

VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS
MATEMATINĖS INFORMATIKOS KATEDRA

Baigiamasis bakalauro darbas

Pirminė metagenominių duomenų analizė
(Primary Analysis of Metagenomic Data)

Atliko: IV kurso Bioinformatikos studentė
Dovilė Preidytė
Darbo vadovas:
Lekt. Irus Grinis
Recenzentas:
Andrius Merkys

Vilnius
2018

Turinys

Ivadas	2
1. Literatūros apžvalga	3
1.1. Metagenomika.....	3
1.2. Metagenominių duomenų modelis	3
1.3. Sekvenavimas	4
1.4. SAM ir BAM failų formatas	5
1.5. Bowtie2	5
1.5.1. Bowtie2 ir indeksavimas	6
1.5.2. Bowtie2 ir išlyginimas	6
1.6. Samtools	6
1.6.1. SAM failų konvertavimas į BAM.....	6
1.6.2. BAM failų rūšiavimas	6
1.6.3. BAM failų indeksavimas	6
1.7. IGV	7
1.8. Needleman-Wunsch algoritmas	7
1.8.1. needleall	7
1.9. Levenshtein atstumo algoritmas	8
1.10.EMBOSS water	8
1.10.1. Algoritmas	8
1.11.EDLIB biblioteka	8
2. Praktinė tiriamoji dalis	10
2.1. Pirmasis eksperimentas ir motyvacija.....	10
2.2. Darbo metodika	12
2.3. Greitas didelės apimties metagenomo skaidymas	12
2.4. Needleall įrankio panaudojimas	13
2.5. EDLIB biblioteka	14
2.5.1. Bibliotekos pritaikymas	14
2.6. Water įrankio panaudojimas	15
2.7. Tyrimas su virusais.....	16
2.8. Tolimesni metodikos bandymai	19
2.9. Tolimesni metodikos bandymai su organizmais	27
Rezultatai ir švados	31
Santrauka	33
Summary	34
Literatūra	35

Ivadas

Metagenomika apibūdina mokslinių tyrimų sritį ir rinkinių veiksmų, kurie leidžia atlikti nepriklausomą mikrobinės bendruomenės analizę iš bet kokių aplinkos mėginių. Per paskutinius 10 – 15 metų, ši sritis labai išaugo, leisdamą tyrėjams prieiti prie didelio kiekio metagenominių duomenų ir plačiau pažvelgti į mus supantį mikrobų pasaulį. Šiuo metu iš plačiai naudojamų duomenų bazių kaip NCBI SRA ar NCBI BioProject galima atsisiųsti didelius kiekius metagenominių rinkinių. Pradėjus tyrimą ir skyrus jam laiko, gali paaiškėti, kad rasti duomenys yra netinkami arba blogai įvertinti. Galbūt jie buvo prastai dokumentuoti, įkeltas netikslingas failas, ar duomenys yra kitokie, nei buvo manyta. Tam, kad būtų galima išvengti paminėtų klaidų ir sutaupyti laiko, galima, pasitelkiant plačiai naudojamus algoritmus, atlikti pirminę metagenominę analizę, kuri dar bus minima kaip *paviršutinė* metagenominė analizė. Tai leistų prieš pradedant eksperimentą atlikti metagenominių duomenų įvertinimą ir susiplanuoti darbo eigą, eksperimentui skirtą laiką ir resursus. Tai taip pat parodytų pradinę metagenominę informaciją apie nuskaitymų poras ir turėtų galimybę padėti identifikuoti kai kuriuos metagenominiame faile esančius organizmus. Tokia analizė neatliktų tikslaus tyrimo, bet leistų bendrai pažvelgti į duomenis ir atlikti jų įvertinimą. Šiame darbe naudojant Needleman-Wunsch (įrankis *emboss:needleall*), redagavimo atstumo (panaudotas įrankis *EBL* biblioteka), Smith-Waterman (įrankis *emboss:water*) algoritmus, atlikta pirminė septynių metagenominių rinkinių duomenų analizė.

Pagrindinis šio darbo tikslas yra sukurti metodiką paviršutiniam metagenomo tyrimui. Pirmoje darbo dalyje apžvelgiama teorinė dalis apie metagenomiką, metagenominę analizę ir darbe panaudotus įrankius bei algoritmus. Antroje darbo dalyje pateikiama informacija apie pirmuosius eksperimentus, kurių rezultatai paskatino pereiti prie pirminės metagenominės analizės tyrimo. Taip pat aprašyta darbo metodika ir eksperimentų, atliktų su skirtingais metagenomais, rezultatai ir analizė.

1. Literatūros apžvalga

1.1. Metagenomika

Vienas žymiausių įvykių mikrobinės ekologijos srityje per paskutinį dešimtmetį buvo metagenomikos pasirodymas ir plėtojimas. Metagenomika yra apibūdinama kaip tiesioginė genomų, gautų su aplinkos mėginiais, genetinė analizė. Tokia genomų visuma dar yra vadinama metagenomu. Metagenomika – terminas apibūdinantis mokslinių tyrimų sritį ir taikomų metodologijų rinkinį, kurios leidžia atlikti metagenominę analizę. Ši sritis pradėjo vystytis nuo aplinkos DNR klonavimo, kurį sekė funkcinė ekspresijos atranka, kuri buvo greitai papildyta atsitiktiniu aplinkos DNR mėginių automatinio (shotgun) sekvenavimu. Šie pradiniai projektai ne tik įrodė metagenominio metodo principus, bet ir atskleidė didelę funkcinį genų įvairovę mikrobiniame pasaulyje. [11]

Metagenomika suteikia priėjimą prie mikrobinės kolonijos funkcinės genų kompozicijos ir platesnį apibūdinimą nei filogenetinis tyrimas, kuris dažniausiai yra paremtas tik vieno geno įvairove, pavyzdžiui, 16S rRNR genu. Pati metagenomika suteikia genetinę informaciją apie potencialiai naujus biokatalizatorius ir fermentus, genominius ryšius tarp neklasifikuotų organizmų funkcijų ir filogenijos, taip pat kolonijų funkcinis ir struktūrinius evoliucinius profilius. Tai, gali būti papildyta metatranskriptiniais arba metaproteominiais metodais, tam, kad būtų galima apibūdinti išreikštus veiksmus. Metagenomika yra galingas įrankis padedantis generuoti naujas mikrobinių funkcijų hipotezes.

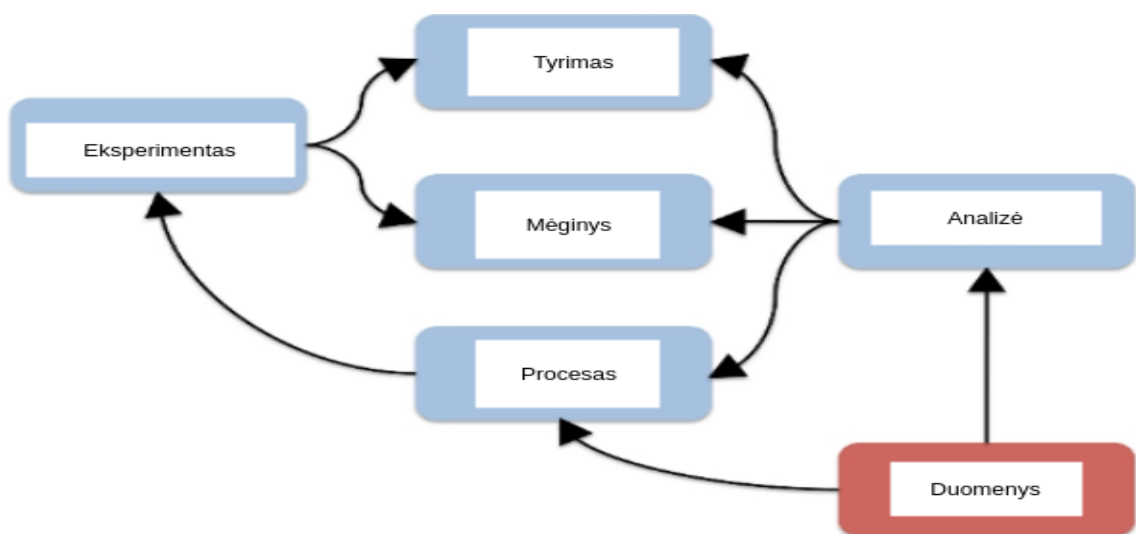
Staigus naujos kartos sekvenavimo kainos kritimas, labai pagreitino sekvenavimu paremtos metagenomikos plėtrą. Metagenominiai automatinio (shotgun) sekvenavimo duomenų rinkiniai labai išpopuliarėjo per kelerius paskutinius metus. [12]

Nepaisant didelio mokslinės bendruomenės susidomėjimo metagenominiais tyrimais, jų analizė kelia iššūkius dėl didelio kiekio sekvenavimo duomenų ir jų sudėtingos prigimties. Vieno metagenomo automatinio sekvenavimo procesai (runs) gali duoti šimtus milijonų sekų, kurios bus pateiktos kaip dešimtys gigabaitų užimantys duomenys. Atliekant daug eksperimentų, kurių kiekvieną sudaro dešimtys ar šimtai tokių procesų, duomenys gali greitai perpildyti duomenų talpyklas, taip sumažindami kiekvieno tyrėjo analizių galimybes. Pačios sekos dažniausiai būna trumpos, nuo 100 iki 500 bazių porų (bp). Dažniausiai nuskaitymai reprezentuoja tik mažą fragmentą iš duomenų rinkinyje besislepiančio pilno ilgio geno sekos. [2] Nepaisant savo trūkumų, dabar metagenomika tampa standartiniu įrankiu daugumai laboratorijų ir mokslininkų, dirbančių su mikrobine ekologija. [12]

1.2. Metagenominių duomenų modelis

Naujos kartos sekvenavimo technologijų pristatymas leidžia net ir mažoms tyrėjų grupėms išgauti didelės apimties sekvenavimo duomenis. Gaunamos DNR sekos ir su jomis susijusi informacija yra paprastai užfiksuojama keliuose tarpusavyje susijusiuose objektuose, kurie pristato tokias idėjas:

- Tyrimas: informacija apie sekvenavimo pastangų dydį, kuris yra reikalingas sugrupuoti visas projekto duomenų dalis.
- Mėginys: informacija apie sekvenavimo mėginių kilmę ir charakteristiką.
- Eksperimentas: informacija apie sekvenavimo eksperimentus. Į šią informaciją įeina ir bibliotekos bei informacija apie instrumentus.
- Procesas (Run): sekvenavimo eksperimento išvestis, kurią sudaro vadinamieji nuskaitymai, atvaizduoti duomenų faile.
- Analizė: išvesties rinkinys, gautas iš pirminių sekvenavimo rezultatų, tokių kaip sekvenavimo surinkimai (assemblies), funkcinės ir taksonominės anotacijos. [8]



1 pav. Bendras nuskaitymo duomenų struktūros modelis [8]

1.3. Sekvenavimas

Tradicionis metagenominis sekvenavimas rėmėsi *Sanger* sekvenavimu, kuris pagamindavo gan ilgus nuskaitymus, didesnius nei 700 bazių porų (bp) su mažu klaidų dažniu. Tačiau, naujos kartos sekvenavimo technologijos, sujungtos su žymiu sekvenavimo kainos sumažinimu, privertė pajudėti nuo Sanger sekvenavimo kaip pačio pagrindinio standarto, dėl jo didelės kainos. Vienas pirmųjų didelės apimties Sangers sekvenavimo metagenominių projektų įvykdytas, kad būtų charakterizuotos mikrobinės populiacijos gyvuojančios Sargaso jūroje. Buvo sekvenuojama daugiau nei 1 milijardas DNR bp ir identifikuota daugiau nei 1.2 milijonai naujų genų.

Vis tobulėjančios sekvenavimo technologijos, sujungtos su dramatiškai mažėjančia kaina, reiškia, kad „sekvenuoti viską“ požiūris gali būti pritaikytas prie įvairių aplinkos mėginių. Tai seka surinkimas ir anotacija, kad būtų galima gauti išsamų mikrobinės bendruomenės gausos vaizdą ir užkoduotą funkcinį potencialą. Vienos plačiausiai naudojamų sekvenavimo platformų yra GS-FLX

454 (Roche), MiSeq, HiSeq, ir genomų analizavimo II platforma (Illumina), SOLiD sistema, Ion Torrent, Ion Proton ir PacBio RS II.

