

VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
MATEMATINĖS INFORMATIKOS KATEDRA

Eimantas Paspirgėlis

Bioinformatikos studijų programa

Matematinės informatikos šaka

**Baltijos jūros vandens mikroorganizmų genominių sekų
tyrimas**

Baltic sea microorganisms genomic sequences research

Bakalauro baigiamasis darbas

Vadovas: lektorius Iruš Grinis

Vilnius 2015

Turinys

Įvadas	3
1. Teorinė dalis	4
1.1. Tiriamo organizmo trumpa charakteristika	4
1.1.1. Baltijos jūroje atliktas tyrimas	4
1.1.2. Sekvenuotų genomų apžvalga	4
1.1.3. Sulfurimonas denitrificans charakteristika	5
1.2. Naudojami įrankiai	6
1.2.1. Python	6
1.2.2. MyRast	7
1.2.3. GeneMark	13
1.2.4. Prokka	14
2. Praktinė dalis	16
2.1. MyRast	16
2.2. Prokka	20
2.3. GeneMark	23
2.4. Baltymų sekų palyginimas	25
2.5. Fermentų paieška	27
2.6. Genomo naršyklė	29
Išvados	33
Literatūra	34

Įvadas

Baltijos jūroje buvo atlikti tyrimai, kurių metu buvo paimtas mėginys Baltijos jūros vandens. Jame buvo rasta įvairių mikroorganizmų, 5 iš jų t.y. 3 archėjos ir 2 bakterijos buvo sekvenuotos ir gauti jų genomai. Išsiaiškinus kokie tai mikroorganizmai, buvo nuspręsta vieną iš jų išanalizuoti.

Bakalauro darbe bus tiriami ir analizuojama bakterijos *Sulfurimonas denitrificans* genomai su trimis skirtingais bioinformatikos įrankiais, tai *MyRast*, *Prokka* ir *GeneMark*. Tai anotacijos įrankiai, su kuriais *Sulfurimonas denitrificans* genomai bus identifikuojami ir anotuojami. Gavus visų trijų įrankių rezultatus jie bus lyginami tarpusavyje tam, kad surasti ir išrinkti tiksliausias baltymų sekas, kurios buvo gautos naudojant anotacijos įrankius. Taip pat, bus surasti baltymų fermentai ir jų keliai(angl. pathway). Surinkus visus duomenis apie bakterijos *Sulfurimonas denitrificans* identifiкуotas baltymų sekas bus parašytas skriptas, *Python* programavimo kalba, kuris generuos html failą ir atvaizduos šios bakterijos genomo naršyklę.

Darbo tikslas – sukurti genomo naršyklę įprastinėmis priemonėmis naudojant skripto kalbas ir html.

Darbo tikslui pasiekti keliama uždaviniai:

- panagrinėti iš vartotojo pusės bioinformatikos anotacijos įrankius: *MyRast*, *Prokka* ir *GeneMark*;
- gauti automatinę *Sulfurimonas denitrificans* genomo anotaciją naudojant šiuos įrankius;
- sukurti skriptus, kurie:
 - palygintų tarpusavyje visų trijų įrankių gautus rezultatus;
 - surastų baltymų fermentus ir jų kelius;
 - pagal *MyRast* išvesties failą automatiškai generuotų *html* kodą, kuris vaizduoja genomo naršyklę.

Darbo teorinėje dalyje bus apžvelgiama tiriamo organizmo trumpa charakteristika, naudojami įrankiai: *Python*, *MyRast*, *Prokka* ir *GeneMark*. Praktinėje dalyje bus apžvelgiama naudojamų įrankių gauti rezultatai, sukurti skriptai ir jų rezultatai, taip pat sugeneruota *Sulfurimonas denitrificans* genomo naršyklė.

1. Teorinė dalis

1.1. Tiriamo organizmo trumpa charakteristika

1.1.1. Baltijos jūroje atliktas tyrimas

Baltijos jūros vandenyse atlikti tyrimai, kurie nurodo faktorius apie paviršutiniuose vandenyse gyvenančius mikroorganizmus, jų populiaciją[9]. Taip pat atlikti tyrimai[8], kuriuose tiriamas bakterijų pasiskirstymas horizontaliai ir vertikalčiai Baltijos jūros druskingumo gradientų.

Šiame tyrime buvo bandoma ištirti mikroorganizmus gyvenančius ne paviršutiniuose vandenyse, o gilesniame, sūresniame vandenyje įsikūrusius prokariotus. Todėl Baltijos jūroje iš maždaug 60 metrų gylio buvo paimtas jos vandens mėginys. Atlikus tyrimus, mėginyje buvo rasta įvairiausių bakterijų ir archėjų. 5 iš jų t.y. 2 bakterijos ir 3 archėjos buvo siunčiamos į Jungtinių Amerikos Valstijų "Bigelow Laboratory for Ocean Sciences " laboratoriją, kad būtų sekvenuoti jų genomai. Atlikus sekvenavimą genomai buvo parsiųsti atgal į Lietuvą ir čia tiriami.

1.1.2. Sekvenuotų genomų apžvalga

Šiame tyrime buvo išsiaiškinta, kad minėti prokariotai manoma yra tokie:

- *Candidatus nitrosopumilus maritimus*(archėja);
- *Halobaculum gomorrense*(archėja);
- *Halobaculum genties*(archėja);
- *Phycisphaera mikurensis*(bakterija);
- *Sulfurimonas denitrificans*(bakterija).

Candidatus nitrosopumilus maritimus[9] yra labai paplitusios archėjos, kurios gyvena jūros vandenyje. Tai pirmasis grupės narys *Ia Crenarchaeota*, kuris turėtų būti išskirtas grynakraujėje kultūroje. Genų sekos rodo, kad grupės *Ia Crenarchaeota* galima rasti daugelyje jūrų pakrančių aplink planetą. Jis yra vienas iš mažiausių mikroorganizmų – 0,2 mikronų skersmens. Jis gyvuoja oksiduodamas amoniaką į nitritus. *Candidatus nitrosopumilus* yra iš *Nitrosopumilaceae* genties.

Halobaculum gomorrense[4] tai archėja, rasta Negyvojoje jūroje. Tai vienintelė *Halobaculum* genties rūšis, kuri yra "strypo/lazdelės" formos.

Halobaculum genties archėja – iš sekvenuoto genomo pavyko nustatyti tik tai jog ši archėja yra iš *Halobalucum* genties.

Phycisphaera mikurensis[3]– rasta ant jūros dumblių. Priklauso *Planctomycetes* tipui(skyriui), *Planctomyces* grupei. *Phycisphaera mikurensis* suteikia įžvalgų į *Planctomyces* gyvavimo ciklą.

Sulfurimonas denitrificans[5] yra *Sulfurimono* genties bakterija. Ši bakterija naudoja sulfidą arba tiosulfatą, kaip elektrono donorą, o nitratą arba nitritus, kaip akceptorius. Šis mikroorganizmas buvo nustatytas hidroterminėse ventiliacijos srityse ir naftos telkiniuose, jis šiose aplinkose gali vaidinti svarbų vaidmenį sieros apytakoje.

Buvo nuspręsta analizuoti bakteriją – *Sulfurimonas denitrificans*.

1.1.3. *Sulfurimonas denitrificans* charakteristika

Sulfurimonas yra bakterijų gentis iš klasės *Epsilonproteobacteria*. Šios grupės nariai naudoja sulfidą, tiosulfatą ir elementinę sierą, kaip elektronų donorus, o CO₂ kaip jų anglies šaltinį. Šios genties nariai yra keturi:

- *Sulfurimonas autotrophica*;
- *Sulfurimonas denitrificans*;
- *Sulfurimonas gotlandica*;
- *Sulfurimonas paralvinellae*.

Sulfurimonas denitrificans kvalifikacija buvo pakeista iš *Thiomicrospira denitrificans*. *Sulfurimonas denitrificans* klasifikacija(1 lentelė):

Mokslinė klasifikacija	
Karalystė:	Bacteria
Skyrius:	Proteobacteria
Klasė:	Epsilonproteobacteria
Būrys(Eilė):	Campylobacterales
Šeima:	Helicobacteraceae
Gentis:	Sulfurimonas

1 lentelė. *Sulfurimonas denitrificans* mokslinė klasifikacija.

Sulfurimonas denitrificans iš pradžių buvo rastas ant jūrų pakrančių nuosėdų. Ši bakterija atlieka svarbų vaidmenį ekosistemoje, vykdo sieros transformaciją ir azoto ciklą. *Sulfurimonas denitrificans* transformuoja sierą per sieros oksidacijos procesą ir paverčia nitritus į diazoto dujas per denitrifikaciją. Ši bakterija sukėlė nemažai susidomėjimo mokslo bendruomenėje, dėl savo unikalios medžiagų apykaitos, kuri gali oksiduoti sierą ir sumažinti nitratų[3]. Žemiau pateiktame paveikslėlyje pavaizduotas *Sulfurimonas denitrificans* bakterija(1 pav.).



1 pav. *Sulfurimonas denitrificans* bakterija.

Sulfurimonas denitrificans[3] genomas turi pakankamai didelius matmenis (2.2 Mbp), tai suteikia didesnę metabolinę universalumą arba reagavimą į aplinką, nei daugumai kitų sekvenuotų klasės *Epsilonproteobacteria* bakterijų. Šios bakterijos genai turi pilną autotrofinį redukcinį Krebso ciklą. Taip pat, turi didelį arsenalą jutimo ir reguliavimo baltymų koduojančių genų, kurie svarbūs, kad būtų galima užkirsti kelią oksidaciniam stresui.

1.2 Naudojami įrankiai

1.2.1. Python

Python yra sukurta Guido van Rossumo 1990 metais. Pirmiausiai ji buvo scenarijų kalba AmoebaOS operacinei sistemai. Python dažniausiai lyginama su Tcl, Perl, Scheme, Java ir Ruby. Python kuriamas kaip atviro kodo projektas.

Python yra daugiaparaigimė programavimo kalba – ji leidžia naudoti keletą programavimo stilių: objektinį, struktūrinį, funkcinį, aspektinį. Python naudoja dinaminį tipų tikrinimą.

Python kūrėjų tikslai buvo sukurti kalbą, kuri yra lengvai skaitoma, išraiškinga, išreikštinė, paprasta (tinkama neprofesionaliems programuotojams). Nors pradžioje ji buvo kuriama kaip scenarijų kalba, dabar ji naudojama ir dideliems programiniams projektams. Taip pat labai paplitusi Linux sistemose.

Python - galinga ir patogi programavimo kalba, sparčiai populiarėjanti pasaulyje. Bene akivaizdžiausias šios kalbos privalumas - didelis įtrauktų bibliotekų kiekis. Tai leidžia

supaprastinti ir pagreitinti programų kūrimą, kurti universalias, pagal poreikius pritaikytas programas. Python taip pat labai intuityvi programavimo kalba.

Su kitomis programavimo kalbomis anksčiau susidūręs žmogus iškart pastebės - tą pačią užduotį galima atlikti šimtu skirtingų būdų. Taip yra todėl, kad Python savyje talpina tiek kitų populiariausių programavimų kalbų sintaksių ypatybes, tiek tūkstančius įvairiausių vidinių funkcijų.

Python - interpretuojama programavimo kalba, t.y. kodas analizuojamas programos vykdymo metu. Be didelio kiekio vidinių bibliotekų yra įtraukta ir daugybė duomenų struktūrų. Minėti aspektai programuotojui suteikia didelę programavimo laisvę, leidžia sutrumpinti kodą bei programos rašymo laiką.

1.2.2. MyRast

Prokariotų genomų sekų skaičius auga nuolat ir auga greičiau nei mokslininkai geba jas visas anotuoti[4]. Tam sukurta visiškai automatizuota paslauga, kuri anotuoja bakterijų ir archėjų genomus. Servisas nustato baltymų koduotę, rRNR ir tRNR genus, priskiria funkcijas genams, prognozuoja kuris posistemis yra atstovaujamas geno, naudoja šią informaciją atkurti metabolinį tinklą ir sukuria lengvai parsisiunčiamą išvesties failą. Servisas, normaliai, anotuoja genomą per 12-24 val. nuo pateikimo, o esamas įgyvendinimas leidžia nuo 50 iki 100 genomų pralaidumą. Svarbu paminėti, jog greitis MyRast nėra svarbiausias aspektas, svarbiausia yra tikslumas, išsamumas ir suderinamumas.

Rast serveryje įgyvendintas automatinis gaminimas dviejų klasių genų funkcijų teiginių(angl. asserted):

- remiantis posistemėmis – teiginiai grindžiami funkcijų atpažinimu posistemėse;
- remiantis ne posistemėmis – teiginiai užpildomi daugiau bendrais metodais, kurie grindžiami integracija.

Faktas, kad Rast išskiria šias dvi anotacijos klases ir naudoja patikimas posistemas kaip pagrindą išsamiai medžiagų metaboliniai rekonstrukcijai, Rast anotacijos tampa, kaip išskirtinai geras atspirties taškas tolimesnei anotacijai. Be to, gaminant pradines genų funkcijas ir metabolinę rekonstrukciją, Rast serveris suteikia aplinką naršymui po anototą genomą ir lyginimą su šimtais kitų genomų per SEED integraciją. Rast turi galimybę rodyti geno kontekstą aplink specifinius genus ir galimybę atsisiųsti atitinkamą informaciją ir anotacijas.

