Izrada računalnih alata za analizu podataka o crijevnom mikrobiomu

Prskalo, Katarina

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Food Technology and Biotechnology / Sveučilište u Zagrebu, Prehrambeno-biotehnološki fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:159:579511

SVEUČILIŠTE U ZAGREBU

PREHRAMBENO-BIOTEHNOLOŠKI FAKULTET

Diplomski rad

Katarina Prskalo

Zagreb, srpanj 2020. 1023/MB

IZRADA RAČUNALNIH ALATA ZA ANALIZU PODATAKA O CRIJEVNOM MIKROBIOMU

Rad je izrađen u Laboratoriju za bioinformatiku na Zavodu za biokemijsko inženjerstvo Prehrambeno-biotehnološkog fakulteta Sveučilišta u Zagrebu pod mentorstvom dr. sc. Janka Diminića, docenta Prehrambeno-biotehnološkog fakulteta Sveučilišta u Zagrebu.

ZAHVALA

Zahvaljujem se mentoru doc. dr. sc. Janku Diminiću na neiscrpnoj strpljivosti i pomoći pri pisanju i izradi diplomskog rada. Također, hvala čitavom Laboratoriju za bioinformatiku na ukazanom povjerenju i podršci.

Zahvaljujem se svim divnim kolegama koje sam upoznala tijekom studija, posebice kolegicama Andrei Plec i Izabeli Maros na mnogobrojnim savjetima i motivacijama uz kuglice sladoleda zbog kojih je ovaj studij prebrzo prošao.

Posebno se zahvaljujem svojoj obitelji i partneru na velikoj strpljivosti te bezuvjetnoj potpori i brizi.

TEMELJNA DOKUMENTACIJSKA KARTICA

Diplomski rad

Sveučilište u Zagrebu
Prehrambeno-biotehnološki fakultet
Zavod za Biokemijsko inženjerstvo
Laboratorij za bioinformatiku

Znanstveno područje: Biotehničke znanosti
Znanstveno polje: Biotehnologija

IZRADA RAČUNALNIH ALATA ZA ANALIZU PODATAKA O CRIJEVNOM MIKROBIOMU

Katarina Prskalo, 1023/MB

Sažetak: Zbog napretka tehnologije sekvenciranja došlo je do eksponencijalnog rasta broja mikrobnih sekvenci u bazama podataka, a metagenomskom analizom dobivaju se kompleksni podaci o taksonomskim profilima i metaboličkim funkcijama. Obrada podataka je sve veći izazov stoga postoji potreba za razvojem računalnih alata koji će olakšati i skratiti vrijeme utrošeno na analize koje je moguće automatizirati. Programski alat razvijen u sklopu ovog diplomskog rada omogućuje grafički prikaz i statističku obradu podataka dobivenih metagenomskom analizom pomoću MetaPhlAn alata. Programski alat napisan je programskim jezikom Python te sadrži pet programa koji vrše obradu podataka i jedan testni program u kojem je moguće testirati funkcionalnosti ostalih. Programi uzimaju različite parametre s kojima je moguće mijenjati prikaz rezultata čime se dobiva na varijaciji u prikazu čime se olakšava interpretacija podataka.

Ključne riječi: crijevni mikrobiom, metagenomika, taksonomski profil, bioraznolikost

Rad sadrži: 44 stranice, 11 slika, 3 tablice, 70 literaturnih navoda, 0 priloga

Jezik izvornika: hrvatski

Rad je u tiskanom i elektroničkom (pdf format) **obliku pohranjen u:** Knjižnica Prehrambeno-biotehnološkog fakulteta, Kačićeva 23, Zagreb

Mentor: *doc. dr. sc. Janko Diminić*

Stručno povjerenstvo za ocjenu i obranu:

1. Izv.prof.dr.sc. Jurica Žučko
2. Doc.dr.sc. Janko Diminić
3. Doc.dr.sc. Andreja Leboš Pavunc
4. Doc.dr.sc. Anamarija Štafa (zamjena)

Datum obrane: 13.07.2020.

BASIC DOCUMENTATION CARD

Graduate Thesis

University of Zagreb
Faculty of Food Technology and Biotechnology
Department of Biochemical Engineering
Laboratory for Bioinformatics

Scientific area: Biotechnical Sciences
Scientific field: Biotechnology

DEVELOPMENT OF COMPUTER TOOLS FOR GUT MICROBIOME DATA ANALYSIS

Katarina Prskalo, 1023/MB

Abstract: Advancements in sequencing technology gave way to exponential growth in number of microbial sequences in databases, whereas metagenomic analysis results are complex data that include taxonomic profiles and metabolic functions. Data processing is gradually becoming a bigger challenge, therefore development of computer tools is required to facilitate and reduce time needed for analyses that are possible to automate. The software tool developed as a part of this master thesis enables graphical display and statistical analysis of data gained through metagenomic analysis with the MetaPhlAn tool. The software tool is coded with Python programing language and contains five programs that process data and one testing program with which is possible to test functions of others. Programs take different parameters which permit changes in showing results that enables variation in displaying which facilitates data interpretation.

Keywords: gut microbiome, metagenomics, taxonomic profile, biodiversity
Thesis contains: 44 pages, 11 figures, 3 tables, 70 references, 0 supplements
Original in: Croatian
Graduate Thesis in printed and electronic (pdf format) **version is deposited in**: Library of the Faculty of Food Technology and Biotechnology, Kačićeva 23, Zagreb.

Mentor: *PhD, Janko Diminić, Assistant professor*

Reviewers:
1. PhD. Jurica Žučko Associate professor
2. PhD. Janko Diminić Assistant professor
3. PhD. Andreja Leboš Pavunc Assistant professor
4. PhD. Anamarija Štafa Assistant professor (substitute)

Thesis defended: 13.07.2020.

SADRŽAJ **stranica**

1. UVOD

Ljudi žive u simbiozi s ogromnom i raznolikom skupinom mikroorganizama koja se sastoji od bakterija, arheja, virusa i eukariotskih mikroorganizama. Ti mikroorganizmi nastanjuju brojna mjesta na ljudskom tijelu poput kože, usne šupljine i gastrointestinalnog trakta, a skupni naziv im je mikrobiota. Većina mikrobiote nalazi se u gastrointestinalnom sustavu ljudi gdje donosi domaćinu brojne koristi poput fermentacije neprobavljene hrane, sinteze esencijalnih vitamina, zaštite od patogena i stimulacije imunosnog sustava (Heintz-Buschart i Wilmes, 2018). Napredak u tehnologiji sekvenciranja omogućio je istraživanje mikrobiote neovisno o kulturi stanica. Nije potrebno sekvencirati i analizirati pojedine vrste mikroorganizama već se sekvenciraju svi organizmi prisutni u uzorku pri čemu se dobije metagenom (Turnbaugh i sur., 2007). Rezultati raznih metagenomskih analiza pružili su nam saznanja o crijevnoj mikrobioti koja donedavno nisu bila poznata. Primjerice, bolesti poput ulceroznog kolitisa i Crohnove bolesti povezane su sa smanjenjem mikrobne raznolikosti crijevne mikrobiote (Carding i sur., 2015). Bioraznolikost i stabilan sastav su glavna svojstva crijevne mikrobiote te je njihovo narušavanje povezano sa raznim bolestima poput upalne bolesti crijeva, metaboličkih poremećaja, dijabetesa i pretilosti (Seksik i Landman, 2015).

S napretkom tehnologije dolazi do eksponencijalnog rasta broja mikrobnih sekvenci u bazama podataka te obrada takvih podataka predstavlja sve veći izazov (Singh i sur., 2017). Razvoj bioinformatičkih alata za obradu metagenomskih podataka ključan je za obradu sekvenciranih podataka u svrhu lociranja novih taksonomskih skupina, gena i biokemijskih funkcija. Nadalje, analiza dobivenih podataka, statistička obrada i vizualizacija važne su kako bi se mogli donijeti nedvosmisleni zaključci (Birkel i sur., 2017; Quince i sur., 2017). Bioinformatički alati za statističku obradu i vizualizaciju omogućuju bolji pregled velike količine podataka, što je potrebno za određivanje korelacije između sastava i funkcija crijevnog mikrobioma te utjecaja mikrobioma na ljudsko zdravlje (Falony i sur., 2016). Međutim, jedan od limitirajućih faktora u znanosti je obučavanje znanstvenika u kompjuterskim znanostima kako bi mogli efikasno analizirati kompleksne metagenomske podatke (Quince i sur., 2017).

U ovom radu korišten je programski jezik Python za izradu programskog alata koji učitava i prikazuje taksonomske podatke dobivenih metagenomskom analizom. Nadalje, alat računa indekse bioraznolikosti te izrađuje grafičke prikaze statističkih podataka.

2. TEORIJSKI DIO

2.1. Programski jezik Python

Python je programski jezik koji je vrlo popularan u biološkim znanostima zbog svoje jednostavnosti te ima opsežan izbor biblioteka, naročito za bioinformatiku (Ekmekci i sur., 2016). Osnivač Pythona je Nizozemac Guido van Rossum koji je prvi put objavio programski jezik 1991. godine (van Rossum, 2009). Python je programski jezik visokog nivoa što znači da koristi elemente prirodnih jezika u programiranju kako bi olakšao korištenje te skriva dijelove kompjuterskog računanja kako bi razvoj programa bio što jednostavniji i razumljiviji (Beal i sur., 2011). Kao i većina drugih današnjih jezika Python podržava objektno-orijentiranu paradigmu što omogućuje obavljanje kompleksnih zadataka na jednostavan i čovjeku prirodniji način te ima veliku, raznoliku knjižnicu biblioteka stoga se koristi u različitim područjima kao programski jezik opće namjene (Lutz, 2013). Jedna od najpoznatijih biblioteka naziva se NumPy, a služi za implementaciju mnogih matematičkih funkcija te ima široku primjenu za akademske svrhe, u laboratorijima i industriji (van der Walt i sur., 2011). Primjeri područja gdje se često koristi Python programski jezik su: sistemsko programiranje, GUI (grafičko korisničko sučelje), web programiranje, programiranje baza podataka te numeričko i znanstveno programiranje (Lutz, 2013).

Kako bi izvršili kod napisan u programskom jeziku Python potreban nam je interpreter koji prevodi i pokreće programski kôd. Programski kôd se prevodi u strojni kôd, kojeg razumije računalo, a zatim se strojni kôd izvršava budući da su to zapravo naredbe za procesor (engl. *central processing unit,* CPU) (Lutz, 2013). Python interpreter i veliki skup standardnih bioblioteka je uključen u osnovnom paketu Python paketa, kojeg je moguće preuzeti sa službene stranice (Python, 2020).

2.1.1. Razvojno okruženje PyCharm

PyCharm je integrirano razvojno okruženje (engl. *Integrated Development Environment,* IDE) za pisanje Python programskog jezika, a razvija ga firma JetBrains od 2010. godine (Taft, 2010). Pomoću IDE olakšan je razvoj programa koji povezuje različite aspekte u razvoju programa na jednom mjestu te je prikazan pomoću grafičkog sučelja (Lutz, 2013). Osim pokretanja programa i uklanjanja pogrešaka PyCharm ima dodatne mogućnosti poput automatskog formatiranja kôda, dovršavanja kôda, olakšane navigacije kroz kôd,

organizacije programa u projekte i drugo (Taft, 2010). PyCharm postoji u dvije verzije, osnovni i profesionalni paket (Haagsman, 2017). Osnovni paket je “otvoren“ u smislu da je moguće preuzeti programski kôd na mrežnom repozitoriju GitHub (GitHub, 2020), te na taj način sudjelovati razvoju tog alata (Jemerov, 2013). Profesionalna verzija PyCharma nudi dodatne mogućnosti kao što su uklanjanje pogrešaka na udaljenom računalu, podršku za web razvoj i podršku za korištenje baza podataka (Haagsman, 2017).

2.1.2. GitHub

GitHub je mrežna usluga koja omogućuje postavljanje internetskog repozitorija i kontroliranje verzija tijekom razvoja softvera pomoću Git-a distribuiranog sustava za upravljanje izvornim kôdom (GitHub, 2020). Git sustav je razvio Linus Torvals tijekom stvaranja Linux kernela. Bilo koji direktorij može postati Git repozitorij stoga nije potreban udaljeni server kako bi se repozitorij dijelio između korisnika. Ako programer želi raditi na nekom projektu onda preuzima kopiju repozitorija pri čemu ga račva (engl. *fork*). Zatim svaki korisnik repozitorija radi svoju verziju i mogu odlučiti koje će promjene preuzeti od drugih korisnika (Krajina, 2019).

GitHub se sastoji od programskih projekta koji su uglavnom otvorenog koda što znači da su svima dostupni te programerima pružaju mogućnost da zajednički pridonose razvoju projekata. GitHub ima grafičko sučelje što olakšava korištenje te omogućuje račvanje izvornog koda kako bi programeri napravili vlastite promjene. Zatim, moguće je napraviti zahtjev za povlačenjem (engl. *pull request*) kako bi se promjene u kodu razmatrale i eventualno uklopile u izvorni kod. GitHub također ima mogućnosti praćenja razvoja projekta (engl. *watch*) ili aktivnosti nekih prinosnika (engl. *follow*) te automatski obavještava o njihovim aktivnostima (Bleiel, 2016).