Didelės apimties sekvenavimo projektai, naudojantys tokias technologijas, generuoja didelius duomenų rinkinius, kuriuose gali būti milijonai genų ir, žinant metagenomų prigimtį, didelė šių genų proporcija bus neįprasta ir turės nežinomą funkciją. Šis, pradinis, netikslinis sekvenavimo būdas gali vėliau būti naudojamas kartu su labiau apibrėžtais sekvenavimo būdais, kurių tikslas yra nustatyti specifinį genų tipą arba šeimą. Pabrėžiama, kad netikslinis yra naudojamas kaip būdas, kai yra labiau gilinamasi į bendrą bendruomenės kompoziciją, filogenetiką ir funkcinį pajėgumą, o ne į tikslus genus ar jų šeimas. Taip pat, metagenominis sekvenavimas, naudojant naujo kartos sekvenavimo technologijas, kartu su anotacija, homologijos paieškomis ir filogenetiniu klasterizavimu, gali atskleisti genus, kurie labiau skiriasi ir yra įdomesni, nei vienodi genai su sekomis, kurios yra žinomos ir naudojamos kurti modelius pradiniais genų tyrimams [4].

1.4. SAM ir BAM failų formatas

Jeigu metagenomo turinys yra žinomas arba spėjamas, galima duomenis atvaizduoti konfigūruojant SAM(The Sequence Alignment/Map) ir BAM failus ir atliekant išlyginimą pagal referentinius genomus. Failų formatas SAM yra bendras formatas skirtas užrašyti nuskaitymų lygiavimus pagal referentinę seką. Šis formatas palaiko skirtingų platformų pagamintus ilgus ir trumpus nuskaitymus. [6]. SAM formatas yra tekstinis formatas, kuriame galima patalpinti duomenis apie sekas eilutėse, kurių elementai į ASCII stulpelius yra atskiriami „tab“ skirtuku.

Dažniausiai yra generuojamas žmogui skaitomas SAM formatas, o ne seserinis atitikmuo, BAM. Antrasis formatas palaiko tuos pačius duomenis, bet jie yra binarinės formos, suspausti ir suindeksuoti. Šiuo metu didžioji dalis SAM formato duomenų yra gaunami iš lygiavimo įrankių, kurie skaito fastq failus ir atsižvelgiant į referentinį genomą priskiria sekas pozicijoms. [10]

1.5. Bowtie2

Bowtie2 yra greitas, nedaug atminties naudojantis įrankis sekų lygiavimui pagal referentinį genomą. Jis tinkamas sulygiuoti nuo 50 iki 1000 simbolių nuskaitymų sekas. Bowtie2 palaiko lokalias, su tarpais ir dvipuses lygiavimo funkcijas. Galima naudoti kelis procesorius, kad būtų pasiektas didesnis greitis. Šis įrankis sugeneruoja SAM formato failus. Įrankio panaudojimas dažniausiai yra pirmasis žingsnis norint palyginti genominius duomenis. Bowtie2 taip pat yra optimizuotas nuskaitymų ilgams ir klaidų režimams, nustatytiems naujausių sekvenavimo aparatų, kaip Illumina HiSeq 2000, Roche 454 ir Ion Torrent. [14]

1.5.1. Bowtie2 ir indeksavimas

Įrankis bowtie2-build išveda Bowtie indeksą pagal duotas DNR sekas. Yra gaunami 6 failai su plėtiniais .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2 ir .rev.2.bt2. Šie failai reikalingi išlygiuoti nuskaitymus prie genomo. Originali seka, kuri yra fasta formato failas, yra nebenaudojama kai yra sudaryti indeksai.

Panaudojimas:

```
bowtie2-build [opcijos]* <reference_in> <bt2_base>
```

- <reference_in> - kableliu atskirtas fasta failų sąrašas, kuriame yra referentinės sekos, pagal kurias vyks išlyginimai. Naudojant papildomą -c opciją yra galima paduoti pačias sekas.
- <bt2_base> - indeksavimo failų bazinis vardas.

1.5.2. Bowtie2 ir išlyginimas

Panaudojimas:

```
bowtie2 [opcijos]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} -S [<hit>]
```

1.6. Samtools

Įrankių rinkinys SAMtools realizuoja įvairias pagalbines funkcijas SAM formate, kaip indeksavimas, lygiavimų peržiūrėjimas ir pristato universalius įrankius nuskaitymų lygiavimo apdorojimui. Samtools taip pat apdoroja ir binarinį BAM formatą, kuris yra ekvivalentus SAM failui [6].

1.6.1. SAM failų konvertavimas į BAM

Komanda:

```
samtools view [opcijos] in.sam|in.bam|in.cram [region...]
```

1.6.2. BAM failų rūšiavimas

Komanda:

```
samtools sort [-l level] [-m maxMem] [-o out.bam] [-O format] [-n] [in]
```

1.6.3. BAM failų indeksavimas

Komanda:

```
samtools index [-bc] aln.bam|aln.cram [out.index]
```

1.7. IGV

IGV (Integrated Genomics Viewer) yra lengvai naudojamas vizualizacijos įrankis, kuris leidžia intuityviai ir realiu laiku tyrinėti įvairius ir didelius genominių duomenų rinkinius. Ši programa palaiko lankstų įvairių genominių duomenų integravimą, taip pat ir failus su informacija apie sekų išlyginimą, mutacijas, genų ekspresijas, metilinią, kopijų skaičių ir genomines anotacijas. IGV naudoja efektyvius, įvairius duomenis palaikančius failų formatus, kurie leidžia stebėti duomenis, nepriklausomai nuo jų dydžio ar rezoliucijos. Įrankis reikalauja mažai resursų.

Programos keičiamo dydžio architektūra padaro ją tinkamą viso geno tyrinėjimui ir naujos kartos sekvenavimo duomenims, taip pat tokiems duomenims, kaip paprastas nuskaitymų padengimas. [9]

1.8. Needleman-Wunsch algoritmas

Needleman-Wunsch (NW) algoritmas (Needleman Wunsch 1970) yra netiesinis globalus optimizavimo metodas, kuris buvo sukurtas amino-rūgščių sekų išlyginimui baltymuose. Tai buvo pirmoji iš daugumos svarbių išlyginimo technikų.

Needleman-Wunsch naudoja dinaminę programavimą tam, kad būtų gaunamas globalus išlyginimas tarp dviejų sekų. Globalus išlyginimas siūlo atsižvelgti į visus dviejų sekų elementus, kol atlieka jų išlyginimą. Jį taip pat galima vadinti „end to end“ išlyginimu. Needleman-Wunsch algoritme, pirmiausia yra sudaroma $m * n$ dydžio (m yra ilgesnioji seka, n trumpesnė) skaičiavimo matrica. Optimalus kiekvienos matricos balas yra apskaičiuojamas pridedant esamo atitikimo balą prie ankščiau apskaičiuotos pozicijos ir atimant tarpo kainas. Kiekvienos matricos pozicija gali turėti teigiamą, neigiamą arba nulinę reikšmę.

Po to kai matrica yra užpildyta, tam, kad būtų galima nustatyti optimalų sekos išlyginimą, yra naudojamas metodas vadinamas atsekimu (trace back). Atsekimas seka pozicijas skaičiavimo matricoje, kuri prisidėjo prie didžiausio balo suradimo. Pozicijos gali susilygiuoti arba būti šalia tarpo, priklausomai nuo informacijos esančios atsekimo matricoje. Taip pat gali egzistuoti keli maksimalūs išlyginimai [5].

1.8.1. needleall

Needleall skaito sekų rinkinį, palygina jas su kitu rinkiniu ir užrašo optimalų išlyginimą į failą. *Needleall* suranda išlyginimą su didžiausiu galimu balu, kuris yra lygus atitikimų sumai (paimama iš taškų matricos), atimant baudas (atsiranda iš atidarymo ir tarpų lyginamose sekose). Pakeitimo matrica, tarpų atidarymai ir praplėtimo kainos yra nusakomos vartotojo. [13]

Pagrindiniai komandinės eilutės argumentai:

```
needleall -asequence [data.fa] -bsequence [data.fa] -gapopen [10.0]  
-gapextend [0.5] -outfile [file.out]
```

1.9. Levenshtein atstumo algoritmas

Levenshtein atstumas tarp dviejų sekų yra gaunamas pagal minimalų skaičių operacijų reikalingų, kad būtų galima pakeisti vieną seką kita. Operacijos šiame kontekste yra insercija, delecija ir vieno ženklo pakeitimas. Levenshtein atstumas pamatuoja panašumus tarp dviejų įvesčių: šaltinio s ir tikslo t . Atstumas, tai skaičius delecijų, insercijų ir pakeitimų, kurie yra reikalingi s pakeisti į t . Kuo didesnis skirtumas tarp įvesčių, tuo didesnis atstumas.

1.10. EMBOSS water

Water įrankis naudoja Smith-Waterman algoritmą (modifikuotą, kad būtų galima padidinti greitį), naudojamą apskaičiuoti lokalių išlyginimą tarp sekų. Tarpo įterpimo kaina, tarpo išplėtimo kaina ir pakeitimo matrica yra naudojamos apskaičiuoti nurodytus išlyginimus. Įrankio išvedimas yra standartinis EMBOSS išlyginimo failas.

1.10.1. Algoritmas

Smith-Waterman algoritmas priklauso algoritmų klasei, kurie gali apskaičiuoti geriausią rezultatą ir lokalių išlyginimą tam tikra tvarka per $m * n$ žingsnių, kur n ir m yra dviejų sekų ilgiai. Šie dinaminio programavimo algoritmai buvo sukurti baltymų sekų palyginimui Smith'o ir Waterman'o.

Dinaminio programavimo metodai užtikrina optimalų lokalių išlyginimą tam, kad būtų išnagrinėjami visi galimi išlyginimai ir pasirinktas pats geriausias. Tai atliekama perskaitant taškų matricą, kurioje yra reikšmės kiekvienam galimam liekanos arba nukleotido sutapimui. *Water* sukuria išlyginimą su maksimaliu galimu balu, kur išlyginimo balas yra lygus sumai sutapimų gautų iš taškų matricos.

Svarbi yra susidorojimo su tarpais problema. Bauda yra atimama iš taškų už kiekvieną tarpo atidarymą, ir iš rezultatų už bendrą tarpų skaičių, padaugintą iš kainos. Dažniausiai, kaina už tarpo išplėtimą yra nustatyta, kad būtų 5-10 kartų mažesnė už tarpo atidarymo kainą. [3]

Algoritmo panaudojimas ir pagrindiniai parametrai:

```
water -asequence [data.fa] -bsequence [data.fa] -gapopen [10.0]  
-gapextend [0.5] -outfile [file.out]
```

1.11. EDLIB biblioteka

Myers bitų vektoriaus algoritmas transformuoja dinaminio programavimo matricą. Transformacija leidžia laikyti kelis elementus kaip bitų vektorių viename procesoriaus registre ir pasiekti paralelizavimą. Mayers taiko papildomus algoritmus, kad būtų galima sumažinti apskaičiuotų bitų vektorių skaičių.

Originalus algoritmas buvo sukurtas, kad būtų galima atlikti pusiau globalų išlyginimą, kur už tarpus eilutės pradžioje ir pabaigoje nėra baudžiama. O *Edlib* bibliotekoje, Mayers algoritmas yra išplėstas taip, kad palaikytų globalų ir pusiau globalų išlyginimą, kai už tarpus eilutės pabaigoje nėra baudžiama.

Mayers algoritmas negrąžina jokios informacijos apie optimalų išlyginimo kelią (optimali redagavimo operacijų seka, kurią reikia atlikti užklausos (query) sekoje, kad būtų galima ją paversti į užduoties (target) seką). *Edlib* bibliotekoje algoritmas yra išplėstas taip, kad būtų galima surasti optimalų išlyginimo kelią trimis įgyvendintiems išlyginimo metodams. Globalus (NW) – standartinis metodas, taikomas naudojant šią biblioteką kai yra norima apskaičiuoti redagavimo atstumą. Šis metodas tiksliausiai pasako kaip viena seka yra panaši į kitą. Priešdelio (SHW) – šis metodas labai panašus į NW, skirtumas tas, kad už neatitikimus sekų pradžioje nėra skiriama bauda. Intarpo (HW) – nuo pirmųjų dviejų būdų šis skiriasi tuo, kad už tarpus pradžioje ir pabaigoje nėra skiriama bauda. [1]

2. Praktinė tiriamoji dalis

2.1. Pirmasis eksperimentas ir motyvacija

Pirmieji bandymai buvo atliekami su duomenimis gautais iš straipsnio „*Genome diversity of marine phages recovered from Mediterranean metagenomes: Size matters*“ publikuoto 2017 metų rugsėjo mėnesį [7]. Šių duomenų SRA SRR926399, dydis fastq formatu 52.2 GB, o fasta formatą sudarė 26.1 GB. Pirminis tyrimo tikslas išanalizuoti straipsnyje pateiktus duomenis ir, remiantis ten aprašyta metodika, atlikti duomenų analizę. Pirmasis žingsnis buvo BLAST panaudojimas. Iš NCBI atsisiųsti 9568 virusų genomai ir iš nuskaitymų failo sudaryta lokali duomenų bazė. Tyrimas buvo atliekamas turint standartinį, ofiso tipo, kompiuterį: Intel(R) Core(TM) i7-3537U ir 8GB RAM. Pradėjus eksperimentą pastebėta, kad su turimais resursais kyla problemų dėl failų apdorojimo ir laiko. Todėl pasirinkta turimą failą padalinti į dalis. Visas duomenų rinkinys buvo suskirstytas į 20 dalių ir atliekama analizė naudojant BLAST įrankį.

