Biosintetski potencijal šest novo sekvencioniranih bakterija roda Streptomyces

Vuković, Andro

Master's thesis / Diplomski rad

2023

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:159:893740

Rights / Prava: Attribution-NoDerivatives 4.0 International/Imenovanje-Bez prerada 4.0 međunarodna

Download date / Datum preuzimanja: **2025-03-31**

prehrambeno biotehnološki fakultet

Repository / Repozitorij:

Repository of the Faculty of Food Technology and Biotechnology

zir.nsk.hr

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

SVEUČILIŠTE U ZAGREBU
PREHRAMBENO-BIOTEHNOLOŠKI FAKULTET

DIPLOMSKI RAD

Zagreb, mjesec godina	Andro Vuković

BIOSINTETSKI POTENCIJAL ŠEST NOVO SEKVENCIONIRANIH BAKTERIJA RODA *Streptomyces*

Rad je izrađen u Laboratoriju za bioinformatiku na Zavodu za biokemijsko inženjerstvo Prehrambeno-biotehnološkog fakulteta Sveučilišta u Zagrebu pod mentorstvom prof.dr.sc. Antonia Starčevića u sklopu projekta ZCI BioProspecting Jadranskog mora (KK.01.1.1.01)

Projekt: Znanstveni centar izvrsnosti za bioprospecting mora BioProCro

Broj projekta: KK.01.1.1.01

Izvor financiranja: ERDF-EU fond

Voditeljica: dr. sc. Rozelindra Čož-Rakovac

ZAHVALA

Zahvaljujem se mentoru prof. dr. sc. Antoniu Starčeviću na pomoći i savjetima pri izradi diplomskog rada. Također se zahvaljujem svojoj obitelji i prijateljima na njihovoj podršci i motivaciji jer su to ljudi koji su uvijek bili uz mene i vjerovali u mene :)

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

Diplomski sveučilišni studij: Molekularna biotehnologija

BIOSINTETSKI POTENCIJAL ŠEST NOVO SEKVENCIONIRANIH BAKTERIJA RODA *Streptomyces*

Andro Vuković, univ. bacc.ing. biotechn.
0058210824

Sažetak: Bakterije roda *Streptomyces* su Gram-pozitivne aerobne bakterije koje tijekom vegetativne faze svog života proizvode mnogobrojne sekundarne metabolite. Oni nisu nužni za rast stanice, ali joj daju određene prednosti nad drugim organizmima u staništu. Mnogi sekundarni metaboliti korisni su za čovječanstvo jer se upotrebljavaju u raznim područjima života. Cilj ovog rada bio je analizirati genome šest bakterija iz roda *Streptomyces* koje su izolirane iz Jadranskog mora te utvrditi sadrže li one sekundarne metabolite koji pokazuju biotehnološki potencijal. Rezultat AntiSMASH analize pokazao je da ovi genomi sadrže brojne genske nakupine koje sudjeluju u sintezi važnih sekundarnih metabolita te da bi se ove bakterijske vrste mogle koristiti za industrijsku proizvodnju potencijalno korisnih proizvoda.

Ključne riječi: *sekundarni metaboliti, anotacija, sastavljanje genoma, usporedna analiza*

Rad sadrži: 42 stranica, 15 slika, 7 tablica, 25 literaturnih navoda, 6 priloga
Jezik izvornika: hrvatski
Rad je u tiskanom i elektroničkom (pdf format) obliku pohranjen u: Knjižnica Prehrambeno-biotehnološkog fakulteta, Kačićeva 23, Zagreb
Mentor: prof. dr. sc. Antonio Starčević

Stručno povjerenstvo za ocjenu i obranu:

1. prof. dr. sc.Anita Slavica (predsjednik)
2. prof. dr. sc. Antonio Starčević (mentor)
3. prof. dr. sc. Igor Slivac (član)*
4. izv. prof. dr. sc. Jurica Žučko (zamjenski član)

Datum obrane:

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

Graduate university study programme: Molecular Biotechnology

BIOSYNTHETIC POTENTIAL OF SIX NEWLY SEQUENCED BACTERIA OF THE *Streptomyces* GENUS

Andro Vuković, univ. bacc.ing. biotechn.
0058210824

Abstract: Bacteria of the *Streptomyces* genus are Gram-positive aerobic bacteria that produce secondary metabolites during the vegetative phase of their life cycle. These metabolites are not necessary for cell growth, but they give the organism certain advantages over other organisms in their habitat. Numerous secondary metabolites are useful for mankind because they are used in various industries. The aim of this work was to analyze the genomes of six bacteria from the *Streptomyces* genus that were isolated from the Adriatic Sea and to determine whether they contain secondary metabolites that exhibit potential to be utilized in biotechnology. The result of the AntiSMASH analysis showed that these genomes contain gene clusters that participate in the synthesis of important secondary metabolites and that these bacterial species could be used for the industrial production of potentially very important products.

