Ataskaita

2 laboratorinis darbas

Darbą atliko: Aleksandra Kondratjeva

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

FASTQ įrašas (record) susidaro iš keturių eilučių:

- Pirma eilutė – prasideda „@" simboliu, sekos identifikatoriu ir gali turėti aprašymą (kaip ir FASTA pavadinimo eilutė).
- Antra eilutė – neapdorotos (raw) sekos raidės.
- Trečia eilutė – prasideda simboliu + ir gali turėti tokį pat sekos identifikatorių ir aprašymą.
- Ketvirta eilutė – koduoja kokybės reikšmes (quality values) sekai iš antros eilutes. Simbolių skaičius turi sutapti su 2 eilutėje esančios sekos simbolių skaičiumi. Kokybes reikšmės yra užkoduotos ASCII formatu su tam tikru poslinkiu (dažniausiai 33 ar 64).

Palyginimas su FASTA formatu:

- Abu formatai atvaizduoja sekas.
- Pagrindinis skirtumas tarp FASTQ ir FASTA yra tai, kad FASTQ taip pat pateikia kokybės reikšmes (quality values) kiekvienai sekos pozicijai, kas leidžia įvertinti sekvencijų duomenų patikimumą. FASTA šios galimybės neturi.
- FASTQ failai del kokybės reikšmių yra šiek tiek didesni.
- FASTA įrašas susidaro iš vienos eilutės, FASTQ iš keturių.

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

   Aš gimiau **20** dieną. ASCII simbolis – **'5'.**

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

   Pirmi ASCII simboliai tai operacinės sistemos specialieji simboliai, jie nespausdinami ir naudojami kad kontroliuoti įvairiais kompiuterio funkcijas. Reikšmių pvz.:

   00 - NUL (Null)

   01 - SOH (Start of Header)

   02 - STX (Start of Text)

03 - ETX (End of Text)

09 - TAB (Horizontal Tab)

10 - LF (Line Feed, New Line)

13 - CR (Carriage Return)

Todėl tai apsunkintų FASTQ failų skaitymą, atsirastų rizika neteisingai interpretuoti simbolius.

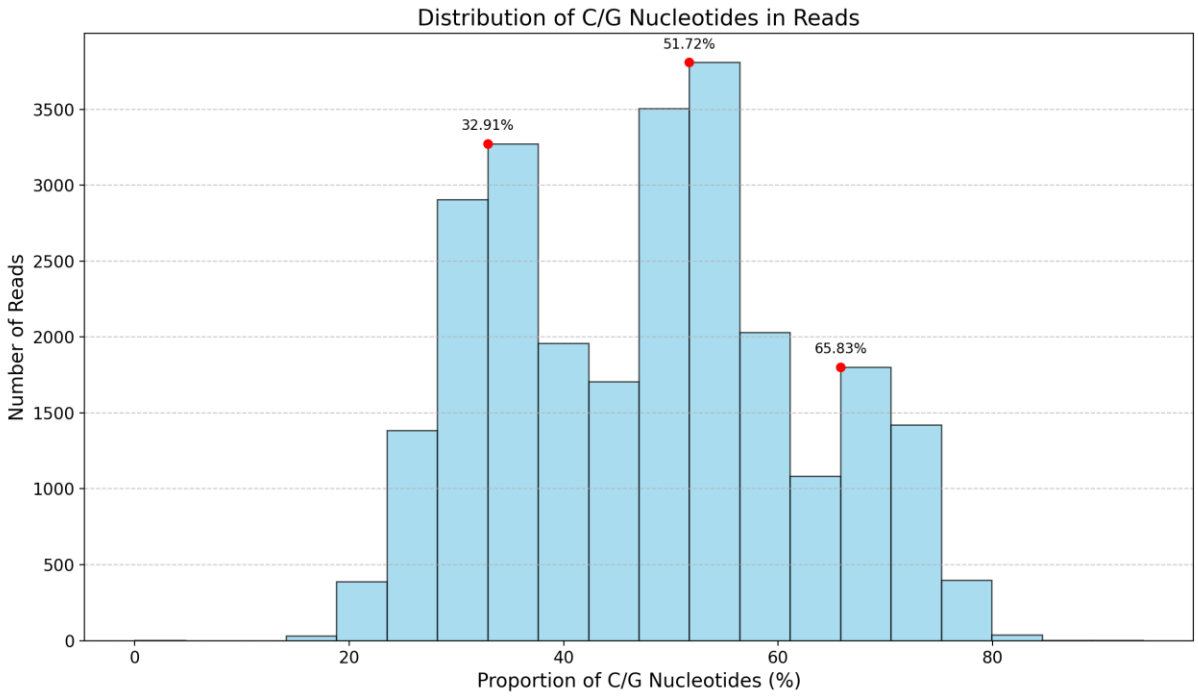
4. Skriptas įkeltas į mano github https://github.com/artnaxel/Bioinformatika/tree/main/2lab