Lokalių duomenų bazės kūrimas:

```
makeblastdb -in data.fasta -dbtype nucl
```

BLAST paieška:

```
blastn -db database -query data.fasta -out data.out
```

Gauti rezultatai atskleidė, kad tarp turimo metagenominio rinkinio yra randama virusų genomų.

Rezultatų pavyzdys:

```
Query= refseq|NC_006883|252401nt|Prochlorococcus phage P-SSM2, complete
genome
Length=252401
Sequences producing significant alignments:
```

	Score (Bits)	E Value
SRR926399.1295772 HWI-ST1218:202:D1J10ACXX:5:1101:11611:151293 ...	193	1e-44
SRR926399.3353860 HWI-ST1218:202:D1J10ACXX:5:1102:19012:181423 ...	191	5e-44
SRR926399.2183021 HWI-ST1218:202:D1J10ACXX:5:1102:13331:52713 l...	191	5e-44
SRR926399.2138552 HWI-ST1218:202:D1J10ACXX:5:1102:18306:47555 l...	191	5e-44
SRR926399.1458377 HWI-ST1218:202:D1J10ACXX:5:1101:5436:169768 l...	191	5e-44
SRR926399.3704024 HWI-ST1218:202:D1J10ACXX:5:1103:17755:20963 l...	189	2e-43
SRR926399.3703698 HWI-ST1218:202:D1J10ACXX:5:1103:15190:20795 l...	189	2e-43
SRR926399.3563061 HWI-ST1218:202:D1J10ACXX:5:1103:11471:5542 le...	189	2e-43
SRR926399.2798038 HWI-ST1218:202:D1J10ACXX:5:1102:16984:120723 ...	189	2e-43
SRR926399.2034443 HWI-ST1218:202:D1J10ACXX:5:1102:5669:36105 le...	189	2e-43
SRR926399.1305536 HWI-ST1218:202:D1J10ACXX:5:1101:19714:152355 ...	189	2e-43
SRR926399.691023 HWI-ST1218:202:D1J10ACXX:5:1101:14629:82103 le...	189	2e-43
SRR926399.457928 HWI-ST1218:202:D1J10ACXX:5:1101:15325:55088 le...	189	2e-43

2 pav. BLAST rezultatų pavyzdys

Pagal gautus rezultatus pastebėta, kad yra randamas didelis skaičius atitikimų su įvairias *Prochlorococcus* fago ir *Synechococcus* fago genomais. Turint tokius rezultatus nuspręsta atlikti tolimesnę analizę. Tam, kad būtų įsitikinta gautų rezultatų teisingumu, gauti duomenys turėjo būti atvaizduoti. Atrinkti genomi, kurių rezultatai buvo vieni geriausių (daugiausia atitikimų). Naujas genominių duomenų rinkinys buvo sudarytas iš 75 cianofagų genomų (vietoje 9568). Šį rinkinį sudarė virusų, puolančių *Prochlorococcus* ir *Synechococcus* genties cianobakterijas, genomi. Rezultatų atvaizdavimui sugeneruoti SAM ir BAM failai (naudojant Bowtie2) ir panaudotas IGV įrankis.

Failų indeksavimas:

```
bowtie2-build -f virus.fa virus_indexes
```

SAM failų gavimas:

```
bowtie2 -x virus_indexes -U data.fa -S data.sam
```

SAM konvertavimas į BAM:

```
samtools view -b -S -o data.sam data.bam
```

BAM failų rūšiavimas ir indeksavimas:

```
samtools sort data.bam data.sorted
```

```
samtools index data.sorted.bam
```

Turint visus reikiamus failus, kelis kartus buvo bandoma atlikti duomenų vizualizavimą naudojant IGV įrankį. Tačiau, bandymai buvo nesėkmingi. Kiekvieną kartą išvesties pranešimas teigė, kad atitinkančių nuskaitymo sekų yra tik kelios. Atvaizduojant tokius duomenis nebuvo gauta jokių rezultatų.

Po pirmojo eksperimento buvo pažvelgta į duomenis kitaip. Nuspręsta, kad šių duomenų analizei pasirinktas būdas galėjo būti netinkamas ir pradėta abejoti pačių duomenų kokybe. Kiekviename eksperimente pasitaiko klaidų dėl įkeltų tuščių duomenų nuorodų, duomenys gali būti neteisingi, per mažai dokumentuoti. Suteikiamas nepakankamas informacijos kiekis gali privesti prie klaidų, o duomenys gali pasirodyti ne tokie, kokie atrodė iš pradžių. Apsvarstant jau turimus duomenis, buvo prieita išvada, kad šiuo metu neturima reikalingų resursų tyrimui tęsti. Iškelta idėja, kad prieš pradedant eksperimentą duomenis reikia kaip nors patikrinti. Toks tyrimas neatliks pilnos duomenų analizės, bet suteiks galimybę pažvelgti į juos prieš investuojant į darbą.

Taip nuspręsta atlikti pirminę metagenominių duomenų analizę pasitelkiant jau žinomus algoritmus ir įrankius. Tokia pirminė analizė suteikia pradinę informaciją apie metagenominius duomenis. Svarbiausia, kad nors tokia analizė reikalauja laiko, bet teisingai pritaikius įrankius jis gali būti sutaupytas vėliau, tyrimo eigoje.

2.2. Darbo metodika

Norint atlikti paviršutinę metagenominių duomenų analizę buvo naudojama tokia metodika: Panaudoti įrankiai *emboss:needleall*, *emboss:water* ir *EBLIB* biblioteką. *Needleall* įrankis naudojamas atlikti globalųjį, o *water* lokalųjį išlyginimą. *EBLIB* biblioteka panaudota kaip įrankis redagavimo atstumui apskaičiuoti. Norint atlikti duomenų analizę, kai lyginami nuskaitymai (angl. reads) *all-versus-all* metodu, yra pradedama nuo C++ kalba parašytos programos panaudojimo. Ši programa leidžia iš metagenominio failo atsitiktinai paimti dalį nuskaitymų ir sudaryti du skirtingus duomenų failus, kurie bus naudojami tyrimo eigėje. Turint šiuos duomenis, darbas yra atliekamas naudojant *needleall* įrankį, kuris apskaičiuoja išlyginimo balus tarp visų sekų ir užrašo juos į failą. Toliau, panaudojus tuos pačius duomenis, atliekami bandymai su *water* įrankių ir pritaikyta pagal poreikius *EBLIB* biblioteka. Ši biblioteka yra panaudojama taip, kad, gavus du failus su nuskaitymais, *all-versus-all* principu būtų apskaičiuojamas visų sekų porų redagavimo atstumas. Visa informacija, gauta panaudojus įrankius, yra surašoma į atitinkamus išvesties failus. Iš šių duomenų yra ruošiamos histogramos ir skaičiuojamas vidurkis, mediana, didžiausios ir mažiausios reikšmės. Taip pat, apskaičiuojamas skaičius porų, kurių išlyginimo balas yra didesnis nei 80% nuo didžiausio galimo balų skaičiaus.

Tokiu būdu, lyginant nuskaitymus tarpusavyje, yra sužinoma platesnė informacija apie duomenis. Sužinomas panašumas tarp metagenominiame faile esančių sekų. Jeigu panašumas yra didelis, gali būti viena iš kelių priežasčių. Organizmas galėjo būti susekvenuotas kelis kartus arba tame mėginyje buvo daugiau nei vienas tos pačios rūšies organizmas. Tai gali reikšti, kad tam tikroje terpėje yra labiau paplitusi viena ar kelios rūšys. Taip pat, didelis panašumas gali parodyti tai, kad duomenų rinkinyje yra organizmų, kurie yra genetiškai panašūs.

Toks pat tyrimas yra atliekamas ir su atrinktais organizmais, kurie yra atrenkami pagal informaciją apie metagenomą. Iš failo su organizmų sekomis yra atsitiktinai paimama jų dalis, kuri yra lyginama su duomenimis iš metagenomo. Taip lyginant dvejus duomenų rinkinius yra bandoma surasti panašumų, kurie parodytų ar metagenominiame faile yra randama tam tikrų organizmų. Tokio tyrimo metu yra sunkiau rasti poras, kurių balas peržengia 80% procentų ribą. Dalis nuskaitymo sekų gali turėti ne pilną tam tikro organizmo informaciją, todėl lyginant jas su sekomis iš genomo gali būti gautas dalinis atitikimas, dėl kurio būtų gaunamas mažesnis išlyginimo balas.

2.3. Greitas didelės apimties metagenomo skaidymas

Žinant, kad teks dirbti su didelės apimties metagenominiais duomenimis, buvo svarstoma kaip būtų galima juos greitai padalinti atliekant eksperimentą, nes darbas su visu failu labai prailgintų tyrimo laiką. Pirminė analizė neturi atlikti viso eksperimento, o tik suteikti pradinę informaciją, todėl nuspręsta rašyti programą, kuri duomenų faile atsitiktinai nustato vietą ir paima dalį sekų. Norint šiuo metodu gauti įvairesnius ir tikslesnius rezultatus, galima eksperimentą kartoti daug kartų. Programa parašyta C++ kalba.

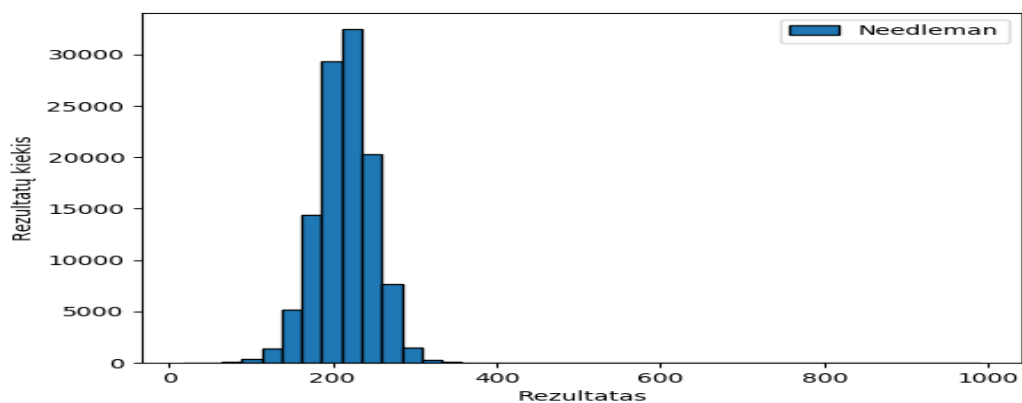
2.4. Needleall įrankio panaudojimas

Šis įrankis panaudojamas tam, kad būtų galima sužinoti globalųjį sekų išlyginimo balą. Naudojant šį įrankį įvestis buvo dveji failai su nuskaitymų sekomis, atsitiktinai atrinktomis iš metagenominio failo. Kiekvieną failą sudarė 200 sekų. Pirmasis bandymas vyko su ankščiau minėtu 26.1 gb metagenominiu failu, kuris gautas iš straipsnio „*Genome diversity of marine phages recovered from Mediterranean metagenomes: Size matters*“. Tarpo atidarymo ir išplėtimo kaina pasirinkta 4. Panaudota numatytoji *ednafull* matrica.