2.2. Mikrobiota

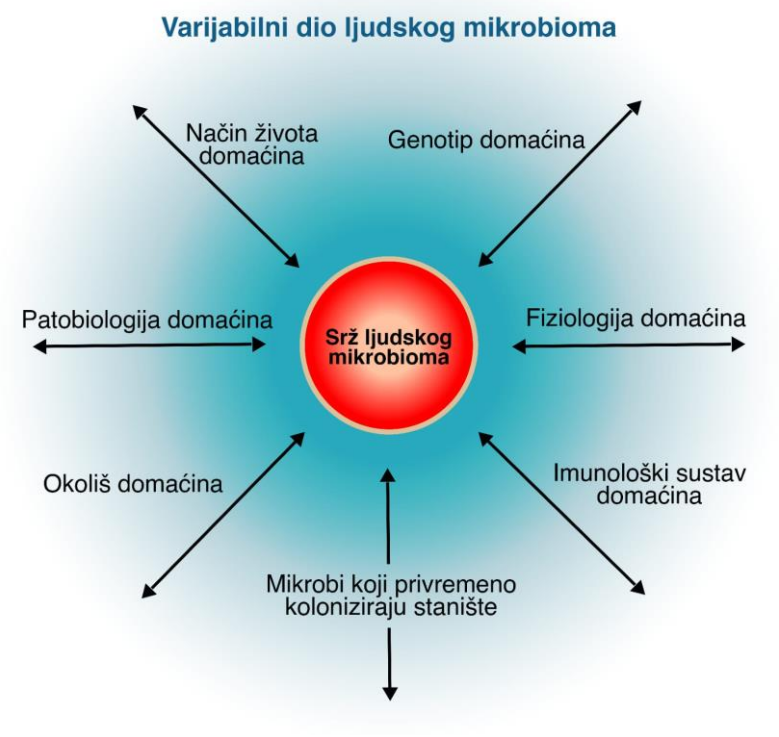
Mikrobiota su mikroorganizmi iz skupina bakterija, arheja i eukariota koje se nalaze unutar i na čovjeku. Sastav mikrobiote je dinamičan i drugačiji kod svakog čovjeka (Turnbaugh i sur., 2007). Mikrobne zajednice se razlikuju ovisno o dijelu tijela u kojem obitavaju poput kožne mikrobiote, oralne ili crijevne mikrobiote (Huttenhower i sur., 2012). Mikrobni ekosustavi koji se nalaze u ljudima potpuno su različiti od ekosustava koje nalazimo u prirodi, a neke bakterijske vrste prisutne su samo u mikrobioti iz čega možemo zaključiti da su mikrobni simbionti ko-evoluirali s domaćinom (Donaldson i sur., 2015).

Kao nastavak na Projekt ljudskog genoma (engl. *Human Genome Project*, HGP) pokrenut je Projekt ljudskog mikrobioma (engl. *Human Microbiome Project*, HMP) čiji je cilj identificirati sve mikrobne vrste mikrobiote, odrediti kako mikrobiota utječe na zdravlje i nastanak bolesti te odrediti da li je moguće manipulirati mikrobiotu kako bi se optimizirala za fiziologiju različitih osoba (Turnbaugh i sur., 2007). Broj bakterijskih stanica koje se nalaze u ljudskom tijelu iznosi okvirno $3{,}8\times10^{13}$ što čini omjer ljudskih i bakterijskih stanica 1:1, a ukupnu masu bakterijskih stanica 0,2 kg. Gastrointestinalni sustav ima visoku koncentraciju bakterijskih stanica i veliki volumen što ga čini sustavom s najbrojnijim i najkompleksnijim mikrobnim zajednicama (Sender i sur., 2016). Mikrobiota gastrointestinalnog sustava smatra se "skrivenim organom" zbog mnogobrojnih uloga i funkcija koje ima u organizmu (Liu, 2016; Clarke i sur., 2014). Primjeri uloga i funkcije crijevne mikrobiote su:

- mikrobnim metabolizmom dolazi do razgradnje hranjivih tvari čiji nusprodukti mogu biti vitamini ili drugi spojevi koji imaju nutritivnu vrijednost za domaćina
- mikrobiota sudjeluje u razgradnji i metabolizmu ksenobiotika te na taj način može neutralizirati toksikant
- mikrobiota potiče obnovu epitelnih stanica gastrointestinalnog sustava
- mikrobiota utječe na urođeni i prirođeni imunološki sustav (Turnbaugh i sur., 2007).

Podaci prikupljeni iz projekata poput HMP daju opsežan pregled mikroba koji su povezani s ljudima. Identificirano je 2172 vrsta bakterija koje su klasificirane u 12 različitih koljena, od koji 93,5 % vrsta pripadaju koljenima *Proteobacteria, Firmicutes, Actinobacteria* i *Bacteroidetes*. Većina bakterija u gastrointestinalnom sustavu nalazi se u debelom crijevu te više od 90 % bakterija spada u koljena bakterija *Firmicutes* i *Bacteroidetes*. Kod ljudi, 386

identificiranih vrsta su striktni anaerobi te uglavnom obitavaju u mukoznim slojevima usne šupljine i gastrointestinalnog trakta (Thursby i Juge, 2017). Turnbaugh i suradnici, u radu objavljenom 2007. godine u sklopu HMP-a, postavili su koncept srži humanog mikrobioma (slika 1) koji nalaže da je uvijek određeni set gena prisutan u određenom staništu kod svih ili većine ljudi. Međutim, u recentnom radu predloženo je da se srž mikrobioma razmatra na razini funkcije koja je prisutna kod svih ljudi (Thursby i Juge, 2017). Varijabilni dio ljudskog genoma može biti rezultat kombinacije različitih faktora poput genotipa domaćina, fiziološkog statusa, patobiologija, načina života, okoliša ili prisutnosti populacije mikroorganizama koja privremeno kolonizira stanište (Turnbaugh i sur., 2007).



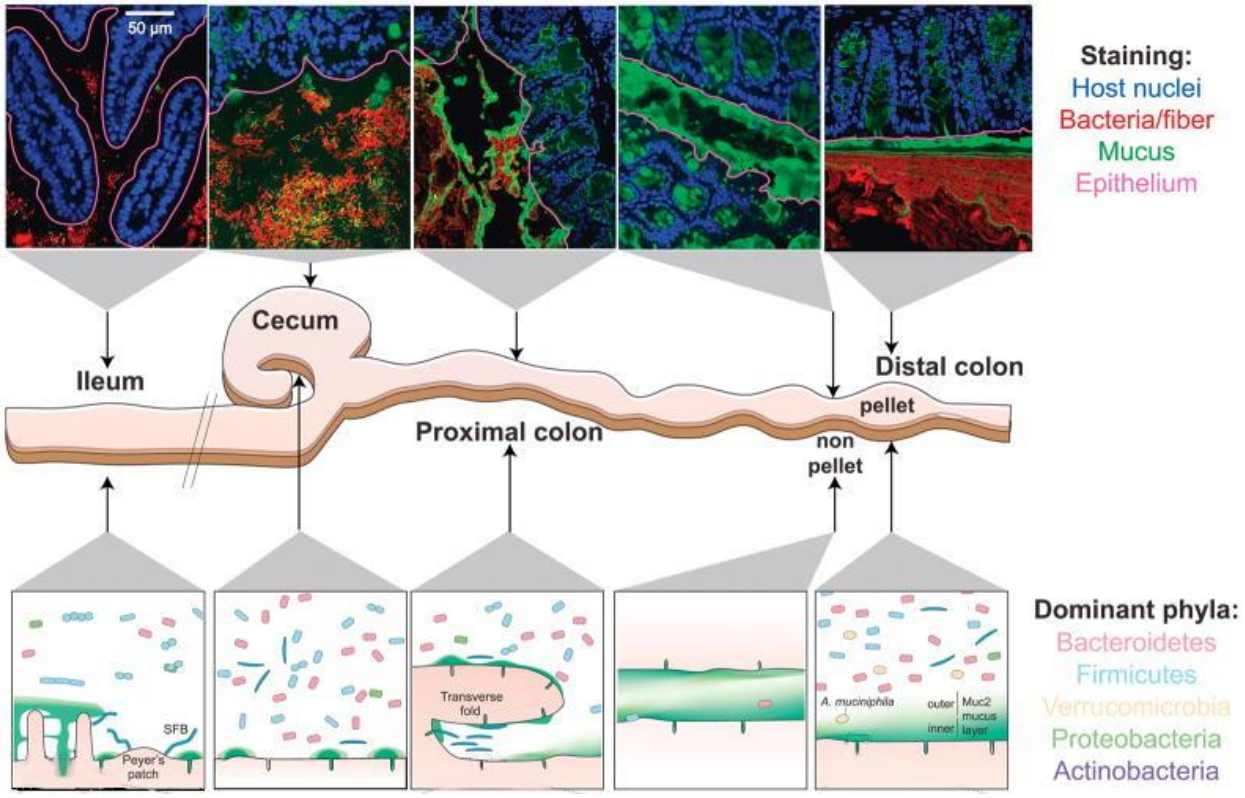
Slika 1. Prikaz srži i varijabilnih faktora ljudskog mikrobioma (Prilagođeno prema Turnbaugh i sur., 2007)

2.2.1. Sastav gastrointestinalnog sustava

Sastav mikrobiote u gastrointestinalnom sustavu ovisi o fiziološkim svojstvima određene regije te se mijenja uzduž i poprijeko gastrointestinalnog sustava. Gustoća i sastav

mikrobiote ovisi o kemijskom, hranjivom i imunološkom gradijentom. U tankom crijevu su uglavnom uvjeti visokih koncentracija kiselina, kisika i antimikrobnih spojeva te brzi protok tvari. U takvim uvjetima mogu preživjeti brzo rastuće bakterije koje su fakultativni anaerobi te imaju sposobnost adhezije na epitel ili mukus (Thursby i Juge, 2017). Iako su uvjeti u tankom crijevu nepovoljni za rast većine bakterija, one koje uspješno rastu imaju pristup jednostavnim ugljikohidratima. U tankom crijevu bakterijski diverzitet značajno je manji nego u debelom crijevu, a prisutne su *Clostridium* spp. te nekoliko vrsta iz koljena *Proteobacteria*. Analizom transkriptoma uzorka iz tankog crijeva zamijećena je povećana ekspresija gena za metabolizam i biokemijske puteve razgradnje jednostavnih šećera u odnosu na uzorke iz debelog crijeva (Donaldson i sur., 2015).

Na početku debelog crijeva nalazi se slijepo crijevo koje sadrži najgušće i najraznovrsnije mikrobne populacije te tamo počinje razgradnja rezistentnih polisaharida koji se nisu metabolizirali u tankom crijevu. Biljojedi imaju veće slijepo crijevo što omogućava mikrobioti da provodi dugotrajnu razgradnju biljnih vlakana. Uvjeti u debelom crijevu značajno su različiti nego u tankom crijevu kao što je: sporiji protok tvari, niže koncentracije antimikrobnih spojeva i nedostatak jednostavnih izvora ugljika. Stoga dolazi do rasta fermentativnih anaeroba koji razgrađuju polisaharide, a to su uglavnom bakterije iz porodice *Bactriodaceae* i *Clostridia* (Donaldson i sur., 2015). Bakterijski sastav nije uniforman uzduž presjeka po poprečnoj osi debelog crijeva (slika 2) zbog prisutnog sloja mukusa. Dio mukusa koji se nalazi uz lumen je rjeđi, vanjski sloj te se tamo nalaze bakterije koje mogu koristiti mukus kao izvor ugljika, a većinom su bakterije vrsta *B. acidifaciens* te *Akkermansia* spp. Unutarnji sloj mukusa je gušći i predstavlja fizikalnu barijeru između visoke koncentracije raznolikih bakterija i epitela domaćina. Unutarnji sloj mukusa također sadrži antimikrobne peptide i sekretorne imunoglobuline te služi kao imunološka barijera (Tropini i sur., 2017; Berry i sur., 2013).



Slika 2. Prikaz karakterističnih crijevnih struktura gastrointestnialnog sustava konfokalnom mikrografijom i shematski te prikaz bakterijskog sastava (prilagođeno prema Tropini i sur., 2017)

2.2.2. Metabolička uloga crijevnog mikrobioma

Procjenjuje se da ukupan zbroj gena koje nalazimo u mikrobioti nadmašuje broj gena u ljudskom genomu za više od 100 puta (Gill i sur., 2006). Veliki genom crijevne mikrobiote domaćinu daje veliku metaboličku raznolikost u debelom crijevu. Humani mikrobiom proširuje metabolizam glikana, aminokiselina i ksenobiotika te sudjeluje u metanogenezi i biosinetzi vitamina i izoprenoida (Marchesi i sur., 2016; Gill i sur., 2006). Hrana koju ljudi konzumiraju ne može se u potpunosti razgraditi pomoću enzima kodiranih u humanom genomu već se dio razgrađuje pomoću crijevne mikrobiote. Stoga, različiti glikani biljnog i životinjskog podrijetla razgrađuju se fermentacijom pomoću mikroorganizama. Fermentacijom nerazgradivih ugljikohidrata dolazi do povećane proizvodnje kiselina koje smanjuju pH lumena. Takva promjena utječe na sastav mikrobiote gdje populacije *Bacteroidetes* teže rastu, dok populacije *Firmicutes*, koje proizvode butirat, prevladavaju u blago kiselim uvjetima (Flint i sur., 2012; Koropatkin i sur., 2012).

2.2.2.1. Kratkolančane masne kiseline

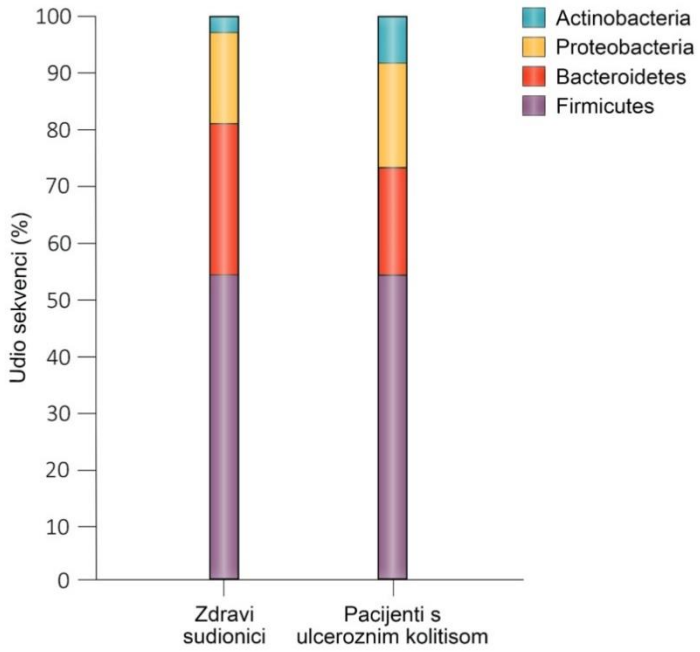
Mikrobnom fermentacijom glikana nastaju kratkolančane masne kiseline (engl. *Short Chain Fatty Acids,* SCFA) koje služe kao nutrijenti koloncitima i drugim stanicama crijevnog epitela. Osim kao izvor nutrijenata, zamijećeno je da SCFA poput butirata, propionata i acetata imaju više uloga kod domaćina. Butirat je preferirani izvor energije za kolonocite, ali također pokazuje tumor supresorsku ulogu kod rasta tumora u debelom crijevu. Acetat se može apsorbirati u krvotok i uklopiti u metabolizam lipida i glukoze, ali također u debelom crijevu potiče proizvodnju butirata kod nekih vrsta i sprječava kolonizaciju nekih patogena (Koropatkin i sur., 2012). Sve više istraživanja o utjecaju SCFA na ljudsko zdravlje pokazuje njihovu regulatornu ulogu u ljudskoj fiziologiji i metabolizmu. SCFA molekule imaju potencijalni utjecaj na homeostazu glukoze, metabolizam lipida, regulaciji apetita te regulaciji imunološkog odgovora i upalnih procesa (Morrison i Preston, 2016).