Keywords: *secondary metabolites, annotation, assembly, analysis*
Thesis contains: 42 pages, 15 figures, 7 tables, 25 references, 6 supplements
Original in: Croatian
Graduate Thesis in printed and electronic (pdf format) form is deposited in: The Library of the Faculty of Food Technology and Biotechnology**,** Kačićeva 23, Zagreb.
Mentor: Antonio Starčević, PhD, Full professor

Reviewers:

1. Anita Slavica, PhD, Full professor (president)
2. Antonio Starčević, PhD, Full professor (mentor)
3. Igor Slivac, PhD, Full professor (member)
4. Jurica Žučko, PhD, Associate professor (substitute)

Thesis defended:

Sadržaj

1. UVOD

Bakterije roda Streptomyces Gram-pozitivne su aerobne bakterije koje rastu na velikom broju različitih staništa. Tijekom svog života one prolaze kroz vegetativnu fazu i fazu stvaranja sekundarnih metabolita. Tijekom vegetativne faze one rastu grananjem vrhova hifa i formiraju micelij, zbog čega su slične filamentoznim fungima. Sekundarni metaboliti se stvaraju kao odgovor stanice na stres. Oni nisu nužni za rast stanice nego joj daju neke prednosti u odnosu na ostale organizme. Sa biotehnološkog stajališta, ove bakterije su bitne jer tijekom vegetativne faze proizvode vitamine i enzime za komercijalnu upotrebu, a tijekom faze proizvodnje sekundarnih metabolita proizvode različite korisne spojeve poput antibiotika i pigmenata (Ferraiuolo i sur., 2021). Sekundarni metaboliti mogu se podijeliti u dvije velike skupine: poliketidi (PKS) i neribosomalno sintetizirani polipeptidi (NRPS). Svaka od ovih skupina sadrži spojeve koji su sastavljeni od velikog broja manjih građevnih jedinica koje se mogu ugraditi u konačni produkt različitim redoslijedom te postoje različite modifikacije istih, što rezultira velikim brojem mogućih konačnih produkata (Rimac, 2013).

Pretjerana upotreba antibiotika dovela je do pojave problema rezistencije patogenih bakterija na antibiotike. Jedno od predloženih rješenja tog problema je sekvencioniranje i anotiranje genoma u svrhu pronalaženja gena koji kodiraju za enzime koji sudjeluju u sintezi novih antibiotika (Rimac, 2013). Dugo vremena je jedina metoda sekvencioniranja bila Sangerova metoda koja se temelji na ugradnji florescentno obilježenih dideoksinukleotida u rastuću molekulu DNA. Pošto ddNTP nema 3'-OH skupinu slijedeći dNTP se ne može vezati na njega i sinteza DNA lanca se tada prekida. Svaki od 4 postojećih ddNTP-ova obilježen je drugačijom florescencijom i na temelju toga se može iščitati sekvenca DNA. Potreba za boljom i ekonomičnijom metodom sekvencioniranja dovela je do razvoja NGS (Next Generation Sequencing). NGS ima mogućnost paralelnog procesiranja milijuna fragmenata sekvenci umjesto 96 njih odjednom, što je značajno smanjilo cijenu sekvencioniranja. Postoje različite platforme za NGS kao što su Roche 454, Illumina Solexa i Pacific Biosciences SMRT (Quail i sur., 2012).

Cilj ovog rada bio je istražiti sekundarni metabolizam šest novo sekvencioniranih bakterija iz roda *Streptomyces* te utvrditi stvaraju li one sekundarne metabolite koji bi mogli biti korisni u biotehnologiji.

2. TEORIJSKI DIO

2.1. SEKUNDARNI METABOLITI MIKROORGANIZAMA

Mikrobni sekundarni metaboliti produkti su sekundarnog metabolizma male molekulske mase koji se obično proizvode tijekom kasne faze razvoja (idiofaze). Sekundarni metaboliti nisu nužni za rast kultura koje ih proizvode, nego služe za različite funkcije preživljavanja u prirodi. Iako nisu nužni za mikrobni rast, sekundarni metaboliti jako su važni za zdravlje, prehranu i ekonomiju našeg društva.

Vjerojatno najvažnija upotreba sekundarnih metabolita je upotreba u anti-infektivnim lijekovima. Ako moderna medicina želi nastaviti u svom sadašnjem obliku, nove obitelji antibiotika moraju dolaziti na tržište u redovitim intervalima. U zadnje vrijeme agresivniji *screening* programi za odabir novih prirodnih i kemijskih spojeva nužni su za proizvodnju novih antibiotika protiv rezistentnih bakterija. Uz to, farmaceutska industrija je tijekom posljednjih desetljeća proširila *screening* programe na ostale bolesti poput lijekova za sniženje kolesterola i lijekova protiv raka.

