VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

PROGRAMŲ SISTEMŲ BAKALAURO STUDIJŲ PROGRAMA

ANTROJO LABORATORINIO DARBO ATASKAITA

Bioinformatika

II – asis laboratorinis darbas

Atliko: 4 kurso 3 grupės studentas Lukas Orliukas

Vilnius – 2023

Turinys

[1. Darbo kodas 3](#_Toc150181477)

[2. FASTQ formatas 3](#_Toc150181478)

[3. Gimimo diena 3](#_Toc150181479)

[4. Nenaudojami kodai 3](#_Toc150181480)

[5. Programa 3](#_Toc150181481)

[6. Bakterijų rūšys 6](#_Toc150181482)

# Darbo kodas

Atlikto darbo kodą galima rasti čia:

<https://colab.research.google.com/drive/1U5z66-KJZWu1RjvpZX-GFY9TH94VWkgW?usp=sharing>

# FASTQ formatas

FASTQ yra tekstinis formatas, skirtas saugoti biologinėms sekoms ir atitinkamai jų kokybės įverčiams, kurie ir yra papildoma informacija lyginant su FASTA formatu. FASTQ formatas ir buvo sukurtas norint sujungti FASTA formato seka su jos kokybės duomenimis. Sekos raidės ir kokybės įvertis yra koduojami ASCII simboliais. FASTQ formatas susideda iš keturių eilučių vienai sekai:

* Pirmoji eilutė prasideda simboliu ‘@’ ir po to esančiu sekos identifikatoriumi bei neprivalomu aprašymu (kaip FASTA formato pavadinimo eilutė)
* Antroji eilutė yra pačios sekos raidės.
* Trečioji eilutė prasideda simboliu ‘+’ ir po to neprivalomi tas pats sekos identifikatorius ir bet koks aprašymas.
* Ketvirtoji eilutė yra sekos kokybės reikšmės. Šios eilutės simbolių skaičius turi sutapti su antros eilutės (pačios sekos) simbolių skaičiumi.

# Gimimo diena

Esu gimęs 12 mėnesio diena. 12 + 33 = 45. Šį skaičių atitinka ASCII simbolis: ‘-’.

# Nenaudojami kodai

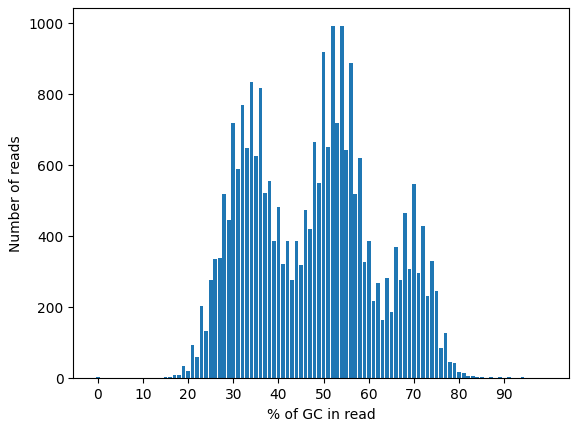
Šie kodai negali būti naudojami, nes jie koduoja kontrolinius simbolius, kurie neturi vizualinės išraiškos, tokius kaip: „Backspace“ ar „Carriage return“. Taip pat ne visi šie simboliai gali būti saugomi tekstinio formato faile, todėl nebūtų įmanoma tinkamai koduoti sekos kokybės informaciją tekstiniu formatu, kuriuo ir yra FASTQ.

# Programa

1. Pateiktame faile yra naudojamas „Sanger Phred+33“ kodavimas, kuris nustatytas tokiu būdu: surastos visų nuskaitymų mažiausia ir didžiausia kokybės įverčio reikšmės paverčiant jas į atitinkamas reikšmes skaičiais. Tada pagal kodavimo poslinkį (33 arba 64) surastas minimalus skirtumas tarp gautų mažiausios ir didžiausios reikšmių ir tame kodavime galimų minimalios ir maksimalios reikšmių. Kievienam kodavimui gautas skirtumo įvertis ir parinktas kodavimas su mažiausiu įverčiu. Galimos kodavimo reikšmės rastos internete (rasta keletas variantų, todėl pasirinktas vienas iš jų):

* Sanger Phred+33": (0, 40)
* Solexa Solexa+64": (-5, 40)
* Illumina 1.3+ Phred+64": (0, 40)
* Illumina 1.5+ Phred+64": (3, 40)
* Illumina 1.8+ Phred+33": (0, 41)

Gauta, kad mažiausias įvertis yra „Sanger Phred+33“ kodavimo.