Rast naudoja naują baltymų šeimų kolekciją. Ši kolekcija yra nurodyta kaip FIGfams rinkinys, išsami publikacija apie jų detalumą yra rengiama. Kiekvienas FIGfam yra sudarytas iš 3 dalių: baltymų rinkinio, šeimos funkcijų ir sprendimų priėmimo procedūrų. Baltymų rinkinys yra manoma, kad yra globaliai panašus ir turbūt homologinis, visi nariai turėtų turėti bendras funkcijas. Sprendimo procedūra paima kaip įvestį baltymų seką ir gražina sprendimą ar baltymai gali būti įtraukti į šeimą. FIGfams konstrukcija vyksta konservatyviai: labai stengiamasi užtikrinti, kad du baltymai įtraukti į tą pačią šeimą turi bendras funkcijas, jei abejojama, baltymai yra dedami į skirtingas šeimas. Du baltymai dedami į tą pačią šeimą jei :

- jei abu įgyvendina tą patį funkcinį vaidmenį ir panašumo regionai tarp sekų apima bent 70% ;
- jei abu yra iš glaudžiai susijusių genomų ir jų panašumas yra didelis.

Yra du atvejai, kai galima teigti, kad baltymai turi bendras funkcijas: pirmasis atspindi ekspertų teiginius, o antrasis, kai nukrypimas yra minimalus. Kuriant FIGfams ir naudojant šiuos du principus, buvo sudaryta kolekcija iš maždaug 17,000 FIGfams, kurie apima baltymų giminystę su posisteme(t.y. FIGfams kuriuos, mes vadiname posistemės pagrindu) ir daugiau kaip 80,000 kuriuose baltymai grupuojami pagal antrąjį principą(t.y. ne posistemės pagrindu FIGfams). Daugelyje pagal antrąjį principą sugrupuotų FIGfams yra tik 2,3 arba 4 baltymai. Laikui bėgant planuojama sumažinti antrojo principo FIGfams. Tai bus daroma kuriant naujus, rankiniu būdu kuruojamus posistemius. Tikėtina, kad didelė dalis naujai sekvenuojamų genomų bus panašūs į esamus genomus ir FIGfams jau bus veiksminga atpažinimo sistema tokiais atvejais.

Pagrindiniai žingsniai anotuojant genomą naudojant RAST:

- kviečiami *tRNR* ir *rRNR* genai;
- kviečiama baltymus koduojanti genų ekspresija;
- sukuriamas filogenetinis kontekstas;
- tikslinė paieška remiantis FIGfams glaudžiai susijusiuose genomuose;
- dar kartą kviečiama baltymus koduojanti genų ekspresija;
- apdorojami likusieji genai dar kartą su visa FIGfams kolekcija;
- išvalomi likę genų kvietimai(t.y. ištrinami pasikartojimai ir nustatoma pradinė pozicija);
- apdorojami likusieji neanotuoti baltymų koduotės genai;
- konstruojama pradinė medžiagų apykaitos rekonstrukcija.

Šiuos žingsnius paanalizuosime išsamiai.

Kviečiami *tRNR* ir *rRNR* genai

Naudojamos kitų mokslinių tyrimų grupių sukurtos priemonės, kad pirmiausia nustatyti *rRNA* ir *tRNA* koduojančių genų ekspresiją. *tRNA* genams nustatyti naudojama *tRNAscan-SE*, o *rRNA* naudojamas įrankis "*search_for_rnas*" kuris sukurtas *Niels Larsen*. Pradedamas procesas kviečiant tuos genus, kuriuos manoma, kad galima nustatyti patikimai. Tada serveris nesaugo baltymus koduojančių genų, kurie reikšmingai sutampa su nors vienu iš šių regionų..

Kviečiama baltymus koduojanti genų ekspresija

Kai *tRNA* ir *rRNA* genų kodavimo regionai yra pašalinami, atliekamas pirminis kvietimas naudojant *GLIMMER2*. Šiuo metu ieškoma pagrįstai apskaičiuotų galimų genų ir *GLIMMER2* yra puiki priemonė šiam tikslui pasiekti. Šiame etape aktualu baltymus koduojantys genai.

Sukuriamas filogenetinis kontekstas

Kai pradinis rinkinys baltymus koduojančių genų buvo nustatytas, imamos tipinės sekos iš nedidelio FIGfams rinkinio, kurios turi savybę būti universaliomis arba beveik universaliomis prokariotuose. Naudojant šį mažą rinkinį sekų ieškoma baltymus koduojančių genų naujame genome. Reikia paminėti, kad šis žingsnis yra labai greitas, nes ieškoma tik naujame genome ir naudojant nedidelį rinkinį atstovaujančių baltymų sekų. Šio žingsnio rezultatas yra mažas rinkinys genų (paprastai 10-15), kuris gali būti naudojamas įvertinti arčiausius filogenetinius kaimynus naujame sekvenuotame genome. Kiekvienam aptiktam genui, nustatoma pradinė padėtis ir jis perkeliamas iš tariamų genų aibės į jau nustatytų genų aibę ir funkcija, kuri priskiriama genui yra paimama iš FIGfams.

Tikslinė paieška remiantis FIGfams glaudžiai susijusiuose genomuose

Kai "kaimyniniai genomai" buvo nustatyti, galima formuoti rinkinį FIGfams, kurie yra šiuose genomuose. Tai FIGfams rinkinys, kurios gali būti rastos naujame genome, ši paieška turi didelę sėkmės galimybę. Kai randamas genas, pakoreguojama jo pradinė padėtis ir jis perkeliamas iš tariamų į jau nustatytų genų aibę. Kaina rasti šiuos genus yra maža.

Dar kartą kviečiama baltymus koduojančių genų ekspresija

Šiuo metu yra sudarytas naujo genomo nustatytų genų rinkinys ir dabar galima jį naudoti kaip mokymo rinkinį iš naujo kviečiant baltymus koduojančius genus. Genomo, kuris yra glaudžiai susijęs su vienu ar daugiau esamų genomų, šis mokymo rinkinys gali apimti daugiau kaip 90% baltymus koduojančių genų.

Apdorojami likusieji genai dar kartą su visa FIGfams kolekcija

Tariamieji genai, kurie lieka, gali būti ieškomi dar kartą visoje *FIGfams* kolekcijoje. Šiuo metu kolekcija atstovaujamų baltymų sekų iš *FIGfams* naudojama apskaičiuoti potencialius ir aktualius *FIGfams*, kurie iš viso apima šiek tiek daugiau nei 100,000 baltymų sekų. Šis žingsnis sudaro galimybę paieškai *FIGfams* kiekvienam iš likusiųjų tariamų genų. Kai jis baigiamas, visi genai kurie gali būti apdorojami naudojant *FIGfams* yra apdorojami.

Išvalomi likę genų kvietimai (t.y. ištrinami pasikartojimai ir nustatoma pradinė pozicija)

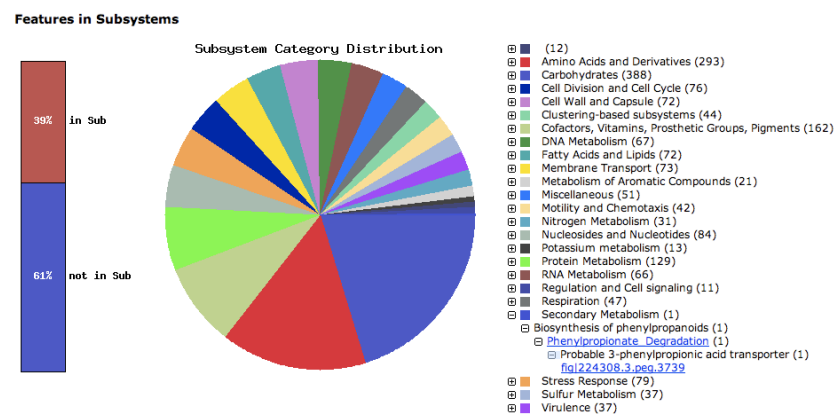
Tariamieji genai, kurie lieka, yra apdorojami ir bandoma išspręsti problemas susijusias su sutampančių genų kvietimu, jie yra koreguojami ir taip toliau. Atsižvelgiant į RAST serverį visi kiti likusieji tariami genai naikinami.

Apdorojami likusieji neanotuoti baltymų koduotės genai

Šiame žingsnyje vyksta paskutiniai funkcijų priskyrimai tariamiems genams. Jei panašumas buvo apskaičiuotas ankstesniame žingsnyje, gali būti tvirtinamos šių panašumų funkcijos. Galima paleisti, bet kurią paprastai dirbančią technologiją, kad būtų naudojamas rinkinys įrankių ir gauti tikslesnį įvertinimą. Naudojant šį metodą apdoroti genai sudaro didžiąją dalį RAST anotacijos. Pirminiame apdorojime dauguma genų naudoja *FIGfams* technologiją ir orientuotas paieškas.

Konstruojama pradinė medžiagų apykaitos rekonstrukcija

Kai funkcijų priskyrimas buvo atliktas, tuomet formuojama pradinė metabolinė rekonstrukcija. Tikslas yra sujungti genus į naujo genomo funkcinį vaidmenų posistemės. Kai pačios posistemės yra sukurtos neapdirbtose kategorijose, kurios atspindi pagrindinių padalinių funkcijas, galima pagaminti išsamią genomo turinio sąmatą, kad sėkmingai prijungti prie posistemės (2 pav.).



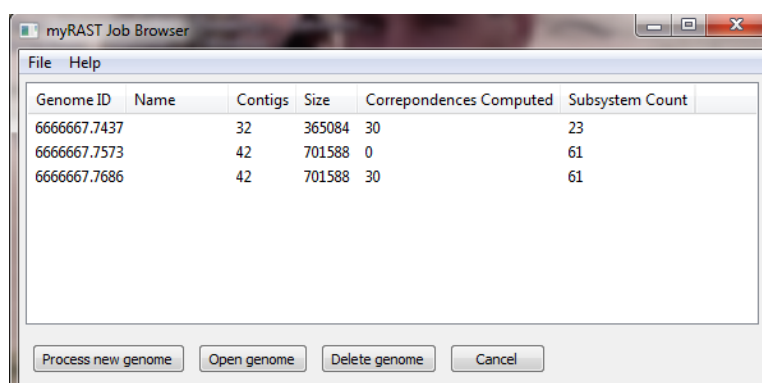
2 pav. Vaizduoja prijungtus genus prie posistemų ir jų pasiskirstymą įvairiose kategorijose.

Reikėtų pabrėžti, kad posistemės apima visus ląstelių modulius, ne tik medžiagų apykaitos kelius. Vadinasi tai kas yra vadinama metaboline rekonstrukcija, yra labiau suvokiama kaip genų grupavimas į modulius.

MyRast nauda

Ankstesniame skyrelyje aprašyta *RAST* pagrindinė technologija ir pagrindžiamas veikimas. Įdėta daug pastangų, kad sukurti paprastą sąsają, kuri siūlo pateikti genomus, stebėti anotacijos progresą, peržiūrėti rezultatus, palyginti juos su šimtais kitų genomų ir parsisiųsti rezultatus bent keliais formatais. Trumpai apie *RAST* naudojimą[5];

1. Kai paleidžiamas *MyRast* matoma jau ankščiau anotuotų genomų sąrašas(3 pav.).



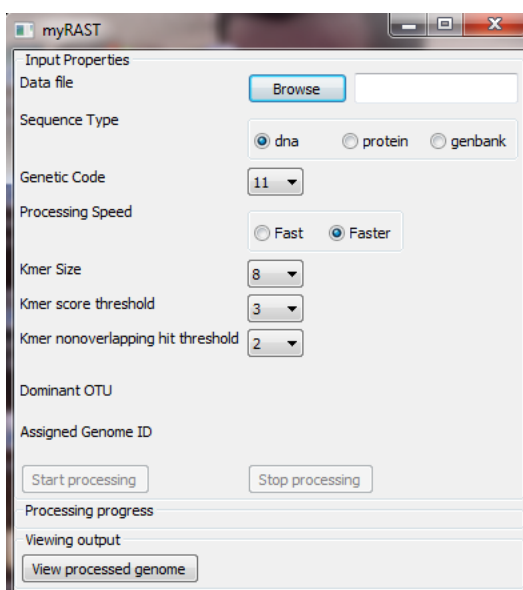
The screenshot shows the 'myRAST Job Browser' window. It contains a table with the following data:

Genome ID	Name	Contigs	Size	Correspondences Computed	Subsystem Count
6666667.7437		32	365084	30	23
6666667.7573		42	701588	0	61
6666667.7686		42	701588	30	61

At the bottom of the window are four buttons: 'Process new genome', 'Open genome', 'Delete genome', and 'Cancel'.

3 pav. Pradinis MyRast langas.

2. Pasirinkus "Process new genome" pasiūloma pasirinkti failą, kuris turi būti anotuojamas(4 pav.).

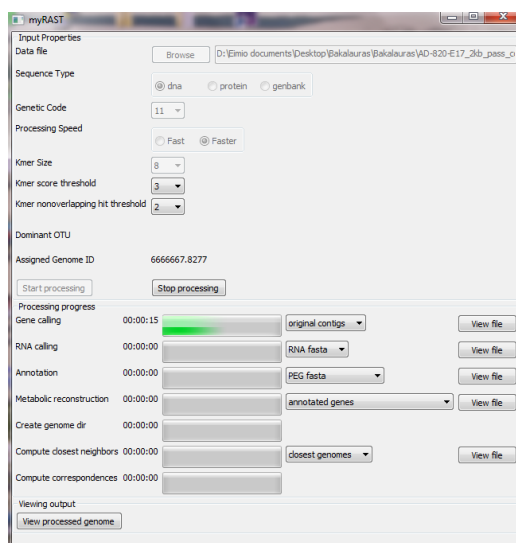


The screenshot shows the 'myRAST' dialog box with the following settings:

- Input Properties
- Data file: [Browse button]
- Sequence Type: ☒ dna, ☐ protein, ☐ genbank
- Genetic Code: 11
- Processing Speed: ☐ Fast, ☒ Faster
- Kmer Size: 8
- Kmer score threshold: 3
- Kmer nonoverlapping hit threshold: 2
- Dominant OTU
- Assigned Genome ID
- [Start processing] [Stop processing]
- Processing progress
- Viewing output
- [View processed genome]

4 pav. Failo ir nustatymų pasirinkimas.