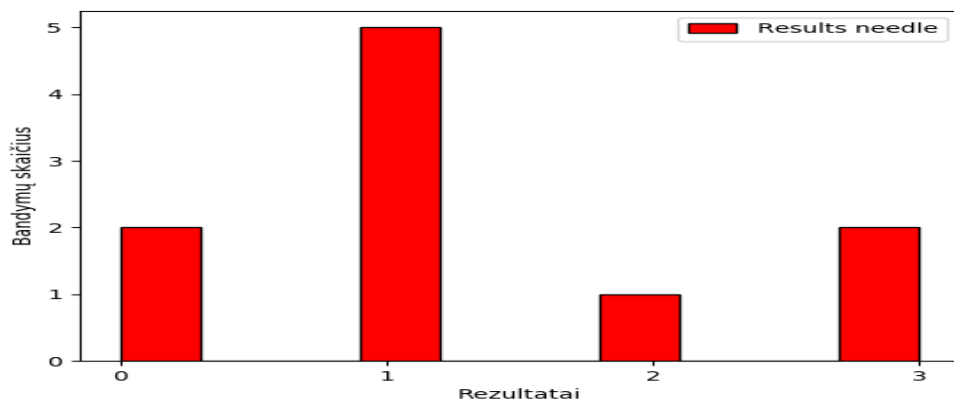
Panaudota komanda:

```
needleall -asequence first.fa -bsequence second.fa -gapopen 4  
-gapextend 4 -outfile results.out
```

Atlikus bandymą su *needleall* įrankių gauti tokie rezultatai:



3 pav. Pirminė analizė naudojant *needleall* įrankį



4 pav. Pirminės analizės rezultatai naudojant *needleall*

Rezultatai <i>needleall</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
215.7043	216	16	992	13

1 lentelė. *needleall* rezultatų lentelė

Šio metagenomo tyrime *needleall* panaudotas 10 kartų, duomenys kaupti rezultatų faile. Metagenominio duomenų failo nuskaitymų ilgis yra 202, todėl naudojant prieš tai minėtus parametrus didžiausias išlyginimo balas galėjo būti 1010. Didžioji dalis rezultatų grupuojasi ties 150-300 taškų riba. Rezultatų lentelėje (1 lentelė), kurioje pateiktas porų, peržengusių 80% ribą, skaičius yra 13. Rezultatų lentelė patvirtina histogramų rezultatus. Visų gautų balų vidurkis yra 215, o maksimali vertė 992, kuri peržengė prieš tai nustatytą ribą. Kaip jau buvo minėta, kiekvieno eksperimento metu į du failus buvo atrenkama po 200 sekų. Tai sudaro labai nedidelę metagenomo dalį. Palyginus šiuos duomenų dydžio skirtumus, iš to, kad pavyko rasti nors ir mažą didelį balą turinčių porų skaičių, galima daryti išvadas apie metagenominius duomenis. Atlikus tokią paviršutinę duomenų analizę ir radus panašių sekų porų, galima teigti, kad metagenominiame faile yra daugiau nei vienas toks pat arba keli panašaus genomo organizmai.

2.5. EDLIB biblioteka

Tolimesnei analizei, tam, kad būtų galima apskaičiuoti redagavimo atstumą, panaudota *EDLIB* biblioteka. Biblioteka yra pritaikyta C++ ir Python kalboms. Naudojant šią biblioteką redagavimo atstumui apskaičiuoti galima pasirinkt vieną iš trijų metodų. Šiame tyrime panaudotas HW metodas.

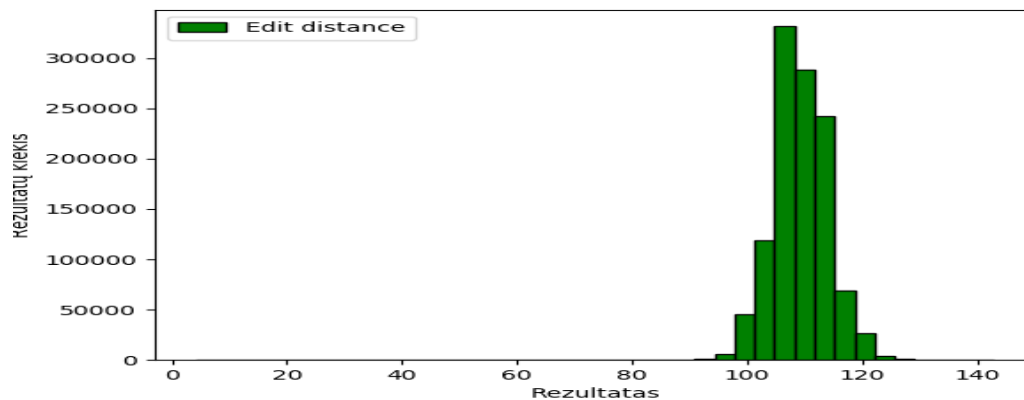
2.5.1. Bibliotekos pritaikymas

Biblioteka, panaudojus Python programavimo kalbą, pritaikyta taip, jai būtų galima paduoti jau turimus du sekų failus ir gauti visų sekų porų redagavimo atstumą.

Vienas iš būdų panaudoti *EDLIB* biblioteką Python kalba:

```
result = edlib.align(seq1, seq2, mode="HW")
print(result["editDistance"])
```

Gautų rezultatų pavyzdys:



5 pav. Pirminė analizė naudojant *edlib*

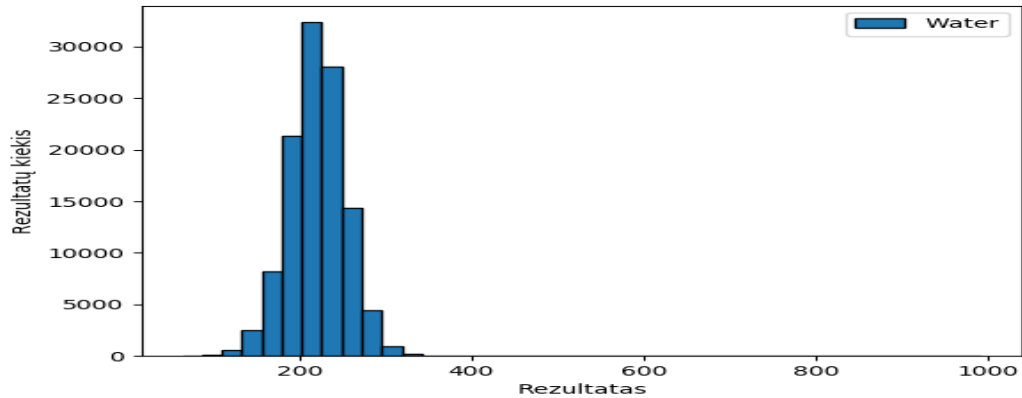
Rezultatai redagavimo atstumo			
vidurkis	mediana	minimumas	maksimumas
109.1882	109	4	143

2 lentelė. Redagavimo atstumo rezultatų lentelė

Šis tyrimas taip pat buvo atliekamas 10 kartų, su tokiais pat failais kaip ir naudojant *needleall* įrankį. Iš histogramos matoma, kad redagavimo atstumo rezultatai labiausiai pasiskirstę ties 100 – 120 balų riba. Mažiausias atstumas 4, didžiausias 143 (2 lentelė), kas parodo, kad šio tyrimo metu buvo rasta labai panašių sekų. Sprendžiant pagal gauta vidurkį, 109, galima pasakyti, kad metagenominiame faile yra panašių sekų, tačiau iš šių rezultatų negalima nustatyti ar yra tam tikrų regionų, kur sekos skiriasi, ir ar skirtumai tarp šių sekų yra išsibarstę atsitiktinai.

2.6. Water įrankio panaudojimas

Tolimesnė bandymo eiga buvo *water* įrankio panaudojimas. Šio įrankio pagalba buvo gaunami rezultatai naudojant lokalųjį išlyginimo būdą. Lokalusis išlyginimas yra tinkamiausias naudoti tada, kai sekos gali būti nevienodo ilgio ir dalinai panašios, kai tik dalis sekos gali atitikti kitą seką. Įrankis buvo pritaikytas taip, kad būtų galima gauti du failus su nuskaitymais ir atlikti išlyginimą naudojant *all-versus-all* principą, kai visi nuskaitymai yra palyginami vieni su kitais. Tarpo atidarymo ir išplėtimo kaina naudota tokia pat kaip ir taikant *needleall* įrankį – 4.



6 pav. Pirminė analizė naudojant water įrankį

Rezultatai <i>water</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
221.2079	221	64	992	13

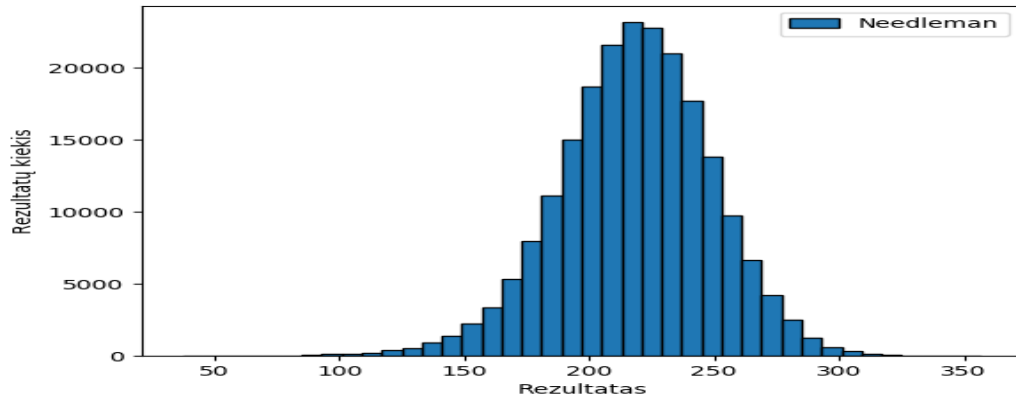
3 lentelė. *water* rezultatų lentelė

Iš duomenų (6 pav.) yra matoma, kad dauguma rezultatų susigrupavo ties 200 - 300 riba. Naudojant šį metodą, sekų su aukštais balais buvo rasta tiek pat. Nors rezultatų vidurkis yra aukštesnis, todėl galima teigti, kad duomenų failuose buvo dalinai persidengiančių sekų, ko negalima nustatyti naudojant *needleall* įrankį. Tokie rezultatai parodo, kad duomenų faile yra panašių arba vienodų sekų, bet tai taip pat parodo sekas, kurios yra panašios ne per visą sekos ilgį.

2.7. Tyrimas su virusais

Panaudojus prieš tai aprašytą metodiką, su tais pačiais duomenimis, buvo atliktas tyrimas su virusų genomais. Šiam bandymui iš NCBI puslapio buvo atsisiųstas 201 cianofagų genomai. Tyrimas, nuo prieš tai atlikto, skyrėsi tuo, kad šį kartą vieną iš duomenų failų sudarė duomenys gauti iš genomio failo, imant atsitiktines nustatyto ilgio sekas.

Pirmą kartą atliekant tokį bandymą buvo gaunamos 202 bp ilgio sekos iš genomio failo. Gauti rezultatai:

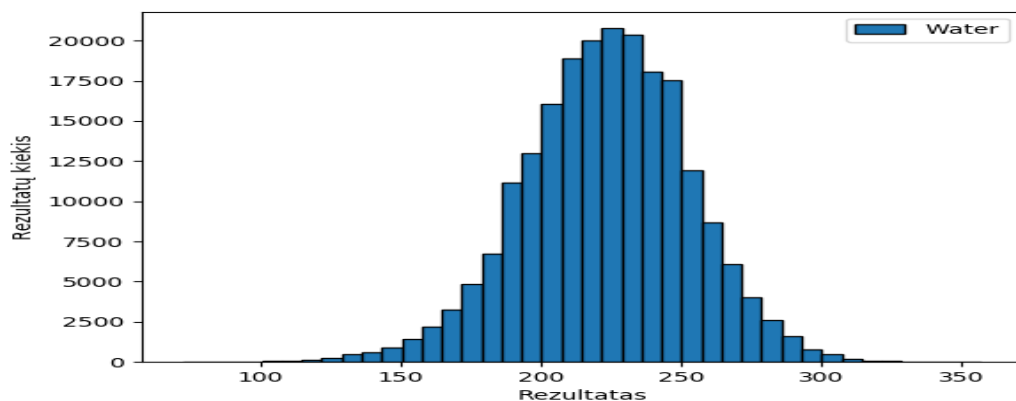


7 pav. Bandymas su virusais. Rezultatai pagal *needleall*

Rezultatai <i>needleall</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
216.8983	218	16	412	0

4 lentelė. *needleall* rezultatų lentelė

Šio tyrimo metu *needleall* įrankis neparodė gerai (daugiau už 80%) įvertintų rezultatų. Maksimalus galimas taškų skaičius, kaip ir pirmojo bandymo metu, išliko 1010, bet šį kartą didžiausia surasta reikšmė buvo tik 412. Kas yra pakankamai geras rezultatas žinant tai, kad ne kiekviename nuskaityme bus paslėpta tik tam tikro viruso genetinė informacija. Gali būti, kad tik dalis sekos turės savyje ieškomą genomo dalį. Taip pat šioje histogramoje rezultatų pasiskirstymas yra platesnis, tai reiškia, kad yra galima tikėtis dalinių panašumų tarp sekų porų.

