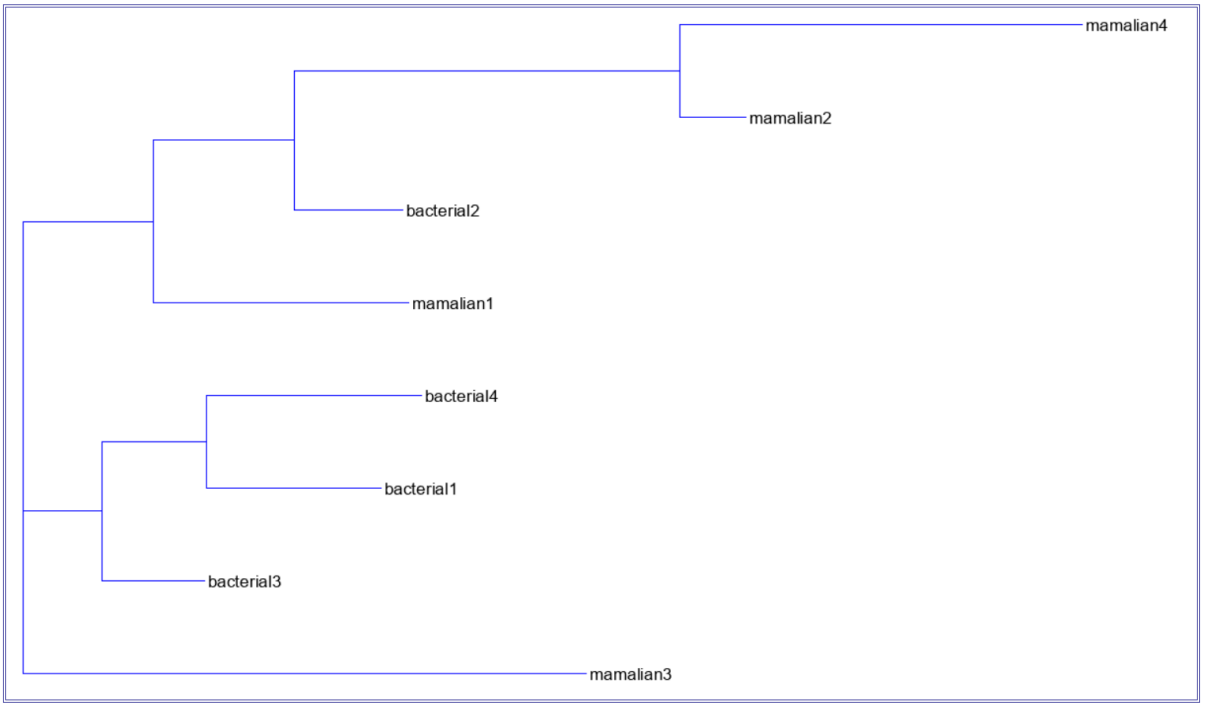
Rezultatai

kodonų matrica (codon_distance.phy):
8
bacterial1 0.000 0.056 0.029 0.030 0.056 0.086 0.065 0.114
bacterial2 0.056 0.000 0.039 0.059 0.050 0.043 0.067 0.069
bacterial3 0.029 0.039 0.000 0.033 0.042 0.066 0.068 0.094
bacterial4 0.030 0.059 0.033 0.000 0.062 0.088 0.069 0.114
mamalian1 0.056 0.050 0.042 0.062 0.000 0.060 0.074 0.085
mamalian2 0.086 0.043 0.066 0.088 0.060 0.000 0.103 0.036
mamalian3 0.065 0.067 0.068 0.069 0.074 0.103 0.000 0.125
mamalian4 0.114 0.069 0.094 0.114 0.085 0.036 0.125 0.000

