Vilniaus universitetas
Matematikos ir informatikos fakultetas

Bioinformatika

Pirmojo laboratorinio darbo ataskaita

Dėstytojas: Gediminas Alzbutas
Studentas: Andrius Bertašius

Turinys

Tikslas

Šio laboratinio tikslas yra susipažinti su fasta formatu. Pateiktose 4 žinduolių ir 4 bakterijų virusų sekose rasti visas START ir STOP kodonų poras, tarp kurių nebūtų kito STOP kodono. Kiekvienam STOP kodonui rasti jam priklausantį ir toliausiai nutousį START kodoną. Atsirinkti tik tas sekas kuios yra netrumpesnės nei 100 simbolių. Iš turimų sekų suskaičiuoti visų kodono, dikodonų, amino rūgčio, diaminų dažnius. Turint visus dažnius palyginti juos tarp skiritngų sekų – naudojant kažkokią formulę, kuri galėtu nusakyti skirtumus, bei iš gautų rezultatų sudaryti atstumų matricą, kodonams, dikodonams, amino rūgštims ir diaminams. Galiausiai įvertinti ar bakteriniai ir žinduoliniai virusai sudaro atskirus klasterius, vertinant ankščiau gautas matricas.

Darbo eiga

Start ir stop kodonų poros

Jog rasti visas START ir STOP kodonų poras, visų pirmiausiai radau pozicijas kuriose yra START (ATG) ir STOP (TAA, TAG, TGA) kodonai. Tuomet priklausomai nuo to kokia liekana gaunama kodono poziciją padalinus iš trejeto, pridėjau minėtą kodoną į atskirą grupę:

- Nulinę – jei liekana yra 0;
- Pirmą – jei liekana yra 1;
- Antrą – jei liekana yra 2;

Ir tuomet eidamas per kiekvieną grupę kiekvienam STOP kodonui priskyriau visus galimus START kodonus, taip kad tarp šios poros nebūtų kito STOP kodono.

Toliausia start ir stop kodonų pora

Kadangi praeitame žingsnyje kiekvienam STOP kodonui priskyriau visus galimus START kodonus. Šiame etape reikėjo praeitį per visų porų sąrašą ir kiekvienam STOP kodonui, parinkti mažiausią skaičių – kuris reiškia – toliausiai nutolusį START kodoną. Ir taip sudarytos toliausios START ir STOP kodonų poros kiekvienai grupei.

Netrumpesnės sekos

Šis žingsnis yra salyginai paprastas – praeiti pro kiekvienos grupės kodonų poras ir išmesti tas poras, tarp kurių atstumas yra mažesnis už 100 simbolių.

Dažniai

Norint surasti dažnį reikia žinoti visų elementų kiekį, ir ieškomo elemento pasikartojimų skaičių. Vadinasi man reikėjo praeiti pro kiekvienos grupės START ir STOP kodonų poras, susiskaičiuoti kiek kartų pasikartojo tam tikras kodonas, dikodonas, amino rūgtis ir diaminas. Buvo galima tuo pačiu skaičiuoti kiek yra iš viso elementų, tačiau aš nusprendžiau sužinojęs kiekvieno elemento pasikartojimų skaičių juos visus sudėti ir taip gauti bendrą kiekį visų kodonų, dikonų, amino rūgščių bei diamonų/ Tad paskutinis žingsnis šiame etape buvo tiesiog kiekvieno kodono, dikodono, amino

rūgšties ir diamino pasikartojimo kiekį padalinti iš atitinkamo bendro elmementų kiekio. Dažnių rezultatai gali būti rasti laboratorinio darbo GitHub repositorijoje[1].

Sekų palyginimas

Palyginti visas sekas yra irgi ganėtinai trivialus dalykas, lyginimui aš naudojau paprasčiausią ir tikriausiai visiem labiausiai pažįstama – Euklido atstumo formule (Fig. 1). Ši atstumo formulė labiausiai sutinkama ieškant atstumo tarp dviejų taškų dvimatėje arba trimatėje erdvėje.

$$d = |x - y| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

Fig. 1: Euklido atstumo bendrinė funkcija

Laboratorinio darbo atvėjų atstumas buvo ieškomas tarp kiekvienos sekos dažnių atitinkamose kategorijose – kodonai, dikodonai, amino rūgštys ir diaminai.