Galima įkelti failą Genbank formatu, contigs failą FASTA formatu arba baltymų sekų failą FASTA formatu. Paprastai nurodoma DNR, tai nurodo, jog norima, kad būtų anotuoti contigs. Kai viskas pasirinkta spaudžiama "Start processing"(5 pav.).



5 pav. Vykdoma genomo anotacija.

Kai pradedamas apdorojimas, matomas valdymo pultas, kuriame rodomi anotacijos žingsniai.

Norint peržiūrėti genomą spaudžiama " View processed genome"(6 pav.).



6 pav. Anotuoto genomo vizualus pavaizdavimas.

Šis vaizdas rodo regioną naujai sekvenuoto genomo. Genai kurie turi tą pačią funkciją yra nudažyti ta pačia spalva. Genai yra vaizduojami kaip strėlės. Užvedus ant bet kurios iš jų yra matoma informacija: ID, funkcija, contig, kur prasideda ir baigiasi, taip pat ilgis. Spaudžiant rodyklės galima pereiti į vieną ar kitą pusę, per vieną geną, pusę ekrano, visą ekraną arba pereiti iš vieno contigs į kitą. Yra trys pagrindinės funkcijos: keisti geno anotaciją, ištrinti arba įterpti naują geną.

1.2.3. GeneMark

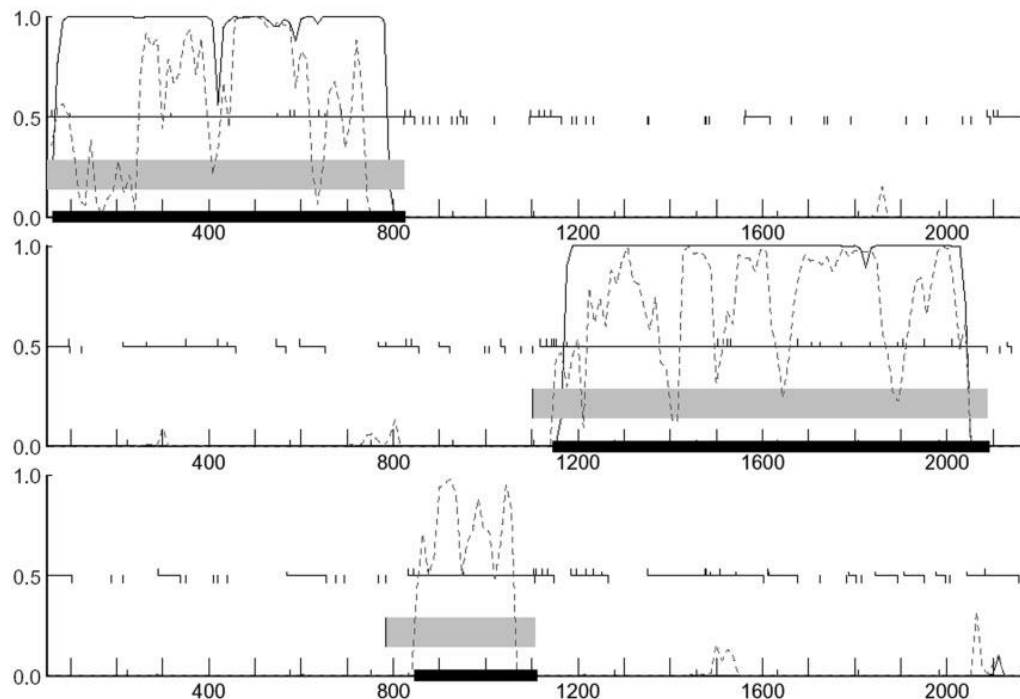
Genų identifikavimo uždavinys dažnai sprendžiamas tyrėjų, kurie susiduria su naujais ir gerai išnagrinėtais genomais, jiems šias užduotis patogiai ir patikimai padeda išspręsti *GeneMark*[6] - interneto programinė įranga (<http://opal.biology.gatech.edu/GeneMark/>). Svetainė suteikia sąsajas su *GeneMark* šeimos programomis, kurios skirtos genų prognozavimui: prokariotų, eukariotų ir virusų genomų. Šiuo metu serveris leidžia analizuoti apie 200 prokariotų ir daugiau kaip 10 eukariotų genomų naudojant kiekvienai rūšiai būdingas programinės įrangos versijas. *GeneMark* svetainė dažnai atnaujinama, pateikiamos naujausios programinės įrangos versijos ir genų modeliai.

GeneMark ir *GeneMark.hmm* gali būti naudojami per *GeneMark* svetainę skirtą priokariotų DNR analizei. DNR analizė, bet kokių prokariotų rūšių yra palaikoma specialios versijos – *GeneMark.hmm*, naudojant euristicinį modelį apskaičiuotą iš nukleotidų dažnio ir sekų įvesties, kurių mažiausias ilgis gali būti 400nt. *GeneMarkS* gali naudoti ilgesnes 1Mb sekas. Kaip dauguma įrankių, *GeneMark* svetainė naudoja panašių sąsajų programas. *GeneMark.hmm* leidžia įkelti failą su DNR sekomis arba įklijuoti jas į laukelį. Jei *FASTA* failo aprašymas prasideda simboliu ">" tai ši eilutė naudojama, kaip pavadinimas. Visos kitos raidės išskyrus A, C, G ir T yra ignoruojamos ir konvertuojamos į N raides. Sąsaja reikalauja pasirinkti rūšies pavadinimą. Pasirinkimas *RBS* modelio yra neprivalomas. *GeneMark.hmm* praneša visus prognozuojamus genus formate, kuriame nurodoma geno kryptis, jo ribos, nukleotidų ilgis ir genų klasės. Klasė nurodo, kuri iš dviejų *Markovo* modelių naudoja *GeneMark.hmm*, tipinį ar netipinį. Galimybė kuri leidžia generuoti *GeneMark* prognozes kartu su *GeneMark.hmm* analize suteikia svarbios papildomos informacijos. Šiuo atveju *GeneMark* yra nustatytas naudoti tuos pačius mokymo duomenis kaip ir *GeneMark.hmm*. Verta paminėti, kad *GeneMark.hmm* ir *GeneMark* papildoma vienas kitą. Grafinės analizės išvestis yra prieinama PDF arba PostScript formatu. Šios išvesties fragmentas, iliustruoja tiek *GeneMark.hmm* tiek *GeneMark* prognozes(7 pav.).

Grafinis išvedimas aiškiai vaizduojamas naudojant keletą *Markovo* grandinės modelių, atstovaujančių skirtingų klasių genus. Kodavimo potencialo grafikas gaunamas naudojant tipinių genų modelius parengtus pagal *GeneMarkS*– žymima juoda linija, o kodavimo potencialas gaunamas naudojant netipinį geno modelį– žymimas punktyrine linija.

Analizė DNR sekų, kurios nėra iš anksto apskaičiuotos specifiniam modeliui, galima naudoti programos versiją, kuri skirta sekoms didesnėms už 400nt. Įrodyta, kad šis metodas naudingas nehomogeninių genomų analizei. Jei modeliai turi būti apskaičiuoti nežinomoms

sekoms, kurios yra 1Mb ar ilgesnio ilgio, gali būti naudojama GeneMarkS programa. Ši programa turi daugiau skaičiavimo išteklių, taigi jos išvestis yra teikiama elektroniniu paštu.



7 pav. Diagrama vaizduojanti GeneMark.hmm ir GeneMark prognozes.

Šiuo metu serveris palaiko sekas užmaskuotas pagal tRNRscan arba panašių programų analizę. GeneMark programos neranda genų aklosose srityse(sekos 'N' simbolių). Aptikimas tikslių genų yra sudėtinga problema genų paieškoje, todėl tobulinami RBS ir Kozak modeliai.

1.2.4. Prokka

Genomo anotacija yra procesas identifikuoti ir ženklinti visus svarbius genomo sekos procesus. Mažiausiai tai turėtų apimti koordinatės identifikuojamų koduojamų regionų ir jų tariamų produktų. Čia pristatomas *Prokka*– komandinės eilutės programinės įrangos įrankis, kuris gali būti įdiegtas, bet kurioje *Unix* sistemoje. *Prokka* koordinuoja esamus programinės įrangos įrankius, kad pasiekti išsamų ir patikimą bakterijų genomų sekų anotaciją. Jei įmanoma, *Prokka* panaudos kelis procesoriaus branduolius, anotacija įvykdytų greičiau. Tipiškas bakterijos genomas, ant keturių branduolių kompiuterio, būtų anotuotas maždaug per 10 minučių. Tai puikiai tinka kartotiniai sekų modelių analizei.

Įvedimas

Prokka tikisi gauti genomo *DNR* sekas *FASTA* formatu. Sekos be tarpų būtų ideali įvestis. Šis sekos failas yra vienintelis privalomas parametras programinei įrangai.

Anotacija

Prokka remiasi išoriniais prognozavimo įrankiais, kad identifikuoti genomo funkcijų koordinates per contigs. Šios priemonės išvardytos 2 lentelėje ir visi jie išskyrus *Prodigal* teikia koordinates ir aprašo atitinkamas funkcijas. Baltymus koduojantys genai yra anotuojami dviem etapais. *Prodigal* identifikuoja galimų genų koordinates, bet neaprašo tariamų genų. Tradicinis būdas prognozuoti ką genas koduoja, jį palyginti su didelės duomenų bazės žinomomis sekomis. Prokka naudoja šį metodą, tik hierarchine tvarka, pradedant nuo mažų patikimų duomenų bazių, po to pereinant prie vidutinių ir galiausiai prie didelių duomenų bazių.

Įrankis	Funkcijos
Prodigal (Hyatt 2010)	coding sequence (CDS)
RNAmmer (Lagesen 2007)	ribosomal RNA genes (rRNA)
Aragorn (Laslett 2004)	transfer RNA and tmRNA genes
SignalP (Petersen 2011)	signal peptides (at N-term of CDS)
Infernal (Kolbe 2011)	non-coding RNA

2 lentelė. Prokka naudojami įrankiai.

Išvedimas

Prokka sukuria dešimties formatų failus į nurodytą išvesties aplanką(3 lentelė).

Galūnė	Failo turinio aprašymas
.fna	FASTA failas su originaliais įvesties contigs (nukleotidai)
.faa	FASTA failas su išverstais kodavimo genais (baltymai)
.ffn	FASTA failas visų genomo funkcijų (nukleotidai)
.fsa	Contig sekos pateikimui (nukleotidai)
.tbl	Funkcijų lentelė pateikimui
.sqn	Redaguotas failas pateikimui
.gbk	Genbank failas su sekomis ir komentarais
.gff	GFF v3 failas su sekomis ir komentarais
.log	Prokka prisijungimo failas apdorojant išvedimą
.txt	Santraukos anotacijos statistika

3 lentelė. Aprašymas Prokka išvedimo failų

Prokka buvo sukurtas siekiant, kad tai būtų tikslus ir greitas įrankis. Pavyzdžiui, anotuojant *Escherichia coli* genomą su tipišku keturių branduolių kompiuteriu užtrunka apie 6 minutes.

2. Praktinė dalis

Praktinėje dalyje sprendžiamas pagrindinis uždavinys yra sukurti genomo naršyklę įprastinėmis priemonėmis naudojant skripto kalbas ir html. Šio uždavinio sprendimas susideda iš kelių dalių:

- *Sulfurimonas denitrificans* bakterijos sekvenuoto genomo analizavimas bioinformatikos įrankiais:
 - *MyRast*;
 - *Prokka*;
 - *GeneMark*.
- įrankių išvesties sekų patikrinimas *Blast*(*Basic Local Alignment Search Tool*);
- bioinformatikos įrankiais anotuočių genomų kryžminis palyginimas tarpusavyje;
- parašytas skriptas baltymų fermentams rasti;
- sukurtas *Python* skriptas skirtas generuoti genomo naršyklę, t.y. html failą.

Toliau aprašysiu šiuos praktinėje dalyje atliktus uždavinius, kokie buvo gauti rezultatai ir kas buvo sukurta.

2.1. MyRast

MyRast tai yra bioinformatikos įrankis skirtas anotuoti prokariotų genomus. Šis įrankis anotuoja genomus ir leidžia parsisiųsti išeities failus, taip pat vizualiai parodo genomo anotaciją. Šio įrankio pagalba mes anotavome analizuojamą *Sulfurimonas denitrificans* (toliau: *Sulfurimonas*) genomą. Gavome anotuotą *FASTA* failą su mums reikalinga informacija.

Pirmiausiai, mes turėjome sekvenuotą bakterijos *Sulfurimonas* genomą. Šis genomas t.y. *FASTA* formato failas, kuriame yra visa *Sulfurimono DNR*. Ši sekvenuota *DNR* susideda iš daugybės A, C, G, T raidžių. Visas genomas yra suskirstytas į 42 dalis, kiekviena dalis turi pavadinimą, ilgį ir identifikacinį numerį(8 pav.).

```
1 >AD-820-E17_NODE_1_length_88502_cov_101.2_ID_1
2 TTTTGTCTCAGTCATCCAGGAGGCAAACAAGAGTTTGATAATAATATAGAAAGTATTA
3 AAAGTGCTTTAAATTTCTACAAGCCGGAAGAATGTAGCGAACTGTAATAAATGACATCA
4 AAGCATTTGTAACATACGCCCATAAAACCTGCGATAACAGCACAAAAAAGAACTTC
5 GATACATGGAAGAGCTACAGGAGATTTCCCAATAAATATTCATGATGAAAGCTAAATT
6 CTCTTGTAAGAGCATCCAAAAATCATAAAAGATGTAATGAGCAAAAGTCAGCTACTT
7 TTACCCAAATAAAAAAGCAATAATGACACTAGAACAAACAATAAAACAAGTCCCGCAAGG
8 TGCAGGAATATATCAATACTTTGATATTAATGGGCATCTCTTTACATTGGGAAAGCAAA
9 AAATCTCTCAAACAGAGTTAAAAGTTATTTTAATTTTACACCTGAATTAAAACCCAATTC
10 TAATCTTTCAAACAGAATCACAAAAATGATTTTGCAAAGTCTTCACTCAGTTATATAGT
11 AGTAAATTCTGAACATGACGCCCTCATCTTAGAAAATTCTCTTATCAAACAGCTGGCTCC
```

8 pav. Fragmentas iš *Sulfurimono* sekvenuoto genomo.

