VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

MATEMATINE S INFORMATIKOS KATEDRA

Eimantas Paspirgėlis

Bioinformatikos studijų programa

Matematines informatikos šaka

**Baltijos jūros  vandens mikroorganizmų genominių sekų tyrimas**

Bakalauro baigiamasis darbas

Vadovas: lektorius Irus Grinis

Vilnius 2015

Turinys

[1.Teorinė dalis 3](#_Toc420577745)

[1.1. Bakterijos rastos Baltijos jūroje. 3](#_Toc420577746)

[1.2. Baltijos jūroje rastos 2 bakterijos ir 3 archėjos 4](#_Toc420577747)

[1.3. Sulfurimonas denitrificans 5](#_Toc420577748)

[1.4. Python 6](#_Toc420577749)

[1.5. MyRast 7](#_Toc420577750)

[1.6. GeneMark 13](#_Toc420577751)

[1.7. Prokka 14](#_Toc420577752)

[2. Praktinė dalis 17](#_Toc420577753)

[2.1. MyRast 17](#_Toc420577754)

[2.2. Prokka 21](#_Toc420577755)

[2.3. GeneMark 24](#_Toc420577756)

[2.4. Baltymų sekų palyginimas 26](#_Toc420577757)

[2.5. Fermentų paieška 28](#_Toc420577758)

[2.6. Genomo naršyklė 30](#_Toc420577759)

[Literatūra: 35](#_Toc420577760)

# Įvadas

Baltijos jūroje buvo atlikti tyrimai, kurių metu buvo paimtas vienas lašas 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 paanalizuoti.

Bakalauro darbe bus tiriama ir analizuojama bakterijos Sulfurimonas denitrificans genomas trimis skirtingais bioinformatikiniais įrankiais, tai MyRast, Prokka ir GeneMark. Tai anotacijos įrankiai, su kuriais Sulfurimonas denitrificans genomas bus transliuojamas į baltymų sekas ir jos anotuojamos. Gavus visų trijų įrankių rezultatus jie bus lyginami tarpusavyje, kad surasti ir išrinkti tiksliausias baltymų sekas, kurios buvo ištransliuotos visų trijų bioinformatikinių įrankių. Taip pat, bus surasti baltymų fermentai ir jų keliai(angl. pathway). Surinkus visus duomenis apie bakterijos Sulfurimonas denitrificans ištransliuotas baltymų sekas bus parašytas skriptas Python programavimo kalba, kuris generuos html failą, kuris atvaizduos šios bakterijos genomo naršyklę.

Darbo problema -- surasti tiksliausias Sulfurimonas denitrificans baltymų sekas, naudojantis bioinformatikiniais įrankiais.

Darbo tikslas -- sukurti genomo naršyklę įprastinėm priemonėm naudojant skripto kalbas ir html.

Darbo tikslui pasiekti iškeliami uždaviniai:

* išsiaiškinti tris bioinformatikinius anotacijos įrankius: MyRast, Prokka ir GeneMark;
* anotuoti bakterijos Sulfurimonas denitrificans genomą į baltymų sekas naudojantis šiais įrankiais;
* parašyti skriptą, kuris palygintų tarpusavyje visų trijų įrankių gautus rezultatus;
* parašyti skriptą, kuris surastų baltymų fermentus ir jų kelius;
* parašyti skriptą, kuris pagal MyRast išvesties failą automatiškai generuotų html kodą, kuris atvaizduoja genomo naršyklę.

Tyrimo metu sukurti ir naudoti įrankiai:

* įrankiai -- bioinformatikiniai genomų anotacijos įrankiai: MyRast, Prokka, GeneMark;
* skriptai -- naudojant Python kalbą bus sukurti skriptai kurie: lygins baltymų sekas, suras fermentus ir jų kelius ir generuos html kodą, kuris vaizduos genomo naršyklę.

# 1.Teorinė dalis

## 1.1. Bakterijos rastos Baltijos jūroje.

Baltijos jūra yra antras pagal dydį sūraus vandens telkinys žemėje, tarnauja kaip drenažo plotas apie 90 milijonų žmonių 14 skirtingų šalių. Jos paviršiaus vanduo pasižymi druskingumo gradientu nuo ~2 iki 8 praktinių druskingumo vienetų(PSU) iš šiaurės į pietus ir vandens stulpelis Baltijos jūroje yra sluoksninis, nes turi stiprų vandens sluoksnį, kur tankio gradientas yra didžiausias per visą vandens telkinį. Pavyzdžiui, upėse yra mišrių zonų, šviežaus ir jūros vandens, kuriose yra trumpa išlikimo trukmė, todėl apriboja galimybes įsikurti autochtoninėms bendruomenėms. Priešingai yra centrinėje Baltijos jūroje, joje yra palyginti stabili aplinka, potvynio galimybė yra nežymi ir išlikimo laikas yra ilgas iki 45 metų ir tikima, kad joje dominuoja mikrobai kurie prisitaikę būtent šiai konkrečiai aplinkai. Nors mažas druskingumas Baltijos jūros paviršiuje atrodo palanki aplinka gėlavandenių bakterijų augimui, tipiškos jūrų bakterijų populiacijos gali suvaržyti reikalavimus aukštam druskingumui. Baltijos jūros paviršiniuose vandenyse atlikta nemažai tyrimų, kurie sąlygoja apie paviršiniuose vandenyse gyvenančius mikrobus, jų populiaciją, taip pat atlikti tyrimai[8] kuriuose tiriamas bakterijų pasiskirstymas iš paimtų 213 mėginių iš 60 stočių išilgai horizontaliai ir vertikaliai druskingumo gradientų Baltijos jūroje.

Šiame tyrime buvo norima ištirti mikrobus gyvenančius ne paviršutiniuose vandenyse, o giliau, sūresniame vandenyje gyvenančias bakterijas. Todėl iš Baltijos jūros maždaug 60 metrų gylio buvo paimtas mėginys tai yra lašas Baltijos jūros vandens. Atlikus tyrimus, jame buvo rasta įvairiausių bakterijų ir archėjų. 5 iš jų t.y. 2 bakterijos ir 3 archėjos buvo siunčiamos į Jungtines Amerikos Valstijas į "Bigelow Laboratory for Ocean Sciences " sekvenuoti. Sekvenuotų bakterijų ir archėjų genomai buvo parsiųsti atgal į Lietuvą ir čia tiriami, kadangi apie šiuos mikrobus yra mažai žinoma informacijos.

## 1.2. Baltijos jūroje rastos 2 bakterijos ir 3 archėjos

Baltijos jūroje paimtame mėginyje buvo rastos trys archėjos ir 2 bakterijos. Visos jos buvo sekvenuotos ir gauti šių bakterijų ir archėjų genomai. Šiame tyrime man reikėjo pasirinkti kurią bakteriją ar archėją analizuoti, todėl pirmiausiai reikėjo išsiaiškinti kokie tai organizmai:

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

Candidatus nitrosopumilus maritimus yra labai paplitusios archėjos, kurios gyvena jūros vandenyje. Tai pirmasis grupės narys 1a Crenarchaeota, kuris turėtų būti išskirtas grynakraujėje kultūroje. Genų sekos rodo, kad grupės 1a Crenarchaeota galima rasti daugelyje jūrų pakrančių aplink planetą. Jis yra vienas iš mažiausių organizmų esantis 0,2 mikronų skersmens. Jis gyvena oksiduodamas amoniaką į nitritus. Nitrosopumilus maritimus yra pajėgi oksiduoti amoniaką kai jo lygis yra mažesnis kaip 10 nanomolių. Taksonomijoje, Candidatus nitrosopumilus, yra iš Nitrosopumilaceae genties.

Halobaculum gomorrense[1] tai labai ekstremali archėja, ji izoliuota iš negyvosios jūros. Tai vienintelė Halobaculum genties rūšis kuri yra strypo/lazdelės formos.

Halobaculum genties archėja - iš sekvenuoto genomo nepavyko nustatyti tiksliau, kokia tiksliai tai archėja, pavyko nustatyti tik jos gentį.

Phycisphaera mikurensis priklauso phylum Planctomyces, jie pirmieji izoliuoti nekultūrinės planctomycete grupės. Phycisphaera mikurensis suteikia ižvalgų į gyvavimo ciklą Planctomyces.

Sulfurimonas denitrificans yra sulfurimono genties bakterija. Ši bakterija naudoja sulfidą arba tiosulfatą ir nitratą arba nitritus kaip elektrono donorą ir akceptorių. Šis mikroorganizmas buvo nustatytas hidroterminėse ventiliacijos srityse ir naftos telkiniuose, jis gali vaidinti svarbų vaidmenį sieros gyvavimą šiose aplinkose.