2.2.3. Utjecaj crijevnog mikrobioma na zdravlje

Bolesti poput upalne bolesti crijeva i metaboličkih poremećaja povezane su s crijevnom mikrobiotom. Dolazi do promjene sastava mikrobiote, a narušena ravnoteža naziva se disbioza. Kod Crohnove bolesti i ulceroznog kolitisa zamijećene su velike promjene u raznolikosti i stabilnosti mikrobiote, ali etiologija nije u potpunosti razjašnjena (Carding i sur., 2015). Pacijenti s dijabetesom tipa 2 pokazuju blagu disbiozu i smanjen im je broj bakterija koje proizvode butirat, a povećan je broj oportunističkih patogena (Qin i sur., 2012).

Smatra se da je pojava upalne bolesti crijeva povezana s industrijalizacijom u zemljama sjeverne Europe i Sjeverne Amerike gdje je i najviša stopa pojave te bolesti (Molodecky i sur., 2012). Moguće je da moderni način života kao što je korištenje antibiotika, pretjerana higijena i procesirana hrana povezan s kolonizacijom crijevnog mikrobioma. Utjecaj kolonizacije crijevnog mikrobioma važan je za regulaciju imunosnog sustava, a abnormalni sastav crijevne mikrobiote je glavno svojstvo upalnih bolesti crijeva. Kod ulceroznog kolitisa dolazi do upale mukoznog dijela debelog crijeva gdje je zamijećena disbioza zaštitničkih i štetnih bakterija (Manichanh i sur., 2012). U istraživanju provedenom na blizancima uočeno je da bolesni blizanac ima značajno manju bioraznolikost te su više zastupljena koljena *Actinobacteria* i *Proteobacteria*, a manje zastupljena su *Bacteroidetes* (slika 3) (Lepage i sur., 2011).

Crohnova bolest je heterogena bolest koja može zahvatiti bilo koji dio gastrointestinalnog trakta i izazvati razne upalne lezije. Provedeno je mnogo istraživanja o ovoj bolesti, ali nije uspješno otkriven patogen koji izaziva pojavu bolesti. Sastav mikrobiote pacijenata koji boluju od Crohnove bolesti značajno se razlikuje od zdravih individualaca. Na primjer, udio bakterije *Faecalibacterium prausnitzii,* koja ima protuupalni učinak, kod pacijenata je značajno manja, dok bakterije iz porodice *Enterobacteriaceae,* pogotovo *Escherichia coli*, ima u značajno većem broju (Manichanh i sur., 2012).



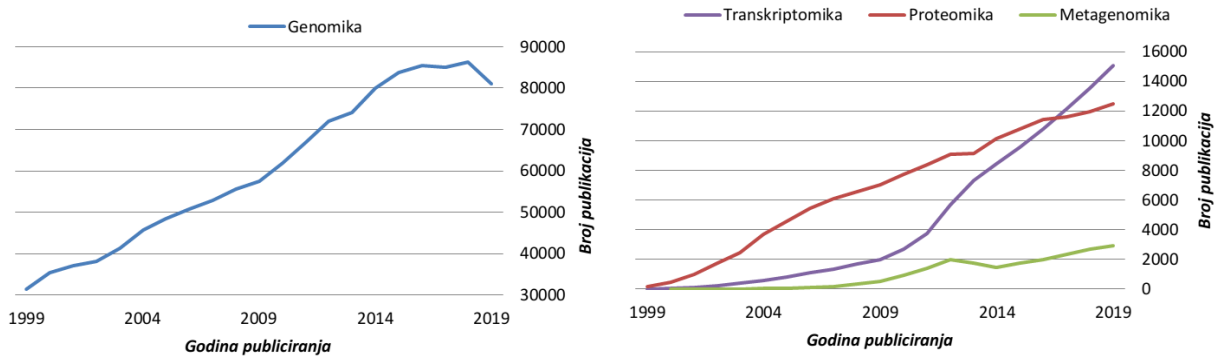
Slika 3. Prikaz udijela bakterijskih sekvenci kod zdravih i bolesnih blizanaca (Prilagođeno prema Lepage i sur., 2011)

U posljednjih nekoliko godina istraživanja pokazuju povezanost crijevnog mikrobioma s neurogenerativnim procesima. Crijevni mikrobiom može direktno ili indirektno utjecati na živčani sustav primjerice bakterijskim sekundarnim metabolitima, metaboličkim prekursorima ili imunosnom signalizacijom te tako utjecati na različite neurogenerativnve procese poput razvoja krvno-moždane barijere, neurogeneza, mijelinacija, sazrijevanje mikroglija te ekspresija neurotransmitera i odgovarajućih receptora (Sharon i sur., 2016). Istraživanja na miševima bez mikroorganizama (engl. *germ free,* GF) pokazuju povećanu permeabilnost krvno-moždane barijere, a kada su GF miševi izloženi mikrobioti bez patogena dolazi do smanjenje permeabilnosti (Braniste i sur., 2014). Nadalje, bolesti gastrointestinalnog sustava i alergije na hranu često se pojavljuju prilikom neurorazvojnih poremećaja što upućuje na utjecaj crijevnog mikrobioma. Istraživanja na miševima i GF

miševima pokazuju da stres utječe na sastav mikrobiote te da je komunikacija između mikrobiote i centralnog živčanog sustava dvosmjerna (Sharon i sur., 2016; Foster i McVey Neufeld, 2013).

2.3. Metagenomika

Metagenomika je interdisciplinarno područje koja spaja molekularnu genetiku, mikrobnu ekologiju i analizu podataka. Predmet istraživanja je metagenom, odnosno ukupan genetski materijal organizama i virusa koji su prisutni u danom uzorku (Sudarikov i sur., 2017). U posljednja dva desetljeća došlo je do porasta broja istraživanja na područjima genomike, metabolomike, transkriptomike, proteomike i drugih "*omika"* (engl. "*omics*"), a porast u broju publikacija objavljenih na tu temu od 1999. prikazana je na slici 4 (Noor i sur., 2019).



Slika 4. Prikaz broja publikacija određenih "omika" po godinama od 1999.

Metagenomika je područje koje proučava kompleksne mikrobne zajednice koje su uzorkovane direktno iz prirodnog okoliša (Ghosh i sur., 2019). Metode sekvenciranja takvih uzoraka nije ograničavajući faktor već upravljanje s ogromnom količinom sekvenci dobivenih takvom analizom. Sukladno tome, broj mikrobnih sekvenci u javnim bazama podataka raste eksponencijalno, stoga to predstavlja znanstvenicima veliki izazov (Singh i sur., 2017). Napredak u metagenomici omogućuje nam bolje razumijevanje ljudskog crijevnog mikrobioma te daje uvid u probleme koje je do nedavno bilo teško istraživati poput uloge mikrobiote u patogenezi IBD-a (Manichanh i sur., 2012).

2.3.1. Bioraznolikost

Ljudski crijevni mikrobiom je iznimno kompleksan i raznolik te varijacije nalazimo između jedinki i fluktuacije u sastavu kroz vrijeme. Razmatranje mikrobiote s ekološkog stajališta daje uvid u povezanost raznolikosti mikrobne zajednice s ljudskim zdravljem i fiziologijom (Lozupone i sur., 2012). Bioraznolikost je pojam koji se koristi kako bi se objasnila razlika u broju vrsta između različitih staništa (Magurran, 1998). U nekom staništu, vrste zauzimaju različite niše s obzirom na raspoložive resurse. Gradijent nekog resursa i interakcije između vrsta mogu omogućiti različite niše koje za posljedicu uvode još vrsta što doprinosi kompleksnosti nekog staništa i bogatstvu (engl. *richness*) vrsta (Whittaker, 1972). Za interpretaciju bioraznolikosti najčešće se koriste dvije komponente: *richness* i *evenness* (hrv. podjednakost). *Richness* predstavlja broj različitih vrsta, a *evenness* govori koliko se udio tih vrsta razlikuje (Magurran, 1998). *Richness* je najčešće korištena mjera te se može koristiti u različite svrhe kao na primjer usporedba prisutnosti nekih vrsta između dva uzorka. Za bolje razumjevanje potrebno je razmatrati i neke druge aspekte poput koliko je vrsta rijetka ili udio vrste u uzorku (Humphries i sur., 1995).

2.3.2. Prikupljanje metagenomskih podataka

Prvi korak u dobivanju metagenomskih podataka je sakupljanje uzoraka, a zatim se uzorci zamrzavaju i skladište. Nadalje, DNA se izolira iz uzorka te slijedi sekvenciranje. Postupci prije sekvenciranja DNA iz uzorka mogu utjecati na kvalitetu i točnost podataka. Temperatura i vrijeme skladištenja mogu utjecati na sastav uzorka, također lijekovi koji se možda mogu naći u probavnom sustavu ljudi imaju utjecaj na mikrobiom (Quince i sur., 2017). Metode sekvenciranja su u zadnjih 10 godina mnogo napredovale i značajno se smanjila cijena sekvenciranja što čini ovu tehnologiju pristupačnijom. Primjeri *Next Generation Sequencing* (hrv. sekvenciranje nove generacije, NGS) tehnologije su 454 Roche i Illumina Solexsa. 454 Roche je pirosekvenciranje koje koristi PCR u emulziji, dodaju se redom sva četiri deoksinukleozid trifosfata. Pri komplementarnom sparivanju dolazi do oslobađanja pirofosfata koji ulazi u daljnje enzimske reakcije čiji je konačan produkt svijetlo koje se zatim detektira. Ova metoda je vrlo popularna za metagenomske analize, ali nedostatak su greške u koncentraciji gena koje se javljaju prilikom umnožavanja DNA PCR-om. Illumina Solexa tehnologija imobilizira fragmente DNA na čvrstu površinu. Zatim se

fragmenti umnažaju PCR-om pri čemu nastaju nakupine identičnih DNA fragmenata. Zatim se fragmenti sekvenciraju pomoću revezibilnih terminatora procesom sekvenciranje kroz sintezu (Thomas i sur., 2012).

Sekvenciranje mikroorganizama u uzorku omogućuje analizu svih mikroorganizama u uzorku bez potrebe uzgajanja na umjetnoj podlozi. Može se sekvencirati čitav genom metodom *Whole Genome Shotgun* (hrv. sačmarica čitavog genoma, WGS) ili samo dio genoma metodom 16S rRNA sekvenciranje (Quince i sur., 2017). 16S rRNA metoda temelji se na sekvenciranju jednog lokusa u genomu - 16S rRNA gena odnosno 16S rDNA. Gen kodira za malu ribosomsku podjedinicu i prisutan je u svim bakterijama i arhejama. 16S rDNA adekvatan je izbor za ovu metodu jer sadrži visoko konzervirane regije, prema kojem se konstruiraju početnice i varijabilne regije pomoću kojih se radi taksonomska identifikacija. Koriste se referentne baze podataka za poravnanje sekvence te se za sekvenciranje mogu koristiti degradirani uzorci jer gen postoji u više kopija u genomu (Singh i sur., 2017). Međutim, 16S rRNA metoda daje samo taksonomski profil uzorka, dok WGS metoda daje informacije o čitavom genomu te omogućuje precizniji i detaljniji opis mikrobne zajednice (Laudadio i sur., 2018). WGS metoda (slika 5) može se opisati u pet koraka:

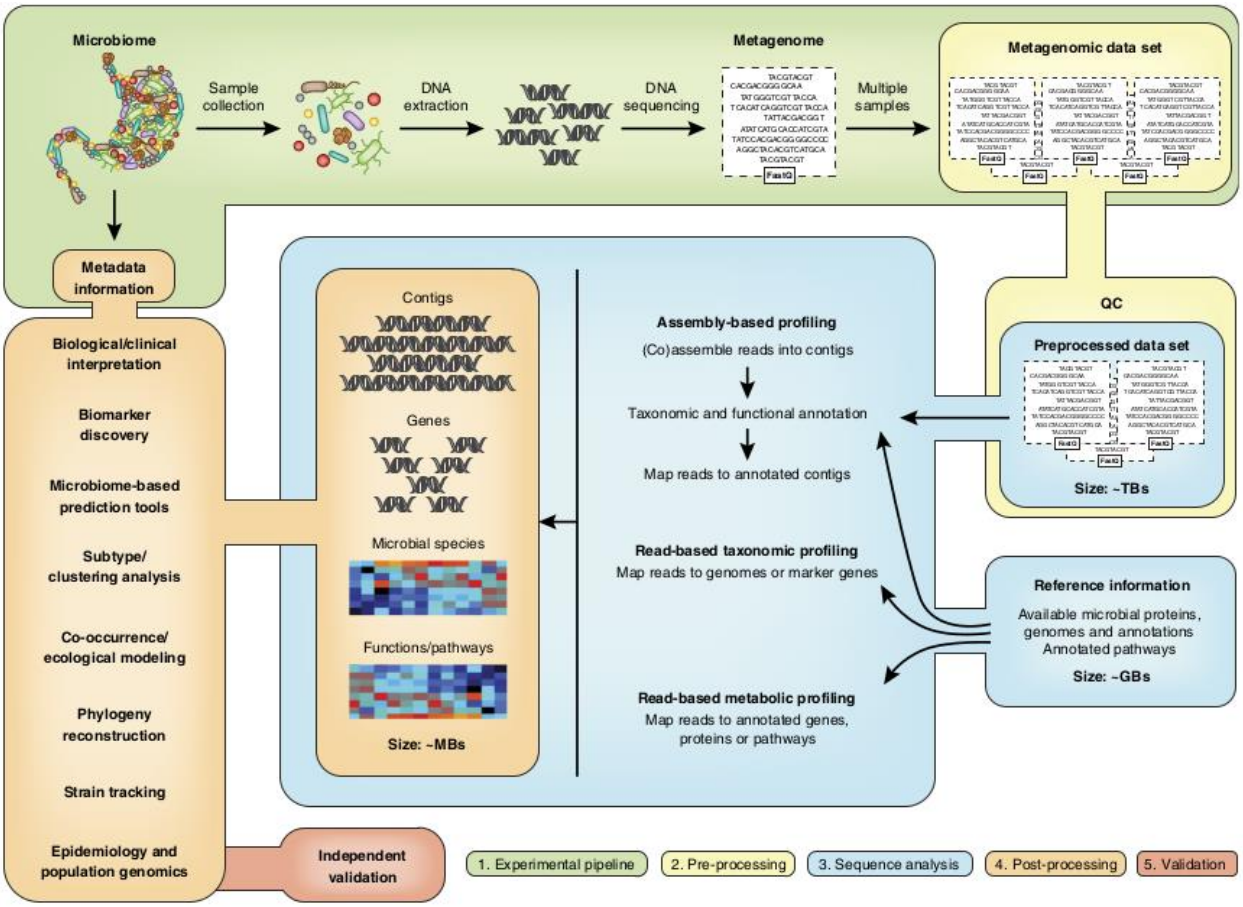
1. eksperimentalni protokol koji se sastoji od sakupljanja uzorka, ekstrakcije DNA i sekvenciranja