Turint sekvenuotą bakterijos *DNR FASTA* formato failą jau galima naudotis bioinformatikos įrankiais tokiais kaip *MyRast*. Naudojantis *MyRast* galima anotuoti trijų rūšių sekų tipus: *DNR*, baltymų ir *Genbank*. Šiuo atveju mums reikalingas *DNR* sekų tipas. Pasirinkti reikia dar kelis nustatymus, tokius kaip anotavimo greitis– pasirinktas greičiausias, o likusieji parametrai palikti numatytieji: genetinis kodas– 11, Kmer dydis– 8, Kmer rezultato riba(angl. score threshold)– 3, nesutampančių hitų riba– 2. Paskutinis žingsnis yra įkelti savo failą su *DNR* sekomis ir spausti *Start* mygtuką, kuris pradeda vykdyti genomo anotaciją. Sulfurimono genomo anotacija naudojant *MyRast* truko apie 1 valandą, nors genomas ir nėra didelis apie 710kb *FASTA* failas, tačiau, manau tai yra pakankamai greitai, nes atliekama ne tik genomo anotacija, bet ji pavaizduojama ir vizualiai. Atliekamą arba jau atliktą anotaciją galima peržiūrėti paspaudus "Open genome" iš karto yra atidaromas vizualus pavaizdavimas anotuoto *Sulfurimono* genomo(9 pav.).



9 pav. MyRast vizualus genomo pavaizdavimas.

Vizualiaame genomo anotacijos pavaizdavime galima atlikti įvairias funkcijas, nagrinėti ir išsiaiškinti kiekvieną geną individualiai. Šiame atvaizdavime matome daug įvairių spalvų rodyklių, kiekviena iš jų vaizduoja skirtingą geną. Kryptis yra į kairę jei baltymo dydis eina nuo didesnės į mažesnę pusę pavyzdžiui, jei genas prasideda 1863 baze ir baigiasi 1342 baze ir atvirkščiai jei į dešinę pusę. Rodyklių spalvos skiriasi, kiekviena spalva žymi genų funkcinę grupę: šviesiai žalia spalva žymi jog tai fermentų grupė, mėlyna spalva žymimas nežinomas(angl. hypothetical) baltymas ir t.t. Taigi, jei domina konkreti geno funkcinė grupė, jas galima surasti pagal spalvas. Užvedus pelę ant kiekvienos rodyklės yra rodomas atskiras langas su to geno informacija: ID, funkcija, contig, pradžia, pabaiga ir baltymo ilgiu. Pažymėjus, bet kurį geną ir paspaudus dešinę pelės mygtuką, galima sužinoti būtent to geno *DNR* ar baltymo

seką, taip pat yra parinktis, kuri leidžia patikrinti konkretų geną *NCBI Blast*, jei manoma, jog tai gali būti klaidingai anotuotas genas. Taip pat yra galimybė ištrinti geną. Pasirinkus, bet kurį geną, *MyRast* lango viršuje atsiranda informacija apie šį geną, galima jį pataisyti pavyzdžiui įvesta neteisinga geno funkcija, spaudžiame "*edit*" mygtuką ir pataisome funkciją į tinkamą. Taip pat yra paieška genome, jei mums reikalingas atitinkamas genas pavyzdžiui, įvedus 600 parodoma genomo dalis, kurioje yra 600 genas, tai labai patogiu ir naudinga analizuojant genomą, jei žinoma koks konkretus genas yra reikalingas. Taip pat galima pasirinkti atvaizduojamą regionų skaičių ir jų dydį, patogiu jei vienu metu reikia dirbti su didesniu kiekiu genų.

Svarbiausia *MyRast* funkcija šiai analizei– duomenų eksportavimas. *MyRast* leidžia eksportuoti duomenis trimis formatais:

- skirtukais atskirtas tekstas;
- kableliais atskirtas tekstas ;
- *FASTA* formatas.

Taip pat, galima pasirinkti funkcijų tipus, galimi variantai yra *peg* ir *rna*. *Peg* tai yra baltymų numeravimas, o *rna* tai *RNR* funkcijų rodymas. Mūsų atveju abu funkcijų tipai mums reikalingi, todėl pažymimi abudu. Sekanti galimybė yra pasirinkti kokios sekos turi būti įkeliamos į eksportuojamą failą, gali būti įtrauktos tiktai ištransliuotos baltymų sekos arba kartu su *DNR* sekomis. Kaip atrodo baltymų ir *DNR* sekų *FASTA* formato išvesties failas žiūrėti 3 paveikslėlyje.

Kaip matyti paveikslėlyje pirmiausiai yra išvedama *DNR* seka, o tik po to jau ištransliuota baltymo seka, abiejų sekų pavadinimai tokie patys, tai parodo, jog tai ta pati seka. Kaip matome *DNR* seka yra akivaizdžiai ilgesnė nei jau išversta baltymo seka(10 pav.).

```
>fig|6666667.7686.peg.28 AD-820-E17_NODE_1_length_88502_cov_101.2_ID_1_26436_26786 C-type cytochrome, putative
ttgaggtatctttatccggttttatttttcataacttcattgcatgctgtagacatgaga
cctttactctttaaaggaaactgcgtcacatgtcaccatacaagcagaacaatatctgca
ccctctattgtagagattaagaaaaattttaagagcattccctcaaaaagggtttt
gttcggtacatgtcgacatgggttgctaaaccaaatatagagactccataatgcatgat
gcaatagacaagtatgaagtgatgcctgacttaggattgacataagtacttcaagagat
atctctgcttacatttacgaaacagatttttagtcagattgaaccataactaa
>fig|6666667.7686.peg.28 AD-820-E17_NODE_1_length_88502_cov_101.2_ID_1_26436_26786 C-type cytochrome, putative
MRYLPVLFITSLHAVDMRPLLFGNVCVTHHTSRTISAPSIVEIKENYLRAFPQKEAF
VAYMSTWVAKPNIETSIMHDAIDKYEVMPLDGLDISTRDISAYIYETDFSQIEPY
>fig|6666667.7686.peg.29 AD-820-E17_NODE_1_length_88502_cov_101.2_ID_1_27225_26764 Phosphoribosyltransferase
atgtttatgaaatattatgcatatgaagattttaacaagacacaaacaactcttgaaa
cagataaaagagtttcagccgagatgatagtggttcattgctagaggtgggttactttg
tctcatgcatggctgaggggattaacataagagatgttcaaacgctaagaactgaactt
tatgatgacacgcacaaaagggatgagataagcatataacagatgtctattcggtgat
attaagagagtttggttgttgatgacatagctgacagcggtgatacgtcaaagctgtt
```

10 pav. Pavaizduota, kaip atrodo *MyRast* išvesties failas, kuriame įtrauktos ir *DNR* sekos.

Kadangi tolimesnei analizei reikalingos tik baltymų sekos, todėl yra pasirenkama, kad nereikia įtraukti DNR sekų į eksportuojamą failą, taip gaunamas FASTA formato failas tik su baltymų sekomis(11 pav.).

```
>fig|6666667.7686.peg.704 AD-820-E17_NODE_40_length_3792_cov_6.578_ID_79_2157_1591 hypothetical protein
MEDTFSNLATYGYIGFLYSLGGGFVALIGAGVLSFLGKMDLTYSIAIAFFANALGDVLL
FYMARYHKSTMMDGIRKRRKLALSHIMMKYGSWILLIQKYIYGIKTLPIAIGLTKYD
FKKFAILNVFSAGVWALTFGLGSYYSGNALVKFAEIIIGDKPWIAPLVVLGGTLWFYLT
HATKKRTK
>fig|6666667.7686.peg.705 AD-820-E17_NODE_40_length_3792_cov_6.578_ID_79_3778_2165 Excinuclease ABC subunit B
MRLSSTANLLSYDDVIVIASVSANYGLGDPQEYENMVQSVAVGDVIAQKKLLRLVEMGY
SRNDTYFDSGHIRVNGETLDIYPPYFEQEAIREFFGDEIEAIYTFDIIDNKRLEDHKQF
TIYATSQFSVSQEKMAVAIKRIEELDERLAFFQAEGLLEHQLKQRFVDFLEMLQTTG
MCKGVENYSRLLTNKKPGEAPYTLDDYFELHKKDYLIVIDESHVSLPQYRDMYAGDRARK
EVLVEYGFRLPSALDNRLKADYINKAPHYLFVSATPSVYELEMSSVTAKQIIRPTGLL
DPIIEIKSSDNQVEDIHDEIKKITVNDERVLITVLTKKMAEALTKYLADLGKQVQYMHSD
IDTIERNQIIRALRLGEFDVLIGINLLREGLDLPEVSLVALDADKEGFLRSETALVQTI
GRGARNSKGRVILYANKITGSMQRAIEKTTARREIQEAHNKKHNITPTTTKRSLDENLKL
EDYGDLYQHKHKMDKIPASERKAITKELMLRMKQAAKELNFEEAARLRDEIMKIKQL
```

11 pav. Fasta formato failas tik su baltymų sekomis.

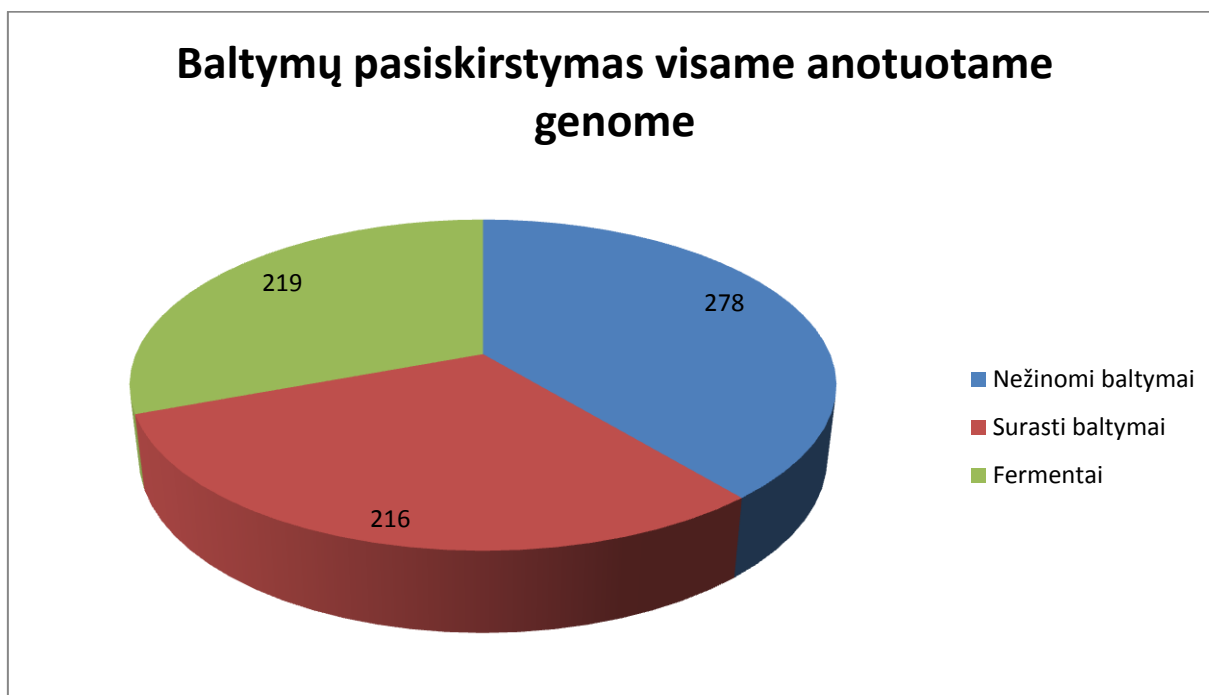
Atlikus bakterijos genomo *Sulfurimono denitrificans* anotaciją su *MyRast* buvo surasta 713 baltymų sekų. Anotuotame genome kiekvienai sekai yra priskiriama viena eilutė su jos anotacija, ji išskiriama pradžioje įterpian varnelės ženklą, panagrinėsiu kas toje eilutėje yra nurodoma, pavyzdžiui, paimsime anotacijos eilutę iš 4 paveikslėlio:

- fig|6666667.7686 - genomo ID, kuris yra priskirtas *MyRast* programos;
- peg.705 - sunumeruojamos sekos pagal eilę;
- AD-820-E17 - sekvenavimo metu genomui suteiktas žymėjimas;
- NODE_40 - žymi iš kurios sekvenuoto genomo dalies buvo ištransliuota baltymo seka;
- length_3792 - dar netransliuoto genomo dalies ilgis;
- ID_79 - taip pat dar netransliuoto genomo ID;
- 3778_2165 - baltymo sekos pradžia ir pabaiga, taip pat galime apskaičiuoti transliuoto baltymo ilgį iš šių skaičių;
- Excinuclease ABC subunit B - baltymo sekos pavadinimas.