1. Atlikus skaičiavimus gautas grafikas, pavaizduotas žemiau (žr. 1 pav.).

1 pav. C/G pasiskirstymas nuskaitymuose

Grafike yra matomi trys „stambūs“ pikai.

1. Žemiau pateikiama lentelė (žr. 1 lentelė), kurioje pateikiama po penkis nuskaitymus kiekvienam „stambiam“ pikui, iš viso penkiolika nuskaitymų, su nuskaitymo ID, seka ir rasta bakterija.

1 lentelė. Nuskaitymų lentelė

|  |  |  |
| --- | --- | --- |
| ID | Seka | Bakterija |
| 66 | GGACGCGCACCTGTGGCGTAGAACTGAATCTCATCACGAACAGCGCCGCCTAAAAGTTTATAAACCGGAAGCCCGACCACTTTGCCGAACAGATCCCACAGAGCCAGATCGACACAAGAAATCGTATTCATCACCAGGCCACCAGAGCCGG | gi|2604885558|emb|OY754442.1| Escherichia coli isolate S142 genome assembly, chromosome: 1 |
| 44 | GAACATCCAGCTCGCGCTCCGTCAGCACGCTGAAGGGATCTTCTTCCGCGCCAAACATTTCACGTTCACGTAAGTACTGATTGACGCGTTCGCTAAAGACTTTGCTGCCTTTCGCTCCGGCACGAATCGCTTCCAGCAATACTTCCGGGTC | gi|2604885558|emb|OY754442.1| Escherichia coli isolate S142 genome assembly, chromosome: 1 |
| 42 | CCCATTGTTGATGAAGGCAGTAACGCCGCTTTTGACATCGTTTATAGCGATGCGCAAGGCGTGAAAAAAGCCGTGTCGGGCTTGCAGGTGCGCCTGATTCGCGAACGCCGCGATTACTACTGGAACTGGTCAGAAGATGAAGGCTGGCAGT | gi|2604885558|emb|OY754442.1| Escherichia coli isolate S142 genome assembly, chromosome: 1 |
| 110 | ACAGAAGAAACCCATTGATGGTGTTCACAATGCCCGCCGAACGCATTCGTGTTGACCAGCGGGCATCCTTTCTCACTCCCCGACCAGAATCACTTCAACCCCAGCCTTTCGCAGTCCTTCCAGGCTATCCGCAGGAATGCCTTCATCAAC | gi|2604885558|emb|OY754442.1| Escherichia coli isolate S142 genome assembly, chromosome: 1 |
| 20 | AACGGCATTCGGTGGACCGGGTTAAAGAAGTGCATCCCGCAGAAGTTTTCCGGGCGTTCCAGCGCGTTGGCCAGTTCGCTGATAGGAATGGTTGAAGTGTTAGACGCCAGCACGGTATCCTGGCGTACTTTTTGTTCGGTTTCTGCCAGT | gi|2604885558|emb|OY754442.1| Escherichia coli isolate S142 genome assembly, chromosome: 1 |
| 79 | CTTCCGGAGTCGTTACAACAACATCAAGGACAAATTCATTAACATCGCTAGAAATTTCAGGTGTAACAGTATAAACCGTTTTCTTCGAAATGCCGCCTTTTCTAAATTCTTTTAGGTAAGTCTGCAATAAGAAATTGATTTTACCATTTTG | gi|2593210064|gb|CP136524.1| Staphylococcus aureus strain BAA-1556\_Catania\_SCV chromosome, complete genome |
| 54 | GACATATTAGATGAGCGCCAATATAATGTTAATAATTTGTTGAATATGCCTTTTGGAGCATGTCATTTTGCGGTTGCAGCGAAACCTGAAACGACCAATTATCGTAAAATCGCAACGAGTTATGTTCATACTGCTGAAACATATTTTAAAT | gi|2593210064|gb|CP136524.1| Staphylococcus aureus strain BAA-1556\_Catania\_SCV chromosome, complete genome |