Išsiaiškinus kokios bakterijos ir archėjos yra rastos Baltijos jūroje, pasitarus su bakalaurinio darbo vadovu buvo nuspręsta analizuoti bakteriją - Sulfurimonas denitrificans.

## 1.3. Sulfurimonas denitrificans

Sulfurimonas yra bakterijų gentis iš klasės Epsilonproteobacteria. Šios grupės nariai auga naudojant nulinę valentę sierą, molekulinį vandenilį ar sumažintus sieros junginius, kaip elektronų donorus ir nitratus, nitritus ir deguonį kaip elektronų akceptorius. CO2 yra naudojamas, kaip vienintelis anglies šaltinis. Šios genties nariai yra keturi:

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

Sulfurimonas denitrificans buvo perkvalifikuotas iš Thiomicrospira denitrificans. Sulfurimonas denitrificans klasifikacija:

|  |  |
| --- | --- |
| Mokslinė klasifikacija | |
| Karalystė: | Bacteria |
| Skyrius: | Proteobacteria |
| Klasė: | Epsilonproteobacteria |
| Eilė: | Campylobacterales |
| Šeima: | Helicobacteraceae |
| Gentis: | Sulfurimonas |

Sulfurimonas denitrificans iš pradžių buvo izoliuotas nuo jūrų pakrančių nuosėdų. Jis gali augti kartu su tiosulfatu ir nitratu arba sulfidu ir deguonimi. Nesenai atlikti tyrimai[2] parodė, jog Sulfurimonas denitrificans gali augti gyvuodamas tik vandeniliu ir jis išauga greičiau, bei tankesnis, nei vien gyvuodamas tiosulfate, net kai vandenilyje ir visai nėra sieros junginių. Žemiau pateiktame paveikslėlyje pavaizduotas Sulfurimonas denitrificans.



Sulfurimonas denitrificans[3] genomas turi daug funkcijų, įskaitant didelius matmenis (2.2 Mbp), tai rodo didesnį metabolinį universalumą arba reagavimą į aplinką. Genai pateikia užbaigtą autotrofinį redukcinį Krebso ciklą. Daugelis atsparumo šeimos transporterių genų (10 iš viso) yra pateikiami, iš jų keletas koduoja sunkiųjų metalų emanacijos transporterius. Įmantrus arsenalas jutimo ir reguliavimo baltymų ir genų, kurie svarbūs, kad reaguoti ir užkirsti kelią oksidaciniam stresui.

## 1.4. Python

Python yra sukurta Guido van Rossumo [1990](http://lt.wikipedia.org/wiki/1990) metais. Pirmiausiai ji buvo scenarijų kalba AmoebaOS [operacinei sistemai](http://lt.wikipedia.org/wiki/Operacin%C4%97_sistema). Python dažniausiai lyginama su [Tcl](http://lt.wikipedia.org/w/index.php?title=Tcl&action=edit&redlink=1), [Perl](http://lt.wikipedia.org/wiki/Perl), [Scheme](http://lt.wikipedia.org/wiki/Scheme), [Java](http://lt.wikipedia.org/wiki/Java_(kalba)) ir [Ruby](http://lt.wikipedia.org/wiki/Ruby). Python kuriama kaip [atviro kodo](http://lt.wikipedia.org/wiki/Atviras_kodas) projektas.

Python yra daugiaparadigmė 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, tokiems kaip [Zope](http://lt.wikipedia.org/w/index.php?title=Zope&action=edit&redlink=1). Taip pat labai paplitusi Linux sistemose.

Python - galinga ir patogi programavimo kalba, sparčiai populiarėjanti pasaulyje. Bene akivaizdžiausias šios kalbos privalumas - milžiniškas 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 leidžia paprastai bei intuityviai užsiimti ir objektiniu programavimu.

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

## 1.5. 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 genomo, naudoja šią informaciją atkurti metabolinį tinklą ir sukuria lengvai parsisiunčiama galutinį failą. Servisas normaliai padaro prieinamą genomą per 12-24 val. nuo pateikimo, o esamas įgyvendinimas leidžia nuo 50 iki 100 genomų pralaidumą. Svarbu paminėti, jog greitis nėra svarbiausias aspektas, svarbiausia yra tikslumas, išsamumas ir suderinamumas.

Rast tikslas yra pasiekti tikslumą, nuoseklumą ir detalumą. Rast serveris įgyvendina automatinį gaminimą dviejų klasių genų funkcijų:

* remiantis posistemių pagrindu yra remiamasi atpažinimo funkcijomis posistemėse;
* remiantis ne posistemių pagrindu yra užpildoma daugiau bendrais metodais grindžiamais integracija.

Faktas, kad Rast išskiria šias dvi anotacijos klases ir naudoja patikimas posistemes kaip pagrindą išsamiai medžiagų metaboliniai rekonstrukcijai daro Rast anotacijas, kaip išskirtinai gerą atspirties tašką tolimesnei anotacijai. Be to, gaminant pradines užduotis genų funkcijoms ir metaboliniai rekonstrukcijai, Rast serveris suteikia aplinką naršymui po anotuotą genomą ir lyginimą su šimtais kitų genomų per SEED integraciją. Rast turi galimybę rodyti genomo 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 publikaciją apie jų detalumą yra rengiama. Kiekvienas FIGfam yra manoma yra sudarytas iš 3 dalių: rinkinio baltymų, šeimos funkcijos ir sprendimų priėmimo procedūros. 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% iš kiekvienos sekos;
* jei abu yra iš glaudžiai susijusių genomų ir jų panašumas yra didelis.

Tai yra du atvejai, kai mes ramūs teigdami, kad baltymai turi bendras funkcijas: pirmasis atspindi ekspertų tvirtinimus, o antrasis, kai nukrypimas yra minimalus. FIGfams konstrukcijai naudoja šiuos du principus, kurie lėmė kolekciją 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 yra baltymai grupuojami pagal antrąjį principą(t.y. ne posistemės pagrindu FIGfams). Daugelyje ne posistemės FIGfams yra tik 2,3 arba 4 baltymai. Laikui bėgant yra planuojama sutraukti ne posistemės FIGgams. Tai bus daroma kuriant naujus, rankiniu būdu kuruojamus posistemius. Tikėtina, kad didelė dalis naujai sekvenuojamų genomų bus arti esamų genomų ir FIGfams jau yra veiksminga atpažinimo sistema tokiais atvejais.

Pagrindiniai žingsniai anotuojant genomą naudojant RAST:

* kviečiami tRNR ir rRNR genai;
* kviečiama baltymus koduojančių genų ekspresija;
* sukuriamas filogenetinis kontekstas;
* tikslinė paieška remiantis FIGfams glaudžiai susijusiuose genomuose;
* dar kartą kviečiama baltymus koduojančių 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 esamos priemonės kitų mokslinių tyrimų grupių, kad pirmiausia nustatyti rRNR ir tRNR koduojančių genų ekspresiją. tRNR genams naudojama tRNAscan-SE ir nustatyti rRNR koduojančius genus naudojamas įrankis " search\_for\_rnas " kuris sukurtas Niels Larsen. Pradedamas procesas kviečiant tuos genus, kuriuos manoma, kad galima patikimai nustatyti. Tada serveris nesaugo baltymų kodavimo genų, kurie reikšmingai sutampa su nors vienu iš šių regionų. Šių genų kvietimas yra beveik neabejotini artefaktai ir RAST bando išvengti daugindamas šias klaidas.

Kviečiama baltymus koduojančių genų ekspresija

Kai tRNR ir rRNR genų kodavimo regionai yra pašalinami, atliekamas pirminis kvietimas naudojant GLIMMER2. Šiuo metu ieškoma pagrįsto įvertinimo tikėtinų genų ir GLIMMER2 yra puiki priemonė šiam tikslui pasiekti. Šiame etape aktualu baltymų koduojančių genų atstovavimas pirminio įverčio tariamuose genuose.

Sukuriamas filogenetinis kontekstas

Kai pradinis rinkinys baltymų kodavimo genų buvo nustatytas , imamos tipinės sekos iš mažo rinkinio FIGfams, kurie turi savybę, kad jie yra universalūs arba beveik universalūs prokariotuose. Šis rinkinys apima, pavyzdžiui, tRNR ligazę. Naudojant šį mažą rinkinį atstovų ieškoma baltymus koduojančių genų iš naujo genomo šiuose FIGfams. Reikia paminėti, kad šis žingsnis yra labai greitas, nes ieškomas tik naujasis genomas ir ieškoma 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, sureguliuojama pradinė padėtis ir perkeliamas iš tariamų genų aibės į nustatytų genų aibę ir funkcija skiriama 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 rinkinys FIGfams kurie gali būti rasti naujame genome ir tikimasi, kad ši paieška turi didelę sėkmės galimybę. Kai randamas genas, pakoreguojama jo pradinė padėtis ir jis perkeliamas iš tariamų į nustatytų genų aibę. Skaičiavimo kaina, kad rasti šiuos genus yra maža.