2. kompjutorska kontrola kvalitete - koraci koji minimiziraju količinu sekvencirane DNA koja nije dio uzorka poput adaptora sekvenciranja ili duplih sekvenci. Također, filtrira se strana DNA ili DNA koja nije ciljana

3. analiza sekvence – dva su pristupa analizi: temeljeno na sastavljanju (engl. *assembly based*) i temeljeno na očitanju (engl. *read based*). Pristup temeljen na sastavljanju sastavlja očitanja i sortira granična očitanja u genome, zatim analizira gene i dijelove genoma. Pristup temeljen na očitanju analizira nesortirana očitanja

4. post procesiranje - korištenje različitih multivarijantnih statističkih analiza za interpretaciju podataka

5. validacija - ponovna analiza podataka neovisne organizacije (Quince i sur., 2017).



Slika 5. Sumarni prikaz tijeka rada WGS metode (Quince i sur., 2017)

2.3.3. Analiza sekvenci i kompjutorska kontrola kvalitete

Sekvence dobivene WGS metodom obrađuju se bioinformatičkim alatima za metagenomiku. Sekvence se analiziraju i uspoređuju s bazama podataka mikrobnih sekvenci kako bi se kvantificirala prisutnost i koncentracija mikrobnih karakteristika. Rezultati tih analiza su, primjerice, broj i koncentracija bakterijskih koljena, vrsta, sojeva ili metaboličkih puteva (McIver i sur., 2018). Prije nego što se izvode zaključci iz bioloških podataka dobro je napraviti kontrolu kvalitete tih podataka. Važno je da su podaci točni i nepristrani jer kvaliteta podataka značajno može utjecati na rezultate. FastQC (Andrews, 2010) je jedan od alata za kontrolu kvalitete koji uočava greške nastale sekvenciranje ili greške u pripremi knjižnica. KneadData (kneadData, 2019) je alat za kontrolu kvalitete metagenomskih sekvenci kojim se mogu obrađivati podaci iz mikrobioma. Međutim ti podaci sadrže visoki udio sekvenci od domaćina naspram bakterijskih sekvenci. Ovaj alat provodi *in silico* razdvajanje bakterijskih sekvenci od "kontaminanata" (kneadData, 2019). Rezultati ovakvih

analiza su tablice koje se nadalje mogu statistički analizirati i vizualizirati kako bi se mogli izvesti zaključci eksperimenata (McIver i sur., 2018).

2.3.4. Određivanje taksonomskog profila

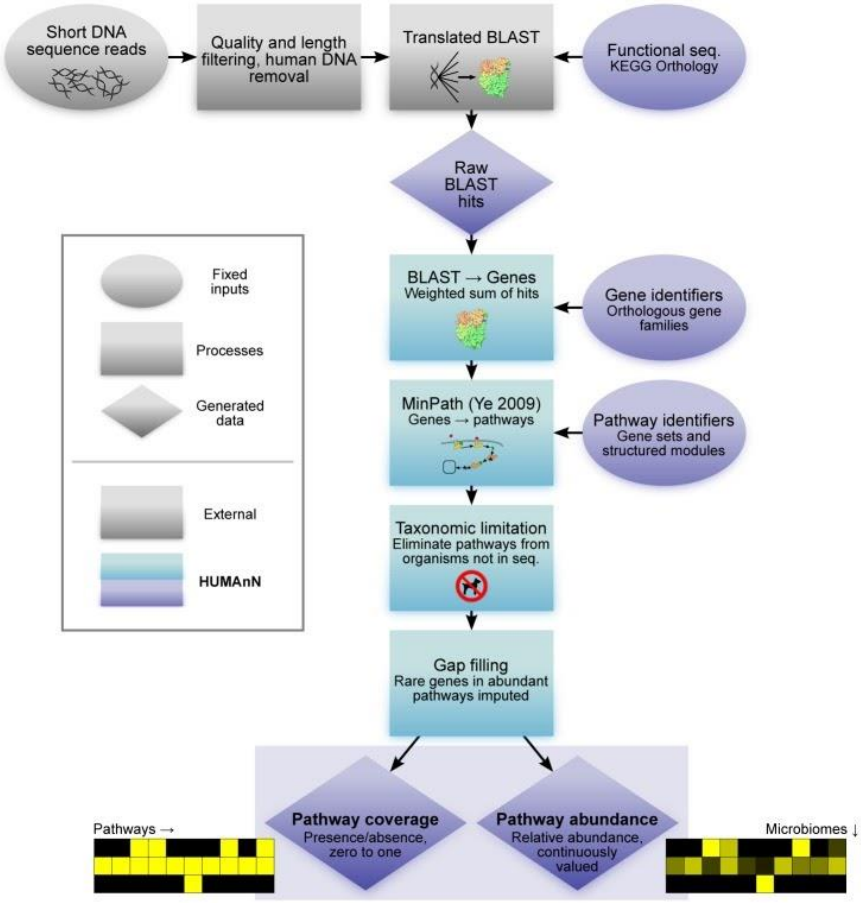
Za analizu metagenoma potrebno je odrediti taksonomski i filogenetski sastav uzoraka. Taksonomska identifikacija se može napraviti pomoću poravnanja sekvenci ili analizom sastava sekvenci pri čemu je važna visoka efikasnost i dobra rezolucija na razini vrste. Prilikom identifikacije pomoću poravnanja sekvenci zahtjeva se velika računalna snaga i usporedba s mnogo mikrobnih referentnih baza podataka (Segata i sur., 2012). Analizom sastava sekvenci DNA se sortira po svojstvima sekvence poput GC sastava, sklonost kodonima i učestalost oligonukleotida (Carola i Rolf, 2011). Također, nastaju problemi kod preciznog određivanja vrsta zbog kratkih očitanja dobivenih Illumina sekvenciranjem. MetaPhlAn (Metagenomic Phylogenetic Analysis) (Segata i sur., 2012) je alat koji ima mogućnost vrlo precizno odrediti profile mikrobnih zajednica u kratkom vremenu. Alat uspoređuje očitanja s relativno malom bazom podataka koja se sastoji samo od markera specifičnih za mikrobnu skupinu. Specifični markeri nedvojbeno identificiraju mikrobnu skupinu, a proces je efikasan obzirom da se za analizu ne koriste čitavi genomi (Segata i sur., 2012). MetaPhlAn 2 (verzija 2.7.2.) ima dodane podatke za bakterijske markere što omogućuje razlikovanje bakterijskih sojeva. Također, u bazu podataka uvedeni su markeri koji omogućuju identifikaciju eukariotskih i viralnih organizama (Truong i sur., 2015).

2.3.5. Određivanje metaboličkih funkcija

Podaci dobiveni sekvenciranjem mogu analizom pokazati koje sve bakterijske vrste se nalaze u uzorku što nam govori o bioraznolikosti. S time upotpunjujemo sliku o tome koja je metabolička uloga crijevne mikrobiote, kakva je biološka aktivnost i kako mikrobiom djeluje na cjelokupni fenotip. HUMAnN (HMP Unified Metabolic Analysis Network) je metodologija koja profilira metaboličke funkcije u mikrobnim zajednicama i prikazuje ih kao ortologne obitelji gena i njihovu relativnu koncentraciju. HUMAnN metodologija razvijena je u sklopu projekta HMP te omogućuje određivanje prisutnosti i količine određenog biokemijskog puta (slika 6). Koristi podatke dobivene WGS sekvenciranjem čitavog genoma

(Abubucker i sur., 2012), te nakon obrade i kontrole kvalitete sekvence se poravnavaju pomoću KEGG Orthology sustava (Kanehisa i sur., 2019). KEGG (Kanehisa i Goto, 2000) baza podataka sadrži organizirane podatke o staničnim funkcijama. Prikazuje se na dva načina: kao mapa biokemijskih puteva te popis gena i molekula. Često se koristi kao referentna baza podataka za interpretaciju bioloških podataka dobivenih novim visoko propusnim tehnologijama (poput sekvenciranja). KEGG Orthology je sustav koji svrstava sekvence u ručno formirane grupe ortolognih gena koje se nalaze u KEGG bazi podataka (Kanehisa i sur., 2010). Nakon poravnanja sekvenci identificiraju se geni i računa se relativna koncentracija gena. Zatim se rekonstruiraju biokemijski putevi pomoću rezultata te se popunjavaju rupe (uzrokovane rijetkim genima u visoko koncentriranim putevima). Rezultat su matrice, jedna koja prikazuje prisutnost ili odsutnost nekog puta i druga koja prikazuje relativnu koncentraciju puta (Abubucker i sur., 2012).

Nova verzija alata HUMAnN2 ima drugačiji pristup pri funkcionalnom profiliranju sekvenciranih mikrobnih uzoraka. Novi algoritam se provodi u tri koraka te nudi precizne profile i provodi analizu u značajno manje vremena (Franzosa i sur., 2018).



Slika 6. Shematski prikaz HUMAnN metodologije. Prikazan je rezultat analize u obliku matrice gdje se vidi prisutnost biokemijskog puta u mikrobiomu određene tjelesne lokacije (Abubucker i sur., 2012)

3. EKSPERIMENTALNI DIO

3.1. Materijali

Za izradu diplomskog rada korišteno je prijenosno računalo s operativnim sustavom Linux. Dodatni programi koji su korišteni naknadno su instalirani na prijenosno računalo. Za izradu programskih alata koristili su se primjeri bioloških podataka iz projekta 1000IBD (Imhan i sur., 2019).

3.1.1. Sklopovlje

Korišteno je prijenosno računalo Lenovo serije G500 model 59-380860 slijedećeg sklopovlja (engl. *hardware*): procesor Intel© Core™ i3-3110M od 2.40 Ghz, radna memorija od 2 GB, čvrstog diska HDD (engl. *hard-disk drive*) od 1 TB.

3.1.2. Operativni sustav

Na čvrstom disku prijenosnog računala instaliran je operativni sustav Linux Mint 19.1 Cinnamon verzija 4.0.10

3.1.3. Programska podrška

U ovom radu korišten je programski jezik Python i razvojno okruženje PyCharm za učitavanje, računalnu analizu i grafički prikaz podataka.

3.1.3.1. Računalna analiza

Python je objektno orijentirani programski jezik opće namjene (Python, 2020). Pomoću Python programskog jezika napisani su svi programi iz ovog rada. Operativni sustav u kojem je izrađen ovaj rad ima prethodno instaliran programski jezik Python, ali stariju verziju 2.7. Nova verzija, Python 3.0 preuzeta je sa službenih stranica Python-a te je instalirana pomoću terminala. Dodatne biblioteke koje su korištene u radu također su instalirane pomoću terminala naredbom `pip install`. Korištene biblioteke su Numpy (Numpy, 2020), Matplotlib (Matplotlib, 2020) te biblioteke CSV i OS koje spadaju u

standardnu Python knjižnicu te uglavnom dolaze u sklopu Python programskog jezika. Biblioteka Numpy je korištena za izvođenje matematičkih operacija, a biblioteka Matplotlib za izradu i prikaz grafova. Biblioteka CSV je korištena za učitavanje podataka koji su CSV (engl. *comma-separarted values*, hrv. vrijednosti razdvojene zarezom) formata pri čemu razdvaja podatke prema graničniku. Biblioteka OS omogućava upravljanje raznim sučeljima operacijskog sustava te je korištena za dohvaćanje informacija poput adrese u kojoj se nalazi program.

3.2. Metode

U ovom diplomskom radu razvijeno je pet računalnih programa koji vrše sljedeće operacije nad podacima:

- učitavaju i raščlanjuju podatke,
- računaju parametre bioraznolikosti,
- provode statističke analize,
- i na koncu grafički prikazuju rezultate.

Također izrađen je pomoćni program s kojim se pokreću ostali programi, a koristi se i kao sustav za testiranje. Svi programi izrađeni su prateći preporučene konvencije pisanja Python koda (van Rossum, 2001).

3.2.1. Program za učitavanje taksonomskih podataka

Program za učitavanje taksonomskih podataka naziva se BiomTable_class te izvršava učitavanje taksonomskih podataka koji su u CSV formatu. Ime formata podataka govori da su to vrijednosti razdvojeni zarezom, ali mogu biti razdvojene bilo kojim graničnikom. Podaci spremljeni u CSV formatu sadrže linije koje predstavljaju redove, a vrijednosti razdvojene graničnikom predstavljaju kolone te čine strukturu tablice.

U taksonomskim podacima kolone sadržavaju taksone primjerice carstvo *Bacteria* ili rod *Bacteroides*. Zadnja kolona uvijek sadrži brojnost odgovarajućeg taksona. Prvi red u tablici prikazuju najvišu taksonomsku razinu, carstvo, a zatim redom niže: koljeno, razred, red, porodica, rod, vrsta i soj.