   a) *Nustatykite koks kokybės kodavimas yra naudojamas pateiktame faile. Parašykite, kokią koduotę nustatėte ir kuo remiantis?*

   Radau, naudojamas kokybės kodavimas pateiktose sekose yra **Sanger Phred+33**. Iš pradžių radau visu kokybes sekų mažiausią ir didžiausią reikšmes (2, 40), paversdama į ASCII reikšmes (35, 73). Toliau kiekvienam kodavimui iš žodyno skaičiavau skirtumą tarp rastų min ir max reikšmių ir žodyne esančių kodavimo schemai galiojančių min max reikšmių. Kodavimo schema pasirinkau pagal šį mažiausį skirtumą.

   b) *Pateikite grafiką, kurio y ašyje būtų read'ų skaičius, x – ašyje – C/G nukleotidų dalis read'o sekoje. Parašykite, koks „stambių" pikų skaičius yra gautame grafike?*

   Gavau tris stambius pikus. GC % rėžiuose (32.9, 37.6), (51.7, 56.4), (65.8, 70.5)



Read Id	**Mikroorganizmas**
M00827:12:000000000-AEUNW:1:1101:21543:2685 1:N:0:6	Staphylococcus aureus strain BAA-1556_Catania_SCV chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:20192:2848 1:N:0:6	Staphylococcus aureus strain BAA-1556_Catania_SCV chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:20889:3446 1:N:0:6	Staphylococcus aureus strain BAA-1556_Catania_SCV chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:12898:3746 1:N:0:6	Staphylococcus aureus strain SA73_2 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:23128:3788 1:N:0:6	Escherichia coli strain JCL301 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:11683:2509 1:N:0:6	Escherichia coli str. K-12 substr. MG1655 strain K-12 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:14568:2958 1:N:0:6	Escherichia coli str. K-12 substr. MG1655 strain K-12 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:13276:3014 1:N:0:6	Escherichia coli strain JCL301 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:15296:3158 1:N:0:6	Escherichia coli strain JCL301 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:11742:3950 1:N:0:6	Escherichia coli strain JCL301 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:18070:3392 1:N:0:6	Thermus thermophilus HB8_001 DNA, complete genome
M00827:12:000000000-AEUNW:1:1101:14615:4214 1:N:0:6	Thermus thermophilus HB8_001 DNA, complete genome

M00827:12:000000000-AEUNW:1:1101:23350:4251 1:N:0:6	Thermus thermophilus HB8_001 DNA, complete genome
M00827:12:000000000-AEUNW:1:1101:11574:4514 1:N:0:6	Thermus thermophilus HB27 chromosome, complete genome
M00827:12:000000000-AEUNW:1:1101:9274:5136 1:N:0:6	Thermus thermophilus HB27 chromosome, complete genome

c) Mikroorganizmų lentelė, kurie buvo rasti BLAST paieškos metu, po 5 kiekvienam pikui.

5. *Kokių rūšių bakterijų buvo mėginyje?*

Mėginyje buvo **Escherichia coli**, **Staphylococcus aureus** ir **Thermus thermophilus**.