Vilniaus Universitetas  
Matematikos ir Informatikos fakultetas

**III Bioinformatikos laboratorinis darbas**

**Paruošė:**

**Edvinas Krupovnickas  
Programų Sistemų IV kurso studentas**

1. Apibūdinkite fastq formatą. (<https://en.wikipedia.org/wiki/FASTQ_format>). Kokia papildoma informacija pateikiam lyginant su FASTA formatu?  
  
**FASTQ** formatas yra sekų duomenų saugojimo ir dalijimosi formatas, kuris paprastai naudojamas NGS (Next Generation Sequencing) sekoskaitos duomenims. Kiekvienas įrašas FASTQ faile susideda iš keturių eilučių:

1. **Pavadinimas**: Eilutė prasideda simboliu "@" ir nurodo sekos atpažinimo informaciją (pvz., sekos ID, papildomos etiketės).
2. **Sekos duomenys**: Nukleotidų sekos atstovai, paprastai naudojami kaip A, T, G, C simboliai.
3. **Papildoma informacija**: Eilutė, prasidedanti simboliu "+", gali būti tuščia arba pateikti papildomą informaciją.
4. **Sekos kokybės**: ASCII koduoti kokybės parametrai, kurie atspindi kiekvienos bazės tikslumą.

**FASTA** formatas apima tik pirmas dvi eilutes: pavadinimą ir seką. **FASTQ** be šių dalykų taip pat pateikia **kokybės informaciją**, kuri leidžia įvertinti, kiek tikslūs yra duomenys kiekvienam nukleotidui.

2. Kurią mėnesio dieną Jūs gimėte? Prie dienos pridėkite 33. Koks ASCII simbolis

atitinka šį skaičių?  
  
Gimiau **spalio 3-ią** dieną.  
  
3 + 33 = **36**.  
  
**36-ių** ASCII atitikmuo yra **$**.

3. Kodėl pirmi 32 ASCII kodai negali būti naudojami sekos kokybei koduoti?

Pirmieji 32 ASCII kodai (nuo 0 iki 31) yra **kontroliniai simboliai**, kurie neturi apibrėžtų spausdinamų simbolių. Jie dažniausiai naudojami įvairiems valdymo veiksmams, pavyzdžiui, teksto eilutės pabaigai ar kitoms reikšmėms. Dėl šios priežasties jie negali būti naudojami sekos kokybės kodavimui, nes kokybės skalė turi naudoti spausdinamus simbolius, kurie gali būti aiškiai interpretuojami.

4. Parašykite skriptą, kuris:

a) nustatytu koks kokybės kodavimas yra naudojamas pateiktame faile. Galimos koduotės:

i. Sanger Phred+33

ii. Solexa Solexa+64

iii. Illumina 1.3+ Phred+64

iv. Illumina 1.5+ Phred+64

v. Illumina 1.8+ Phred+33

Parašykite, kokią koduotę nustatėte ir kuo remiantis?

A screen shot of a computer program

Description automatically generated

Rėmiausi kokybės reikšmių .fastq faile intervalais (kiekvienas kodavimas turi savo intervalą.)

Gavau kodavimą **Phred+33** (gali būti tiek **Sanger**, tiek **Illumina 1.8+**, kadangi turi tą patį intervalą).  
  
Bet žinant, jog plačiau šiais laikais naudojamas **Illumina 1.8+ Phred+33**, spėčiau šis kodavimas ir naudojamas.

b) analizuotų C/G nukleotidų pasiskirstymą read’uose. Pateikite grafiką, kurio y ašyje būtų read’ų skaičius, x ašyje - C/G nukletidų dalis read’o sekoje (100 proc.

Reikštų, kad visi simboliai read’o sekoje yra G ir C)

Parašykite, koks „stambių“ pikų skaičius yra gautame grafike? (tikrai mažiau nei 6) A computer screen shot of a program code

Description automatically generated A graph of a blue column

Description automatically generated with medium confidence

Stambių pikų skaičius grafike: **3** (sužymėti raudonai).

c) paimtų po 5 kiekvieno piko viršūnės sekų ir atliktų blast’o paieškas. Naudokite nr/nt duombazę, paiešką apribokite taip, kad ieškotų atitikmenų tik bakterinės

sekose (organizmas “bacteria”). Analizei naudokite tik patį pirmą atitikmenį. Pateikite lentelę, kurioje būtų read’o id ir rasto mikroorganizmo rūšis A screen shot of a computer program

Description automatically generated

Lentelės pateikti negaliu, nes BLAST‘as nulūžinėja nepasiekęs pabaigos... Geriau būtų naudoti lokalų BLAST‘ą, bet nt duombazė užima 180GB…

5. Kokių rūšių bakterijų buvo mėginyje?

**Escherichia coli (viena iš jų - mutavusi)**

**Thermus thermophilus**

**Klebsiella pneumoniae**

**Staphylococcus aureus**

**Visus Python script’us galima rasti “Bioinformatika3” GitHub repozitorijoje.**