Taksonomski podaci sadrže u sebi dva različita graničnika: okomita crta i tabulator znakom (engl. *tab character*) (slika 7). Prilikom učitavanja podataka BiomTable_class razdvaja podatke prvo po jednom graničniku te u drugom dijelu programa dodatnom operacijom razdvaja po drugom graničniku. Podaci se raščlanjuju pozivom metodom `get_table` i pohranjuju se u varijable tipa lista.

```
k__Bacteria 100.0
k__Bacteria|p__Bacteroidetes    61.53827
k__Bacteria|p__Proteobacteria   20.86671
k__Bacteria|p__Firmicutes   16.19394
k__Bacteria|p__Actinobacteria   0.84814
k__Bacteria|p__Verrucomicrobia  0.55295
```

Slika 7. Primjer dijela CSV tablice koji prikazuje brojnost na razini carstva i koljena. U primjeru se vidi da su prisutna 2 tipa graničnika: okomita crta i tabulator znak

Glavne metode programa BiomTable_class su `tax_abundance` i `tax_richness`. Obje metode za argument uzimaju taksonomsku razinu te vraćaju tablicu u kojoj je prva kolona naziv taksonomske skupine, a druga kolona odgovarajuća vrijednost (brojnost ili bogatstvo).

Metoda `tax_abundance` radi na način da čita red po red tablice koja je dobivena prilikom učitavanja CSV datoteke. U svakom redu provjerava zadnji stupac zatim uspoređuje prvi znak stupca koji označava jednoslovnu kraticu taksonomske razine s unesenim argumentom koji je željena taksonomska razina. Ako se prvi znak stupca i uneseni argument podudaraju slijedi raščlanjivanje te ćelije na temelju tabulatorskog graničnika. Rezultat raščlanjivanja sprema se u novu tablicu u kojoj je prva kolona naziv taksonomske skupine, a druga odgovarajuća brojnost.

Bogatstvo taksonomske razine je broj vrsta koja spada u određenu taksonomsku razinu. Metoda `tax_richness` određuje bogatstvo uzorka iz učitane tablice. Algoritam radi tako da u prvom koraku u novu tablicu spremi podatke o nazivima taksonomskih skupina dane taksonomske razine. Zatim, u drugom koraku prolazi kroz učitanu tablicu te filtrira samo one redove koji sadržavaju vrste. Za svaku vrstu provjerava taksonomsku razinu danu u argumentu te ju uspoređuje s taksonomskom skupinom iz tablice dobivene u prvom koraku. Ako se taksonomske skupine podudaraju onda ih zbraja i pohranjuje u drugu kolonu tablice. Rezultat je tablica u kojoj je prvi stupac taksonomska skupina, a drugi stupac bogatstvo.

Tablice dobivene metodama `tax_abundance` i `tax_richness` koriste se u drugim programima kako bi se odredili indeksi bioraznolikosti, provela statistička analiza i izradili grafovi.

3.2.2. Određivanje bioraznolikosti

Bioraznolikost označava razliku u broju vrsta između uzoraka te se dijeli na alfa i beta diverzitet. Alfa diverzitet označava raznolikost vrsta unutar jednog uzorka, dok beta diverzitet uspoređuje raznolikost jednog uzorka u odnosnu na druge uzorke. Za određivanje indeksa alfa i beta diverziteta razvijen je program Diversity_class. Program računa bogatstvo vrsta, Simpson indeks i Shannon indeks što su kvantitativne mjere alfa diverziteta te Jaccard indeks što je kvantitativna mjera beta diverziteta. U programu Diversity_class naslijeđene su metode za dobivanje tablica s podacima o brojnosti i bogatstvu iz programa BiomTable_class te se pomoću njih računaju parametri alfa i beta diverziteta.

Simpson indeks izražava koliki je stupanj „dominacije" neke vrste u uzorku. Na vrijednost Simpson indeksa utječe važnost najzastupljenijih vrsta. Računa se kao suma kvadrata brojnosti vrste prema formuli:

$$C = 1 - \sum_{i=1}^{S} p_i^2 \qquad [1]$$

Gdje je *S* broj vrsta, a p_i brojnost vrste u uzorku. Metoda za računanje Simpson indeksa uzima redom podatke iz tablice brojnosti. Svaka vrijednost se kvadrira zatim zbraja i sprema u varijablu. Konačno, metoda vraća dobivenu sumu oduzetu od jedan.

Shannon indeks koristi se za prikazivanje stupnja podjednake raspodjele (engl. *evenness)* između vrsta. Pomoću tog indeksa se prikazuje koliko postoji vrsta čiji je broj jedinki isti u danom uzorku. Računa se prema formuli:

$$H' = -\sum_{i=1}^{S} p \cdot \ln(p_i) \qquad [2]$$

Gdje je S broj vrsta, a p_i brojnost taksonomske skupine u uzorku (Whittaker, 1972). Prilikom računanja Shannon indeksa potrebno je pozvati funkciju za logaritmiranje iz dodatne Python biblioteke Numpy.

Za analizu beta diverizteta koristi se Jaccard indeks. Jaccard indeks prikazuje sličnost dva uzorka odnosno uzima u obzir vrste koje su prisutne u oba uzorka. Računa se prema formuli:

$$J = \frac{U \cdot V}{U + V + U \cdot V} \qquad [3]$$

Gdje je U zbroj brojnosti vrsta iz jednog uzorka koji su prisutni u oba uzorka, a V je zbroj brojnosti vrsta iz drugog uzorka koji su prisutni u oba uzorka (Real i Vargas, 1996). Pozivanjem metode za računanje Jaccard indeksa potrebno je u argument navesti tablicu brojnosti za drugi uzorak. Zatim, metoda uspoređuje tablice te računa vrijednosti U i V. Ako dvije tablice nemaju zajedničkih vrsta onda metoda javlja iznimku (`raise Exception`) te ispisuje poruku u konzolu.

3.2.3. Programi za izradu grafičkih prikaza

Vizualizacija podataka je važna jer olakšava interpretaciju podataka, stoga su razvijeni programi Biom_chart i Box_plot koji izrađuju grafove na temelju podataka analizom programa BiomTable_class. Biom_chart izrađuje kružni i stupčasti dijagram jednog uzorka, dok Box_plot učitava više uzoraka, računa interkvartile, medijan i srednju vrijednost te izrađuje kutijasti dijagram.

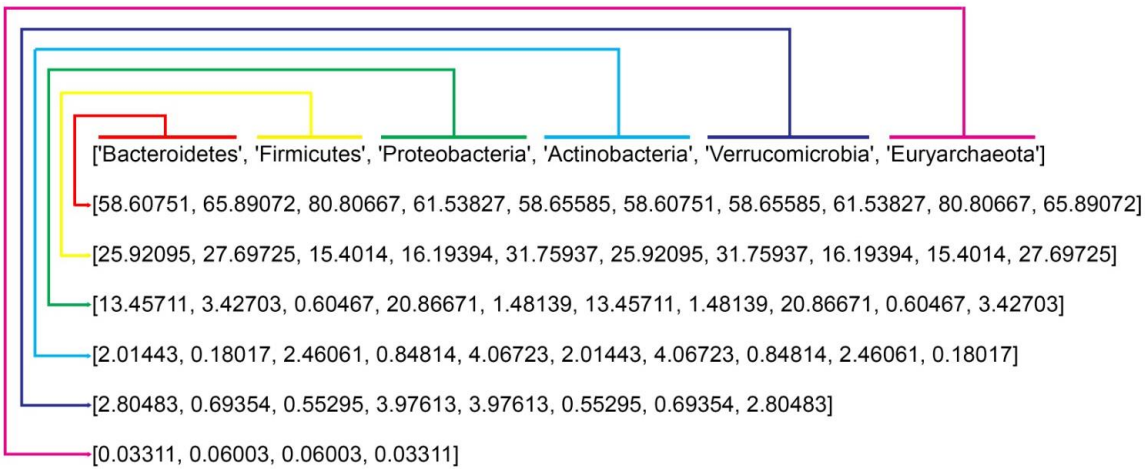
3.2.3.1. Kružni i stupčasti dijagram

Program Biom_chart izrađuje kružni i stupčasti dijagram koji služe za vizualizaciju taksonomskih podataka uzoraka. Program sadrži jednu funkciju koja za argumente uzima tablicu s podacima, vrstu podataka (brojnost ili bogatstvo), taksonomsku razinu i tip grafa (kružni ili stupčasti dijagram). Prilikom učitavanja podataka u program, provjerava se koja vrsta podataka je navedena u argumentu. Shodno tome se poziva metoda BiomTable_class programa `tax_abundance` ako je argument brojnost odnosno metoda `tax_richness` ako je argument bogatstvo. Zatim, dobivena tablica razdvaja se u dvije varijable s podacima i oznakama za dijagram. Prilikom razdvajanja varijabli provjerava se vrijednost brojnosti te ako je vrijednost brojnosti manja od jedan onda se ta vrijednost pridodaje posebnoj varijabli koja se ne uključuje u dijagram. Tako se sprječava da mnogo malih vrijednosti brojnosti se ne prikažu na dijagramima što ih čini preglednijima. Zatim se varijable s podacima i oznakama za dijagram unose kao argumenti metoda iz biblioteke Matplotlib za izradu grafova. Ovisno o

unesenom argumentu za tip grafa poziva se metoda za izradu kružnog odnosno stupčastog dijagrama iz biblioteke Matplotlib. Konačno, poziv na metodu `show` prikazuje izrađeni dijagram.

3.2.3.2. Izrada kutijastog dijagrama

Kutijasti dijagram prikazuje odnos minimuma i maksimuma seta podataka, medijan te pravokutnik („kutiju“) u kojoj se nalazi 50 % podataka. Za izradu kutijastog dijagrama razvijen je program Box_plot koji sadrži funkciju box_plot_drawer koja za argumente uzima naziv mape u kojoj se nalaze uzorci, tip podataka (bogatstvo ili brojnost) te taksonomsku razinu. Prilikom izrade kutijastog dijagrama potrebno je u program učitati više datoteka uzoraka. Funkcija metodama iz biblioteke OS čita sadržaj mape navedene u argumentu zatim provjerava nazive datoteka. Datoteke koje u imenu sadržavaju „metaphlan“ te imaju nastavak „.txt“ učitavaju se jedna po jedna u program pozivom na program BiomTable_class. Zatim, ovisno o taksonomskoj razini i vrsti podataka izrađuje se tablica metodom `tax_abundance` odnosno `tax_richness`. Nadalje, potrebno je sortirati i spremiti podatke iz svakog uzorka. Podaci se sortiraju u novu tablicu pohranjenu u novu varijablu na način da prvi red sadržava nazive svih taksonomskih skupina koje se pojavljuju u uzorcima, a ostali redovi sadržavaju iznose brojnosti ili bogatstva s time da drugi red sadržava sve vrijednosti taksonomske skupine iz prve kolone, treći red sadržava vrijednosti taksonomske skupine iz druge kolone itd. (slika 8). Nakon što su učitane sve datoteke svi potrebni podaci za izradu kutijastog dijagrama spremljeni su u varijablu. Zatim se poziva metoda Matplotlib biblioteke za izradu kutijastog dijagrama koja za argumente uzima oznake za x os, što je prvi red varijable te podatke za izradu dijagrama koji su pohranjeni u ostatku varijable. Konačno, metodom `show` izrađuje se grafički prikaz s paralelnim kutijastim dijagramima koji prikazuje medijan, srednju vrijednost, minimum, maksimum te interkvartilni raspon.



Slika 8. Prikaz primjera tablice te na koji način su organizirani podaci

3.2.4. Analiza brojnosti biokemijskih puteva

Analizom brojnosti biokemijskih puteva može se odrediti metabolička uloga crijevnog mikrobioma, stoga je u sklopu ovog programskog alata razvijen program PathwayTable_class. Program učitava podatke o biokemijskim putevima, raščlanjuje ih te filtrira one podatke koji nisu potrebni za daljnju analizu.

Podaci o biokemijskim putevima mikrobiote dobiveni su HUMAnN alatom te su spremljene u datoteku CSV formata. Program učitava podatke iz datoteke prilikom čega ih raščlanjuje na temelju graničnika te ih pohranjuje u tablicu. Tablica sadrži naziv biokemijskog puta i brojnost puta (engl. *pathway abundance*). Zatim se tablica filtrira prilikom čega se izbacuju podaci o putevima koji nisu mapirani, integrirani ili su specifični za neku bakterijsku taksonomsku skupinu. Putevi koji nisu mapirani i integrirani imaju u svojim nazivima „*UNMAPPED*“ (hrv. nemapiran) odnosno „*UNINTEGRATED*“ (hrv. neintegriran) te se na temelju toga filtriraju. Putevi specifični za bakterijsku taksonomsku skupinu imaju naziv podijeljen na dva dijela kojem je prvi dio naziv puta, a drugi naziv taksonomske skupine bakterije te su odvojeni okomitim znakom. Prilikom filtriranja, program traži okomiti znak te na taj način prepoznaje koji su putevi specifični za bakterijske taksone te ih izbacuje. Nakon filtriranja podaci se spremaju u novu tablicu te se koriste dalje u programu. Nadalje, razvijena je metoda `get_table` koja za rezultat vraća čitavu tablicu i metoda `get_top_pathways` koja za argument uzima broj te za rezultat vraća najzastupljenije biokemijske puteve.

4. REZULTATI I RASPRAVA

Crijevni mikrobiom ima sastavnu ulogu u ljudskoj fiziologiji, zdravlju i bolesti te ima važnu ulogu u mnogim metaboličkim putevima domaćina. Istraživanje crijevnog mikrobioma uvelike je omogućeno razvojem visoko propusnih metoda sekvenciranja. Klasične metode u mikrobiologiji obuhvaćale su uzgoj mikroorganizama na hranjivoj podlozi što nije moguće za većinu vrsta koje čine crijevni mikrobiom. Nadalje, sekvenciranjem uzorka crijevnog mikrobioma određuje se metagenom čime se dobiva slika ekosustava. Iz metagenoma moguće je odrediti taksonomski profil i metaboličke funkcije crijevnog mikrobioma te analizirati utjecaj na domaćina. Primjerice, smanjena bioraznolikost crijevne mikrobiote povezana je s bolestima poput ulceroznog kolitisa i Chronove bolesti. Napredak u metagenomici doveo je do rasta količine podataka o taksonomskim profilima i metaboličkim funkcijama (Carding i sur., 2015; Turnbaugh i sur., 2007). Stoga, obrada podataka je sve veći izazov te postoji potreba za razvojem računalnih alata koji će olakšati i skratiti vrijeme utrošeno na analize. Cilj ovog diplomskog rada je izrada računalnog programa za analizu bioloških podataka crijevne mikrobiote. Programski alat omogućuje grafički prikaz i statističku obradu podataka dobivenih metagenomskom analizom pomoću MetaPhlAn alata. Programski alat napisan je programskim jezikom Python te sadrži 5 programa koji vrše obradu podataka i 1 testni program u kojem je moguće testirati funkcionalnosti ostalih. Programi uzimaju različite parametre s kojima je moguće mijenjati prikaz rezultata čime se dobiva na varijaciji u prikazu i olakšava se interpretacija podataka.

