**Bioinformatika**

2 laboratorinis darbas

Tomas Ukrinas

Programų sistemos 4 kursas 5 grupė

# Apibūdinkite fastq formatą. (https://en.wikipedia.org/wiki/FASTQ\_format). Kokia papildoma informacija pateikiam lyginant su FASTA formatu?

Fastq formatas papildomai turi quality eilutė. Quality eilutėje kiekvienas simbolis atitinka sekos simboliui.

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

Esu gimęs 30 dieną. 30 + 33 = 63. Atitinkantis simbolis yra: „?“.

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

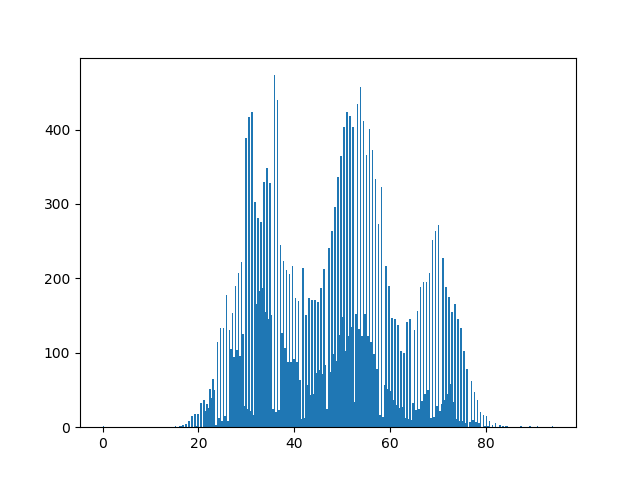
Nes pirmieji 32 ASCII kodai nėra simboliai. Tai yra tam tikros reikšmės (kaip pvz simbolis 0 yra *null*, arba simbolis 3 yra *end of text*). Jų neįrašysi į seką.

# Parašykite skriptą

## Nustatyti koks kokybės kodavimas yra naudojamas pateiktame faile

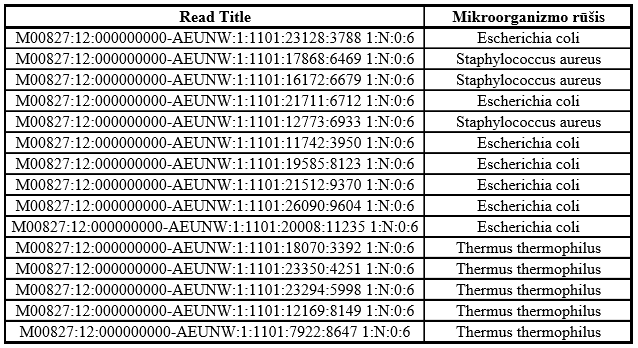
Parašytas skriptas (faile *Script.py*) analizuoja sekas ir ieško naudojamo kodavimo atmetimo būdu. Atsakymų faile matome, kad galimi variantai Sanger bei Illumina1.8. Kadangi simbolio J kokybės eilutėse niekada nesutikome, galime daryti prielaidą, kad naudojama **Sanger** koduotė.

## Analizuoti C/G nukleotidų pasiskirstymą read’uose



Grafike galime matyti peak‘us ties GC procentais 35, 53 bei 70. Todėl toliau bus analizuojamos būtent šios peak viršūnės (t.y. 3\*5=15 sekų).

## Paimti po 5 kiekvieno piko viršūnės sekų ir atlikti blast’o paieškas



# Kokių rūšių bakterijų buvo mėginyje?

* Escherichia coli
* Staphylococcus aureus
* Thermus thermophilus