Dar kartą kviečiama baltymus koduojančių genų ekspresija

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

Apdorojami likusieji genai dar kartą su visa FIGfams kolekcija

Tariami genai, kurie lieka, gali būti ieškomi dar kartą visoje FIGfams kolekcijoje. Šiuo metu kolekcija atstovaujamų baltymų sekų iš FIGfams naudojami apskaičiuoti potencialius, aktualius FIGfams 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 išlieka, yra tvarkomi bandant išspręsti problemas susijusias su sutampančių genų kvietimu, jie yra koreguojami ir taip toliau. Atsižvelgiant į RAST serverį yra naikinami likusieji tariami genai prieš reikalingus baltymus duomenų bazėje bandant nustatyti ar jie galėtų būti naudojami sprendžiant konfliktus.

Apdorojami likusieji neanotuoti baltymų koduotės genai

Šiame žingsnyje yra paskutiniai funkcijų priskyrimai tariamiems genams. Jei panašumas buvo apskaičiuotas ankstesniame žingsnyje, šie panašumai gali būti prieinami ir tvirtinamos funkcijos. Galima įdarbinti bet kurį paprastai dirbančių technologijų, kad paleisti rinkinį įrankių ir gaminti tikslesnį įvertinimą. Apdoroti genai naudojant šį metodą sudaro didžiają dalį RAST anotacijos. Pirminiame perdirbime dauguma genų naudoja FIGfams technologiją ir orientuotas paieškas, ši kaina yra sumažinama iki minimumo naudojant RAST, nemažinant tikslumo.

Konstruojama pradinė medžiagų apykaitos rekonstrukcija

Kai paskyrimas funkcijų buvo atliktas, tada pradinė metabolinė rekonstrukcija yra suformuojama. Tikslas yra sujungti genus į naujo genomo funkcinių vaidmenų posistemes, nustatant kada rinkinys jungčių specifiniai posistemei yra pakankama parama aktyviems variantams posistemės ir užbaigti rinkinį aktyvių variantų. Kai pačios posistemės yra įrengtos neapdirbtose kategorijose, kurios atspindi pagrindinių padalinių funkcijas, galima pagaminti išsamią sąmatą genomo turinio, kad sėkmingai prijungta prie posistemės[žiūrėti pav 1].

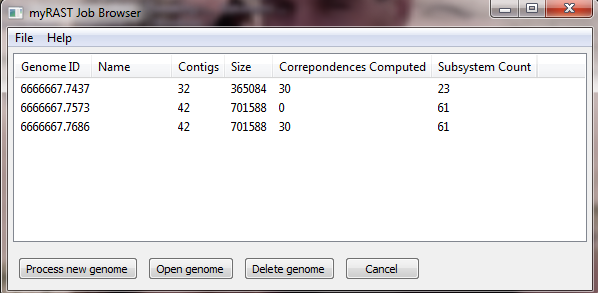


1 pav. Vaizduoja prijungtus genus prie posistemių ir jų pasiskirstymą įvairiose kategorijose.

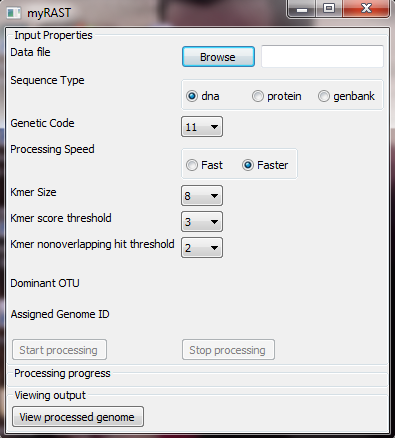
Reikėtų pabrėžti, kad posistemės apima visus modulius ląstelių mašinų, ne tik medžiagų apykaitos kelius. Vadinasi tai kas yra vadinama metaboline rekonstrukcija, yra labiau suvokiama kaip grupavimas genų į modulius.

MyRast nauda

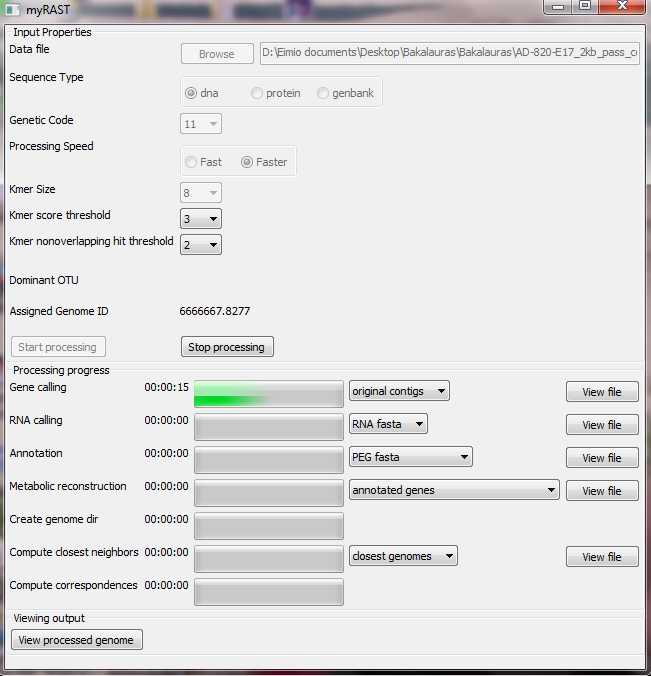
Ankstesniame skyrelyje aprašyta RAST pagrindinė technologija ir pagrindžiamas veikimas. Įdėta daug pastangų, kad sukurti paprastą sąsają, kuri siūlo pateikti genomus, stebėti anotacijos progresą, peržiūrėti rezultatus, palyginti juos su šimtais kitų genomų ir parsisiųsti rezultatus bent keliais formatais.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

1. Kai paleidžiamas myRast matomas toks vaizdas. 

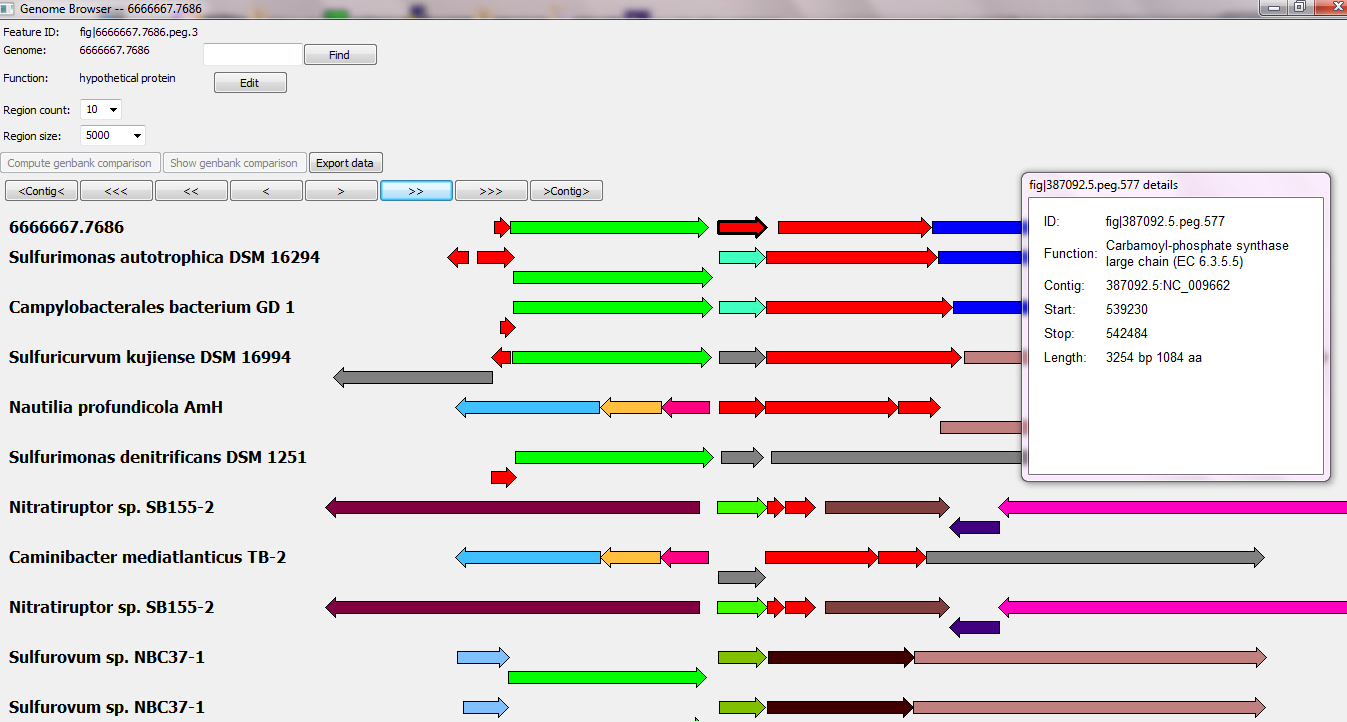
2. Pasirinkus "Process new genome" pasiūloma pasirinkti failą, kuris turi būti anotuojamas.