4.1. Računalni program BiomTable_class

Računalni program BiomTable_class učitava i raščlanjuje podatke o taksonomskom sastavu uzorka. Za učitavanje taksonomskih podataka korištena je CSV biblioteka iz standardne Python knjižnice. Prilikom poziva, program kao argument uzima naziv datoteke koja sadrži podatke o brojnosti taksona iz uzorka. Zatim se datoteka učitava te se podaci o brojnosti taksona raščlanjuju i računa se bogatstvo vrsta za uzorak. Kao rezultat program vraća podatke o brojnosti i bogatstvu uzorka u obliku koji ostali programi iz programskog alata mogu koristiti za prikaz grafova, računanje indeksa i statističku analizu.

Taksonomski podaci uzoraka dobiveni su MetaPhlAn analizom te su pohranjeni u datoteke koje su u CSV formatu. Međutim, budući da su podaci biološke klasifikacije u stablastoj strukturi (roditelj - dijete) tako se i ovdje CSV format proširuje da može primiti takvu stablastu strukturu podataka. Svaki redak sadrži takson i više pripadajućih taksona viših

taksonomskih kategorija i vrijednost brojnosti. U primjeru je prikazano nekoliko linija MetaPhlAn CSV formata koje ilustriraju stablastu strukturu:

```
k__Bacteria    100.0
k__Bacteria|p__Firmicutes     16.19394
k__Bacteria|p__Firmicutes|c__Clostridia      16.05459
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales    16.05459
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Eubacteriaceae 5.11922
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Lachnospiraceae       2.53793
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridiales|f__Oscillospiraceae      1.57889
```

U navedenom primjeru mogu se uočiti 2 tipa graničnika: okomita crta | i tabulator znak. Taksoni su odvojeni okomitom crtom, dok je brojnost taksona odvojena tabulator znakom. Tabulator znak je primjer specijalnog znaka koji se u programu označava pomoću 2 znaka: obrnuta kosa crta i slovo t (`\t`). Računalni program specijalne znakove gleda kao jedan znak, a sadrže obrnutu kosu crtu te drugi znak koji varira. Primjerice u Linux i Mac sustavima `\n` je specijalni znak za novi red, a `\b` za povratnik (engl. *backspace*).

Metoda za učitavanje podataka prilikom učitavanja raščlanjuje podatke po proizvoljno definiranom graničniku, a u programu raščlanjuje po okomitoj crti. Dodatnom operacijom provodi se raščlamba po drugom graničniku. Algoritam pomoću for petlje iz linija izdvaja tekst koji je potrebno dodatno raščlaniti po tabulator znaku (`\t`). Primjer takvog teksta za taksonomsku razinu koljeno:

```
Proteobacteria\t20.86671
```

Prilikom drugog raščlanjivanja program čita tekst `Proteobacteria` koji se nalazi prije specijalnog znaka `\t` i pohranjuje ga u varijablu kao ime taksonomske skupine. Tekst nakon znaka `\t` predstavlja vrijednost brojnosti za takson `Proteobacteria` te se konvertira u brojčani izraz i pohranjuje u varijablu. Na taj način dobiveni su podaci koji su prikladni za daljnju obradu i vizualizaciju te se mogu pozvati u drugim programima.

Podaci o uzorcima sadrže taksonomski profil uzoraka, ali takve tablice su neprikladne za interpretaciju i daljnju obradu stoga je razvijen program BiomTable_class koji uspješno učitava i omogućuje ispis sažete tablice. Također, program BiomTable_class uspješno računa bogatstvo vrsta u uzorku. Podaci o brojnosti i bogatstvu uzorka organizirani su na način da se mogu koristiti u drugm programima u sklopu programskog alata. Iako je moguće sve

programe napisati u jedan dugačak progrm, to nije standardna praksa jer je otežano uklanjanje grešaka i navigiranje kroz programski kôd. Stoga je programski alat organiziran u više programa koji sadrže različite funkcije, a pozivom na program BiomTable_class dobivaju se podaci o uzorcima koji se koriste u ostalim programima.

4.2. Računalni program Diversity_class

Indeksi bioraznolikosti su kvantitativne mjere alfa i beta diverziteta. Za računanje indeksa bioraznolikosti razvijen je program Diversity_class. Metode programa Diversity_class kao ulazni argument koriste taksonomsku kategoriju (taksonomsku razinu) za koju se računa indeks.

Simpson indeks je mjera alfa diverziteta te na vrijednost utječe dominantno zastupljena vrsta. Metoda za računanje Simpson indeksa koristi podatke o brojnosti taksonomske razine pozivom na metodu `tax_abundance`:

```
for data in self.tax_abundance(taxonomy):
    d += np.power(data[1] / 100, 2)
return round(1 - d, 2)
```

Shannon indeks prikazuje stupanj podjednake raspodjele unutar uzorka. Metoda za računanje Shannon indeksa funkcionira na sličan način kao metoda za računanje Simpson indeksa:

```
for data in self.tax_abundance(taxonomy):
    h += data[1] / 100 * np.log(data[1] / 100)
return round(-h, 2)
```

Bogatstvo vrsta također je pokazatelj bioraznolikosti te predstavlja broj vrsta u danom uzorku. Program Diversity_class računa bogatstva vrsta te rezultat ispisuje u konzolu. U tablici 1 su prikazani primjeri izračunatih vrijednosti bogatstva vrsta te Simpson i Shannon indeksa.

Tablica 1. Primjeri izračunatih vrijednosti parametara alfa diverziteta na razini vrste

Redni broj uzorka	Bogatstvo vrsta	Simpson indeks	Shannon indeks
1.	64	0,91	2,89
2.	75	0,93	3,13
3.	55	0,91	2,95
4.	67	0,87	2,60
5.	78	0,89	2,75

Beta diverzitet predstavlja raznolikost između uzoraka, a jedan od pokazatelja je Jaccard indeks. Stoga je razvijena metoda za računanje Jaccard indeksa koja uspoređuje dva uzorka. U ugniježđenoj for petlji uspoređuju se nazivi vrste iz dva uzorka te ako se podudaraju njihove vrijednosti brojnosti se zbrajaju:

```
for species_1 in self.tax_abundance(taxonomy):
    for species_2 in abundance_2:
        if species_1[0] == species_2[0]:
            u += species_1[1] / 100
            v += species_2[1] / 100
        if u == 0 or v == 0:
            raise Exception("There are no shared species from
these two samples")
return round((u * v) / (u + v - u * v), 4)
```

Metoda za računanje Jaccard indeksa rezultat također ispisuje u konzolu te su u tablici 2 prikazani primjeri izračunatih vrijednosti Jaccard indeksa.

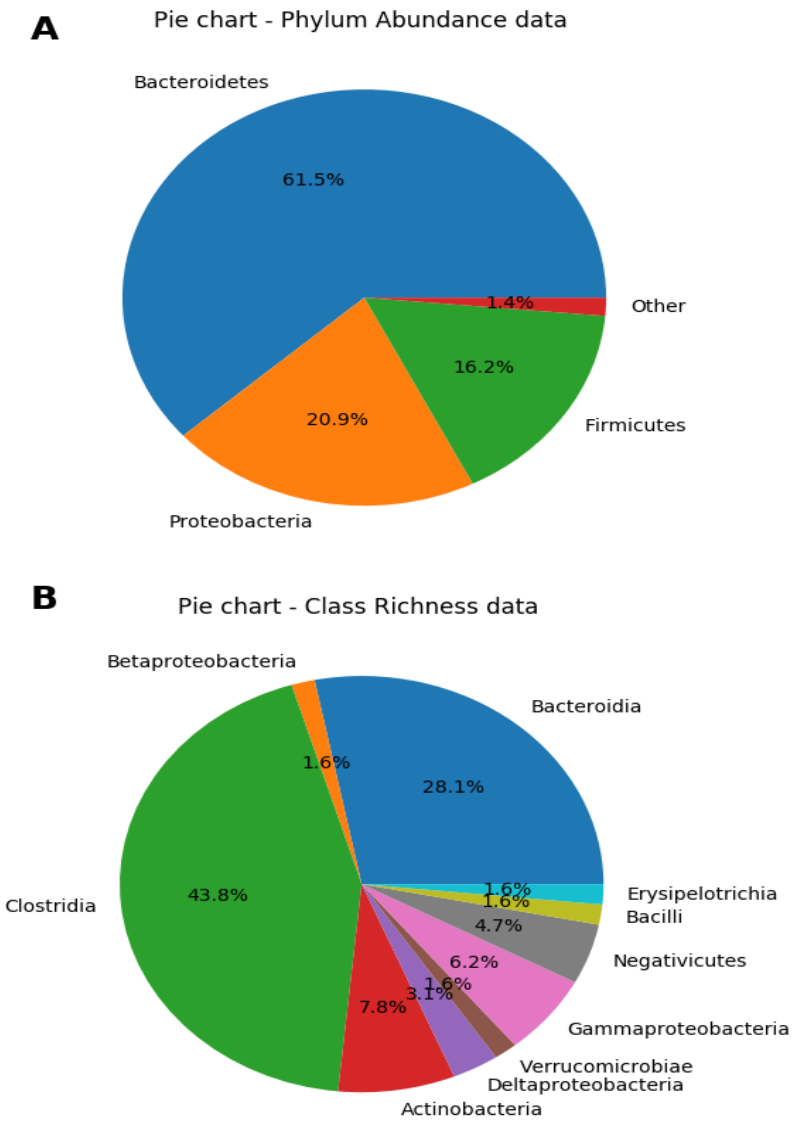
Tablica 2. Primjeri izračunatih vrijednosti Jaccard indeksa. Kolone i redovi su redni brojevi uzoraka za koje je izračunata vrijednost Jaccard indeksa

Redni broj uzorka	1.	2.	3.	4.	5.
1.	1,0000	0,5053	0,5486	0,5393	0,4723
2.	0,5053	1,0000	0,7612	0,7702	0,7593
3.	0,5486	0,7612	1,0000	0,7676	0,7496
4.	0,5393	0,7702	0,7676	1,0000	0,8519
5.	0,4723	0,7593	0,7496	0,8519	1,0000

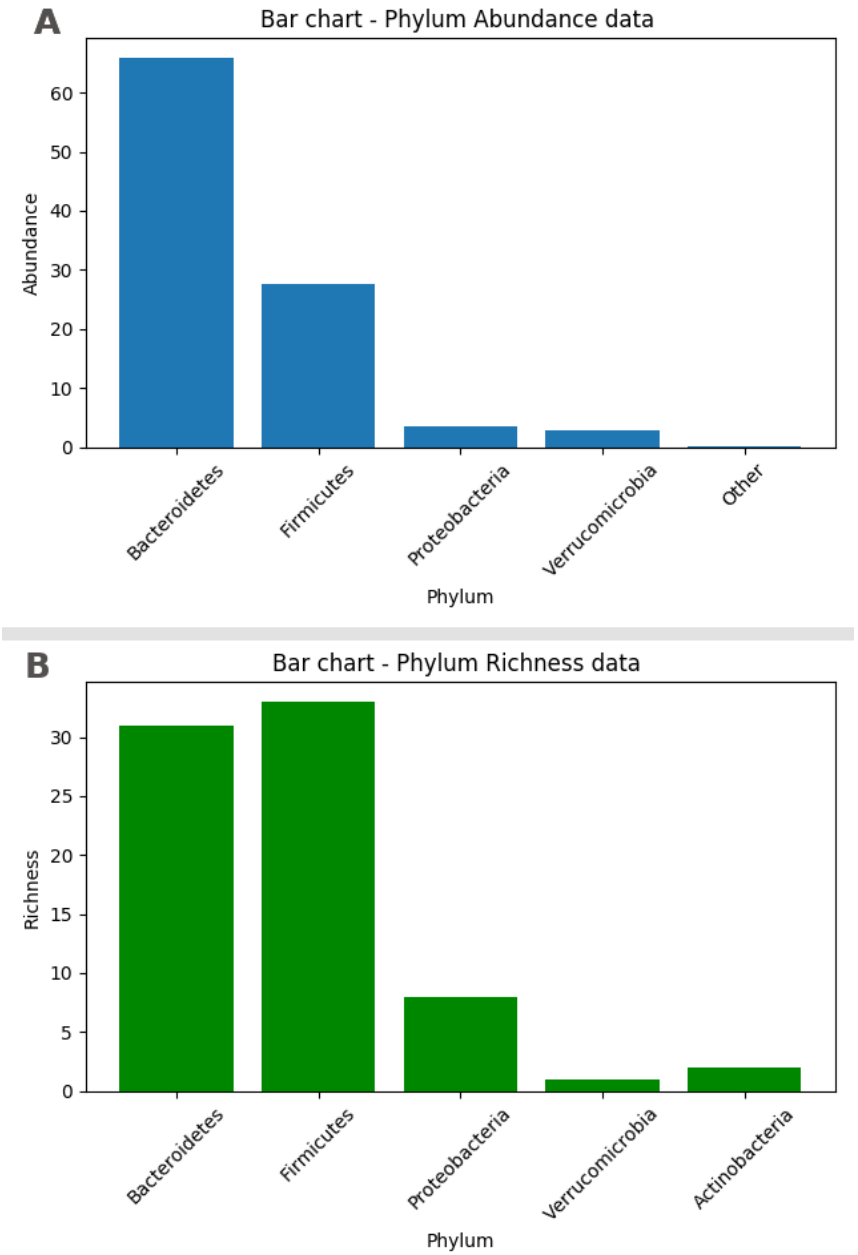
Sastav mikrobiote odnosno njene varijacije često se povezuju s raznim patološkim stanjima, a pomoću indeksa bioraznolikosti moguće je kvantificirati i uspoređivati te varijacije. Program Diversity_class uspješno računa mjere alfa diverziteta Simpson i Shannon indeks koji govore o raznolikosti unutar uzorka te Jaccard indeks koji je mjera beta diverziteta koji uspoređuje dva uzorka. Kako bi se mogao izvesti zaključak o uzorku na temelju indeksa bioraznolikosti potrebno je napraviti usporedbu između uzoraka jer vrijednost indeksa sam po sebi nema značenje. Program Diversity_class ispisuje vrijednost indeksa u konzolu stoga je potreban prijenos u tablični oblik kako bi se mogla provest usporedba (tablice 1 i 2). Nadalje, povezivanjem podataka o taksonomskom sastavu s drugim podacima, primjerice godinama, spolu, prehrambenim navikama, bolesti i slično bilo bi moguće izvesti korelaciju koja bi doprinijela razumijevanju uloge stabilnosti sastava crijevne mikrobiote. Također, moguće je nadograditi program s metodama za računanje dodatnih indeksa, primjerice Pielou, Chao1 ili Heip indeks. Ovisno o zahtjevima korisnika programa moguće je nadodati gotovo bilo koji indeks bioraznolikosti koji se računa pomoću brojnosti i bogatstva taksona.

