# Bioinformatika, II laboratorinis darbas, lasivos formos ataskaita

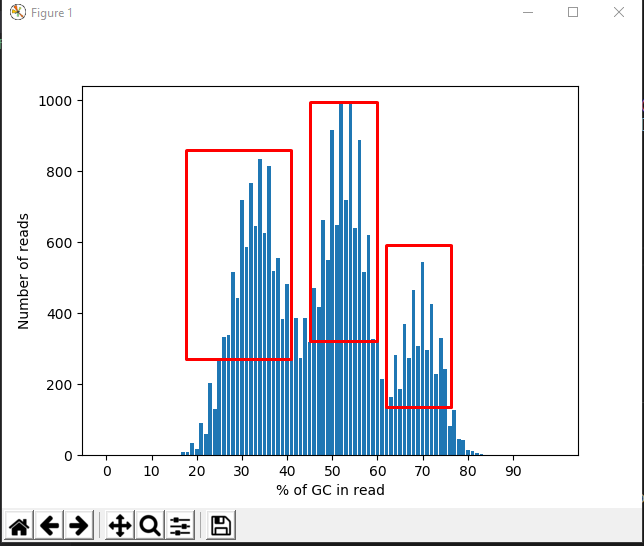
Github nuoroda į darbo kodą: <https://github.com/kasparas12/bioinformatika>

1. FASTQ yra tekstinis formatas, skirtas saugoti biologinėms sekoms. Lyginant šį formatą su įprastu FASTA, galima pastebėti, jog papildomai yra saugomi ir sekos kokybės įverčiai, užkoduoti ASCII simboliais. Formato sudėtis:
   1. Pirma eilutė prasideda ‘@’ simboliu, toliau seka sekos identifikatorius ir neprivalomas aprašymas, pvz.: @SEQ\_ID
   2. Antroje eilutėje yra pati užkoduota biologinė seka
   3. Trečia eilutė prasideda ‘+’ ženklu, toliau eina neprivalomi tie patys sekos identifikatorius ir aprašymas
   4. 4 eilutė yra užkoduotos sekos kokybės reikšmės ASCII simboliais. Simbolių skaičius turi sutapti su 2 eilutėje esančios sekos simbolių skaičiumi.
2. Gimiau Lapkričio 18 d., todėl 18 + 33 = 51 skaičiaus reikšmę atitinka ASCII simbolis ‘3’
3. Pirmi 32 ASCII kodai neturi vizualių simbolinių išraiškų – tai visokios tarpų, tab’ų, operacinės sistemos specialiosios reikšmės (CARRIAGE RETURN, NEW LINE), todėl šios reiškmės nėra tinkamos kodavimui, ne visas jas galime net išsaugoti tekstinio formato faile, ko mums tikrai reikia norint nagrinėti, užkrauti, saugoti su sekomis susijusią informaciją.
4. Mano skriptą rasti galima nurodytoje github repozitorijoje, 2\_laboratorinis aplanke, 2\_lab.py faile:
   1. Nustačiau, jog pateiktų sekų kokybės įverčiai koduojami **Sanger Phred+33** algoritmu. Radau visų read’ų mažiausia ir didžiausią kokybės įverčio reikšmes paversdamas jas į skaitines išraiškas iš ASCII simbolio naudojant ord() funkciją. Tuomet pagal kiekvieno algoritmo poslinkį (33 arba 64) ieškojau minimalaus skirtumo tarp gautų min ir max reiškmių ir tame algoritme galiojančių min ir max reikšmių. Kiekvienam algoritmui gavau skaitinį skirtumo įvertį ir pasirinkau tą algoritmą, kurio skirtumo skaitinis įvertis mažiausias. Ieškodamas algoritmo interpretavau, jog pateiktos sekos yra neapdorotos (angl. Raw sequences) , todėl jų galimi kokybės reikšmių rėžiai, kuriuos radau internete:

* Sanger Phred+33": (0, 40)
* Solexa Solexa+64": (-5, 40)
* Illumina 1.3+ Phred+64": (0, 40),
* Illumina 1.5+ Phred+64": (2, 41), 0,1 reikšmės nenaudojamos, 3 reikšmė ‚B‘ – quality control indicator, t.y. specialios reikšmės sekoje simbolis
* Illumina 1.8+ Phred+33": (0, 41)

Gavau, jog užduotyje pateiktų sekų kokybės min, max – [2,40], todėl artimiausias algoritmas gavosi **Sanger Phred+33**

* 1. Pagal nurodytus reikalavimus skaičiavau GC dalis sekose ir gavau štai tokią diagramą:



Taigi, iš grafiko matome, jog stambių pikų skaičius yra 3. Šie pikai yra [20;40], [40;60] ir [60;80] GC % rėžiuose.