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



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

Norint peržiūrėti genomą spaudžiama "View processed genome". 

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 ir koks ilgis. Spaudžiant rodykles galima pereiti į vieną ar kitą pusę, per vieną geną, pusę ekrano, visą ekraną arba pereiti iš vieno contigs į kitą. Yra trys pagrindinės galimybės: keisti geno anotaciją, ištrinti arba įterpti naują geną.

## 1.6. GeneMark

Genų identifikavimo uždavinys dažnai sprendžiamas tyrėjų, kurie dažnai 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 skirtas genų prognozavimui prokariotuose, eukariotuose ir virusų sekomis. Šiuo metu serveris leidžia analizuoti apie 200 prokariotų ir > 10 eukariotų genomų naudojant kiekvienai rūšiai būdingas programinės įrangos versijas. GeneMark svetainė dažnai atnaujinama, pateikiamos naujausios programinės įrangos versijos ir genų modeliai.

Tiek 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 euristinį modelį apskaičiuotą iš nukleotidų dažnio iš sekų įvesties, kurios mažiausias ilgis gali būti 400nt . GeneMarkS gali naudoti ilgesnes 1Mb sekas. Kaip daugelis programų GeneMark svetainė naudoja panašių sąsajų programas. GeneMark.hmm interneto sąsaja priima kaip įvestį vieną DNR seką iš įkelto failo arba įklijuotą tekstą į laukelį. Jei FASTA failo aprašymas prasideda simboliu ">" tai ta 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 kurį iš dviejų Markovo modelių naudoja GeneMark.hmm, tipinį ar netipinį modelį, su didesne tikimybe genų sekoms. Galimybė generuoti GeneMark prognozes kartu su GeneMark.hmm analize suteikia papildomos svarbios informacijos. Šiuo atveju GeneMark yra nustatytas naudoti tuos pačius mokymo duomenis kaip ir paleidžiamus GeneMark.hmm. Verta paminėti, kad GeneMark.hmm ir GeneMark papildo vienas kitą. Grafinės analizės išvestis yra prieinama PDF arba PostScript formatu. Šios išvesties fragmentas, iliustruojantis tiek GeneMark.hmm tiek GeneMark prognozes pavaizduota 1 paveiksle.



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

Analizė prokariotų DNR sekų, kurios nėra iš anksto apskaičiuotos specifiniam modeliui galima naudoti programos versiją kuri skirta sekoms didesnėms už 400nt. Įrodyta, kad šis metodas naudingas analizei nehomogeninių genomų. Jei modeliai turi būti apskaičiuoti nežinomoms sekoms 1Mb ar ilgesnio ilgio, gali būti naudojama GeneMarkS programa. Ši programa turi daugiau skaičiavimo išteklių, taigi jos išvestis yra teikiama elektroniniu paštu.

Ateities kryptys GeneMark programinės įrangos interneto svetainės apima kelių genomo elementų kurie nėra prognozuojami GeneMark.hmm arba GeneMark, kaip rRNR arba tRNR ir gerina aptikimą genų 50 galuose. 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.

## 1.7. Prokka

Genomo anotacija yra procesas identifikuoti ir ženklinti visus svarbius genomo sekos procesus. Mažiausiai tai turėtų apimti koordinates prognozuojamų 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 turtingą ir patikimą bakterijų genomo sekų anotaciją. Jei įmanoma Prokka panaudos kelis procesoriaus branduolius ir tipiškas bakterijos genomas gali būti anotuotas, ant keturių branduolių kompiuterio, per maždaug 10 minučių. Tai puikiai tinka kartotiniai sekų modelių analizei.

Įvedimas

Prokka tikisi gauti surinktas genomo DNR sekas FASTA formatu. Sekos be tarpų būtų ideali įvestis, tačiau tikimasi, kad tipiškas sekų rinkinys bus pagamintas de novo programinės įrangos. Šis sekos failas yra vienintelis privalomas parametras programinei įrangai.

Anotacija

Prokka remiasi išoriniais prognozavimo įrankiais identifikuoti genomo funkcijų koordinates per contigs. Šios priemonės išvardytos pirmoje lentelėje ir visi jie išskyrus Prodigal teikia koordinates ir aprašo atitinkamas funkcijas. Baltymų kodavimo genai yra anotuojami dviem etapais. Prodigal identifikuoja kandidatų genų koordinates, bet neaprašo tariamų genų. Tradicinis būdas prognozuoti ką genas koduoja, jį palyginti su didelės duomenų bazės žinomomis sekomis, paprastai baltymo sekos lygį ir perduoti anotaciją labiausiai atitinkančiam. Prokka naudoja šį metodą, tik hierarchine tvarka, pradedant nuo mažų patikimų duomenų bazių, po to pereinant prie vidutinių ir galiausiai kuruoti baltymų šeimų modelius.

|  |  |
| --- | --- |
| **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 |

1 lentelė. Prokka naudojami įrankiai.

Išvedimas

Prokka gamina dešimt failų į nurodytą išvesties aplanką. Šie failai aprašyti 2 lentelėje.

|  |  |
| --- | --- |
| Galūnė | Aprašymas failo turinio |
| .fna | FASTA failas su originaliu 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 kuriame sekos ir komentarai |
| .gff | GFF v3 failas kuriame sekos ir komentarai |
| .log | Prisijungimo failas Prokka apdorojant išvedimą |
| .txt | Santraukos anotacijos statistika |

2 lentelė. Aprašymas Prokka išvedimo failų

Prokka buvo sukurtas siekiant būti tiksliu ir greitu įrankiu. Pavyzdžiui anotuojant Escherichia coli genomą su tipišku keturių branduolių kompiuteriu užtruko apie 6 minutes.

# 2. Praktinė dalis

Pirmiausiai praktinėje dalyje buvo nuspręsta kurį bakterijos ar archėjos genomą analizuoti. Paanalizavus visus genomus ir išsiaiškinus kokie tai organizmai buvo pasirinktas bakterijos Sulfurimonas denitrificans genomas.

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

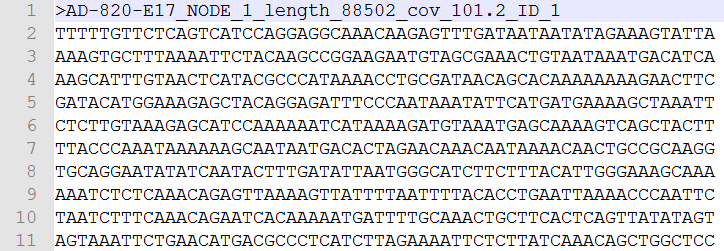
* vienos iš Baltijos jūroje paimto mėginio bakterijos sekvenuoto genomo analizavimas bioinformatikiniais įrankiais:
  + MyRast;
  + Prokka;
  + GeneMark.
* įrankių išvesties patikrinimas Blast(Basic Local Alignment Search Tool);
* bioinformatikiniais įrankiais anotuotų genomų kryžminis palyginimas tarpusavyje;
* parašytas skriptas baltymų fermentams rasti;
* sukurtas python skriptas skirtas generuoti genomo naršyklę, t.y. html failą.