4.3. Računalni program Biom_chart

Kako bi se grafički prikazalo bogatstvo i brojnost uzorka razvijen je program Biom_chart. Grafički prikaz rezultata je lakše interpretirati nego tablični prikaz stoga program ima opcije za izradu dva tipa grafa kružni i stupčasti. Funkcije za crtanje grafa primaju parametre za različito oblikovanje grafičkog prikaza (slika 9 i slika 10). Algoritam na temelju tipa podataka (brojnost ili bogatstvo) i taksonomske razine dohvaća odgovarajuće podatke pomoću BiomTable_class programa. Na taj način moguće je izrađivati različite grafičke prikaze te ih uspoređivati.



Slika 9. Primjeri grafičkog prikaza kružnog dijagrama izrađenih programom Biom_chart. A) Prikaz brojnosti uzroka na razini koljena B) prikaz bogatstva uzorka na razini razreda



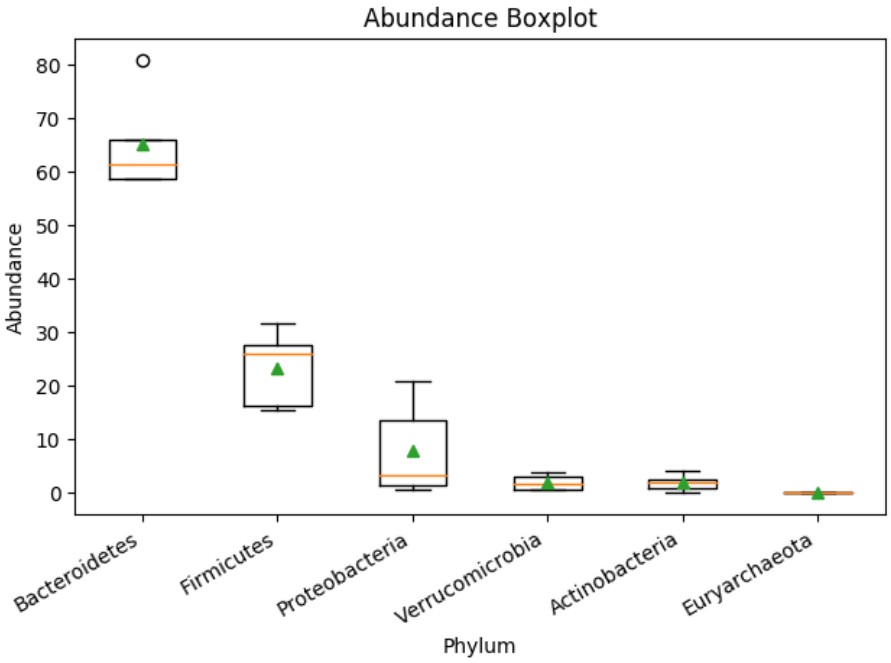
Slika 10. Primjeri grafičkog prikaza stupčastog dijagrama. A) Stupčasti dijagram brojnosti na razini koljena B) Stupčasti dijagram bogatstva na razini koljena. Iz ovog grafičkog prikaza moguće je zaključiti odnos bogatstva vrsta i brojnosti na razini koljena

Grafički prikaz podataka je važan korak prilikom interpretacije podataka te postoji mnogo različitih programa za izradu grafičkih prikaza podataka među kojima je najpoznatiji Excel. Iako programi specijalizirani za izradu grafova imaju mnogo više opcija i funkcionalnosti koji se mogu dodati grafičkom prikazu, u svrhu automatiziranog i

jednostavnog prikazivanja taksonomskih podataka u sklopu programskog alata uspješno je razvijen program Biom_chart. Razmatranjem varijabli brojnosti i bogatstva moglo bi se zaključiti da taksoni poprimaju slične vrijednosti, ali to nije uvijek slučaj. Iz grafičkog prikaza brojnosti taksona uzorka na razini koljena (slika 10) može se primijetiti da *Bacteroidetes* ima značajno veću brojnost od ostalih taksona dok u prikazu bogatstva taksona može se primjetiti da *Firmicutes* poprima najveću vrijednost bogatstva. Iz ovakvog prikaza rezultata mogu se izvesti zaključci o dijeti donora uzorka i stabilnosti crijevnog mikrobioma. Primjerice, visoke vrijednosti brojnosti *Bacteroidetes* taksona, koji su tolerantni na visoke koncentracije žući, povezani su s dijetom temeljenom na životinjskom tkivu (David i sur., 2014).

4.4. Računalni program Box_plot

U sklopu programskog alata razvijen je program Box_plot za statističku obradu taksonomskih podataka te prikaz pomoću kutijastog dijagrama. Kutijasti dijagram izrađuje se na temelju uređene petorke: minimum, maksimum, donji interkvartil i gornji interkvartil te medijan. Linije kutijastog dijagrama (slika 11) predstavljaju sve podatke obuhvaćene u danom uzorku, a pravokutnik (kutija) predstavlja interval u kojem se nalazi 50 % podataka (interkvartilni rapon). Širina interkvartilnog raspona predstavlja varijabilnost uzorka odnosno što je širi raspon to je veća varijabilnost. Program za izradu kutijastog dijagrama Box_plot ima opciju za označavanje srednje vrijednosti te iz odnosa medijana i srednje vrijednosti može se odrediti simetričnost seta podataka. Nadalje, program Box_plot prilikom izrade kutijastog dijagrama izbacuje ekstremne vrijednosti koje su označene kao točke.



Slika 11. Grafički prikaz vrijednosti brojnosti taksona te usporedba medijana i interkvartilnog raspona koristeći kutijasti dijagram. Na apscisi se nalaze taksoni na razini koljena, a na ordinati se nalazi brojnost

Za razliku od programa za računanje indeksa bioraznolikosti, prikaza kružnog i stupčastog dijagrama gdje je potrebno u programe učitati podatke o jednom uzorku, prilikom izrade kutijastog dijagrama potrebno je učitati više od jednog uzorka. Nadalje, potrebno je odrediti medijan i interkvartilne vrijednosti seta uzoraka. Stoga, funkcija za izradu kutijastog dijagrama za argument uzima naziv direktorija u kojem se nalazi set uzoraka. Iako program

može prikazati kutijasti dijagram na temelju jednog uzorka, za izradu statistički opravdanog dijagrama potrebno je imati što veći broj uzoraka. Osim naziva direktorija, funkcija za argument uzima i tip podataka brojnost ili bogatstvo na temelju kojih se izrađuje dijagram.

Podaci učitani u metodu za izradu kutijastog dijagrama organizirani su na način da su nazivi taksona u jednoj listi, a njihove vrijednosti brojnosti ili bogatstva u zasebnoj odgovarajućoj listi. Dio programa koji provodi sortiranje podataka:

```
for data in data_set:
    if data[0] not in possible_labels:
       possible_labels.append(data[0])
       sorted_data.append([])
       sorted_data[0] = possible_labels
       sorted_data[possible_labels.index(data[0])             +
1].append(data[1])
```

Učitani uzorci sadrže ime taksonomske skupine i odgovarajuću brojnost ili bogatstvo. Sortirani podaci u prvom redu sadrže imena taksonomskih skupina, a svaki slijedeći red sadrži vrijednosti s time da svaki red sadrži podatke odgovarajućeg rednog broja taksonomske skupine iz prvog reda (slika 8). Algoritam na temelju imena taksonomske skupine sprema vrijednost u odgovarajući red u sortiranoj tablici. Ako taksonomska skupina nije navedena u prvom redu, onda se ona sprema na zadnje mjesto u prvom redu te se na dnu tablice dodaje novi red. Na taj način u svega nekoliko linija koda moguće je izraditi novu varijablu sa sortiranim podacima iz kojih je dalje jednostavno izraditi kutijasti dijagram.

Važan korak u obradi bioloških podataka je i statistička analiza. Program Box_plot uspješno sortira taksonomske podatke i izrađuje kutijasti dijagram u kojem su prikazani raspon uzorka, medijan te srednja vrijednost (slika 11). U svrhu prikaza načina rada programa korišteno je 5 uzoraka, ali za statistički smislenu interpretaciju potrebno je imati što veći broj uzoraka. Kako bi se taksonomski podaci učitali u program moraju biti pohranjeni u direktoriju koji se nalazi u istom direktoriju kao program. U tu svrhu potrebno je preuzeti opsežnu bazu taksonomskih podataka na računalo. Ovakav način pohrane podataka je jednostavan, ali problematičan prilikom ažuriranja podataka. Za svaku nadopunu baze podataka bilo bi potrebno „ručno“ dodavati nove taksonomske podatke u direktorij. Alternativno, programski alat mogao bi se adaptirati da ima pristup internetskoj bazi

taksonomskih podataka. Na taj način korisnici programskog alata imali bi pristup istim podacima i time bi rezultati analiza bili uniformni. Nadalje, prilikom povećanja količine podataka koje program obrađuje potrebno je voditi računa o vremenu potrebnom za izvršavanje tih operacija. U svrhu izrade alata u sklopu rada koristilo se 5 primjera taksonomskih podataka, stoga se duljina vremena izvršavanja programa nije razmatrala. Prilikom programiranja moguće je napraviti određeni zadatak na više načina te ako je potrebno moguće je provesti optimizaciju kôda u svrhu skraćivanja vremena izvršavanja programa.

4.5. Računalni program PathwayTable_class

Za učitavanje i filtriranje podataka o metaboličkim putevima dobivenim HUMAnN alatom razvijen je program PathwayTable_class. Podaci o metaboličkim putevima su također CSV formata i sadrže jedan tabulatorski graničnik te sadrže ime metaboličkog puta i odgovarajuću vrijednost brojnosti. Stoga, za učitavanje se koristi CSV standardna Python biblioteka koja sadrži metode za učitavanje i raščlambu podataka. Nakon raščlambe, filtriraju se i izbacuju metabolički putevi koji su specifični za bakterijske vrste te putevi koji nisu integrirani i mapirani. Program za rezultat vraća listu metaboličkih puteva, odgovarajuću vrijednost brojnosti i udio metaboličkog puta. Rezultat se ispisuje u konzolu te su u tablici 3 navedeni deset najzastupljenijih metaboličkih puteva za dva uzorka. Program ima mogućnost ispisa čitave tablice, ali čak i nakon filtriranja uzorci sadrže preko 250 metaboličkih puteva koje je nepregledno prikazivati pomoću tablice. Nadalje, može se zamijetiti da su udjeli metaboličkih puteva vrlo mali (< 3 %) te da nema metaboličkog puta koji značajno odstupa u brojnosti.

Tablica 3. Primjeri deset najzastupljenijih metaboličkih puteva za dva uzorka i njihov udio

	Uzorak 1		Uzorak 2	
Redni broj	Naziv metaboličkog puta	%	Naziv metaboličkog puta	%
1.	Adenozin ribonukleotid *de novo* biosinteza	2,61	Adenozin ribonukleotid *de novo* biosinteza	2,49
2.	UDP-N-acetilmuramoil-pentapeptid biosinteza II	2,00	UDP-N-acetilmuramoil-pentapeptid biosinteza II	2,28
3.	UMP biosinteza	1,99	Peptidoglikan biosinteza I	2,28
4.	UDP-N-acetilmuramoil-pentapeptid biosinteza I	1,97	UDP-N-acetilmuramoil-pentapeptid biosinteza I	2,24
5.	Peptidoglikan biosinteza I	1,96	trna punjenje	1,91
6.	Koenzime A biosinteza II	1,81	Gvanozin ribonukleotid *de novo* biosinteza	1,80
7.	Metileritritol fosfatni put I	1,80	Metileritritol fosfatni put I	1,78
8.	Gvanozin ribonukleotid *de novo* biosinteza	1,77	Degradacija škroba V	1,74
9.	S-adenozil-L-metionin ciklus I	1,65	Glikoliza IV	1,74
10.	Urat biosintheza/inozine 5'-fosfat degradacija	1,62	Urat biosintheza/inozine 5'-fosfat degradacija	1,71

Prisutnost i brojnost mikrobnih metaboličkih puteva, kao i taksonomski profil, ustvrđuje se meagenomskom analizom. Razmatranjem metaboličkih puteva na temelju metagenomskih podataka se dodaje nova dimenzija podacima o uzorku. Usporedbom metaboličkih puteva moguće je odrediti koji putevi su karakteristični s određenim patološkim stanjem. Primjerice u radu koji su objavili Jiang i suradnici 2020. godine pokazuje se snažna korelacija povećanja udjela određenih metaboličkih puteva s visokim udjelom humane DNA što se često pojavljuje kod pacijenata s rakom debelog crijeva i Chronove boesti. Program PathwayTable_class razvijen u sklopu programskog alata sadrži osnovne operacije otvaranja, filtriranja i ispis metaboličkih podataka. Takav prikaz podataka nije baš efikasan jer su podaci opsežni i teško ih je proučavati. Dakle, ako želimo nadalje analizirati metaboličke puteve potrebno ih je na neki način grupirati ili dodatno filtrirati kako bi se mogli grafički prikazati ili povezati s većom bazom podataka o metaboličkim putevima kako bi se mogla izvesti

statistička analiza. Također, ako su poznati dodatni podaci o donorima uzorka poput dobi, spola, zdravstvenog stanja ili prehrambene navike moglo bi se zaključiti kako ti čimbenici utječu na zastupljenost metaboličkih puteva.

