Vilniaus universitetas

Matematikos ir informatikos fakultetas

Informatikos katedra

Programų sistemų studijų programa

Bioinformatika

Trečiojo laboratorinio darbo ataskaita

Ataskaitą tikrino: Prof. Dr. Gediminas Alzbutas

Ataskaitą parengė: Simonas Nausėda

Vilnius

Įvadas

Laboratorinio darbo tikslas: Panagrinėti FASTQ failų formatą. Taip pat parašyti script‘ą nagrinėjantį fastq failą, išanalizuoti nukleotidų pasiskirstymą read‘uose.

Užduotys

1. **Apibūdinkite fastq formatą. (https://en.wikipedia.org/wiki/FASTQ_format). Kokia papildoma informacija pateikiam lyginant su FASTA formatu?**

FASTQ yra tekstinio pagrindo (text-based) failų formatas skirtas saugoti biologines sekas (įpratai nukleotidines) ir jų kokybės įverčius. Sekos simbolis ir kokybės įvertis koduojami vienu ASCII simboliu. Šis formatas skiriasi nuo FASTA tuo, kad papildomai saugo sekos kokybės įverčius.

2. **Kurią mėnesio dieną Jūs gimėte? Prie dienos pridėkite 33. Koks ASCII simbolis atitinka šį skaičių?**

27d. 27 + 33 = 60. ASCII lentelėje skaičių 60 atitinka simbolis „<“

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

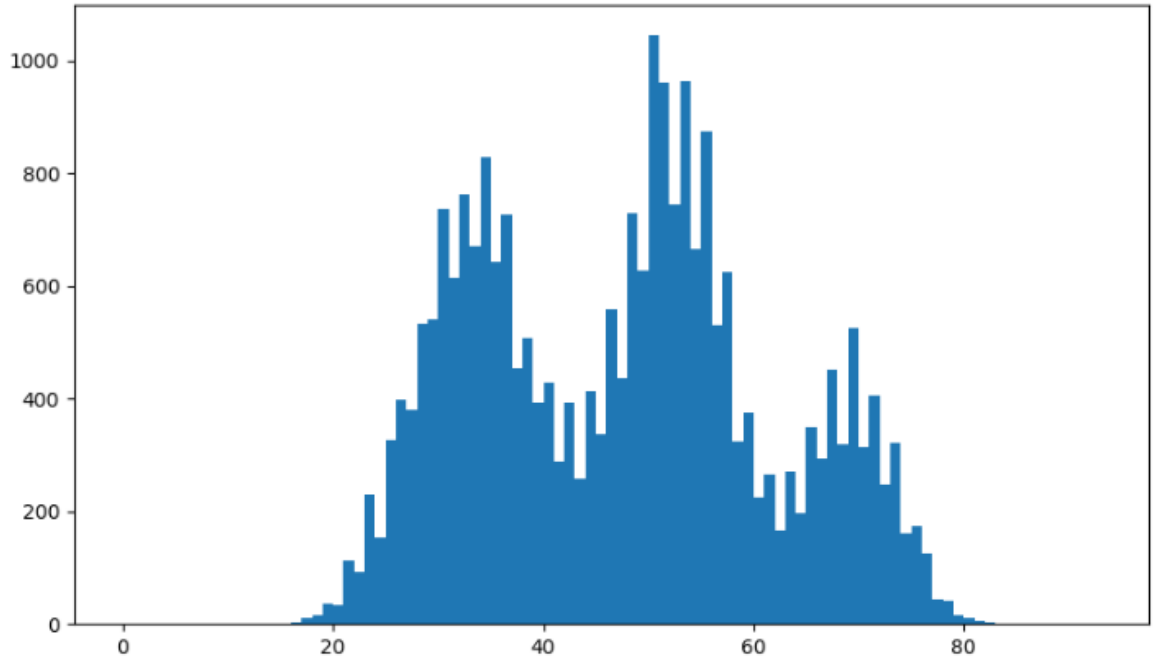
Šie kodai netinka sekos kokybei koduoti, nes jie nėra spausdinami, t.y jų neįmanoma atvaizduoti. Jie skirti įvairių signalų apie veiksmus siuntimui į terminalą.

4. **Parašykite skriptą, kuris:**
    a. **nustatytų 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?

Pateiktame faile naudojamas Sanger Phred+33 kokybės kodavimas. Tai buvo nustatyta naudojantis Python „bioinfokit“ bibliotekos analys.format.fq_qual_var() funkcija.

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)**



Iš grafiko matome, kad „stambių" pikų skaičius yra 3. Intervalai: [25:40], [45:58], [65, 75]

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.**

ID	Organizmas
@M00827:12:000000000-AEUNW:1:1101:15734:4405 1:N:0:6	Staphylococcus aureus strain IVB6168 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:14559:5316 1:N:0:6	Staphylococcus aureus strain IVB6168 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:16172:6679 1:N:0:6	Staphylococcus aureus strain IVB6168 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:21711:6712 1:N:0:6	Escherichia coli isolate KresCPE0301 genome assembly, plasmid: 1
@M00827:12:000000000-AEUNW:1:1101:12773:6933 1:N:0:6	Staphylococcus aureus strain IVB6154 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:18967:1954 1:N:0:6	Escherichia coli strain BM28 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:13519:2271 1:N:0:6	Escherichia coli strain BM28 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:18503:2566 1:N:0:6	Escherichia coli strain BM28 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:16860:5625 1:N:0:6	Escherichia coli strain BM28 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:19754:5804 1:N:0:6	Escherichia coli strain BM28 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:18070:3392 1:N:0:6	Thermus thermophilus HC11 DNA, complete genome
@M00827:12:000000000-AEUNW:1:1101:23350:4251 1:N:0:6	Thermus thermophilus strain N-1 chromosome, complete genome

@M00827:12:000000000-AEUNW:1:1101:23294:5998 1:N:0:6	Thermus thermophilus strain N-1 chromosome, complete genome
@M00827:12:000000000-AEUNW:1:1101:7922:8647 1:N:0:6	Thermus thermophilus HC11 DNA, complete genome
@M00827:12:000000000-AEUNW:1:1101:11245:8780 1:N:0:6	Thermus thermophilus HC11 DNA, complete genome

5. Kokių rūšių bakterijų buvo mėginyje?

Staphylococcus aureus, Escherichia coli, Thermus thermophilus.