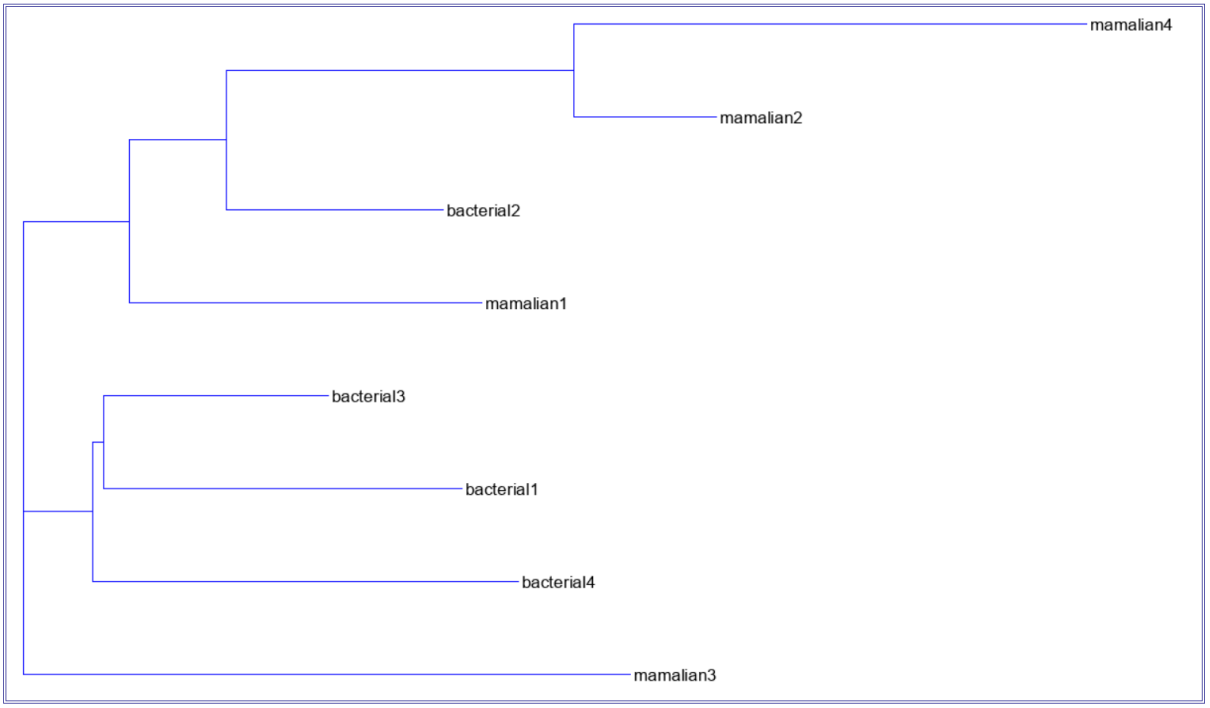
dikodonų matrica (dicodon_distance.phy):
8
bacterial1 0.000 0.023 0.016 0.020 0.025 0.033 0.027 0.043
bacterial2 0.023 0.000 0.018 0.025 0.022 0.019 0.027 0.030
bacterial3 0.016 0.018 0.000 0.020 0.020 0.026 0.028 0.037
bacterial4 0.020 0.025 0.020 0.000 0.027 0.033 0.029 0.043
mamalian1 0.025 0.022 0.020 0.027 0.000 0.024 0.029 0.034
mamalian2 0.033 0.019 0.026 0.033 0.024 0.000 0.037 0.018
mamalian3 0.027 0.027 0.028 0.029 0.029 0.037 0.000 0.046
mamalian4 0.043 0.030 0.037 0.043 0.034 0.018 0.046 0.000

Bakterinių virusų tarpusavio atstumai yra nuo 0,029-0,059, žinduolių – 0,036-0,125. Bakterinių ir žinduolių virusų tarpusavio atstumai vidutiniškai 0,072, kas yra šiek tiek daugiau nei grupės viduje. Nors ir dikodonų matricoje atstumai yra mažesni, dėsningumas tas pats.

kodonų medis:



dikodonų medis:



Medžiai gauti naudojant https://www.trex.uqam.ca/index.php?action=trex&menuD=1&method=2, neighbor joining metodą ir proportional edge lengths.

Kodonų ir dikodonų medžiai šiek tiek skiriasi, šaka su bacterial1, bacterial3 ir bacterial4 grupuojama skirtingai. Kadango dikodonų analizei svarbi ne tik amino rūgščių bendra

visuma, bet ir lokalus kontekstas (kokia rūgštis šalia jo), tokių ne didelių skirtumų tarp medžiu galima buvo tikėtis.

Labiausiai išsiskiriantis virusas – mamalian3, kuris sudaro atskirą šaką abiejuose medžiuose.