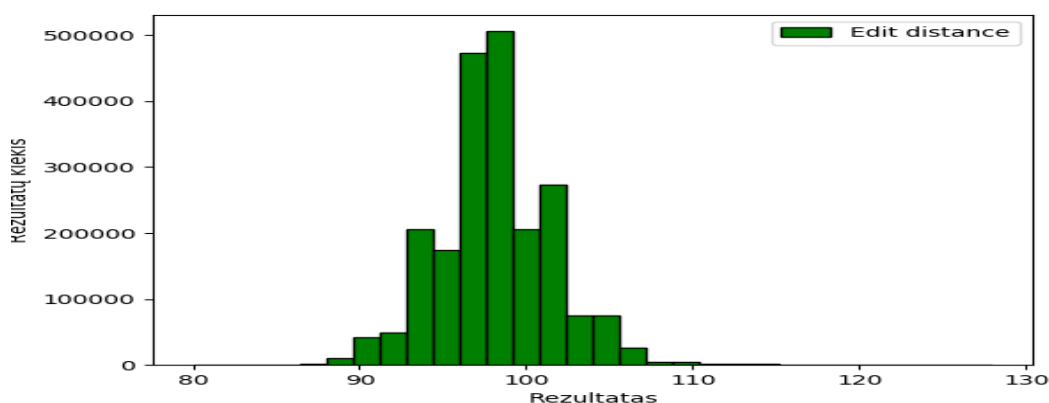


8 pav. Bandymas su virusais. Rezultatai pagal *water*

Rezultatai <i>water</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
222.1401	223	45	415	0

5 lentelė. *water* rezultatų lentelė

Atlikus tyrimą su *water* įrankiu taip pat nebuvo rasta sekų porų turinčių gerą įvertinimą. Duomenys skiriasi nuo gautų naudojant *needleall*. Šiuo atveju vidurkio, maksimalių ir minimalių taškų skaičiai didesni. Taip pat iš histogramos (8 pav.) matosi, kad buvo daugiau sekų porų, kurių rezultatai varijuoja nuo 150 iki 200. Turint tokius rezultatus galima spėti, kad dalis sekų, kurios susilygiavo, buvo persidengusios, todėl rezultatai yra geresni nei naudojant globalųjį išlyginimą.



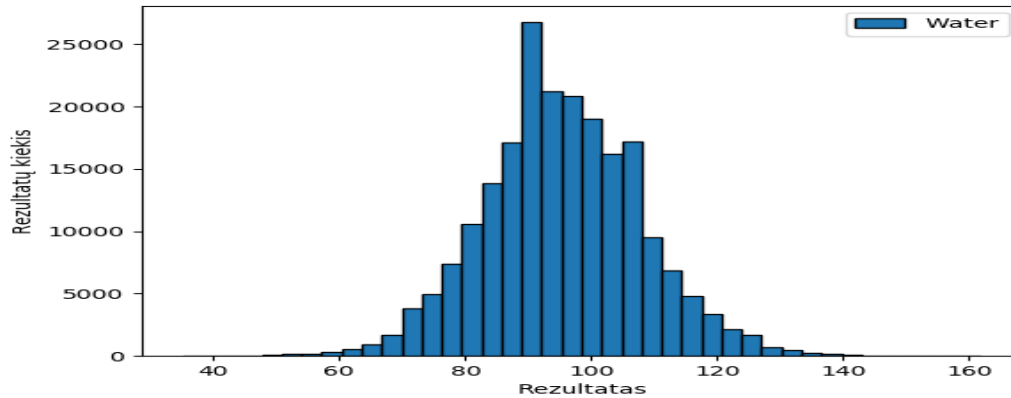
9 pav. Bandymas su virusais. Rezultatai pagal redagavimo atstumą

Rezultatai redagavimo atstumo			
vidurkis	mediana	minimumas	maksimumas
96.3992	96	75	123

6 lentelė. Redagavimo atstumo rezultatų lentelė

Iš redagavimo atstumo rezultatų yra matoma, kad vidurkis yra panašus į prieš tai gautus. Bet šiuo atveju mažiausia buvusi reikšmė yra daug didesnė.

Toks pat bandymas buvo pakartotas dar kartą, tik naujo bandymo metu iš failo su virusų genomis buvo imamos sekos, kurių ilgiai yra 60bp. Gauti rezultatai:



10 pav. Bandymas su virusais. Rezultatai pagal *water*

Rezultatai <i>water</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
95.3371	95	35	162	0

7 lentelė. *water* rezultatų lentelė

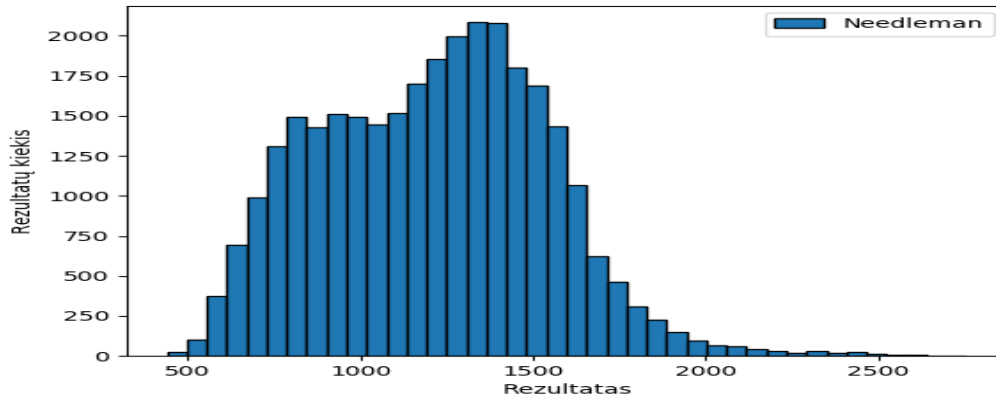
Šio tyrimo metu buvo atsižvelgiama į su *water* įrankiu gautus rezultatus ir tikimasi, kad su mažesnėmis sekomis bus pasiekti geresni rezultatai. Didžiausias galimas taškų skaičius buvo 300, o porų su didžiausiu balu rasta 162. Tai nėra pakankamai geras rezultatas tokioms trumpoms sekoms. Iš tokių rezultatų galima spręsti, kad tyrimui reikia galingesnių įrankių ir tikslingiau atrinktų organizmų.

2.8. Tolimesni metodikos bandymai

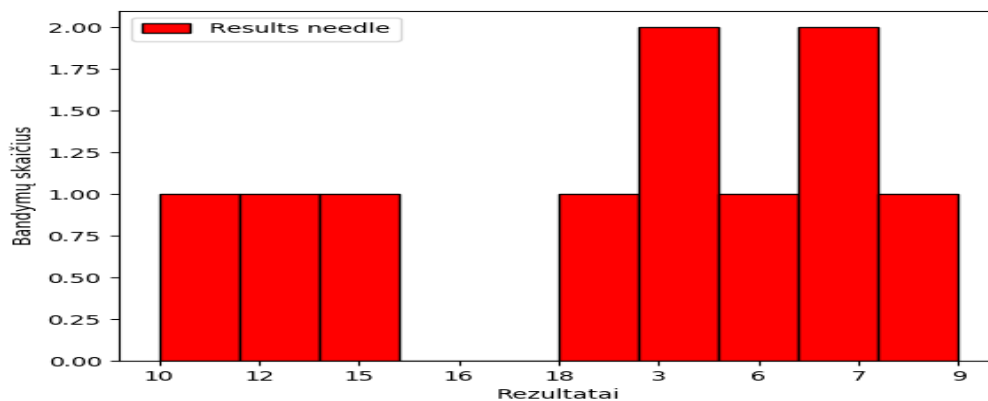
Tolimesnių eksperimentų metu buvo atliekama pirminė metagenominė analizė naudojant jau prieš tai minėtą metodiką, kuri buvo naudota atliekant tyrimą su pirmuoju metagenomu. Bandymai atlikti su dar 6 skirtingais metagenominiais rinkiniais, kurie buvo atsisiųsti iš NCBI SRA. Bandymas su kiekvienu duomenų rinkiniu buvo atliktas 10 kartų, visų bandymų trukmė iki 10 minučių. Bandymai atliekami lyginant tik nuskaitymus. Prie kiekvieno bandymo aprašyta trumpa informacija apie duomenų rinkinį ir gautus rezultatus.

Antrasis metagenomas, tai mėginys paimtas iš dirvožemio. Jį įkėlė *ICAR-National Rice Research Institute*. Šis mėginys buvo paimtas tam, kad būtų galima tirti bakterijų bendruomenes dirvožemyje, kuriame auginami ryžiai ir kuris ilgą laiką buvo purškiamas pesticidais. Failo dydis 180.6 MB fasta formatu, nuskaitymų ilgis apie 560bp, failo SRA SRR7110584.

Gauti rezultatai:



11 pav. Antrojo metagenomo rezultatai naudojant *needleall*

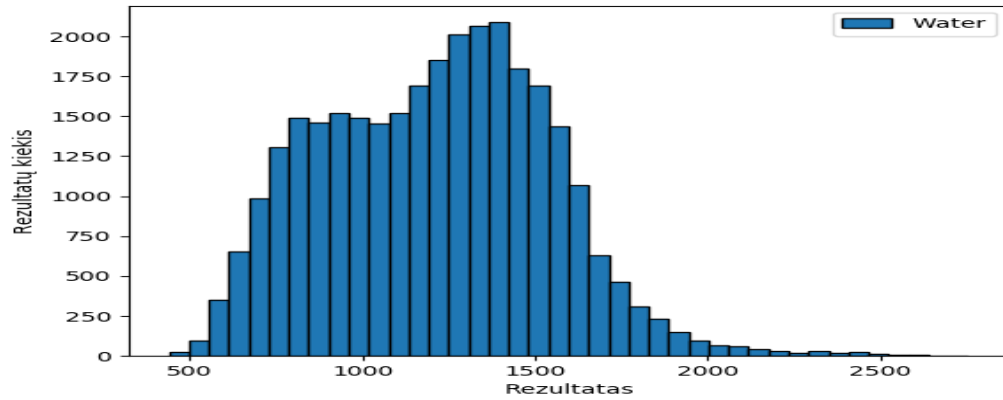


12 pav. Antrojo metagenomo rezultatai

Rezultatai <i>needleall</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
1205.2290	1227	420	2789	103

8 lentelė. *needleall* rezultatų lentelė

Šiame tyrime buvo naudojami tokie patys parametrai kaip ir pirmajame. Pagal nuskaitymų ilgį yra žinoma, kad didžiausias galimas balų skaičius yra apie 2800, kadangi metagenominiame faile nuskaitymu ilgai skiriasi. Pažvelgus į rezultatus yra matomas labai platus pasiskirstymas, kuris varijuoja nuo 600 – 1600 su dideliu porų skaičiumi. Taip pat pat matoma, kad rezultatų pasiskirstymas neatitinka standartinio. Taip galėjo nutikti dėl nevienodo nuskaitymo sekų ilgio, kadangi metagenominiame faile jų ilgiai varijuoja nuo 250 bp iki 560 bp. Patikrinus kiek yra porų, kurios peržengė 80% ribą nuo didžiausio galimo balų skaičiaus, jų rastą 103. Tokie rezultatai parodo, kad duomenų faile yra tokių pat arba panašių organizmų.

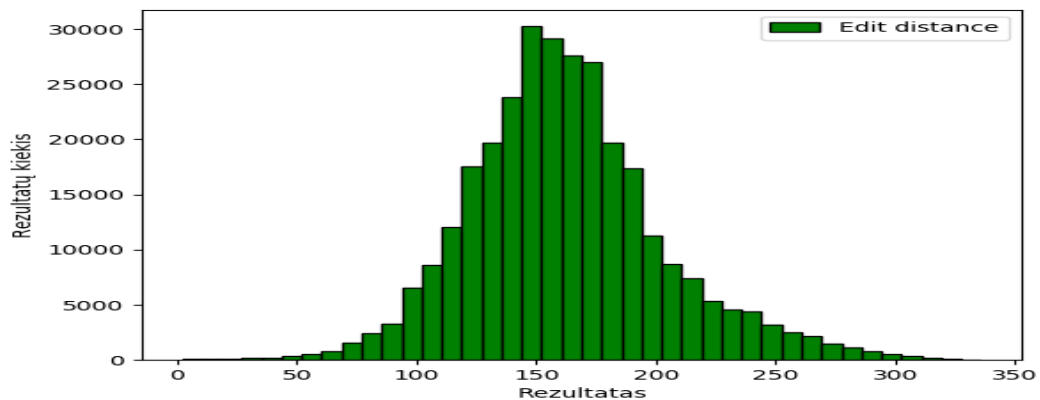


13 pav. Antrojo metagenomo rezultatai naudojant *water*

Rezultatai <i>water</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
1203.58	1225	420	2789	103

9 lentelė. *water* rezultatų lentelė

Panaudojus *water* įrankį rezultatų pasiskirstymas išlieka toks pat platus kaip ir naudojant *ne-edleall*. Rezultatų vidurkis nuo gautų naudojant pirmąjį įrankį beveik nesiskiria, tai parodo, kad sekų, kurios buvo dalinai persidengusios nebuvo surasta šio bandymo metu.