Gauti rezultatai buvo sudėti į atstumų matricas Phylip formatu[2]. Kaip atrodo matrica galima pamatyti 1-oje lentelėje.

1. Lentele: Atstumu matricos pavyzdys

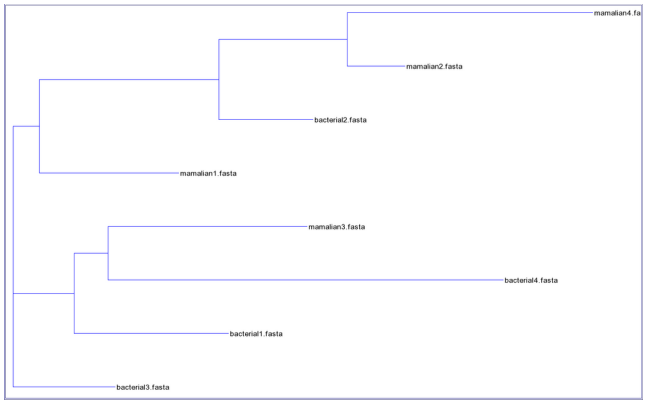
	Bacterial 1	Bacterial 3	Bacterial 2	Bacterial 4	Mamalian1	Mamalian2	Mamalian3	Mamalian4
Bacterial 1	0,0000	0,0692	0,0395	0,0710	0,0497	0,0835	0,0608	0,1112
Bacterial 2	0,0692	0,0000	0,0539	0,1125	0,0600	0,0403	0,0731	0,0607
Bacterial 3	0,0395	0,0539	0,0000	0,0793	0,0379	0,0647	0,0587	0,0931
Bacterial 4	0,0710	0,1125	0,0793	0,0000	0,0873	0,1250	0,0816	0,1483
Mamalian1	0,0497	0,0600	0,0379	0,0873	0,0000	0,0650	0,0645	0,0938
Mamalian2	0,0835	0,0403	0,0647	0,1250	0,0650	0,0000	0,0892	0,0408
Mamalian3	0,0608	0,0731	0,0587	0,0816	0,0645	0,0892	0,0000	0,1122
Mamalian4	0,1112	0,0607	0,0931	0,1483	0,0938	0,0408	0,1122	0,0000

Klasterizavimas

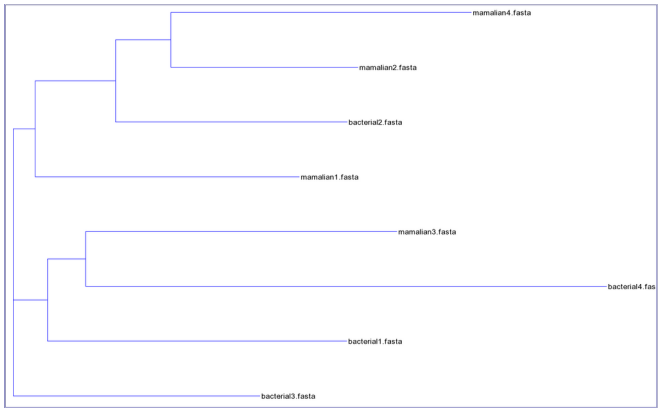
Turint reikiamo formato matricas jas tik reikia sudėti į atitinkamą įrankį kuris sugeneruoją medį, rodantį klasterizavimą neighbour-joining metodu. Buvo pasirinktas dėstytojo pasiulytas internetinis

1 https://github.com/DuckDrick/uni-bioinformatika/
2 https://en.wikipedia.org/wiki/PHYLIP#File_format