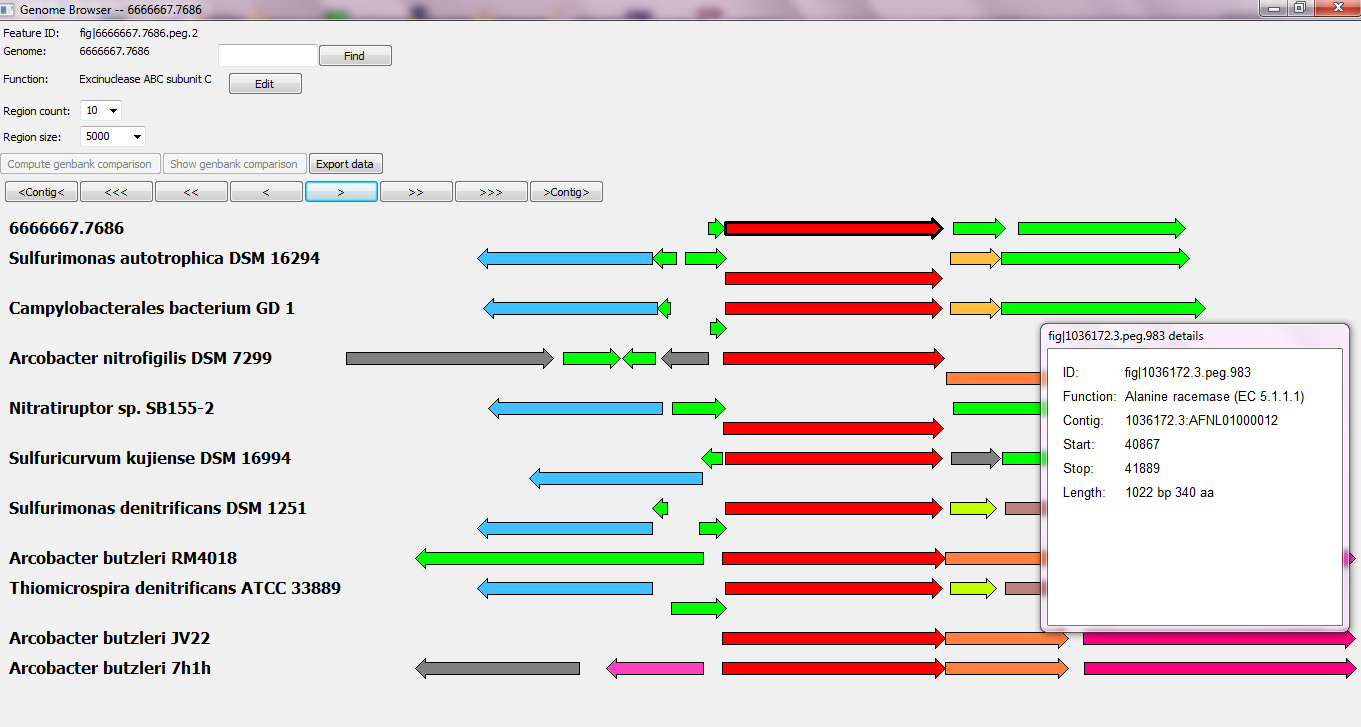
Toliau aprašysiu šiuos praktinėje dalyje atliktus uždavinius, kokie buvo gauti rezultatai ir kas buvo sukurta.

## 2.1. MyRast

MyRast tai yra bioinformatikinis įrankis skirtas anotuoti prokariotų genomus. Šis įrankis anotuoja ir leidžia parsisiųsti tiek išeities failus, tiek 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. Taip atrodo iškarpa DNR iš genomo, kuris tik sekvenuotas, bet dar neanalizuotas:

1 pav. Iškarpa iš Sulfurimono sekvenuoto genomo.

Turint sekvenuotą bakterijos DNR FASTA formato failą jau galima naudotis bioinformatikiniais įrankiais tokiais kaip MyRast. Naudojantis MyRast galima anotuoti trijų rūšių sekų tipus: DNR, baltymų ir genbank. Šiuo atveju mums reikalingas DNR sekų tipas. Pasirinkus 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 peržiūrėti galima paspaudus "Open genome" iš karto yra atidaromas vizualus pavaizdavimas Sulfurimono anotuoto genomo.  2pav. MyRast vizualus genomo pavaizdavimas.

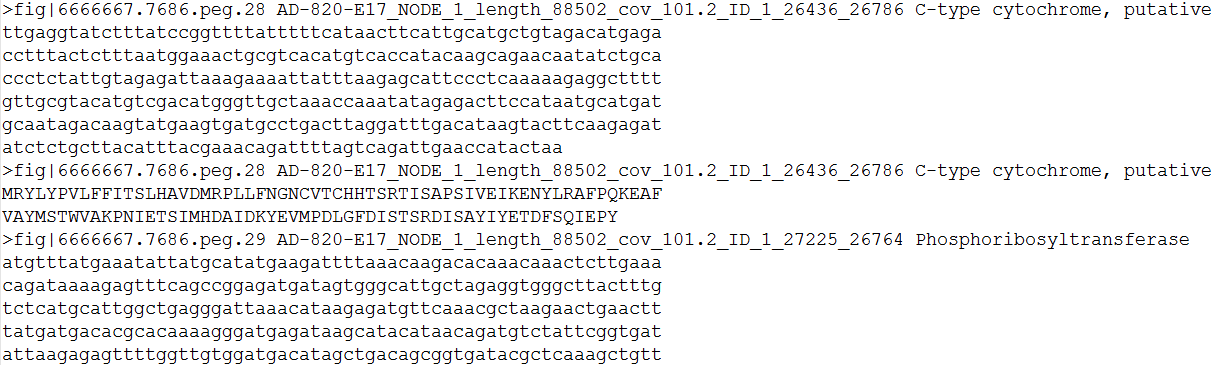
Vizualiame genomo anotacijos pavaizdavime galima atlikti įvairias funkcijas, nagrinėti ir išsiaiškinti kiekvieną geną individualiai. Šiame atvaizdavime matome daug įvairių spalvų rodyklių, kiekviena iš jų vaizduoja skirtingą geną, jo kryptis yra į kairę jei baltymo dydis eina nuo didesnės į mažesnę pusę pavyzdžiui jei genas prasideda 1863 baze ir baigiasi 1342 baze ir atvirkščiai jei į dešinę pusę. Rodyklių spalvos skiriasi, kiekviena spalva žymi genų funkcines grupes: šviesiai žalia spalva žymi jog tai fermentų grupė, mėlyna spalva žymimas nežinomas(angl. hypothetical) baltymas ir t.t. Taigi, jei domina kažkokia konkreti geno funkcinė grupė, jas galima surasti tiesiog 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 dešinį pelės mygtuką, galima sužinoti būtent to geno DNR ar jau išverstą jo kodą, taip pat yra parinktis, kuri leidžia patikrinti konkretų geną NCBI Blast, jei kyla abejonė, 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 parodomas genomo dalis kurioje ir yra 600 genas, tai labai patogu ir praverčia analizuojant genomą, jei žinomas koks konkretus genas yra reikalingas. Taip pat galima pasirinkti atvaizduojamą regionų skaičių ir jų dydį, patogu jei reikia vienu metu dirbti su vienas nuo kito nutolusiais genais.

Svarbiausia MyRast funkcija mano analizavimui t.y. duomenų eksportavimas. MyRast leidžia eksportuoti duomenis trimis formatais:

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

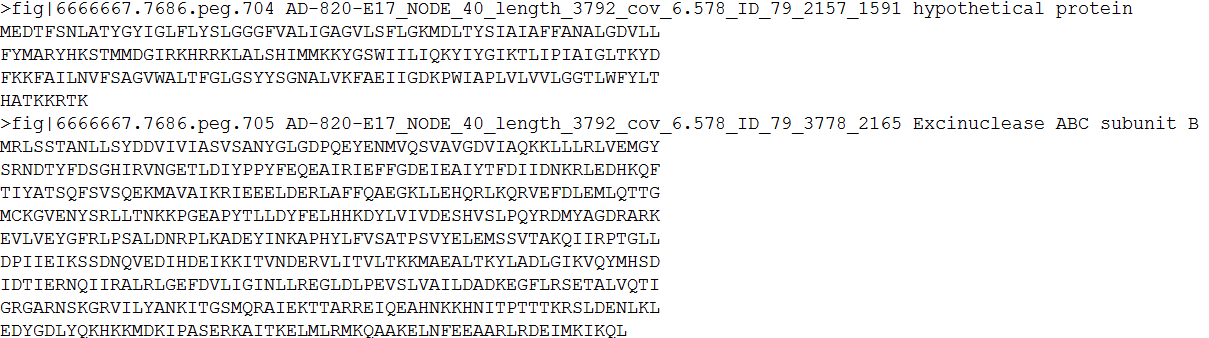
Taip pat galima pasirinkti funkcijų tipus t.y. galimi variantai yra peg ir rna. Peg tai yra baltymų numeravimas, o rna tai RNR funkcijų rodymas. Mūsų atveju abu šie funkcijų tipai mums reikalingi, todėl pažymimi abudu. Sekanti galimybė yra pasirinkti kokios sekos turi būti įdėtos į eksportuojamą failą,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

Kaip matyti paveikslėlyje čia pavaizduota 28 seka, 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.



3 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 ir jis pavaizduotas 4 paveikslėlyje.