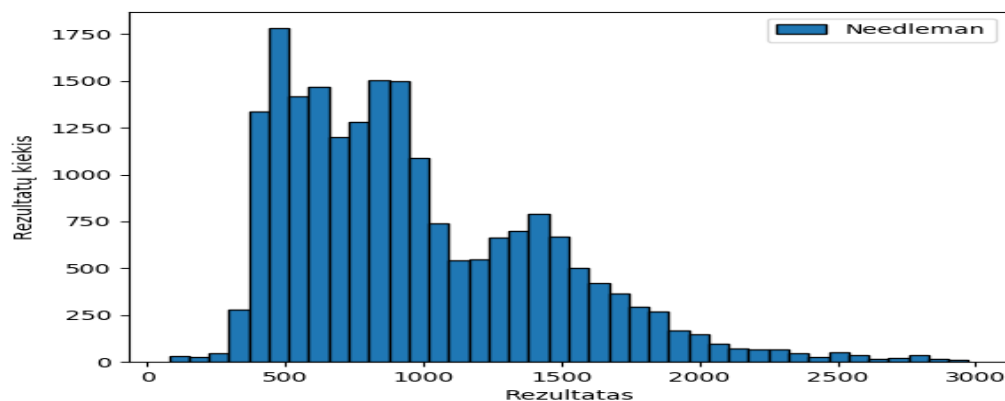
14 pav. Antrojo metagenomo redagavimo atstumo rezultatai

Rezultatai redagavimo atstumo			
vidurkis	mediana	minimumas	maksimumas
160.3832	157	2	366

10 lentelė. Redagavimo atstumo rezultatų lentelė

Apskaičiavus redagavimo atstumą pastebėta, kad mažiausia reikšmė buvo 2. Toks rezultatus leistų teigti, kad buvo surastos beveik visiškai vienodos sekų poros, bet tokiu atveju turėjo būti pasiekta didesnė reikšmė su *needleall* ir *water* įrankiais, nes apytikslė didžiausia reikšmė turėtų būti 2800. Bet kaip jau buvo anksčiau pastebėta, šiame faile nuskaitymų ilgiai skiriasi, todėl panašiausios poros galėjo būti ir vienos iš trumpiausių, kurios yra sutinkamos metagenominiame faile. Tokiu atveju taip pat yra galima daryti išvadas, kad didžiausia galima reikšmė galėtų būti ir mažesnė, nes nėra tikimybės, kad vieni ilgiausių nuskaitymų bus atsitiktinai atrinkti tyrimo metu. Sumažinus didžiausio galimo rezultato ribą, rastų porų skaičius, kurie peržengia 80% būtų dar didesnis. Tai dar kartą parodytų, kad metogianominiame faile yra organizmų, kurių genominė informacija yra panaši arba atlikus tolimesnius tyrimus būtų galima surasti dominuojančius organizmus toje aplinkoje.

Trečiasis duomenų rinkinys buvo naudotas tirti gorgonijų mikrobiomą. Failo SRR3180808, dydis 631.5 MB, nuskaitymų ilgis 600 bp, maksimalus galimas surinkti išlyginimo balų skaičius 3000.

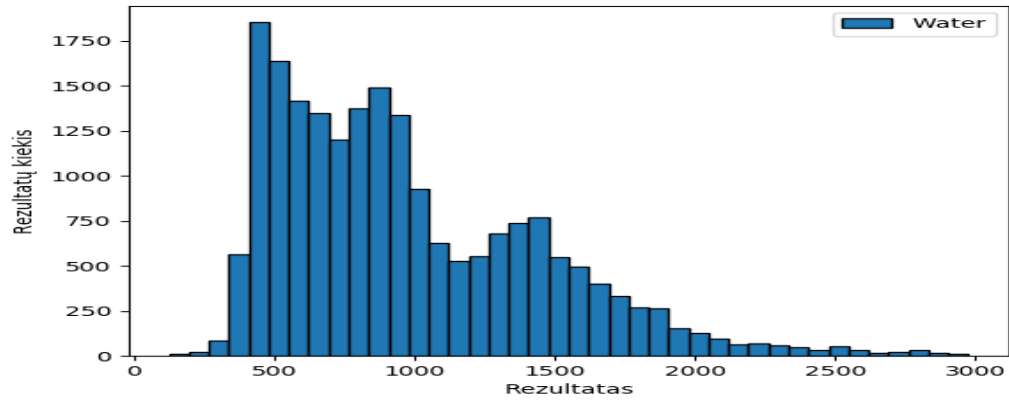


15 pav. Trečiojo metagenomo rezultatai naudojant *needleall*

Rezultatai <i>needleall</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
969.9096	866	78	2975	213

11 lentelė. *needleall* rezultatų lentelė

Pagal gautus rezultatus, panaudojus *needleall* įrankį, yra matoma, kad didelė nuskaitymo sekų porų dalis surinko nuo 500 - 1000 balų įvertinimą, taip pat yra galima pastebėti tris pakilimus. Šiame metagenomiame rinkinyje nuskaitymų ilgis buvo fiksuotas, visuomet, 600 bp. Todėl būtų galima teigti, kad toks rezultatų pasiskirstymas įvyko, nes kai kurios poros viso eksperimento metu buvo atrinktos daugiau nei vieną kartą. Taip pat aiškiai matoma, kad yra nemažai rezultatų, kurie pasiekė aukštus balus. Rasta 213 porų, kurių rezultatai peržengė nustatytą ribą.

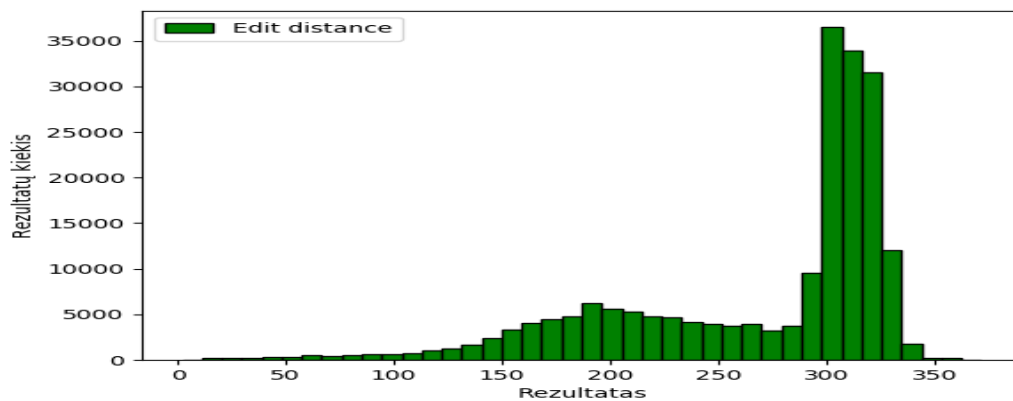


16 pav. Trečiojo metagenomo rezultatai naudojant *water*

Rezultatai <i>water</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
974.9898	868	125	2975	213

12 lentelė. Rezultatai pagal *water* įrankį

Atliekant tyrimą su *water* įrankiu situacija yra labai panaši. Matomas nereguliarus rezultatų pasiskirstymas, kuris galėjo atsirasti dėl to, nes sekos galėjo būti paimamos kelis kartus arba tarp metagenominio rinkinio pasitaikė nuskaitytų sekų, kurių ilgiai skiriasi.



17 pav. Trečiojo metagenomo analizė naudojant redagavimo atstumą

Rezultatai redagavimo atstumo			
vidurkis	mediana	minimumas	maksimumas
269.7643	302	2	372

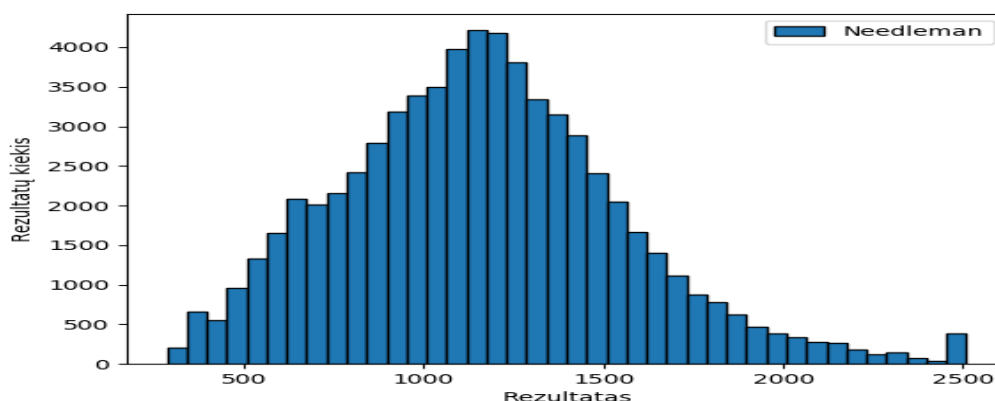
13 lentelė. Redagavimo atstumo rezultatai

Redagavimo atstumo apskaičiavimas tik dar kartą patvirtina prieš tai padarytas išvadas. Galėjo nutikti taip, kad atrinktos sekos kartojasi. Taip pat yra matoma, kad nemaža dalis porų surinko mažus redagavimo balus, kas parodo jų panašumą.

Daryti tiksliai išvadas apie šį duomenų rinkinį yra sunku, kadangi atlikus tyrimą yra matomas labai nereguliarus duomenų pasiskirstymas. Tai leidžia daryti išvadas, kad šiuo metu naudojamas sekų atrinkimo būdas nėra tinkamas ir jį reikėtų tobulinti. Žinoma, gali būti, kad tokie rezultatai yra ne šios klaidos pasekmė. Bet patikrinti kitas teorijas neužtenka paviršutinio metagenominio tyrimo ir reikėtų atlikti tikslesnį bandymą.

Ketvirtasis bandymas atliktas su metagenomu, kuris buvo naudojamas tirti kvėpavimo takų virusus. Šis metagenomas įkeltas 2018-04-12, jį pateikė *CDC Pathogen Discovery Team*. Prie metagenominės informacijos buvo pridėta, kad šiame mėginyje tikėtasi rasti žmogaus respirovirusą 1. Šio duomenų failo SRA SRR6994554, dydis 260.6 mb, nuskaitymų ilgis apie 450 bp, maksimalus galimas surinkti išlyginimo skaičius apie 2250.

Gauti tokie rezultatai:

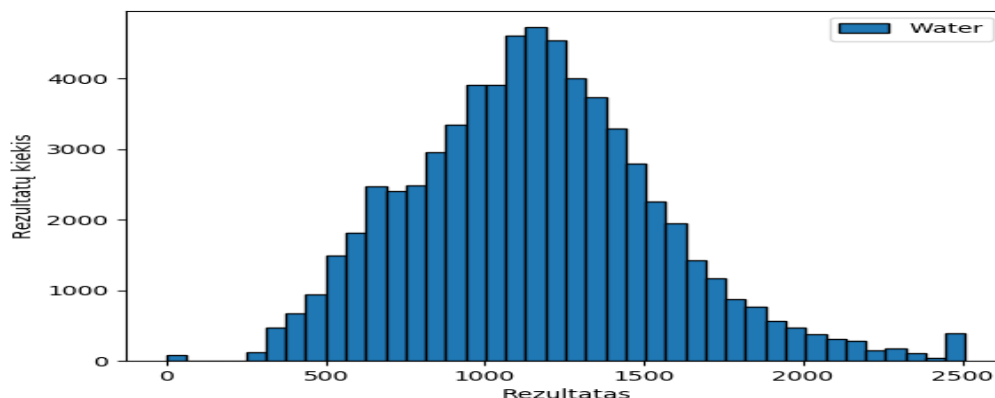


18 pav. Ketvirtojo metagenomo rezultatai naudojant *needleall*

Rezultatai <i>needleall</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
1159.5055	1146	285	2510	4469

14 lentelė. *needleall* rezultatų lentelė

Iš rezultatų galima pamatyti, kad surasta labai daug porų, kurių išlyginimo balas yra labai didelis. Kaip prieš tai buvo minėta, apytikslis sekų ilgis yra 450 bp, bet jis varijuoja nuo 300 iki daugiau nei 500 bp, todėl lentelėje yra matomas didesnis maksimalus surinktas balas, nei manyta prieš pradedant tyrimą. Iš histogramos pastebimas labai ryškus pakilimas ties maksimalia taškų riba. Tai parodo, kad buvo rasta daug vienodų arba beveik vienodų sekų porų. Dėl to, galima teigti, kad yra dominuojanti organizmų rūšis arba kelios rūšys, kurios labai tiksliai bus randamos duomenų faile jeigu bus atliekami tolimesni ir tikslesni tyrimai.