Dalis, nuo sekvenavimo metu genomui suteikto žymėjimo iki ID yra paimama iš failo kuris anotuojamas, tai ne *MyRast* atliktos anotacijos dalis. Kaip matome 4 paveikslėlyje 704 sekoje kartais *MyRast* nepavyksta surasti baltymo sekos ir ji yra pažymima, kaip "*hypothetical protein*". Visus nesurastus baltymus t.y. tuos kurie pažymimi "*hypothetical protein*" skripto pagalba išsikerpame iš pagrindinio *FASTA* failo ir ieškome *NCBI Blast protein* pasinaudojant *Blast* skriptu, kuris leidžia vienu metu nurodyti daug ieškomų sekų ir jas parsisiunčia *html* formatu. Pastebėta, jog tos baltymų sekos, kurios *MyRast* pažymėtos nežinomomis ir palyginus jas su rezultatais gautais iš *Blast* apie 90% sekų *Blaste* pirmu numeriu t.y. su didžiausia tikimybe, kad tai yra būtent ta seka yra "*hypothetical protein*" ir jau antru numeriu einančios sekos dažniausiai

įvardija koks tai galėtų būti konkretus baltymas. Galima daryti išvada, jog *MyRast* įrankis renkasi labiausiai tikėtiną iš gaunamų baltymų, nors dažniausiai jau antru numeriu yra pateikiamas tinkamas atsakymas.

Dabar pavaizduosime, kaip *MyRast* anotuosiose 713 sekose pasiskirstė nežinomi baltymai, surasti baltymai ir fermentai(12 pav.).



12 pav. Baltymų pasiskirstymas visame anotuotame genome.

Kaip matome, truputį daugiau nei trečdalis baltymų buvo nesurasta, tai atrodo didelis skaičius, tačiau reikėtų atsižvelgti į tai, jog *Sulfurimonas* yra labai mažai ištirta bakterija, o tai yra faktorius, jos DNR sekų atpažinimo anotavimo sistemose.

2.2. Prokka

Viso genomo anotacija tai procesas identifikuoti funkcijas ar savybes genomo *DNR* sekų ir žymėti juos naudinga informacija. Tam yra skirta *Prokka* - programinės įrangos įrankis greitam anotavimui bakterijų, archėjų ir virusų genomų, kuris sukuria standartus atitinkančius išvesties failus. Šis įrankis yra valdomas naudojantis komandine eilute. Taip pat šis įrankis yra pakankamai sunkiai instaliuojamas, reikia nemažai priedų, kad būtų galima naudotis *Prokka*.

Naudotis šiuo įrankiu nėra taip paprasta, kaip *MyRast*, tačiau paskaičius vartotojo vadovą ir pasidomėjus apie šį įrankį, bei pabandžius atlikti kelias anotacijas viskas tampa suprantama. Šis įrankis negali sukurti vizualaus genomo pavaizdavimo, jis teikia tik įvairius išvesties failus, o jų

yra išties nemažai. *Prokka* generuoja 10 išvesties failų ir generuoja juos visus atliekant, bet kokią anotaciją. *Prokka* turi nemažai opcijų nuo tokių kaip direktorijos nurodymas į kurią turi būti sukurti visi 10 išvesties failų iki tokių kaip genų ar rūšių vardų nustatymas.

Šiuo bioinformatikos įrankiu taip pat anotavome *Sulfurimonas denitrificans* bakteriją. Pirmiausiai komandinėje eilutėje reikėjo parašyti komandą, kuri įvykdytų bakterijos anotaciją, parašyta paprasta, aiški ir pakankama komanda, kad būtų įvykdyta anotacija:

```
prokka --outdir Prokka --prefix Sulfurimonas AD-820-E17.fa
```

Šia komanda nurodau direktoriją **Prokka** į kurią norėčiau, kad būtų sukurti visi 10 išvesties failai, tų failų priešdėlį *Sulfurimonas* ir FASTA formato failą kuriame yra sekvenuoto genomo DNR sekos **AD-820-E17.fa**. Šis įrankis genomo anotaciją įvykdė tikrai greitai, *Sulfurimono* genomą anotavo per maždaug 10 minučių(13 pav.).

```
>PROKKA_00001 UvrABC system protein C
MTLEQTIKQLPQGAGIYQYFDINGHLLYIGKAKNLSNRVKS YFNFTPELKPN SNLSNRIT
KMILQTASLSYIVVNSEHDALILENSLIKQLAPKYNILLRDDKTYPIYIDNSSEYPRFD
ITRKIIKSADITYFGPYSGARDILDSIYEVCKLVQKKACLKSKKACLYYQIDKCLAPCE
FKVSHVRYKAELDLAQELIQNKKT LISKLTEKMSFYAEEMRFEEAGELRDRIERISRSEI
KSEIDFASNENYDIFVLHNSETRAVAVRIFMRNGKIISSSHDFIQ LNDGYDEDEFYQ RVL
LDFYAKEKPPIIAPILVMKKFSGLEIIAEHLSILFEKKALITAPSRGDKKHLIDLAVLNA
KELLKADKKQDLTKLFT EIKELLSLERIPNRVEIFDN SHMAGVATVGAMIVYENSQFDKK
SYRTYHLDAKDEYAQMRETLTRRVESFSKNSPPDLWIIDGGTLLRLAVDILDSNGIFID
VIAISKEKIDAKSHRAKGKAHDILHTKDESFR LNPNDKRLQWIQNLRDEAHRSAIAFHKK
TKLKLDKASKLLNLHGISEAKIKLLNHFGTFEALKE LSEEEIASVLNIKDAQAIKNIYK
>PROKKA_00002 Aerotaxis receptor
MDKVIPVDEEYIYKGRVIISQTDVKGIITFANRK FYEVSGYALDELIGSSHSIMRHPNMP
KAVFEKIWETISGGQIWTGIIKNLRKDGRYYWVDIEILPIHNENNELTGYISARKPASRK
NVNETMALYKKMLATEQGDKNVNL
```

13 pav. FASTA formato failas generuotas su Prokka.

FASTA faile yra nurodomas tik genų eiliškumas ir surastas baltymas, truputį netoks kokio tikėjau, man labai trūksta informacijos apie baltymo pradžią ir pabaigą genome, taip pat nežinomas kiekvieno geno ilgis. Taip pat nėra nurodoma fermentų pavadinimų, o tai irgi nėra naudinga, nes man fermentai reikalingi genomo naršyklei kurti. Negaliu teigti, jog *Prokka* neanotavo man reikalingų duomenų, jie buvo anotuoti tik kituose formatuose. Visa man reikalinga informacija yra .gbk faile t.y. standartinis *Genbank* failas, kur kiekvienai sekai aprašoma jos informacija(14 pav.).

```

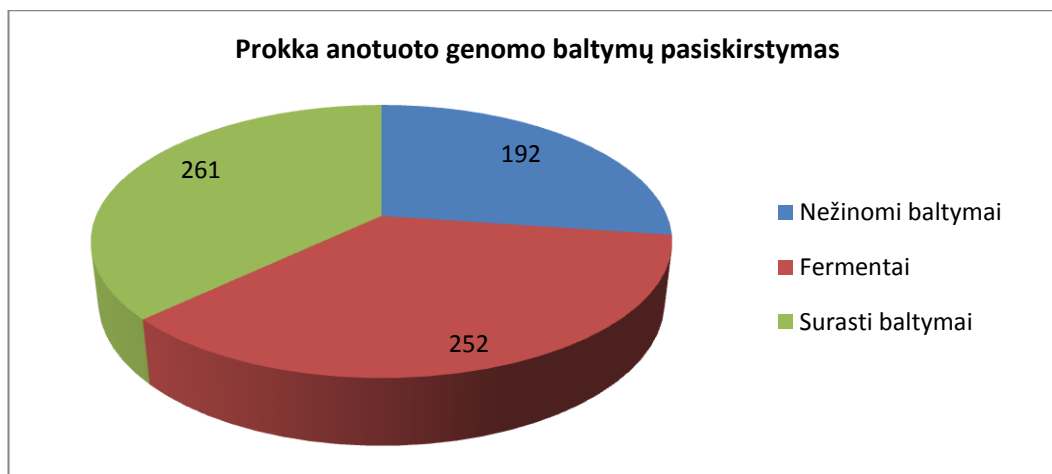
CDS      4158..6530
         /gene="barA_1"
         /locus_tag="PROKKA_00004"
         /EC_number="2.7.13.3"
         /inference="ab initio prediction:Prodigal:2.6"
         /inference="similar to AA sequence:UniProtKB:P0AEC5"
         /codon_start=1
         /transl_table=11
         /product="Signal transduction histidine-protein kinase
         BarA"
         /translation="MQIGLKNRLRLISLFPILILFSLTSYFIYDFYMNQENHIVFLT
         LTFVIWVISIILAILGYLLSNEISTNITKLEDVLRVAEDNKNDFSSGKEMDLHTPEGT
         ADAYNLLEKIIKKTREDKAFAQEASEAKSMFLANMSHEIRTPNGIVGFTELLKDTGL
         GEEQLEFVEIIEKSSNLEIINNILDLSKIESNKIEVEDVIFDPITEFESAVDVYAV
         RASEKHIDLGCFIDPALERPLKGDPTKIKEVIINLLSNAVKFTSNAGALNVEIRKIDS
         GIEGTTRISFEIQDSGIGITSEQKSRIFDAFSQADISITRKYGGTGLGLTISSRFVEL
         MGGKLNLRSKLGKGTTFYFTLDFEEAAEEAVNSSKNIFSKLNALILESPLYTKKQESYL
         REYLDYFGVNYKMFKNLDEIEISQKYSYDYLVDYEFISELLHKYEEFPPKVILLA
         KSSFMKKIDSMTLNIYKTLYEPLNISKLQQILSNYYTEKLTMTKIKKVVKKVEEKDL
         KFNANILVAEDNVINQKLIKRTLEDIGLSVSVASNGLEAFQKRKDNFDLI FMDIQMP
         YLDGIEATQEILDYEKDYNKPHVPIIALTANALKGDRAKFL EAGLDEYTTKPLIRSEI
         VSM LN NF L SDFVVS GSA AISDTKSKPIVIPKKEVASLPESKNDGKNYKADILLAKQSA
         FESKLYAKILTELEYSYEITSNIEELKNLTKEFTYKLVLDDEFNGLELSQFSKDIKE
         SNVATGFKTHLILINSSQKEKILEYKPYVDEI IENVVNKDILQLVFKKFI"

```

14 pav. Ištrauka iš anotuoto Genbank formato failo.

Šio formato faile yra visa reikalinga informacija: DNR sekos pradžia ir pabaiga, jos ilgis, vieta genome, surasti fermentai, baltymo pavadinimas ir ištransliuotas pats baltymas.

Su *Prokka* įrankio pagalba buvo anotuotas *Sulfurimono* genomas, rasta 705 sekos t.y. 8 mažiau nei su *MyRast* įrankiu. Dabar pavaizduosime, kaip šiam įrankiui pavyko anotuoti baltymus ir fermentus(15 pav.).



15 pav. Prokka įrankiu anotuoto genomo baltymų pasiskirstymas.

Kaip matome *Prokka* įrankis identifikavo daugiau baltymų nei *MyRast*, mažiau nei trečdalis, liko neidentifikuoti. Visus nežinomus baltymus taip pat patikrinus *Blast*, buvo apie 90% baltymų sekų kuriose pirmoje vietoje buvo rodomas "*hypothetical protein*". *Prokka* įrankis man patiko dėl greito anotavimo ir didelio pasirinkimo išvesties failų, bet trūko informacijos *FASTA* formato faile ir deja, bet jis vizualiai neatvaizduoja genomo.

2.3. GeneMark

GeneMark yra interneto programinės įrangos svetainė, kuri suteikia sąsajas su *GeneMark* šeimos programomis. Jos skirtos prognozuoti prokariotų, eukariotų ir virusų genomų genus. Programos išvesties failai atsiunčiami į elektroninį paštą. Išbandžiau šias *GeneMark* šeimos programas:

- *GeneMark*;
- *GeneMark.hmm*;
- *GeneMarkS*.

Naudojimasis *GeneMark* svetaine yra labai paprastas, pirmiausiai reikia pasirinkti programą kuria bus naudojama, aš išbandžiau tris jų programas:

Pasirinkus *GeneMark* programą, atsiranda vienintelis pasirinkimas prokariotai, paspaudus ant jo, atsiranda langas leidžiantis pasirinkti kelis nustatymus. Pirmiausiai yra įkeliamas failas su sekomis tik *FASTA* formatu, sekančiame lange pasirenkama prokarioto rūšis, čia pavyko rasti reikalingą *Sulfurimonas denitrificans*. Sekančiame lange keli pasirinkimai susiję su išvesties failais, galimi pasirinkimai baltymų ir genų nukleotidų failai, pasirenkami abu. Taip pat yra leidžiama pasirinkti kodavimo potencialų grafiką, tačiau jis neleidžiamas *multi-FASTA* failams, taigi jis netinkamas, nes *Sulfurimonas denitrificans* genome daug DNR sekų. Paskutiniame žingsnyje yra prašoma įvesti elektroninį paštą, į kurį bus nusiųsti išvesties failai. Nustebino tai, jog vos spėjau paspausti vykdymo mygtuką ir nuėjus į paštą failai jau buvo atsiųsti. Failuose skyrėsi tik sekos vienos išverstos į baltymų sekas, kitos į genų nukleotidų (16 pav.).

```
>orf_1 (AD-820-E17_NODE_1_length_88502_cov_101.2_ID_1, 3 - 323) translated
FVLSHPGGKQEFDDNNIESIKSALKFYKPEECSETVINDIKAFVTHTPIKPAITAQKKELRYMERATGDFP
INIHDEKLNSLVKSIQKIIKDVNEQKSATFTQIKKQ*
>orf_2 (AD-820-E17_NODE_1_length_88502_cov_101.2_ID_1, 323 - 2125) translated
MTLEQTIKQLPQGAGIYQYFDINGHLLYIGKAKNLSNRVKSIFYNFTPELKPNSNLSNRITKMILQTASLS
YIVVNSEHDALILENSLIKQLAPKYNILLRDDKTYPIYIDNSSEYPRFDITRKIIKSADITYFGPYSVG
ARDILDSIYEVCKLVQKKACLKSKKACLYYQIDKCLAPCEFKVSHVRYKAELDLAQELIQNKKTLISKLT
EKMSFYAEEMRFEELRDRIERISRSEIKSEIDFASNENYDIFVLHNSETRAVAVRIFMRNGKIISSS
HDFIQNLNDGYDEDEFYQRVLLDFYAKEKPPIIAPILVMKKFSGLEIIAEHLSILFEKKALITAPSRGDKK
HLIDLAVLNAKELLKADKKQDLTKLFTTEIKELLSLERIPNRVEIFDNSHMAGVATVGAMIVYENSQFDKK
SYRTYHLDADDEYAQMRETLTRRVESFSKNSPDLWIIDGGTTLLRLAVDILDSNGIFIDVIAISKEKID
AKSHRAKGKAHDILHTKDESFRLLNPNDKRLQWIQNLRLDEAHRSAIAFHKKTKLKLDKASKLLNLHGISEA
KIKKLLNHFGTFFALKELSEEEIASVLNIKDAQAIKNIYK*
```

16 pav. GeneMark programos išvesties FASTA failas.