4 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 įterpiant varnelės ženklą, panagrinėsiu kas toje eilutėje yra nurodoma, pavyzdžiui paimsime anotacijos eilutę iš 4 paveikslėlio:

* fig|66666667.7686 - 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 kurio 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 perimama 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" scripto pagalba išsikerpame iš pagrindinio FASTA failo ir ieškome NCBI Blast protein pasinaudojant Blast skriptu kuris leidžia vienu metu nurodyti daug ieškomų sekų ir jas parsiunčia html formatu. Pastebėta, jog tos baltymų sekos, kurios MyRast pažymėtos kaip nežinomos 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 "hipothetical 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 anotuotose 713 sekose pasiskirstė nežinomi baltymai, surasti baltymai ir fermentai. Pavaizduota 5 diagramoje:

5 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 sąlygoja, jos DNR sekų atpažinimą anotavimo sistemose.

## 2.2. Prokka

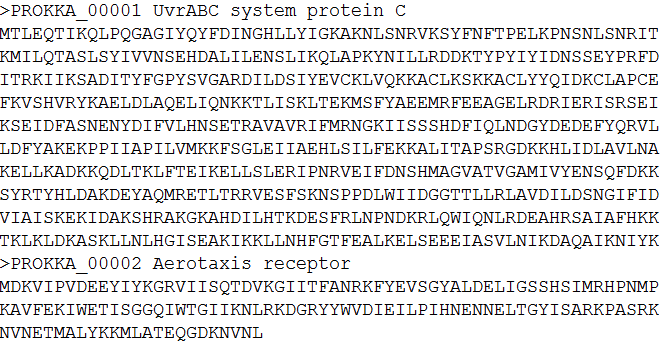
Viso genomo anotacija tai procesas identifikuoti funkcijas ar savybes genomo DNR sekų ir žymėti juos naudinga informacija. Tam yra skirta Prokka - programinės įrangos įrankis greitam anotavimui bakterijų, archėjų ir virusų genomų, kuris sukuria standartus atitinkančius išvesties failus. Šis įrankis yra valdomas naudojantis komandine eilute, o tai nėra suprantama kiekvienam žmogui. Taip pat šis įrankis yra pakankamai sunkiai įrašomas, reikia nemažai priedų, kad būtų galima įsirašyti ir 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 bioinformatikiniu įrankiu taip pat anotavome savo Sulfurimonas denitrificans bakteriją. Pirmiausiai komandinėje eilutėje reikėjo parašyti komandą, kuri įvykdytų bakterijos anotaciją, rašiau paprastą, bet aiškią ir pakankamą komandą :

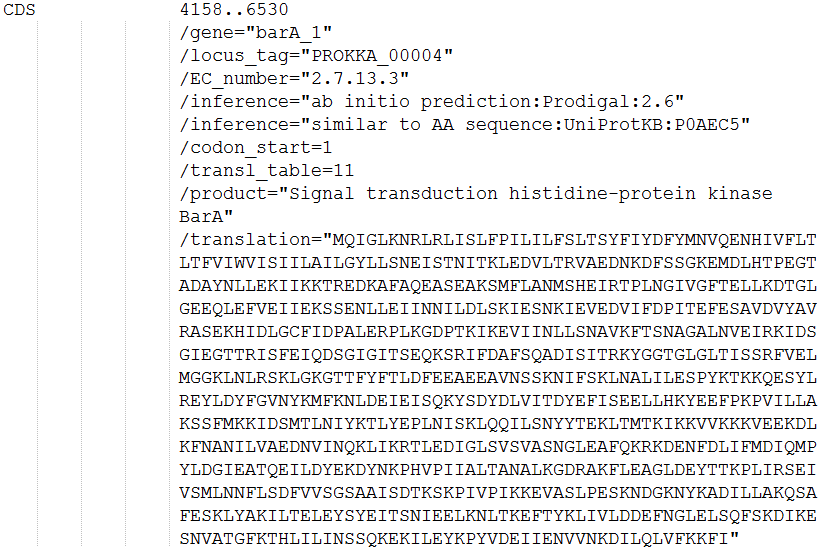
**prokka --outdir Prokka --prefix Sulfurimonas AD-820-E17.fa**

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 ir kuris norėčiau, kad būtų anotuojamas **AD-820-E17.fa.** Šis įrankis genomo anotaciją įvykdė tikrai greitai, Sulfurimono genomą anotavo per maždaug 10 minučių. FASTA anotacijos failas atrodė štai taip:



6 pav. FASTA formato failas generuotas su Prokka.

FASTA faile yra nurodomas tik genų eiliškumas ir surastas baltymas, truputį netoks kokio tikėjausi, 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 man, 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 yra kiekvienai sekai aprašoma jos informacija. 7 paveikslėlyje pavaizduota ištrauka iš .gbk formato failo:



7 pav. Ištrauka iš anotuoto Genbank formato failo.

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

Su Prokka įrankio pagalba buvo anotuotas Sulfurimonas genomas, rasta 705 sekos t.y. 8 mažiau nei su MyRast įrankiu. Dabar pavaizduosime, kaip šiam įrankiui pavyko anotuoti baltymus ir fermentus. 8 diagramoje matomas nežinomų baltymų(angl. hypothetical protein), surastų fermentų ir surastų baltymų pasiskirstymas.

8 diagrama. Prokka įrankiu anotuoto genomo baltymų pasiskirstymas.

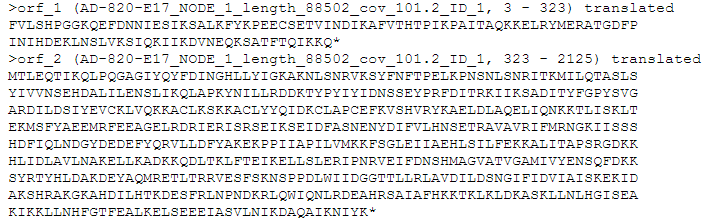
Kaip matome Prokka įrankis identifikavo daugiau baltymų nei MyRast, mažiau nei trečdalis, liko neidentifikuoti. Visus nežinomus baltymus taip pat patikrinus Blast, buvo apie 90% baltymų sekų kuriose pirmoje vietoje buvo rodomas "hypothetical protein". Prokka įrankis man patiko dėl greito anotavimo ir 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 genus prokariotų, eukariotų ir virusų genomų. GeneMark svetainė yra dažnai atnaujinama pateikiamos naujausios programinės įrangos versijos ir genų modeliai. Programos išvesties failai atsiunčiami į elektroninį paštą. Išbandžiau šias GeneMark šeimos programas:

* GeneMark;
* GeneMark.hmm;
* GeneMarkS.

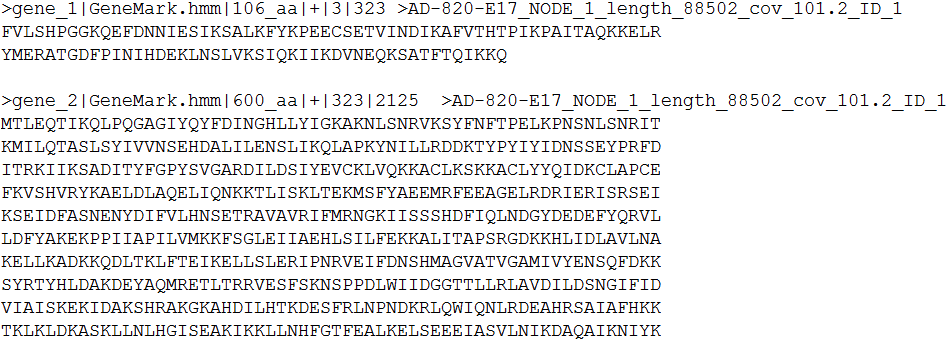
Naudojimasis GeneMark svetaine yra labai paprastas, pirmiausiai reikia pasirinkti programą kuria bus naudojamasi, aš išbandžiau tris jų programas:

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 man reikalingą Sulfurimonas denitrificans. Sekančiame lange keli pasirinkimai susiję su išvesties failais, galimi pasirinkimai baltymų ir genų nukleotidų failai, pasirenku abudu. Taip pat yra leidžiama pasirinkti kodavimo potencialų grafiką, tačiau jis neleidžiamas multi-FASTA failams, taigi jis man netinka, nes pas mane 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, taip atrodo baltymų sekų FASTA failas:

9 pav. GeneMark programos išvesties FASTA failas.