| 195 | TTTTGATACCTAGCATTTTCCCAATATGCTGAATCATCGCCGTTCTGTAATAATCTTCACTTGGCATTCTTGGAATTTGAACAACATCCTGTTCATTTTTATAGACTTGCTTAAATGGTACAGCTGAAAAATCAAAGTAATGTCCGTTTTC | gi|2593210064|gb|CP136524.1| Staphylococcus aureus strain BAA-1556\_Catania\_SCV chromosome, complete genome |
| 37 | CAGAATACCAAGATTGCTTATTATTTTTTAGACTAGGTGATTTCTATGAAATGTTTTATGAAGATGCCAAGGAGGCATCACGTGTACTTGAAATTACTTTGACTAAAAGAGATGCTAAAAAAGAAAATCCAATTCCGATGTGTGGTGTTCC | gi|2593210064|gb|CP136524.1| Staphylococcus aureus strain BAA-1556\_Catania\_SCV chromosome, complete genome |
| 301 | TTGTTTAAAGTGCCTGAAAATGTTAGTCACTTAAAAATTCAATTTGTACCTAAAAATGATGCAATAACAGATGCGCGTGGCATTTATCAACTAAGAGACGGTTATAAATACTATGACTTTGTTGACTCAATTGGTCTACATTCTGGTTCGC | gi|2593210064|gb|CP136524.1| Staphylococcus aureus strain BAA-1556\_Catania\_SCV chromosome, complete genome |
| 29 | GTAGCCCAGGGGAAAGGGAAGGTCCAGGCCCAGAAGGGCCCGCACGCCGAAGCCGGCCAGAAAGGCCAAAACCCCGCCTCCCATGTAGATCGCGGCGAAGAGCCCGGGCCGCTCCTCGCGGGGCACGGCCTTGGCCAACACCTCCCAGT | gi|2432789391|dbj|AP025540.1| Thermus thermophilus HB8\_001 DNA, complete genome |
| 98 | AAAAGGGTTCGTCCGCATCGCCCTGGTGCGGCCCCCGGAAGAGCTCCTCGAGGCGGCCAAGCGCATCCGCGAGGCCCTGGACTAAGCGCCCAGCCTACTGCCAACGCTCCTCGGCGGGAAGCCCCTCAAGGCGGCGGAGAACCTCCCCCAC | gi|2432789391|dbj|AP025540.1| Thermus thermophilus HB8\_001 DNA, complete genome |
| 170 | GAGGGGTAGGGCCAAGGCCAGGGTGAGGAGGGCGGAGCCGAGGCCGTACTCCAGGCTGAAGAGGTAGCGCTCCCAGTAGTAGGGGTTGGCCAGGGCGCGGCCGAACCCCTCCCCCACCCCCAGGACCAGGACCCGGCCTAAGGGGTAGA | gi|2432789391|dbj|AP025540.1| Thermus thermophilus HB8\_001 DNA, complete genome |
| 50 | CTGGAAGATGGCCACGGGGAGGAACTTGGTCTTGGGGGCGAGCCCCGCGATGCCCTGGCCGTCCCTGGGGGCACTCACCGTGGAGGCCACGTAGGTGCCGTGGTAGAGGTCGGCGGGGGGGGTGAGGCCCTGGATAAAGTTCAGCCAGCCG | gi|2432789391|dbj|AP025540.1| Thermus thermophilus HB8\_001 DNA, complete genome |
| 188 | CTGCATGTAGCGGAGGTCCGCCTCGCCGAGCCCGGGAGGCGGGGCCTGGAGGAGGGTGTGGAGGACCACCAGCTTGGGGCCTTGGGCCTCGAGGGGGTCCAGGACGTAGTCGCCCCACTTCCCGCCGTAAAACCCGTACTCGTGCTGGAGG | gi|2432789391|dbj|AP025540.1| Thermus thermophilus HB8\_001 DNA, complete genome |

# Bakterijų rūšys

Mėginyje buvo rasta Escherichia coli, Staphylococcus aureus ir Thermus thermophilus bakterijų.