Išvesties failas atrodo neinformatyviai, tik išverčiamos sekos į baltymų sekas. Eilutėje kuri apibūdina baltymo seką yra tik perkopijuojama informacija iš pradinio failo, pažymima sekos

pradžią ir pabaigą ir galėdama pažymėti, jog baltymo seka yra išversta. Apie tai koks tai galėtų būti baltymas nėra pateikiama jokios informacijos.

Naudojantis *GeneMark.hmm* ir *GeneMarkS* programomis, viskas vyko taip pat kaip ir naudojantis *GeneMark*, tik prisidėjo keli papildomi nustatymų pasirinkimai. Naudojantis *GeneMark.hmm* atsirado papildomas pasirinkimas išvesties failo formato genų prognozavimui, galimi pasirinkimai tarp *LST* ir *GFF* formatų. Naudojantis *GeneMarkS* prie formato pasirinkimo atsirado dar ir sekos tipo pasirinkimas: prokariotas, eukariotas, virusas, fagai, *EST/cDNA*, mes pasirinkome prokarioto seką ir abiem programom failo formatas parinktas *LST*. Visi kiti nustatymai tokie patys, kaip ir naudojantis *GeneMark* programa. Į elektroninį paštą failai atsiunčiami taip pat iš karto. Abu išvesties *FASTA* failai atrodo identiška (17 pav.).

```
>gene_1|GeneMark.hmm|106_aa|+|3|323 >AD-820-E17_NODE_1_length_88502_cov_101.2_ID_1
FVLSHPGGKQEFNNIESIKSALKFYKPEECSETVINDIKAFVTHTPIKPAITAQKKELR
YMERATGDFPINIHDEKLNSLVKSIQKIIKDVNEQKSATFTQIKKQ

>gene_2|GeneMark.hmm|600_aa|+|323|2125 >AD-820-E17_NODE_1_length_88502_cov_101.2_ID_1
MTLEQTIKQLPQGAGIYQYFDINGHLLYIGKAKNLSNRVKSYPNFTPELKPNNSNLSNRIT
KMILQTASLSYIVVNSEHDALILENSLIKQLAPKYNILLRDDKTYPIYIDNSSEYPRFD
ITRKIIKSADITYFGPYSGARDILDSIYEVCKLVQKKACLKSKKACLYYQIDKCLAPCE
FKVSHVRYKAELDLAQELIQNKKTILSKLTEKMSFYAEEMRFEEAGELRDRIERISRSEI
KSEIDFASNENYDIFVLHNSETRAVAVRIFMRNGKIISSSHDFIQLNDGYDEDEFYQRVL
LDFYAKEKPPIIAPILVMKKFSGLEIIAEHLSILFEKKALITAPSRGDKKHLIDLAVLNA
KELLKADKKQDLTKLFTEIKELLSLERIPNRVEIFDNHMGAVATVGAMIVYENSQFDKK
SYRTYHLDAKDEYAQMRETLTRRVESFSKNSPPDLWIIDGGTTLLRLAVDILDSNGIFID
VIAISKEKIDAKSHRAKGKAHDILHTKDESFRLNPNDKRLQWIQNLRDEAHRSAIAFHKK
TKLKLDKASKLLNLHGISEAKIKKLLNHFGTFEALKELSEEEIASVLNIKDAQAIKNIYK
```

17 pav. *GeneMark.hmm* ir *GeneMarkS* išverstos baltymų sekos.

Išvesties faile matoma, jog anotacijos eilutėje prisidėjo skaičius rodantis baltymo sekos ilgį, šios informacijos neteikė nė vienas įrankis. Taip pat rodoma sekos pradžia ir pabaiga ir pradinio genomo dalies pavadinimas.

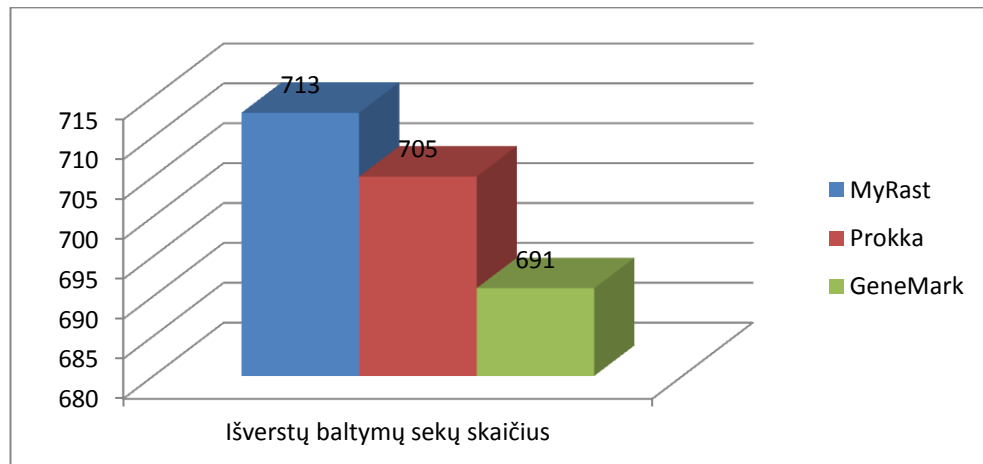
Išvesties failai tarp *GeneMark.hmm* ir *GeneMarkS* visiškai nesiskiria, tačiau palyginus juos su *GeneMark* programos išvesties failu atsiranda vienas didelis skirtumas. Skiriasi rastų sekų skaičius:

- *GeneMark.hmm* - 735 rastos sekos;
- *GeneMarkS* - 735 rastos sekos;
- *GeneMark* - 691 rastos sekos.

Tikslaus paaiškinimo neradau, kodėl rasti skirtingi kiekliai sekų, tačiau manau, kad tai galbūt susiję su *GeneMark.hmm* ir *GeneMarkS* programose pasirinktu genų prognozavimo formatu *LST*, nes šie failai tarpusavyje nesiskiria.

2.4. Baltymų sekų palyginimas

Atlikus *Sulfurimonas denitrificans* genomo anotacijas su trimis skirtingais bioinformatikos įrankiais ir gavus 3 *FASTA* formato failus su identifikuotomis baltymų sekomis reikia palyginti juos tarpusavyje. Šis palyginimas reikalingas, nes atlikus anotacijas su trimis skirtingais įrankiais gauti rezultatai skiriasi (18 pav.).



18 pav. Išverstų baltymų sekų skaičius su kiekvienu įrankiu.

Rezultatai skiriasi nežymiai. Vadinas, tas pats genomas su tomis pačiomis *DNR* sekomis yra identifikuojamas skirtingai. Todėl atliksime palyginimą kiekvieno *FASTA* failo su kiekvienu t.y iš viso buvo atliekami 3 palyginimai:

- *MyRast* su *Prokka*;
- *MyRast* su *GeneMark*;
- *Prokka* su *GeneMark*.

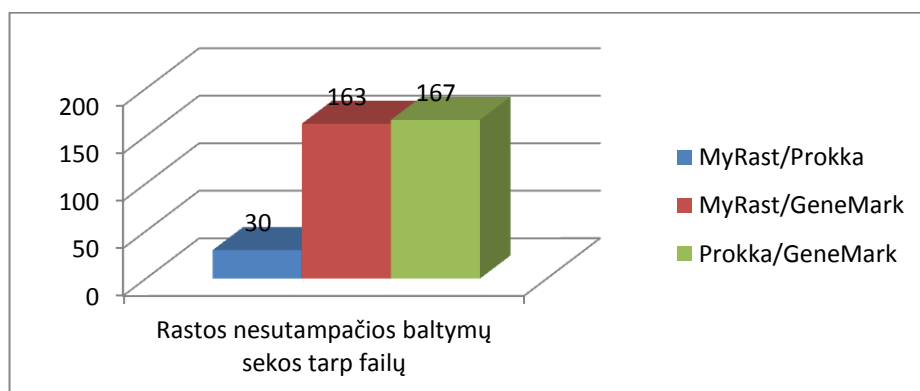
Kadangi failai nėra tokie maži, kad būtų galima atlikti palyginimą rankiniu būdu vienos sekos ieškant kitame faile ir taip ieškant visų sekų, todėl buvo parašytas *Python* skriptas, kuris atlieka visą palyginimą ir rezultate išveda sekas kurios yra nerandamos ieškomame faile t.y. baltymų sekų skirtumas tarp *FASTA* failų.

Trumpai apžvelgsime parašyto *Python* skripto veikimą ir jo atliekamus žingsnius:

- nuskaitomi failai;
- sukaipomas failas į atskiras sekas;
- kiekviena seka suskirstoma į mažesnes sekas po 20 raidžių;
- ieškoma mažesnių sekų kitame baltymų sekų faile;
- spausdinami rezultatai.

Pirmiausiai yra nuskaitomi abu lyginami failai. Nuskaitomi į vientisą tekstą t.y. į vieną eilutę ir iš karto ištrinami naujos eilutės simboliai, kad būtų galima rasti jame ieškomą seką. Nuskaičius failus jie yra sukarpmi pagal anotacijos eilutes, kad liktų tikrai atskiros baltymų sekos, be jokių pašalinių ženklų, nes atlikus anotacijas su skirtingais įrankiais baltymai yra anotuojami skirtingai. Turint visus baltymus, kaip atskiras eilutes, tos eilutes yra dar kartą karpomos į mažesnes eilutes po 20 simbolių, šis veiksmas atliekamas tam, kad išvengti vienos raidės nesutapimo. Vienos raidės nesutapimas tai pavyzdžiui, jei paimsime vieną baltymo seką ir ieškosime kitame faile, jei būtų surasta tokia pati seka kuri skiriasi tik vienu simboliu, programa išvestų ją kaip nesutampančią eilutę. Mus domina visiškai arba bent didžiąja dalimi sutampančios baltymų sekos t.y. mes ieškome mažųjų sekų kitose baltymų sekose, jei ji randama yra tikrinama pilna seka, jei pilnos sekos neradome, toliau tikriname mažąsias sekas iš to pačios baltymo sekos, jei randame daugiau kaip pusę mažųjų sekų, tada baltymas pažymimas kaip sutampančias. Jei baltymo seka nerandama tuomet ji spausdinama kaip nesutampančias.

Atlikus palyginimą su visais trimis *FASTA* failais, gauti skirtingi rezultatai(19 pav.).



19 pav. Nesutampančių baltymų sekų kiekis tarp failų.

Kaip matome diagramoje, rezultatai labai stipriai išsiskyrė lyginant su *GeneMark* įrankiu, tiek lyginant su *MyRast* 163 skirtingos sekos, tiek su *Prokka* 167 skirtingos sekos. Kadangi tokie dideli skirtumai tarp įrankių, didžiąją dalį tų sekų kurios nurodytos kaip skirtingos patikrinau tiesiog rankiniu būdu, kad įsitikinčiau, jog skriptas veikia tikrai gerai ir tai pasitvirtino– jis veikia gerai.

Iš to išplaukia išvada, kad *GeneMark* programa nepatikimai identifikuoja *DNR* sekas į baltymų sekas, bent jau šiam konkrečiam *Sulfurimono* genomui, kadangi didelis neatitikimas tarp sekų ir beveik toks pat neatitikimų skaičius lyginant *GeneMark* su *MyRast* ir *Prokka*. Pastarųjų dviejų įrankių lyginimas manau, davė realius rezultatus, 30 skirtingų baltymų sekų nėra didelis skaičius atsižvelgiant į tai, jog ir pats identifikuotų sekų skaičius skiriasi 8 sekomis, taip pat

skiriasi pačių įrankių metodai identifikuojant baltymus. Visos šios 30 sekų buvo patikrintos *Blast* naudojantis skriptu ir rezultatai buvo pasiskirstę tolygiai žiūrint į pirmoje vietoje esančias prognozes t.y. apie trečdalis buvo nežinomi baltymai, trečdalis priklausė *Sulfurimonas* bakterijai ir trečdalis buvo nurodyti kaip kitų organizmų baltymai. Tolimesniame tyrime remsiuosi tik tomis baltymų sekomis, kurios buvo vienodos lyginant *MyRast* ir *Prokka* įrankių išvestis, *GeneMark* baltymų sekos davė nepatikimus rezultatus, todėl siekiant kuo tikslesnių rezultatų šio įrankio baltymų sekomis nebus remiamasi tolimesniame tyrime.