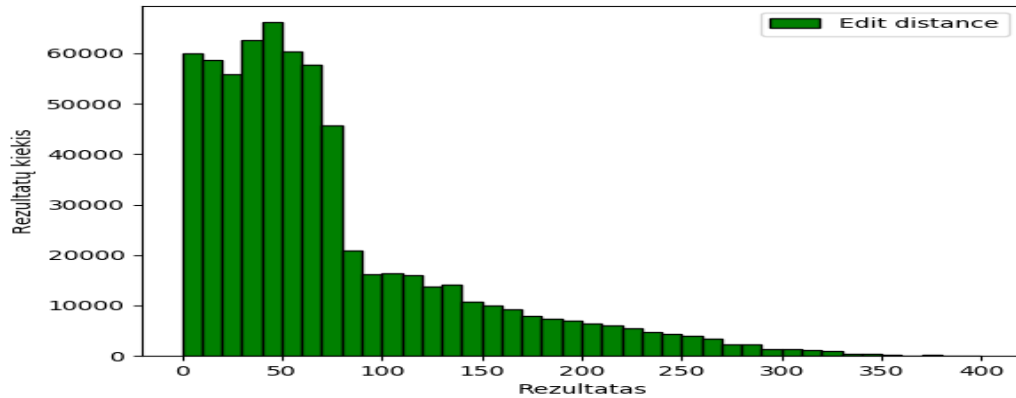


19 pav. Ketvirtąjo metagenomo rezultatai naudojant *water*

Rezultatai <i>water</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
1158.1534	1146	0	2510	4469

15 lentelė. *water* rezultatų lentelė

Kaip ir iš prieš tai gautų rezultatų, matomas labai didelis porų skaičius su aukštais balais. Vienas iš skirtumų, kuris yra pastebimas iš histogramos, tai didesnis skaičius porų su labai mažais balais.



20 pav. Ketvirtojo metagenomo rezultatai naudojant redagavimo atstumą

Rezultatai redagavimo atstumo			
vidurkis	mediana	minimumas	maksimumas
72.9713	54	0	400

16 lentelė. Redagavimo atstumo rezultatų lentelė

Iš paskutinės rezultatų grupės yra aiškiai matoma, kad faile dominuoja viena ar kelios organizmų rūšys. Minimalus balas teigia, kad buvo rastos visiškai vienodos sekos, o maksimumui esant 400 galima teigti, kad tos sekų poros arba sekų porų ilgiai buvo apie 400 bp.

Toliau pateikiami dar 3 metagenominių duomenų aprašymai ir rezultatų lentelės.

SRA ERR2094672: Šiuos duomenis pateikė Pietų Kalifornijos Universitetas. Metagenominis failas buvo gautas iš San Pedro kanalo Kalifornijoje. Tyrimo tikslas buvo ištirti žmogaus poveikį mikrobinei bendruomenei. Duomenų failo dydis 2.9 GB, nuskaitymu ilgis 300 bp, maksimalus galimas taškų skaičius 1500.

SRA ERR647655: Šis metagenominis failas gautas iš Sao Paulo zoologijos sodo ežero. Duomenų failo dydis 757.2 MB, nuskaitymų ilgis apie 500 bp, maksimalus galimas taškų skaičius 2500.

SRA DRR093002: Duomenų rinkinį įkėlė *Japan Agency for Marine-Earth Science and Technology (JAMSTEC)*. Metagenominis mėginys gautas norint atkurti dar neklasifikuotus mikrobinius genomus, kurie buvo gauti iš jūros dugno, kuriame yra didelis sieros kiekis. Failo dydis 1.7 GB, nuskaitymų ilgis 602, maksimalus galimas surinkti skaičius 3010.

Rezultatai <i>needleall</i>					
SRA	vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
ERR2094672	275.7891	288	1	652	0
ERR647655	393.2702	403	1	932	0
DRR093002	585.40287	575	16	1544	0

17 lentelė. *needleall* rezultatų lentelė

Rezultatai <i>water</i>					
SRA	vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
ERR2094672	284.3301	294	29	652	0
ERR647655	400.4870	409	2	936	0
DRR093002	594.5815	582	91	1552	0

18 lentelė. *water* rezultatų lentelė

Rezultatai redagavimo atstumo				
SRA	vidurkis	mediana	minimumas	maksimumas
ERR2094672	150.8683	150	14	218
ERR647655	228.9870	245	31	502
DRR093002	299.9236	301	204	396

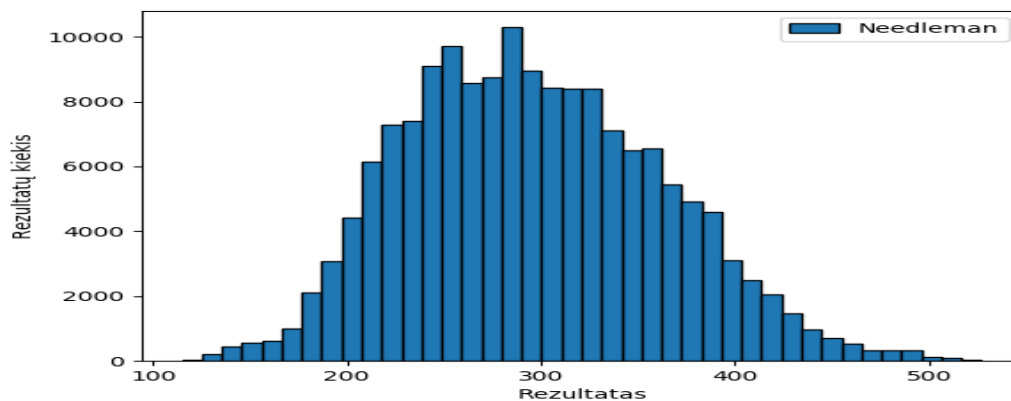
19 lentelė. Redagavimo atstumo rezultatų lentelė

Su visais šiais duomenimis buvo gautos standartinio pasiskirstymo histogramos. Nei viename iš šių tyrimų didžiausias gautas rezultatas neperžengė net 50% nuo didžiausio galimo išlyginimo balo. Todėl galima teigti, kad šiuose matagenominiuose failuose esantys organizmai yra labai skirtingi arba panašių organizmų skaičius buvo per mažas juos aptikti šio bandymo metu. Tokius rezultatus įtakoja ir tai, kad duomenų rinkiniai yra didesni nei prieš tai buvę, su kuriais buvo rasta daug porų su aukštais balais. Gavus tokius rezultatus nėra galima teigti, kad duomenų rinkiniuose nėra panašių organizmų, tiesiog, šių eksperimentų metu, atlikus paviršutinę metagenominę analizę, jų nebuvo surasta.

2.9. Tolimesni metodikos bandymai su organizmais

Pasirinkta su metagenomu SRR6994554 pakartoti prieš tai atliktą tyrimą, bet šį kartą metagenominius duomenis lyginti su atrinktomis sekomis. Kadangi prie duomenų failo buvo nurodyta, kad tikėtasi surasti žmogaus respirovirusą 1, iš NCBI atsisiųstas failas su skirtingomis šio viruso dalinėmis sekomis. Visa darbo eiga buvo beveik tokia pati kaip ir atliekant pirmuosius eksperimentus, tik šį kartą vienas iš tyrimo duomenų failų buvo gautas paimant atsitiktines sekas iš failo su virusais ir sudarant pasirinkto ilgio sekas.

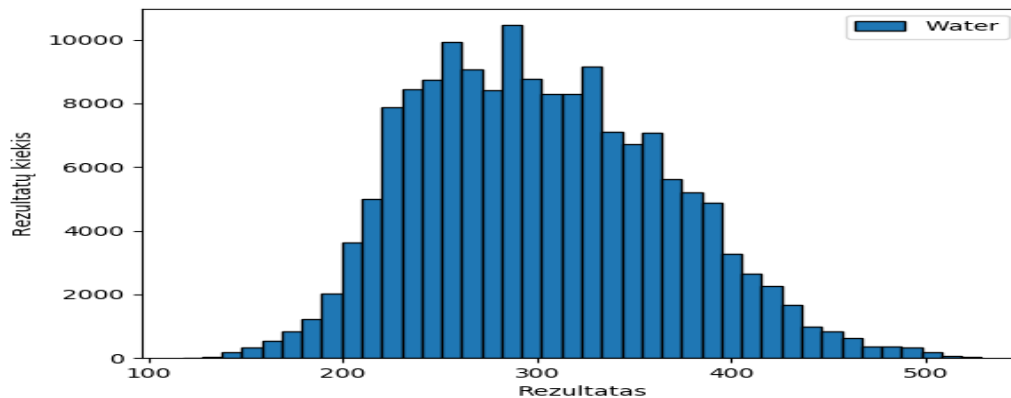
Pirmą kartą bandymas atliktas kai sekų ilgis buvo 450 bp. Gauti tokie rezultatai:



21 pav. Tyrimo su organizmais rezultatai naudojant *needleall*

Rezultatai <i>needleall</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
295.9258	292	115	527	0

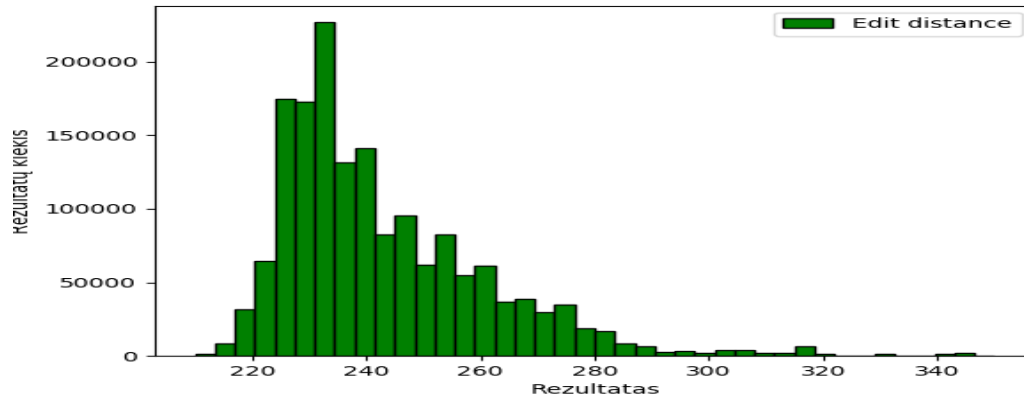
20 lentelė. *needleall* rezultatų lentelė



22 pav. Tyrimo su organizmais rezultatai naudojant *water*

Rezultatai <i>water</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
302.2381	297	117	529	0

21 lentelė. *water* rezultatų lentelė



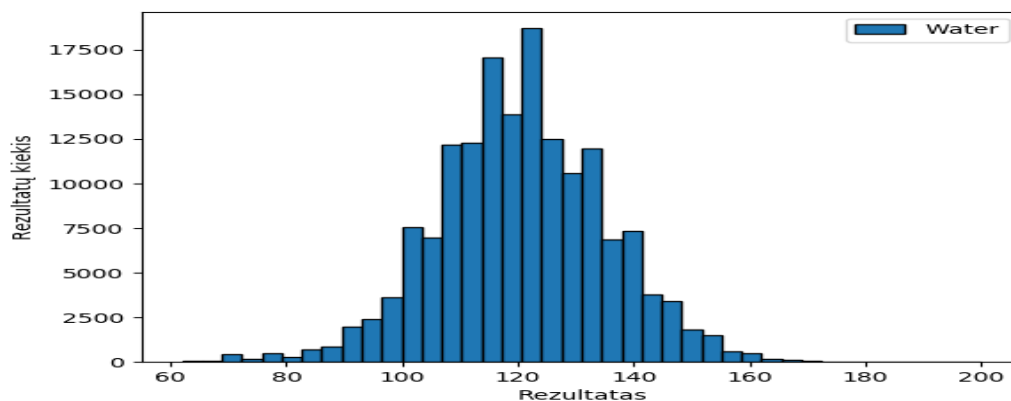
23 pav. Tyrimo su organizmais rezultatai naudojant redagavimo atstumą

Rezultatai redagavimo atstumo			
vidurkis	mediana	minimumas	maksimumas
242.8284	237	210	350

22 lentelė. Redagavimo atstumo rezultatų lentelė

Šio tyrimo metu didžiausias rezultatas galėjo būti 2250. Jeigu lyginsime rezultatus su to paties genomo rezultatais, gautais pirmajame tyrime (18 pav., 19 pav., 20 pav.), tai pamatysime, kad dabar visi rezultatai yra pasislinkę į kairę pusę. Taip pat jų pasiskirstymas nebėra toks platus. Pirmajame tyrime rezultatai varijavo nuo 300 iki 2500, o šio bandymo metu nuo 115 iki 530. Galima spėti, kad aukštų rezultatų šio tyrimo metu nepavyko gauti dėl to, nes sekų ilgis buvo paimtas per didelis. Kadangi kaip yra žinoma šiame metagenominiame rinkinyje pats trumpiausias ilgis gali būti net 250.