4.6. Usporedba programskog alata s postojećim bioinformatičkim alatima

Danas je prisutno mnogo bioinformatičih softvera za analizu mikrobioma koji su otvorenih izvora što znači da su dostupni svima. Neki od alata su QIIME2 (Bolyen i sur., 2019), mothur (Schloss i sur., 2009) i phyloseq (McMurdie i Holmes, 2013). QIIME2 je jedana od poznatijih bioinformatičkih platformi koja se ističe po alatima za interaktivnu vizualizaciju podataka. Mothur alat foksurian je na generiranje podataka koje je jednostavno vizualizirati, a phyloseq alat je fokusiran na statističkoj analizi podataka o mikrobiomu. Programski alat razvijen u sklopu ovog diplomskog rada je također otvorenog kôda te se nalazi na internet repozitoriju GitHub (https://github.com/PKatarina/UMCG-Microbiome) gdje bilo tko može kopirati kôd, tražiti greške i predlagati promjene.

Programski alat razvijen u sklopu ovog diplomskog rada je relativno jednostavan i ne sadrži mnogo funckionalnosti u usporedbi sa softverima koji su prisutni na internetu. Međutim, poznavanjem programskog jezika moguće je razvijati alat u skladu s vlastitim potrebama. Nadalje, povezivanjem različitih programskih jezika moguće je optimizirati kôd. Primjerice, korištenjem programskog jezika SQL koji je specijaliziran za upravljanje tablicama u bazama podataka moglo bi se olakšati sortiranje i filtriranje taksonomskih i metaboličkih podataka. Nadalje, korištenjem programskih jezika specijaliziranih za prikaz u internetskom pregledniku moguće je izraditi web aplikaciju koja bi sadržavala grafičko sučelje koje omogućava korištenje korisnicima koji ne poznaju programske jezike.

5. ZAKLJUČCI

Upotrebom programskog jezika Python razvijen je programski alat koji učitava i obrađuje taksonomske i metaboličke podatke te sadrži pet programa:

1. BiomTable_class program učitava i raščlanjuje taksonomske podatke te ih sprema u oblik kojeg koriste ostali programi.

2. Divesrity_class program računa mjere alfa diverziteta Shannon indeks i Simpson indeks te mjeru beta diverziteta Jaccard indeks.

3. Biom_chart program grafički prikazuje brojnosti i bogatstvo taksonomskih podataka.

4. Box_plot program provodi statističku analizu seta taksonomskih podataka te ih grafički prikazuje kutijastim dijagramom.

5. PathwayTable_class učitava i filtrira podatke o metaboličkim putevima.

6. LITERATURA

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., White, O., Kelley, S. T., Methé, B., Schloss, P. D., Gevers, D., Mitreva, M., Huttenhower, C. (2012) Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLoS Comput. Biol.* **8**.

Andrews, S. (2010) Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Pristupljeno 19. veljače 2020.

Beal, J., Phillips, A., Densmore, D., Cai, Y. (2011) High-Level Programming Languages for Biomolecular Systems. U: Design and Analysis of Biomolecular Circuits, (Koeppl, H., Setti, G., Bernardo M. di, Densmore, D. ured.), Springer New York, New York. str. 225–252.

Berry, D., Stecher, B., Schintlmeister, A., Reichert, J., Brugiroux, S., Wild, B., Wanek, W., Richter, A., Rauch, I., Decker, T., Loy, A., Wagner, M. (2013) Host-compound foraging by intestinal microbiota revealed by single-cell stable isotope probing. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 4720–4725.

Birkel, G. W., Ghosh, A., Kumar, V. S., Weaver, D., Ando, D., Backman, T. W. H., Arkin, A. P., Keasling, J. D., Martín, H. G. (2017) The JBEI quantitative metabolic modeling library (jQMM): a python library for modeling microbial metabolism. *BMC Bioinformatics* **18**, 205.

Bleiel, N. (2016) Collaborating in GitHub. U: 2016 Ieee International Professional Communication Conference (Ipcc) Ieee, New York.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E.,

Rasmussen, L. B., Rivers, A., Robeson, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J. G. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857.

Braniste, V., Al-Asmakh, M., Kowal, C., Anuar, F., Abbaspour, A., Tóth, M., Korecka, A., Bakocevic, N., Ng, L. G., Kundu, P., Gulyás, B., Halldin, C., Hultenby, K., Nilsson, H., Hebert, H., Volpe, B. T., Diamond, B., Pettersson, S. (2014) The gut microbiota influences blood-brain barrier permeability in mice. *Sci. Transl. Med.* **6**, 263ra158-263ra158.

Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M., Owen, L. J. (2015) Dysbiosis of the gut microbiota in disease. *Microb. Ecol. Health Dis.* **26**, 26191.

Carola, S., Rolf, D. (2011) Metagenomic Analyses: Past and Future Trends. *Appl. Environ. Microbiol.* **77**, 1153–1161.

Clarke, G., Stilling, R. M., Kennedy, P. J., Stanton, C., Cryan, J. F., Dinan, T. G. (2014) Minireview: Gut Microbiota: The Neglected Endocrine Organ. *Mol. Endocrinol.* **28**, 1221–1238.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J., Turnbaugh, P. J. (2014) Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563.

Donaldson, G. P., Lee, S. M., Mazmanian, S. K. (2016) Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20–32.

Ekmekci, B., McAnany, C. E., Mura, C. (2016) An Introduction to Programming for Bioscientists: A Python-Based Primer. *PLOS Comput. Biol.* **12**, e1004867.

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., Kurilshikov, A., Bonder, M. J., Valles-Colomer, M., Vandeputte, D., Tito, R. Y., Chaffron, S., Rymenans, L., Verspecht, C., De Sutter, L., Lima-Mendez, G., D'hoe, K., Jonckheere, K., Homola, D., Garcia, R., Tigchelaar, E. F., Eeckhaudt, L., Fu, J., Henckaerts, L., Zhernakova, A.,

Wijmenga, C., Raes, J. (2016) Population-level analysis of gut microbiome variation. *Science* **352**, 560–564.

Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P., Forano, E. (2012) Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* **3**, 289–306.

Foster, J. A., McVey Neufeld, K.-A. (2013) Gut–brain axis: how the microbiome influences anxiety and depression. *Trends Neurosci.* **36**, 305–312.

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., Huttenhower, C. (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968.

Ghosh, A., Mehta, A., Khan, A. M. (2019) Metagenomic Analysis and its Applications. In Encyclopedia of Bioinformatics and Computational Biology, Elsevier, str. 184–193.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., Nelson, K. E. (2006) Metagenomic Analysis of the Human Distal Gut Microbiome. *Science* **312**, 1355–1359.

GitHub (2020) Build software better, together. <https://github.com> Pristupljeno 17. ožujka 2020.

Haagsman, E. (2017) PyCharm Community Edition and Professional Edition Explained: Licenses and More. <https://www.jetbrains.com/> Pristupljeno: 11. kolovoza 2019.

Heintz-Buschart, A., Wilmes, P. (2018) Human Gut Microbiome: Function Matters. *Trends Microbiol.* **26**, 563–574.

Humphries, C. J., Williams, P. H., Wright, R. I. V. (1995) Measuring Biodiversity Value for Conservation. *Annu. Rev. Ecol. Syst.* **26**, 93–111.

Huttenhower, C. (2012) Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214.

Imhann, F., Van der Velde, K. J., Barbieri, R., Alberts, R., Voskuil, M. D., Vich Vila, A., Collij, V., Spekhorst, L. M., Van der Sloot, K. W. J., Peters, V., Van Dullemen, H. M., Visschedijk, M. C., Festen, E. A. M., Swertz, M. A., Dijkstra, G., Weersma, R. K. (2019) The 1000IBD project: multi-omics data of 1000 inflammatory bowel disease patients; data release 1. *BMC Gastroenterol.* **19**, 5.

Jemerov, D. (2013) PyCharm 3.0 Community Edition source code now available. <https://www.jetbrains.com/> Pristupljeno 11. kolovoza 2019.

Jiang, P., Lai, S., Wu, S., Zhao, X.-M., Chen, W.-H. (2020) Host DNA contents in fecal metagenomics as a biomarker for intestinal diseases and effective treatment. *BMC Genomics* **21**.

Kanehisa, M., Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38**, D355–D360.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., Tanabe, M. (2019) New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595.

kneadData (2019) biobakery / kneadData / wiki / Home — Bitbucket. <https://bitbucket.org/biobakery/kneaddata/wiki/Home> Pristupljeno 19. veljače 2020.

Koropatkin, N. M., Cameron, E. A., Martens, E. C. (2012) How glycan metabolism shapes the human gut microbiota. *Nat. Rev. Microbiol.* **10**, 323–335.

Krajina T. (2019) Uvod u git. <https://github.com/tkrajina/uvod-u-git> Pristupljeno 17. ožujka 2020.

Laudadio, I., Fulci, V., Palone, F., Stronati, L., Cucchiara, S., Carissimi, C. (2018) Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. *OMICS.* **22**, 248–254.

Lepage, P., Häsler, R., Spehlmann, M. E., Rehman, A., Zvirbliene, A., Begun, A., Ott, S., Kupcinskas, L., Doré, J., Raedler, A., Schreiber, S. (2011) Twin Study Indicates Loss of Interaction Between Microbiota and Mucosa of Patients With Ulcerative Colitis. *Gastroenterology* **141**, 227–236.

Liu, S. (2016) The Development of Our Organ of Other Kinds—The Gut Microbiota. *Front. Microbiol.* **7**.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., Knight, R. (2012) Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230.

Lutz, M. (2013) Learning Python. O'Reilly, Peking.

Magurran, A. E. (1998) Measuring richness and evenness. *Trends Ecol. Evol.* **13**, 165–166.

Manichanh, C., Borruel, N., Casellas, F., Guarner, F. (2012) The gut microbiota in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 599–608.

Marchesi, J. R., Adams, D. H., Fava, F., Hermes, G. D. A., Hirschfield, G. M., Hold, G., Quraishi, M. N., Kinross, J., Smidt, H., Tuohy, K. M., Thomas, L. V., Zoetendal, E. G., Hart, A. (2016) The gut microbiota and host health: a new clinical frontier. *Gut* **65**, 330–339.

Matplotlib (2020) Matplotlib: Python plotting — Matplotlib 3.2.1 documentation.

McIver, L. J., Abu-Ali, G., Franzosa, E. A., Schwager, R., Morgan, X. C., Waldron, L., Segata, N., Huttenhower, C. (2018) bioBakery: a meta'omic analysis environment. *Bioinformatics* **34**, 1235–1237.

McMurdie, P. J., Holmes, S. (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8**, e61217.

Molodecky, N. A., Soon, I. S., Rabi, D. M., Ghali, W. A., Ferris, M., Chernoff, G., Benchimol, E. I., Panaccione, R., Ghosh, S., Barkema, H. W., Kaplan, G. G. (2012) Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* **142**, 46-54.

Morrison, D. J., Preston, T. (2016) Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* **7**, 189–200.

Noor, E., Cherkaoui, S., Sauer, U. (2019) Biological insights through omics data integration. *Curr. Opin. Systems Biol.* **15**, 39–47.

Numpy (2020) NumPy — NumPy. < https://numpy.org/> Pristupljeno 20. travnja 2020.

Python (2020) Welcome to Python.org. < https://www.python.org/> Pristupljeno 20. travnja 2020.

Qin, J., Li, Y., Cai, Z., Li, Shenghui, Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, Songgang, Yang, H., Wang, Jian, Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K., Wang, J. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., Segata, N. (2017) Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844.

Real, R., Vargas, J. M. (1996) The Probabilistic Basis of Jaccard's Index of Similarity. *Systematic Biol.* **45**, 380–385.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Horn, D. J. V., Weber, C. F. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* **75**, 7537–7541.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814.

Seksik, P., Landman, C. (2015) Understanding Microbiome Data: A Primer for Clinicians. *Digest. Dis.* **33**, 11–16.

Sender, R., Fuchs, S., Milo, R. (2016) Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* **14**.

Sharon, G., Sampson, T. R., Geschwind, D. H., Mazmanian, S. K. (2016) The Central Nervous System and the Gut Microbiome. *Cell* **167**, 915–932.

Singh, B., Crippen, T. L., Tomberlin, J. K. (2017) An introduction to metagenomic data generation, analysis, visualization, and interpretation. U: Forensic Microbiology, (Carter, D. O., Tomberlin, J. K., Benbow M. E., Metcalf J. L.,ured.), John Wiley & Sons, Ltd, Chichester, str. 94–126.

Sudarikov, K., Tyakht, A., Alexeev, D. (2017) Methods for The Metagenomic Data Visualization and Analysis. *Curr. Issues Mol. Biol.* **24**, 37-58.

Taft, D. K. (2010) JetBrains Strikes Python Developers with PyCharm 1.0 IDE. <https://www.eweek.com/> Pristupljeno: 18. listopada 2019.

Thomas, T., Gilbert, J., Meyer, F. (2012) Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* **2**, 3.

Thursby, E., Juge, N. (2017) Introduction to the human gut microbiota. *Biochem. J.* **474**, 1823–1836.

Tropini, C., Earle, K. A., Huang, K. C., Sonnenburg, J. L. (2017) The gut microbiome: Connecting spatial organization to function. *Cell Host Microbe* **21**, 433–442.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., Segata, N. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., Gordon, J. I. (2007) The Human Microbiome Project. *Nature* **449**, 804–810.

van der Walt, S., Colbert, S. C., Varoquaux, G. (2011) The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13**, 22–30.

van Rossum, G. (2009) Guido's Personal Home Page. <https://gvanrossum.github.io//> Pristupljeno: 19. veljače 2020.

van Rossum G., Warsaw B., Coghlan N. (2001) PEP 8 -- Style Guide for Python Code. <https://www.python.org/dev/peps/pep-0008> Pristupljeno: 20. travnja 2020.

Whittaker, R. H. (1972) Evolution and Measurement of Species Diversity. *Taxon.* **21**, 213–251.

IZJAVA O IZVORNOSTI

Izjavljujem da je ovaj diplomski rad izvorni rezultat mojeg rada te da se u njegovoj izradi nisam koristio/la drugim izvorima, osim onih koji su u njemu navedeni.

_____

Ime i prezime studenta