2.5. Fermentų paieška

Fermentas– baltyminis katalizatorius, kuris paspartina organizme vykstančias chemines reakcijas tūkstančius kartų. Be fermentų šios reakcijos nevyktų arba vyktų labai lėtai ir organizmai negalėtų egzistuoti. Fermentai yra sudėtingos organinės medžiagos: vienkomponenčiai– sudaryti tik iš baltymų ir dvikomponenčiai– į kurių sudėtį, be baltymų, įeina ir kitos medžiagos, dažniausiai vitaminai. Šiuo metu iširta daugiau kaip 3000 fermentų.

Fermentai yra svarbios medžiagos organizmui, todėl jas taip pat svarbu surasti ir pažymėti fermento kelius(angl. pathway) kuriant genomo naršyklę. Kurie baltymai turi fermentus atskirai ieškoti nereikės, tai jau atliko *MyRast* įrankis anotuodamas genomą, prie kiekvienos sekos kuri turi fermentus skliausteliuose pažymėta pavyzdžiui (EC 3.4.21.53) kuris nurodo tam tikro fermentą. Mes surasime kiekvieno to fermento kelius(angl. pathway). Kadangi iš viso palygintame faile yra 212 fermentų, pirmiausiai buvo parašytas nedidelis skriptas, kuris iš failo iškerpa visas eilutes kuriose yra fermentų pavadinimai ir iš tų eilučių paima jau tik pačių fermentų pavadinimus ir taip gaunamas sąrašas visų faile esančių fermentų.

Toliau, reikia parsisiųsti visą informaciją apie tuos fermentus, tam yra naudojama *Kegg* duomenų bazė. *Kegg* tai yra rinkinys duomenų bazių susijusių su genomais, biologiniais keliais, ligomis, narkotikais ir cheminėmis medžiagomis. *Kegg* naudojama bioinformatikos tyrimams ir švietimui, įskaitant duomenų analizę genomikoje, metagenomikoje, metabolomikoje. Parsisiųsti informaciją apie kiekvieną fermentą parašytas nedidelis skriptas, kuris automatiškai parsisiunčia visus *html* failus apie kiekvieną nurodytą fermentą(20 pav.).

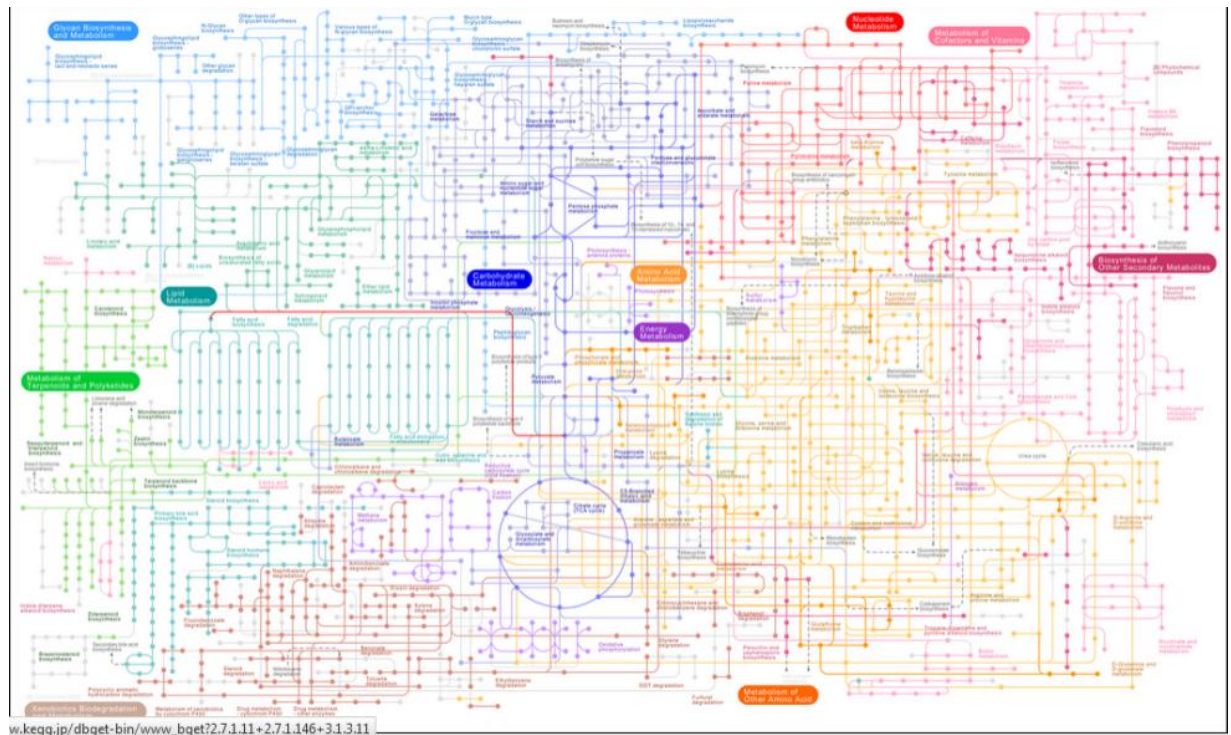
```
#!/usr/local/bin/python

import urllib

seq = open("RastKeggEC1", "r")
seq1 = seq.readlines();
for line in seq1:
    url = "http://www.kegg.jp/dbget-bin/www_bget?ec:" + line
    name = "ec" + line
    print line
    urllib.urlretrieve(url, filename=name)
```

20 pav. Skriptas skirtas parsisiųsti html failus su informacija apie fermentus.

Tame faile yra reikiama informacija apie kiekvieno fermento kelius(angl. pathway). Fermento kelias, tai internetinis adresas į kurį nuėjus yra vizualiai pavaizduojamas tam fermentui priklausantis kelias. Taip atrodo fermentui *ec6.4.1.2* priklausantis metabolinis kelias(21 pav.).



21 pav. Vaizduoja fermento *ec6.4.1.2* metabolinį kelią.

Iš viso parsųsta 212 *html* failų iš kurių reikia paimti adresus, kurie parodo konkretaus fermento kelius. Tam, taip pat buvo parašytas skriptas, kuris nusiskaito visų failų pavadinimus į sąrašą, po to ima po vieną failą, jį atidaro ir visą nusiskaito tuo pačiu ištrindamas naujos eilutės simbolius. Sekančiu žingsniu suskaičiuoja kiek faile yra kelių ir tiek kartų iškerpa iš failo reikalingų kelių pavadinimus, suformuojamas adresas ir rezultatai rašomi į failą (22 pav.).

```
ec2.1.1.207
ec1.2.1.38
http://www.kegg.jp/kegg-bin/show_pathway?ec00330+1.2.1.38
http://www.kegg.jp/kegg-bin/show_pathway?ec01100+1.2.1.38
http://www.kegg.jp/kegg-bin/show_pathway?ec01110+1.2.1.38
http://www.kegg.jp/kegg-bin/show_pathway?ec01130+1.2.1.38
ec2.7.6.3
http://www.kegg.jp/kegg-bin/show_pathway?ec00790+2.7.6.3
http://www.kegg.jp/kegg-bin/show_pathway?ec01100+2.7.6.3
```

22 pav. Pavaizduota, kaip atrodo sąrašas su atrinktais fermentų keliais.

Kaip matome pavaizduoti fermentai ir jiems priklausantys keliai, nuėjus tais adresais matysime vizualų tų kelių pavaizdavimą. Kai kurie fermentai kelių neturi, po jais nieko nerašoma.

2.6. Genomo naršyklė

Bioinformatikoje genomo naršyklė tai grafinė sąsaja pavaizduoti informaciją iš biologinės duomenų bazės genomo duomenims. Genomo naršyklė leidžia tyrėjams vizualizuoti ir naršyti pasirinktą genomą su anotuotais duomenimis, genų prognozavimu, baltymais ir t.t. Anotuoti duomenys dažniausiai yra iš kelių skirtingų šaltinių.

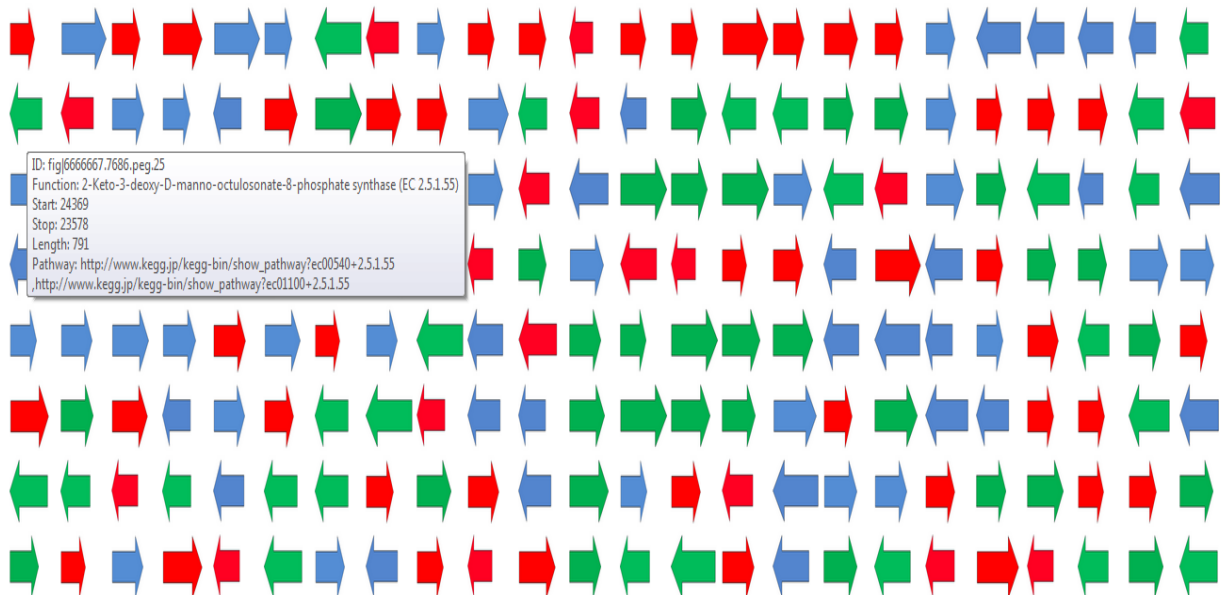
Paskutinis žingsnis yra panaudoti visus gautus rezultatus kuriant genomo naršyklę. Tai yra rezultatus, kurie buvo gauti palyginus *MyRast* ir *Prokka* baltymų sekas ir surastus fermentų kelius. Iš viso palygintame faile yra 683 sutampančios baltymų sekos, todėl sukurti genomo naršyklę su anotuojama informacija tokiam kiekiui sekų daug laiko reikalaujantis darbas. Šiam darbui atlikti buvo nuspręsta parašyti skriptą, kuris generuotų *html* failą su atvaizduojama genomo naršykle. Skriptas parašytas *Python* programavimo kalba, žiūrėti prieduose. Šis skriptas yra universalus, jis gali būti naudojamas ir kitiems su *MyRast* įrankiu identifikuotiems genomams, tačiau jame turėtų būti tik baltymų sekos t.y. neįtrauktos DNR sekos į išvesties failą. Trumpai aptarsime parašytą skriptą. Skripto pagrindiniai atliekami žingsniai:

- nuskaitomi failai(baltymų sekų ir fermentų);
- į failą įrašomos pirmosios eilutės reikalingos *html* failui pradėti;
- nuskaitomi ir apskaičiuojami baltymų sekų ilgiai, id ir pavadinimai;
- surandami baltymai kuriems priskirti fermentai ir priskiriami fermentų keliai;
- patikrinama baltymų sekų kryptis;
- apskaičiuojamas rodyklės dydis, kiekvienam baltymui žymėti;
- generuojamas paprastas *html* failas su reikalinga informacija;
- tuo pačiu generuojamas atskiras *html* failas su baltymų sekomis ir jų anotacija.

Pirmiausiai iš bendro *MyRast* ir *Prokka* failo yra nuskaitomos baltymų sekos su anotacija apie juos ir iš atskiro failo nuskaitomi fermentai su jų keliais(angl. pathway). Pačioje pradžioje įrašomos kelios eilutės į failą kurios reikalingos *html* failo pradžia. Iš nuskaitytų baltymų anotacijų yra nuskaitoma baltymo sekos pradžia ir pabaiga ir apskaičiuojamas jų ilgis, taip pat nuskaitoma id ir pavadinimas. Id tai kiekvienos baltymo sekos anotacijoje pirmieji simboliai pavyzdžiui id = fig|6666667.7686.peg.1, sekančių sekų id skiriasi tik paskutiniu skaičiumi. Pavadinimas tai baltymo sekos pavadinimas pavyzdžiui *Excinnuclease ABC subunit C*. Toliau, kiekvieno baltymo anotacijoje ieškoma ar jam yra priskirtų fermentų, jei taip iškerpamas to fermento pavadinimas ir atskirame faile ieškoma ar tam fermentui yra priskirtų kelių, jei taip jie paimami ir dedami į *html* failą kartu su ilgiu, pavadinimu ir id. Jei baltymo seka neturi jam priskirtų fermentų, tuomet darbas tęsiamas toliau. Sekančiame žingsnyje yra nustatoma kokia

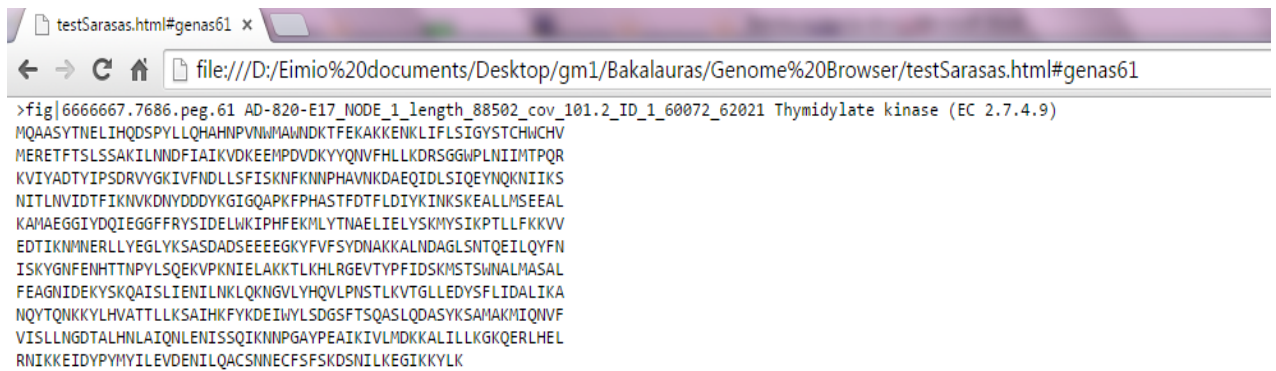
baltymo sekos kryptis, jei iš baltymo pradžios atėmus pabaigą gaunamas minusinis skaičius tai rodyklės, kuri simbolizuoja baltymą, kryptis bus į kairę pusę, jei plusinis tuomet kryptis į dešinę pusę. Rodyklės yra paaimamos kaip paveikslėliai rodančios viena arba kita kryptimi. Rodyklės yra trijų spalvų: raudona– nežinomi baltymai, žalia– fermentai ir mėlyna– likusieji baltymai. Kiekvienam baltymui yra apskaičiuojamas jo ilgis ir nustatomas jo rodyklės dydis. Turint visą reikalingą informaciją apie baltymą galima kiekvienam generuoti *html* kodą, kuris nurodys visus duomenis apie jį. Taip pat kiekvienam baltymui yra suteikiama nuoroda į kitą *html* failą, kuriame yra sąrašas visų baltymų sekų su jų pavadinimais, taip pat jei tai fermentas, po juo rašomos nuorodos į fermentų kelius. Šis sąrašas skirtas tam, kad vartotojas prireikus galėtų pamatyti kiekvieną baltymą individualiai.