Išvesties failas atrodo neinformatyviai, yra tik išverčiamos sekos į baltymų sekas. Eilutėje kuri apibūdina baltymo seką yra tik perkopijuojama informacija iš pradinio failo, pažymima sekos pradžia ir pabaiga ir gale parašoma, 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 formata 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 identiškai: 10 paveikslėlyje pavaizduota išverstos baltymų sekos:



10 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 nesiskiria niekuo, tačiau palyginus juos su GeneMark programos išvesties failu atsiranda vienas didelis skirtumas. Skiriasi rastas sekų skaičius ir pakankamai nemažai:

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

Tikslaus paaiškinimo neradau, kodėl rasti skirtingi kiekiai 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 bioinformatikiniais įrankiais ir gavus 3 FASTA formato failus su išverstomis baltymų sekomis reikia palyginti juos tarpusavyje. Šis palyginimas reikalingas, nes atlikus anotacijas su trimis skirtingais įrankiais gauti rezultatai skiriasi. Kiekvieno įrankio rastos baltymų sekos matomos 11 diagramoje:

11 diagrama. Išverstų baltymų sekų skaičius su kiekvienu įrankiu.

Rezultatai skiriasi nežymiai, tačiau skiriasi. Vadinasi, tas pats genomas su tomis pačiomis DNR sekomis yra išverčiamas 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 vienos sekos ieškant kitame faile ir taip ieškoti visas sekas, 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;
* sukarpomas 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 jie yra 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 sukarpomi pagal anotacijos eilutes, kad liktų tiktai 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 sukarpomos po 20 simbolių, šis veiksmas atliekamas tam, kad išvengti vienos raidės nesutapimo. Vienos raidės nesutapimas yra jei paimsime vieną baltymo seką ir ieškosime kitame faile ar tokia yra ir būtų surasta tokia pati, tik skirtųsi vienas simbolis, 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 skaitome baltymą kaip sutampantį. Jei baltymo seka nerandama tuomet ji spausdinama kaip nesutampanti, taip pat galima spausdinti ir sutampančias sekas.

Atlikus palyginimą su visais trimis FASTA failais, gauti skirtingi rezultatai. Rezultatai matomi 12 diagramoje:

12 diagrama. Nesutampančių baltymų sekų kiekis tarp failų.

Kaip matome diagramoje, rezultatai labai stipriai išsiskyrė lyginant su GeneMark programa, tiek lyginant su MyRast 163 skirtingos sekos, tiek lyginant 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.

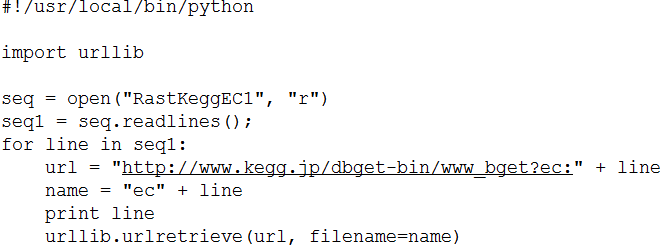
Iš to išplaukia išvada, kad GeneMark programa nepatikimai transliuoja DNR sekas į baltymų sekas, bent jau šiam konkrečiam Sulfurimono genomui, kadangi gana didelis neatitikimas tarp sekų ir beveik toks pat neatitikimų skaičius lyginant GeneMark su MyRast ir Prokka. Pastarųjų dviejų įrankių lyginimas manau davė realius rezultatus, 30 skirtingų baltymų sekų nėra didelis skaičius atsižvelgiant į tai, jog ir pats išverstų sekų skaičius skiriasi 8 sekom ir skiriasi pačių įrankių metodai transliuojant 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 kitų organizmų baltymai. Tolimesniame tyrime remsiuosi tik tomis baltymų sekomis, kurios buvo rastos vienodos lyginant MyRast ir Prokka įrankių failų išvestis, GeneMark baltymų sekos davė nepatikimus rezultatus, todėl siekiant kuo tikslesnių rezultatų šio įrankio baltymų sekomis nesiremsiu tolimesniame tyrime.

## 2.5. Fermentų paieška

Fermentas – [baltyminis](http://lt.wikipedia.org/wiki/Baltymas) [katalizatorius](http://lt.wikipedia.org/wiki/Katalizatorius), paspartinantis [organizme](http://lt.wikipedia.org/wiki/Organizmas) vykstančias chemines [reakcijas](http://lt.wikipedia.org/wiki/Reakcija) 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 ištirta daugiau kaip 3000 fermentų.

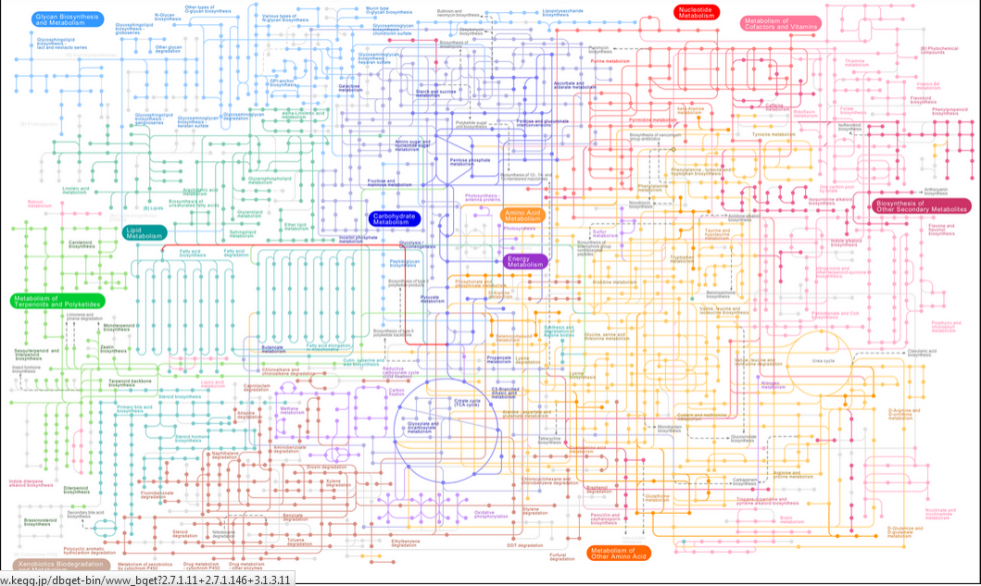
Fermentai yra svarbios medžiagos organizmui, todėl jas taip pat svarbu surasti ir pažymėti prie kiekvienos baltymų sekos to fermento kelius(angl. pathway) kuriant genomo naršyklę. Kurie baltymai turi fermentus atskirai ieškoti nereikės, tai jau atliko MyRast įrankis anotuodamas genomą, prie kiekvienos sekos kuri turi fermentus skliausteliuose pažymėta pavyzdžiui (EC 3.4.21.53) kuris nurodo tam tikro fermento pavadinimą. 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. Sekančiame žingsnyje kitas nedidelis skriptas iš tų eilučių paima jau tik pačių fermentų pavadinimus ir taip gaunamas sąrašas visų faile esančių fermentų.

Toliau, reikia parsiųsti 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 parašiau nedidelį skriptą, kuris automatiškai parsiunčia visus html failus apie kiekvieną nurodytą fermentą. 13 paveikslėlyje pavaizduotas skriptas.



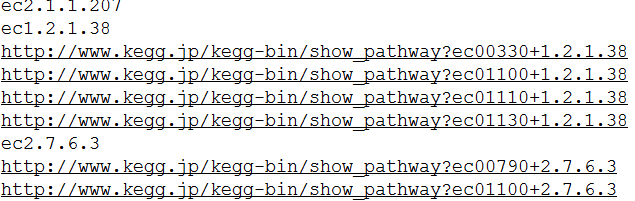
13 pav. Skriptas skirtas parsiųsti html failus su informacija apie fermentus.

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(14 pav).



14 pav. Vaizduoja fermento ec6.4.1.2 metabolinį kelią.

Iš viso parsiųsta 212 html failų iš kurių reikia ištraukti adresus kurie parodo konkretaus fermento kelius. Tam taip pat buvo parašytas skriptas kuris nusiskaito visų failų pavadinimus į sąrašą, poto 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 (15pav).