* 1. Žemiau lentelėje surašiau visus 15 read‘ų, kurie paimti iš pikų viršūnių (po 5 kiekvienam pikui) ir juos atitinkančios bakterijos, rastos BLAST paieškos metu. E value rodo galimo atsitiktinumo reikšmę, kuo arčiau 0, tuo mažesnė atsitiktinumo tikimybė.

|  |  |  |  |
| --- | --- | --- | --- |
| Read id | Read | Bacteria | E value |
| 66 | GGACGCGCACCTGTGGCGTAGAACTGAATCTCATCACGAACAGCGCCGCCTAAAAGTTTATAAACCGGAAGCCCGACCACTTTGCCGAACAGATCCCACAGAGCCAGATCGACACAAGAAATCGTATTCATCACCAGGCCACCAGAGCCGG | gi|1759084408|gb|CP042844.1| Escherichia coli strain JME66 chromosome, complete genome | ~0 |
| 44 | GAACATCCAGCTCGCGCTCCGTCAGCACGCTGAAGGGATCTTCTTCCGCGCCAAACATTTCACGTTCACGTAAGTACTGATTGACGCGTTCGCTAAAGACTTTGCTGCCTTTCGCTCCGGCACGAATCGCTTCCAGCAATACTTCCGGGTC | gi|1759084408|gb|CP042844.1| Escherichia coli strain JME66 chromosome, complete genome | ~0 |
| 42 | CCCATTGTTGATGAAGGCAGTAACGCCGCTTTTGACATCGTTTATAGCGATGCGCAAGGCGTGAAAAAAGCCGTGTCGGGCTTGCAGGTGCGCCTGATTCGCGAACGCCGCGATTACTACTGGAACTGGTCAGAAGATGAAGGCTGGCAGT | gi|1759084408|gb|CP042844.1| Escherichia coli strain JME66 chromosome | ~0 |
| 20 | AACGGCATTCGGTGGACCGGGTTAAAGAAGTGCATCCCGCAGAAGTTTTCCGGGCGTTCCAGCGCGTTGGCCAGTTCGCTGATAGGAATGGTTGAAGTGTTAGACGCCAGCACGGTATCCTGGCGTACTTTTTGTTCGGTTTCTGCCAGT | gi|1759084408|gb|CP042844.1| Escherichia coli strain JME66 chromosome, complete genome | ~0 |
| 110 | ACAGAAGAAACCCATTGATGGTGTTCACAATGCCCGCCGAACGCATTCGTGTTGACCAGCGGGCATCCTTTCTCACTCCCCGACCAGAATCACTTCAACCCCAGCCTTTCGCAGTCCTTCCAGGCTATCCGCAGGAATGCCTTCATCAAC | gi|1759084408|gb|CP042844.1| Escherichia coli strain JME66 chromosome, complete genome | ~0 |
| 79 | CTTCCGGAGTCGTTACAACAACATCAAGGACAAATTCATTAACATCGCTAGAAATTTCAGGTGTAACAGTATAAACCGTTTTCTTCGAAATGCCGCCTTTTCTAAATTCTTTTAGGTAAGTCTGCAATAAGAAATTGATTTTACCATTTTG | gi|1760491771|gb|CP030661.1| Staphylococcus aureus strain ER02826.3 chromosome, complete genome | ~0 |
| 54 | GACATATTAGATGAGCGCCAATATAATGTTAATAATTTGTTGAATATGCCTTTTGGAGCATGTCATTTTGCGGTTGCAGCGAAACCTGAAACGACCAATTATCGTAAAATCGCAACGAGTTATGTTCATACTGCTGAAACATATTTTAAAT | gi|1760491771|gb|CP030661.1| Staphylococcus aureus strain ER02826.3 chromosome, complete genome | ~0 |