Įvykdžius skriptą yra sugeneruojami du *html* failai, vienas iš jų genomo naršyklė, kitas baltymų sekų sąrašas(23 pav.).



23 pav. *Sulfurimonas denitrificans* genomo naršyklės dalis.

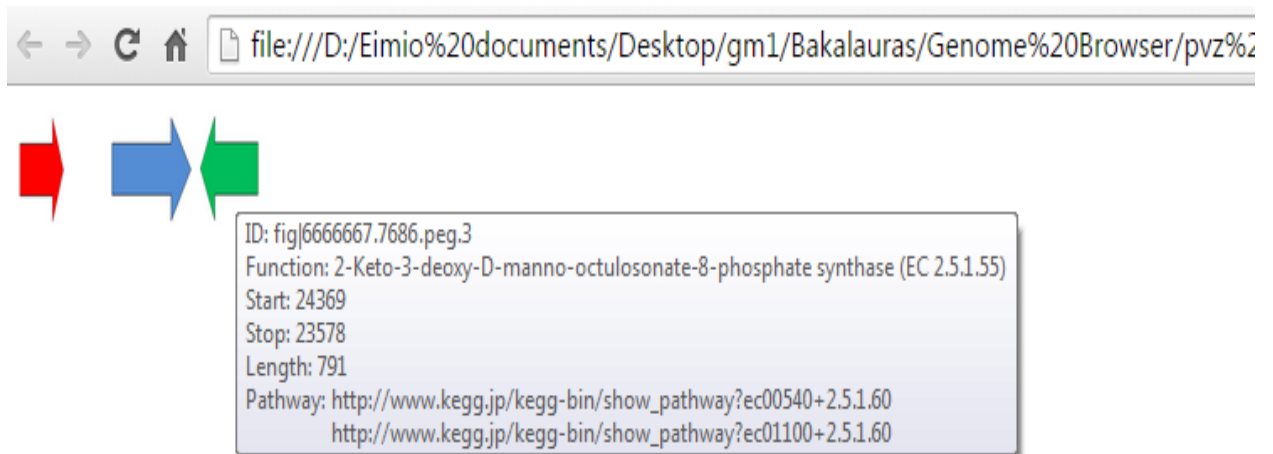
Kaip matyti paveikslėlyje, daug skirtingo dydžio ir krypties rodyklių, kiekviena iš jų atvaizduoja tam tikrą *Sulfurimonas denitrificans* genomo baltymą. Iš viso yra 683 baltymų sekos ir pavaizduota tiek pat rodyklių. Užvedus pelę ant bet kurios rodyklės, kaip matyti 16 paveikslėlyje, atsiranda laukelis kuriame yra informacija apie būtent tą baltymą, šiame laukelyje rodoma informacija: ID, funkcija, baltymo sekos pradžia ir pabaiga, jos ilgis ir fermentų keliai, jei baltymas turi fermentų. Paspaudus ant bet kurios rodyklės vartotojas yra nukeliamas į kitą *html* failą, kuriame yra pavaizduotas būtent tas baltymas ant kurio buvo paspausta(24 pav.).



```
>fig|6666667.7686.peg.61 AD-820-E17_NODE_1_length_88502_cov_101.2_ID_1_60072_62021 Thymidylate kinase (EC 2.7.4.9)
MQAASYTNELIHQDSPYLLQHAHNPVNMAMNDKTFEKAKKENKLIPLSIGYSTCHWCHV
MERETFTLSLSSAKILNNDFAIKVDKEEMPDVKYQNVFHLKDRSGGWPLNIIMTPQR
KVIYADTYIPSDRVYGVKIVFNDLLSFISKNFKNPHAVNKDAEQIDLSIQEYNQKNIKS
NITLNVIDTFIKNVKDNVDDYKIGIQAPKFPASTFDFTLDIYKINKSKEALLMSEAL
KAMAEGGIVDQIEGGFFRYSIDELWKIPHFEMLYTNAELIELYSKNYSIKPTLLFKKVV
EDTIKNMNERLLYEGLYKSASDADSEEEGKYFVFSYDNAAKKALNDAGLSNTQEILQYFN
ISKYGNFENHTTNPYLSQEKVPKNIELAKKTLKHLRGEVTPFIDSKMSTSWNALMASAL
FEAGNIDEKYSKQAI SLIENILNKLQKNGVLYHQVLPNSTLKVTGLLEDYSFLIDALIK
NQYTQNKYLVHATTLLKSAIHKFYKDEIWLSDGFSQSASLQDASYKSAMAKMIQNVF
VISLLNGDTALHNLAIQNLEISSQIKNPGAYPEAIKIVLMDKKALILLKGKQERLHEL
RNIIKEIDVPMYILEVDENILQACSNNECFSSKDSNILLKEGIIKKYK
```

24 pav. Naršyklės langas paspaudus ant rodyklės, rodoma baltymo seka.

Pateiksiu pavyzdį, kaip atrodytų genomo naršyklė jeigu joje būtų tik trys baltymai. Visas genomo naršyklės funkcionalumas išliks, užvedus pelę taip pat bus rodoma informacija apie baltymą, o paspaudus rodoma pati baltymo seka(25 pav.).



25 pav. Genomo naršyklė iš bet kokių trijų *Sulfurimonas denitrificans* baltymų.

Šiuo pavyzdžiu parodoma, jog visiškai nesvarbu kiek baltymų sekų yra genome, ar tai būtų 3 sekos ar 3000 sekų, skriptas sugeneruoja *html* kodą, bet kokiam skaičiui baltymų, neprarandant genomo naršyklės funkcionalumo.

Panagrinėsiu, kaip atrodo sugeneruotas *html* kodas ir kokias funkcijas jis atlieka. Tai iš paprasčiausių elementų sudarytas kodas, kuris yra sugeneruojamas *Python* programavimo kalba parašyto skripto. Generavimo metu į *html* kodą yra sudedama visa reikalinga informacija susijusi su baltymo seka(26 pav.).

```

|<tr>
|<td width = 50px><a href = "testSarasas.html#genas25"> </td>
|<td width = 50px><a href = "testSarasas.html#genas26"> </a></td>

```

26 pav. Ištrauka iš genomo naršyklės html kodo.

Paanalizuosiu iš kokių elementų susideda *html* kodas. Pirmiausiai visas *html* kodas yra dedamas į lentelę, tam kad išlaikyti tvarkingą rodyklių pasiskirstymą. *<tr>* simboliai atskiria eilutes, kiekvienoje eilutėje yra dedamas tik tam tikras skaičius rodyklių, kad nepersidengtų viena rodyklė su kita. Kiekvienas baltymas prasideda simboliu *<td>* jis nurodo *html* lentelės vieną langelį. Toliau pačiame *<td>* nurodoma plotis tarp langelių, kad nebūtų, per dideli ar per maži tarpai. Nuoroda į kitą failą, kad paspaudus ant rodyklės būtų parodyta baltymo seka, tai atliekama pasinaudojant *href* elementu, jo viduje nurodant kito failo *pavadinimą* ir *id*. ** elementu yra nurodomas paveikslėlis, kuris bus dedamas į lentelės langelį, taip pat nustatomas paveikslėlio aukštis ir plotis, jie nustatomi apskaičiuojant pagal baltymo sekos ilgį. Taip pat elementas *title* tam, kad užvedus būtų rodoma informacija apie baltymą.

Šis *html* kodas yra sukurtas paprasčiausiomis priemonėmis, bet turintis reikalingą funkcionalumą. Genomo naršyklė puikiai vaizduojama *Google Chrome* naršyklės, kitų naršyklių gali būti iškraipomas vaizdas.

Išvados

Išanalizavus bakterijos *Sulfurimonas denitrificans* genomą ir įvykdžius anotacijas su trimis skirtingais bioinformatikos įrankiais: *MyRast*, *Prokka* ir *GeneMark*, taip pat atlikus visų įrankių išvesčių palyginimus galiu teigti, jog *MyRast* ir *Prokka* įrankiai anotacija atliko patikimai, greitai ir suteikdami naudingos informacijos t.y. buvo identifikuotas ir anotuotas *Sulfurimonas denitrificans* genomas. *GeneMark* įrankis dideliu sekų skirtumu išsiskyrė iš visų įrankių, todėl šio įrankio išvestis buvo nebenaudojama tolimesniam tyrimui. Taip pat, pasinaudojant *MyRast* anotacija, kuri identifikuoja ir baltymų fermentus, buvo surasti fermentų keliai ir panaudoti genomo naršyklės kūrime. Manau, didžiausią naudą vartotojui suteikia įrankis *MyRast*, suteikdamas aiškią anotaciją apie genomą ir taip pat leisdamas tą pačią anotaciją matyti vizualiai, bei naršyti po ją.

Taip pat, remiantis atliktomis genomo anotacijomis ir išanalizuota informacija apie genomą, buvo sukurtas *Python* skriptas, kuris generuoja *html* failą t.y. genomo naršyklę. Šis skriptas yra tinkamas, bet kokio genomo anotacijai atliktai su *MyRast* įrankiu. Sukurta genomo naršyklė vaizduoja baltymus, kaip rodykles, kurios yra išskirtos keliomis spalvomis, identifikuojamos skirtingas baltymų grupes. Užvedus pelę ant baltymo rodoma informacija apie jį: ID, funkcija, baltymo pradžia, pabaiga ir jo ilgis, bei fermentų keliai. Paspaudus ant baltymo yra rodoma baltymo seka. Genomo naršyklė naudingas įrankis analizuojant genomą, suteikia supratimo, aiškumo ir galimybę nagrinėti kiekvieną baltymą individualiai.

Todėl, apibendrinant gautus rezultatus, galima daryti tokias išvadas:

- *MyRast* įrankis yra paprasčiausias, patogiausias ir aiškiausias iš visų trijų naudotų įrankių;
- visi įrankiai identifikavo ir anotavo genomą, tačiau *Prokka* ir *MyRast* suteikia daugiau ir tikslesnės reikalingos informacijos apie tiriamą genomą;
- sukurti skriptai, kurie:
 - palygina anotacijų įrankių išvesties failus ir leidžia gauti nesutampančias sekas, kaip rezultatą;
 - randa fermentus ir parsiunčia informaciją apie juos, taip pat leidžia tą informaciją apdoroti;
 - generuoja paprastą *html* kodą, kuris vaizduoja tiriamo *Sulfurimonas denitrificans* genomo naršyklę.

Literatūra

- [1] Anders F Andersson, L. R. (2010). Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *The ISME Journal* .
- [2] Daniel PR Herlemann, M. L. (2011). Transitions in bacterial communities along the 2000km salinity gradient of the Baltic Sea. *The ISME Journal* .
- [3] Fukunaga Y, K. M. (2009). Phycisphaera mikurensis gen. nov., sp. nov., isolated from a marine alga, and proposal of Phycisphaeraceae fam. nov., Phycisphaerales ord. nov. and Phycisphaerae classis nov. in the phylum Planctomycetes. *The Journal of general and applied microbiology* .
- [4] Garrity, G. B. (2001). *Bergey's Manual of Systematic Bacteriology :Volume One : The Archaea and the Deeply Branching and Phototrophic Bacteria Editors:.*
- [5] Yuchen Han, M. P. (August 29, 2014). The Role of Hydrogen for Sulfurimonas denitrificans' Metabolism. *Plos* .
- [6] John Besemer, M. B. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses.
- [7] *Microbe Wiki*. (n.d.). Nuskaityta iš https://microbewiki.kenyon.edu/index.php/Sulfurimonas_denitrificans
- [8] Ramy K Aziz8, 9. D. (2008). The RAST Server: Rapid Annotations using Subsystems.
- [9] Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics Advance Access* .
- [10] Stefan Schouten, E. C. (2008). Intact Membrane Lipids of “Candidatus Nitrosopumilus maritimus,” a Cultivated Representative of the Cosmopolitan Mesophilic Group I Crenarchaeota.
- [11] Team, S. (n.d.). Genome Annotation by SEED Team.