15 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

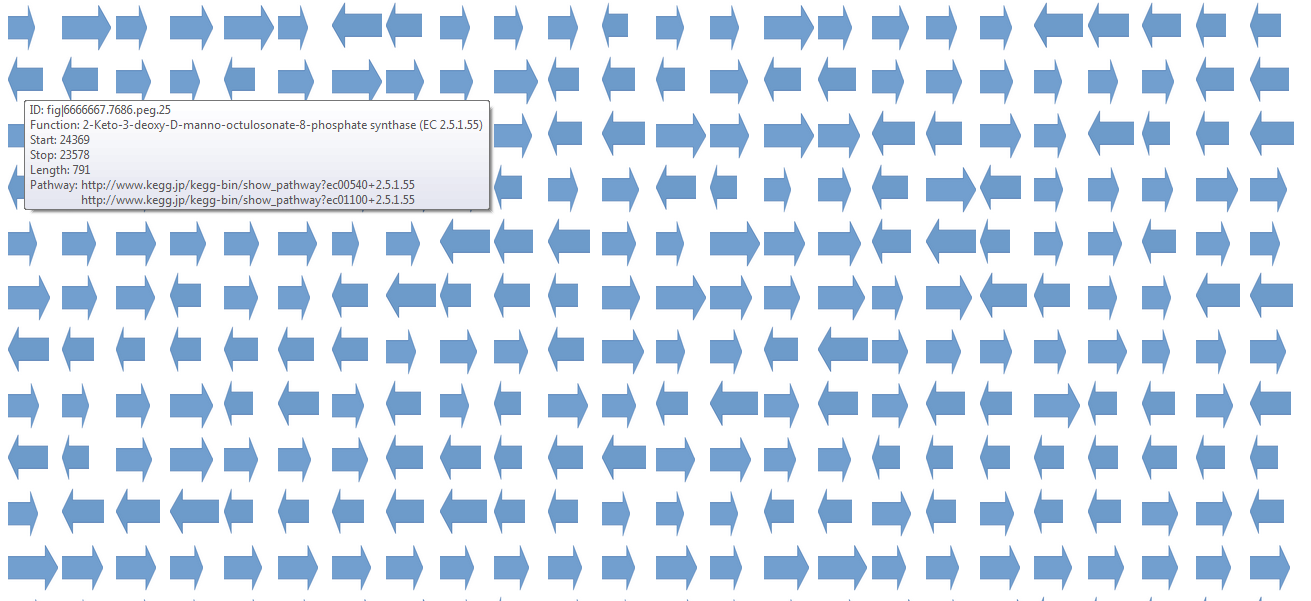
Bioinformatikoje genomo naršyklė tai grafinė sąsaja pavaizduoti informaciją iš biologinės duomenų bazės genomo duomenims. Genomo naršyklė leidžia tyrėjams vizualizuoti ir naršyti pasirinktą genomą su anotuotais duomenimis, genų prognozavimu, baltymais ir t.t. Anotuoti duomenys dažniausiai yra iš kelių 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 fermentų surastus kelius. Iš viso palygintame faile yra 683 sutampančios baltymų sekos, todėl sukurti genomo naršyklę su anotuojama informacija tokiam kiekiui sekų daug laiko reikalaujantis darbas, todėl 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 transliuotiems 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 dydis rodyklės, kiekvienam baltymui žymėti;
* generuojamas paprastas html failas su reikalinga informacija;
* tuo pačiu generuojamas atskiras html failas su baltymų sekomis ir jų anotacija.

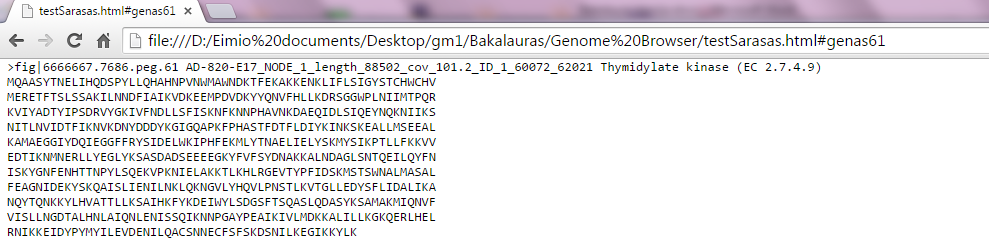
Pirmiausiai iš bendro MyRast ir Prokka failo yra nuskaitomos baltymų sekos su anotacija apie juos ir iš atskiro failo nuskaitomi fermentai su jų keliais(angl. pathway). Pačioje pradžioje įrašomos kelios eilutės į failą kurios simbolizuoja html failo pradžią. Iš nuskaitytų baltymų anotacijų yra nuskaitoma 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 *Excinuclease 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 pliusinis tuomet kryptis į dešinę pusę. Rodyklės yra paimamos kaip paveikslėliai rodančios viena arba kita kryptimi. 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. Šis sąrašas generuojamas kaip atskiras html failas kuriame yra dedamos visos baltymų sekos, 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. 16 paveikslėlyje pavaizduotas Sulfurimonas denitrificans genomo naršyklės dalis.



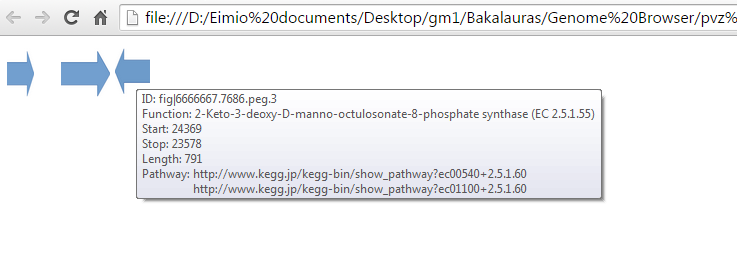
16 pav. Sulfurimonas denitrificans genomo naršyklės dalis.

Kaip matyti paveikslėlyje, daug skirtingo dydžio ir krypčių 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. 17 paveikslėlyje pavaizduota kaip atrodo naršyklės langas paspaudus ant kurios nors rodyklės.



17 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 tą baltymą, o paspaudus rodoma pati baltymo seka. 18 paveikslėlyje pavaizduota genomo naršyklė iš trijų baltymų.

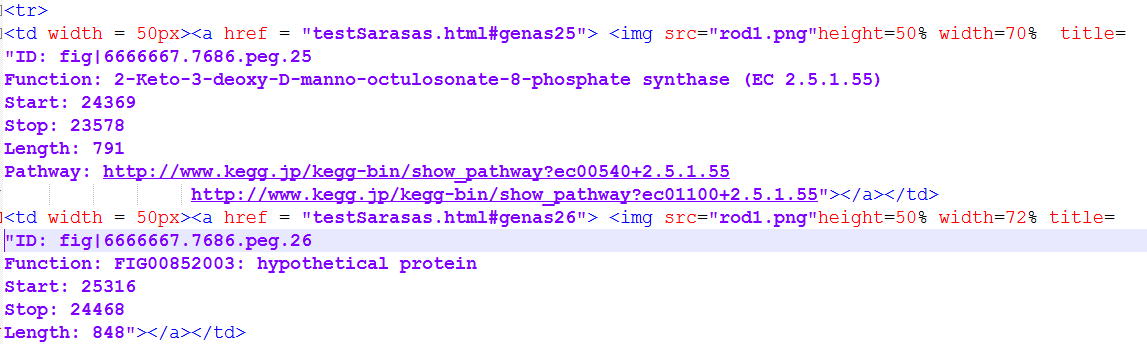


18 pav. Genomo naršyklė iš trijų Sulfurimonas denitrificans baltymų.

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. 19 paveikslėlyje pavaizduota, kaip atrodo ištrauka iš html kodo.

Paanalizuosiu iš ko susideda html kodas. Pirmiausiai visas html kodas yra dedamas į lentelę, tam kad išlaikyti tvarkingą rodyklių pasiskirstymą. *<tr>* simboliai atskiria eilutes,

19 pav. Ištrauka iš genomo naršyklės html kodo.

kiekvienoje eilutėje yra dedamas tik tam tikras skaičius rodyklių, kad nepersidengtų viena rodyklė su kita. Kiekvieno 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, yra nurodoma pasinaudojant *href* elementu, jo viduje nurodant kito failo *pavadinimą* ir *id*. *<img>* elementu yra nurodomas paveikslėlis kuris bus įdedamas 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 gali būti iškraipomas vaizdas.

# Literatūra:

1. Bergey's Manual of Systematic Bacteriology :Volume One : The Archaea and the Deeply Branching and Phototrophic Bacteria Editors: Editor-in-chief: **Garrity**, George **Boone**, David R., **Castenholz**, Richard W 309psl

2. Straipsnis : The Role of Hydrogen for Sulfurimonas denitrificans’ Metabolism. Yuchen Han, Mirjam Perner . Published: August 29, 2014 DOI: 10.1371/journal.pone.0106218.

3. https://microbewiki.kenyon.edu/index.php/Sulfurimonas\_denitrificans apie sulfurimoną

4. The RAST Server: Rapid Annotations using Subsystems Technology

5. Genome Annotation by The SEED Team\*

6. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses

7. Prokka: rapid prokaryotic genome annotation

8. Transitions in bacterial communities along the 2000km salinity gradient of the Baltic Sea