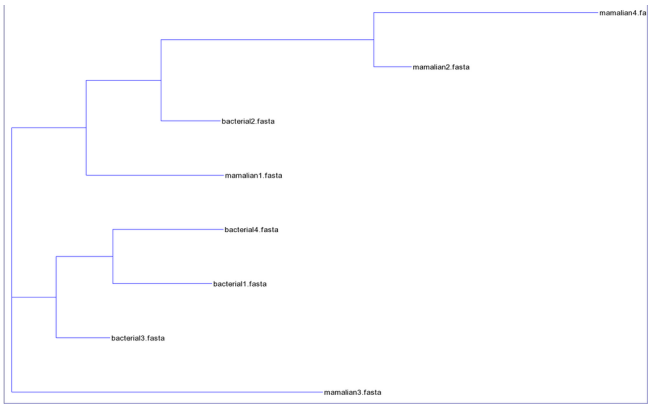
įrankis[3]. Žemiau pateikti visi keturi klasterizacijos medžiai – kodonų, dikodonų, amino rūgščių bei diaminų, atitinkamai 1, 2, 3 ir 4-as paveikslėliai.



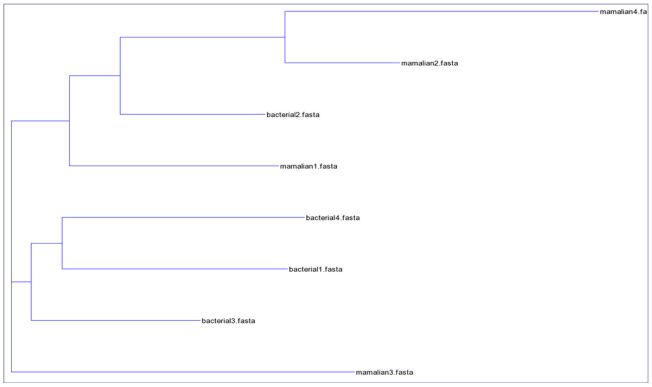
2 pav. Kodonų klasterizacijos medis



1 pav. Dikodonų klasterizacijos medis



3 pav. Aminų klasterizacijos medis



4 pav. Diaminų klasterizacijos medis

Iš medžių matome, kad kodonų ir dikodonų atvėju (toliau – K-atvėjas) į atskirą šaką išsišakoja *bacterial3.fasta*, o amino ir diaminų atvėju (toliau – A-atvėjas) – *mamalia3.fasta*. K-atvėju prie *mamalian* virusų kruvelės (1 šaka) įsiterpus *bacterial2.fasta*, o prie bakterinių virusų (2 šaka*)* įsiterpęs *mamalian3.fasta*. A-atvėju į mamalian krūvelę yra įsiterpęs *bactarial2.fasta,* bakterial šaka, sudaryta vien iš bakterinių virusų. Jeigu apkeistumėm *bacterial2.fasta* ir *mamalian3.fasta* vietomis būtų galima medžius dalinti į du klasterius, vienas būtų viršutinis – kuriame būtų mamalian virusai, o kitas apatinis su bakteriniais virusais.

Kodo sprendimai

Vienintelis dalykas kurį vertėtų paminėti, tai Bio[4] bibliotekos SeqIO klasė, jos dėka sekų nuskaitymas buvo lengvas ir greitas, reverse komplimentų gavimas yra tik nutolęs per vienos funkcijos iškvietimą, o kodonų sekos pavertimui reikia iškviesti tik vieną transfrom metodą. Ši biblioteka gali ir dar daugiau, bet dėja jos neišnagrinėjau, nes pasirodė greičiau rankomis parašyti paprastą ciklą negu eiti per ištisus puslapius dokumentacijos. O kalbant apie kitas kodo vietas, tai yra naudojamos paprastos pythono teikiamos galimybės kaip ciklai, sąrašų manipuliavimas, darbas su simbolių eilutėmis ir regex’ai.

3 http://www.trex.uqam.ca/index.php?action=trex&menuD=1&method=2
4 https://biopython.org

Kodo problemos

Kodas galetų būti labiau suskaidytas į funkcijas, šiuo metu jis kodo gabaliukai kartojasi 4 kartus. Ne visi kintamųjų vardai gali būti aiškiai suprantami. Pradinė funkcija, kuri iškviečiama sekai, nedaro tai ką ji sako – surasti kodonų poras, iš tiesų ji padaro vis ne viską iki dažnių skaičiavimo.