Toks pat bandymas buvo atliktas dar kartą. Šį kartą atrinktų sekų ilgis buvo 100 bp.



24 pav. Tyrimo su organizmais rezultatai naudojant *water*

Rezultatai <i>water</i>				
vidurkis	mediana	minimumas	maksimumas	$\geq 80\%$
120.4952	120	62	200	0

23 lentelė. *water* rezultatų lentelė

Tyrimo metu buvo labiau koncentruojamasi į su *water* įrankiu gautus rezultatus, dėl labai besiskiriančio visų paimtų sekų ilgio. Didžiausias galimas šio tyrimo rezultatas galėjo būti 500, o surastas 200, kas neparodo pakankamai gerų rezultatų, kad būtų galima identifikuoti organizmą.

Atliekant patį pirmąjį tyrimą su šiuo metagenomu, rezultatai parodė labai didelį skaičių sekų, kurios yra vienodos arba panašios, dalis nuskaitymų iš šio failo buvo patikrinti naudojant BLAST įrankį. Kadangi šis failas buvo gautas tirti žmogaus respirovirusą 1, buvo galima rasti daug žmogaus genominių sekų ir ribosominių RNR, kas ir davė neregulius rezultatus pirmuosiuose bandymuose.

Šis tyrimas parodė, kad taikyta metodika dar nėra tinkama organizmų paieškai, nebent yra tikrai žinomas tokio organizmo būvimas metagenominiame faile.

Rezultatai ir švados

Paviršutinė metagenominių duomenų analizė buvo atlikta su septyniais metagenominių duomenų rinkiniais. Pirmiesiems bandymams, su visais duomenų rinkiniais, buvo atlikta pirminė analizė lyginant nuskaitymus. Kiekvienam failui buvo atlikti bandymai naudojant tris įrankius: *needleall*, *water* ir pagal poreikius pritaikytą *EBL*B biblioteką. Šie įrankiai leido pažvelgti į duomenis iš skirtingų pozicijų ir parodė skirtingus rezultatus. Atlikus tyrimą su pirmaisiais keturiais metagenominiais rinkiniais buvo rasta nuskaitymų, kurių surinktų išlyginimų skaičius, panaudojus *needleall* ir *water* įrankius, siekė daugiau nei 80% nuo galimo didžiausios taškų skaičiaus. Tokia riba buvo pasirinkta viso darbo metu, kad būtų galima sužinoti pakankamai aukštų rezultatų kiekį. Gauti duomenys leido daryti išvadas apie tai, kad metagenominiuose failuose yra organizmų, kurie yra panašūs savo struktūra arba sutinkami daugiau nei vieną kartą. Tai leido teigti, kad metagenominiame faile yra dominuojančios organizmų grupės. Pavyzdžiui, dar kartą pažvelgus į tyrimo, kurio duomenų SRA SRR6994554 (tyrimas ieškantis žmogaus respiroviruso 1) rezultatus, aiškiai matoma, kad yra daug tokių pačių arba panašių organizmų grupių. Tyrimai su likusiais trimis metagenominiais rinkiniais parodė kitokius rezultatus. Šių tyrimų metu nebuvo rasta porų, turinčių aukštus balus, kas leido teigti, kad norint su jais atlikti tolimesnius tyrimus reikia galingesnių įrankių. Atlikus šiuos bandymus, tokie patys tyrimai buvo pakartojami bandant surasti metagenominiame faile esančius organizmus. Iš pradžių toks tyrimas buvo atliekamas su pačiu pirmuoju metagenomiu failu (SRA RR926399). Buvo atsisiųsti 201 cianobakterijų genomai. Tyrimas pakartotas du kartus imant skirtingus sekų ilgius.

Pirmąjį kartą, atrenkant sekas iš genomio failo, jų ilgis buvo toks pat, kaip ir pačių nuskaitymų – 202 bp. Bandymo rezultatai neparodė reikšmingų porų atitikimų ir sekų išsilygiuojančių su dideliais balais. Bandymo metu didžiausias galimas balų skaičius buvo 1010, o surastas 415. Tai gali būti pakankamai geras rezultatas, nes į nuskaitymą dažnai gali patekti tik dalis tam tikro genomio elemento. Dėl šių priežasčių bandymas buvo dar kartą pakartotas.

Antrojo tyrimo metu iš genomio failo gautų sekų ilgis buvo 60 bp. Šio tyrimo metu vienas svarbiausių rodiklių buvo lokalusis išlyginimas. Šio eksperimento metu buvo tikimasi surasti sekų poras, kurių surinktų išlyginimo taškų skaičius būtų pakankamai didelis. Tokių sekų surasti nepavyko, didžiausias surinktas taškų skaičius naudojant *water* įrankį buvo 217, kai pati didžiausia reikšmė galėjo būti 400. Atsižvelgiant į straipsnį, kuriame yra publikuotas šis metagenominis duomenų rinkinys, yra žinoma, kad tyrimas buvo sudėtingas, reikalavo galingos įrangos ir daug laiko.

Antrą kartą toks tyrimas, kai yra ieškoma organizmų metagenominiame faile, buvo atliekamas su metagenomu, kuriame ieškoma žmogaus respiroviruso 1 (SRA SRR6994554). Šiam tyrimui iš NCBI duomenų bazės buvo atsisiųstas šio viruso dalinių sekų rinkinys. Toks tyrimas kartotas du kartus. Pirmąjį kartą atrinktų sekų ilgiai buvo 450 bp, antrąjį 100 bp. Šių tyrimų metu nepavyko rasti sekų porų su aukštais išlyginimo rezultatais, bet tai leido padaryti išvadas apie metodikos veiksmingumą.

Atlikus šiuos tyrimus su skirtingais metagenominiais rinkiniais yra galima padaryti tokias pagrindines išvadas:

- Panaudojus tris įrankius buvo atlikta paviršutinė metagenominių duomenų analizė, kuri leido įvertinti tam tikrų metagenomų duomenis. Buvo padarytos išvados apie faile esančių nuskaitymo sekų porų panašumus, kas leistų teigti, kad metagenome yra panašių arba vienodų organizmų.
- Tokius pat tyrimus atliekant lyginant nuskaitymus su atrinktais organizmais, nuspręsta, kad šiuo metu tokia metodiką dar nėra pilnai tinkama. Žinoma, gauti rezultatai priklauso nuo pačio metagenominio failo ir atrinktų organizmų.
- Norint, kad metodika būtų veiksmingesnė reikėtų ją tobulinti. Būtų galima išbandyti daugiau įrankių, leidžiančių tirti metagenominius duomenis. Taip pat reikėtų tobulinti atsiktinių sekų atrinkimo būdą, kad būtų užtikrinta jog tokios pat sekos nebus atrinktos kelis kartus.
- Norint atlikti paviršutinę metagenominių duomenų analizę, naudojant šią metodiką, yra užtrunkama iki 10 minučių. Rezultatai parodo pradinę informaciją apie nuskaitymų panašumus ir leidžia daryti išvadas apie ten esančių organizmų paplitimą.
- Šiuo metu tokia metodika gali būti naudinga norint nustatyti ar yra panašių organizmų metagenominiuose rinkiniuose ir pridėjus papildomą funkcionalumą atlikti tolimesnius tyrimus tik su tokiomis sekomis.

Santrauka

Pagrindinis šio darbo tikslas, sukurti metodiką, kuri leistų atlikti pirminį metagenominių duomenų tyrimą. Kadangi dauguma metagenominių duomenų yra didelės apimties, paviršutinė duomenų analizė gali padėti išsigryninti informaciją apie tolimesnius tyrimo tikslus, darbo eigą ir sudėtingumą. Šiam darbui buvo panaudoti trys pagrindiniai įrankiai, *emboss:needleall* globaliajam išlyginimui, *emboss:water* lokaliajam išlyginimui ir pritaikyta *edlib* biblioteka redagavimo atstumiui nustatyti. Metodika buvo pritaikyta atlikti pirminę duomenų analizę septyniems skirtingiems metagenominių duomenų rinkiniams, kai jų duomenys buvo lyginami su jais pačiais. Tyrimas parodė tam tikrą informaciją apie kiekvieną iš duomenų rinkinių, kas leido daryti išvadas apie ten sutinkamų organizmų panašumus, paplitimą ir nustatyti pradinį metodikos veiksmingumą. Tokia pat metodika buvo pritaikyta bandant nustatyti tikslius organizmus iš duomenų rinkinių, kas leido padaryti naujas išvadas apie metodikos tobulinimą ir galimą tolimesnį pritaikymą.

Summary

The main purpose of this paper was to create a method that would allow to perform primary metagenomic data analysis. Because of the size and quantity of metagenomic data, primary metagenomic analysis can help researcher to identify goals, workflow and complexity of future analysis. Three tools were used for this work. They were *emboss:needleall* for global alignment, *emboss:water* for local alignment and fitted *edlib* library that was used to calculate edit distance. The method was adapted to perform primary metagenomic data analysis on seven different data sets when the data that was compared was taken only from metagenomic file. During the research some information was obtained from every data file. This data allowed to make deductions about the similarity and prevalence of organisms that can be found in metagenomic data. These tests also allowed to determine the efficiency of implemented technique. The same research was repeated with the aim to identify specific organisms. This allowed to make deductions about possible future improvements and further adaptations.

Literatūra

- [1] Martin Šošić and Mile Šikić. Edlib: a c/c++ library for fast, exact sequence alignment using edit distance. *Bioinformatics* 33.9, 1394-1395, 2017. <https://doi.org/10.1093/gigascience/gix047>.
- [2] Hubert Denise Alex L Mitchell Maxim Scheremetjew. Ebi metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Research, Pages D726–D735*, January 2018. <https://doi.org/10.1093/nar/gkx967>.
- [3] Alan Bleasby. Emboss water. <http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/water.html>, 2013. Release 6.6.0.
- [4] Marchesi JR Culligan EP Sleator RD. Metagenomics and novel gene discovery: promise and potential for novel therapeutics. *Applied and Environmental Microbiology*, 77(4):1153-1161, Virulence. 2014;5(3):399-412. doi:10.4161/viru.27208.
- [5] K Jonah Kenai. *Bio-sequence search engine based on Damerau Levenshtein distance algorithm*. Magistrinis darbas, University of Nairobi, 2011, p. 21.
- [6] Wysoker A. Li H. Handsaker B. The sequence alignment/map format and samtools. Tom. 25, 2009, pages 2078–2079 doi:10.1093/bioinformatics/btp352.
- [7] Gonzalez-Serrano R López-Pérez M Haro-Moreno JM. Genome diversity of marine phages recovered from mediterranean metagenomes: size matters. *PLoS Genetics*; 13(9), 2017. doi:10.1371/journal.pgen.1007018.
- [8] Robert D. Finn Petra ten Hoopen. The metagenomic data life-cycle: standards and best practices. *GigaScience*, 2017, Volume 6, Issue 8. doi:10.1093/bioinformatics/btw753.
- [9] Winckler W. Robinson J. Thorvaldsdóttir H. Integrative genomics viewer. In 10 January 2011. doi:10.1038/nbt.1754.
- [10] Sam. <http://genome.sph.umich.edu/wiki/SAM>.
- [11] Daniel R. Simon C. Metagenomic analyses: past and future trends. *Applied and Environmental Microbiology*, 77(4):1153-1161, 2011. doi:10.1128/AEM.02345-10.
- [12] Meyer F. Thomas T Gilbert J. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2012. 2:3. doi:10.1186/2042-5783-2-3.
- [13] Mahmut Uludag. Emboss needleall. <http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/needleall.html>, 2013. Release 6.6.0.
- [14] What is bowtie 2. <http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#what-is-bowtie-2>, 28 Apr 2016.