| 195 | TTTTGATACCTAGCATTTTCCCAATATGCTGAATCATCGCCGTTCTGTAATAATCTTCACTTGGCATTCTTGGAATTTGAACAACATCCTGTTCATTTTTATAGACTTGCTTAAATGGTACAGCTGAAAAATCAAAGTAATGTCCGTTTTC | gi|1760491771|gb|CP030661.1| Staphylococcus aureus strain ER02826.3 chromosome, complete genome | ~0 |
| 37 | CAGAATACCAAGATTGCTTATTATTTTTTAGACTAGGTGATTTCTATGAAATGTTTTATGAAGATGCCAAGGAGGCATCACGTGTACTTGAAATTACTTTGACTAAAAGAGATGCTAAAAAAGAAAATCCAATTCCGATGTGTGGTGTTCC | gi|1760491771|gb|CP030661.1| Staphylococcus aureus strain ER02826.3 chromosome, complete genome | ~0 |
| 301 | TTGTTTAAAGTGCCTGAAAATGTTAGTCACTTAAAAATTCAATTTGTACCTAAAAATGATGCAATAACAGATGCGCGTGGCATTTATCAACTAAGAGACGGTTATAAATACTATGACTTTGTTGACTCAATTGGTCTACATTCTGGTTCGC | gi|1760491771|gb|CP030661.1| Staphylococcus aureus strain ER02826.3 chromosome, complete genome | ~0 |
| 29 | GTAGCCCAGGGGAAAGGGAAGGTCCAGGCCCAGAAGGGCCCGCACGCCGAAGCCGGCCAGAAAGGCCAAAACCCCGCCTCCCATGTAGATCGCGGCGAAGAGCCCGGGCCGCTCCTCGCGGGGCACGGCCTTGGCCAACACCTCCCAGT | gi|1701967042|dbj|AP019801.1| Thermus thermophilus HC11 DNA, complete genome | ~0 |
| 170 | GAGGGGTAGGGCCAAGGCCAGGGTGAGGAGGGCGGAGCCGAGGCCGTACTCCAGGCTGAAGAGGTAGCGCTCCCAGTAGTAGGGGTTGGCCAGGGCGCGGCCGAACCCCTCCCCCACCCCCAGGACCAGGACCCGGCCTAAGGGGTAGA | gi|1701967042|dbj|AP019801.1| Thermus thermophilus HC11 DNA, complete genome | ~0 |
| 98 | AAAAGGGTTCGTCCGCATCGCCCTGGTGCGGCCCCCGGAAGAGCTCCTCGAGGCGGCCAAGCGCATCCGCGAGGCCCTGGACTAAGCGCCCAGCCTACTGCCAACGCTCCTCGGCGGGAAGCCCCTCAAGGCGGCGGAGAACCTCCCCCAC | gi|1701967042|dbj|AP019801.1| Thermus thermophilus HC11 DNA, complete genome | ~0 |
| 188 | CTGCATGTAGCGGAGGTCCGCCTCGCCGAGCCCGGGAGGCGGGGCCTGGAGGAGGGTGTGGAGGACCACCAGCTTGGGGCCTTGGGCCTCGAGGGGGTCCAGGACGTAGTCGCCCCACTTCCCGCCGTAAAACCCGTACTCGTGCTGGAGG | gi|1701967042|dbj|AP019801.1| Thermus thermophilus HC11 DNA, complete genome | ~0 |
| 170 | GAGGGGTAGGGCCAAGGCCAGGGTGAGGAGGGCGGAGCCGAGGCCGTACTCCAGGCTGAAGAGGTAGCGCTCCCAGTAGTAGGGGTTGGCCAGGGCGCGGCCGAACCCCTCCCCCACCCCCAGGACCAGGACCCGGCCTAAGGGGTAGA | gi|1701967042|dbj|AP019801.1| Thermus thermophilus HC11 DNA, complete genome | 7 |

1. Mėginyje buvo **E.Coli**, **Staphylococcus aureus (auksinis stafilokokas)**, **Thermus thermophilus** bakterijos (mikroorganizmai)