Glavni biosintetski putevi uključeni u sekundarni metabolizam su oni koji formiraju aromatske spojeve, izoprene, oligosaharide, peptide, poliketide i β-laktamske prstenove. Znanje o biosintetskim putevima varira od slučajeva u kojima su aminokiselinske sekvence enzima i nukleotidne sekvence gena poznate do onih kod kojih su čak i enzimski koraci još uvijek nepoznati. Sekundarni metaboliti se formiraju preko enzimskih puteva koji se odvijaju preko individualnih proteina, slobodnih ili u kompleksu, ili kroz dijelove velikih multifunkcionalnih polipeptida koji sudjeluju u mnoštvu enzimskih puteva (npr. Poliketid sintetaze).

Geni koji kodiraju za enzime su obično kromosomalni (npr. geni za sintezu streptomicina), a samo nekoliko njih su plazmidni (npr. oni koji kodiraju za sintezu metilenomicina A iz *Streptomyces coelicolor*). Bilo da su kromosomski ili plazmidni, geni sekundarnog metabolizma obično su grupirani u genske nakupine, odnosno klastere, naročito u prokariota, ali ne nužno kao pojedinačni operoni.

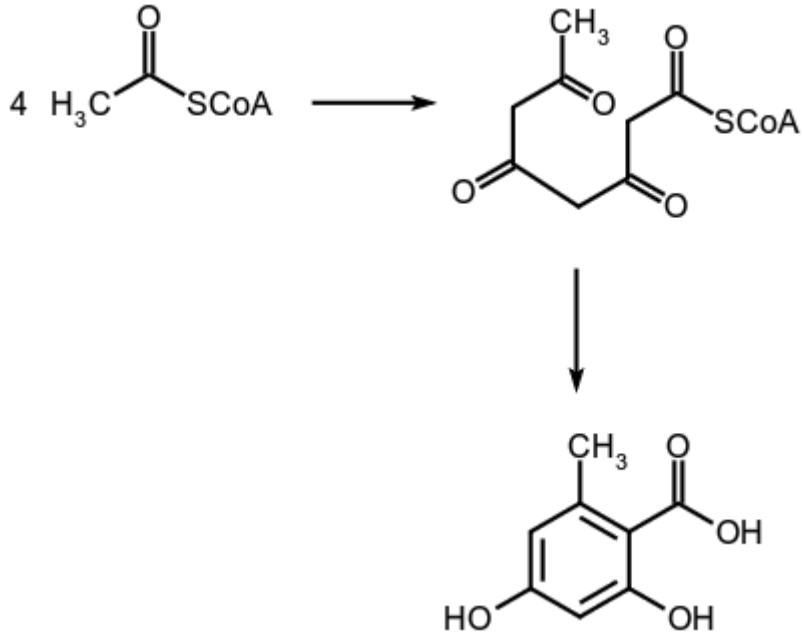
Sekundarni metabolizam se obično događa u kasnoj fazi razvoja mikroorganizama koji ih proizvode. Vremenska priroda sekundarnog metabolizma zasigurno je genetske prirode, ali ekspresija može biti pod velikim utjecajem okolišnih manipulacija. Stoga se sekundarni metabolizam često pokreće iscrpljivanjem nutrijenata, dodatkom induktora i/ili smanjenjem brzine rasta. Ovi događaji stvaraju signale koji uzrokuju kaskadu regulatornih događaja što rezultira u kemijskoj (sekundarni metabolizam) i morfološkoj (morfogeneza) diferencijaciji

mikrobnih proizvođača sekundarnih metabolita. Signal je često induktor butirolaktona niske molekulske mase koji se veže i inaktivira regulatorni protein koji inače prevenira sekundarni metabolizam i morfogenezu u uvjetima brzog rasta i dostatnosti nutrijenata. Formiranje antibiotika također je regulirano nutrijentima (dušikom, fosforom i izvorom ugljika), metalima, brzinom rasta, povratnom kontrolom i enzimskom inaktivacijom (Ruiz i sur., 2010).

Dvije velike skupine sekundarnih metabolita su poliketidi i neribosomalno sintetizirani peptidi. Poliketidi se sintetiziraju pomoću enzimskog sustava poliketid sintaze (engl. „Polyketide Synthase“, PKS), a neribosomalno sintetizirani peptidi pomoću sintetaze enzimskog sustava neribosomalnih peptida (engl. „Non Ribosomal Peptide Synthetase“, NRPS) (Rimac, 2013).

2.1.1. Poliketidi

Poliketidi (acetogenini, ketidi) su prirodni spojevi koji sadrže izmjenične karbonilne i metilenske skupine (β-poliketoni). Potječu od ponovljene kondenzacije (Slika 1) acetil koenzima A (preko malonil koenzima A) te postoje spojevi koji potječu od njih daljnjim kondenzacijama (McNaught, Wilkinson i Union 1997).



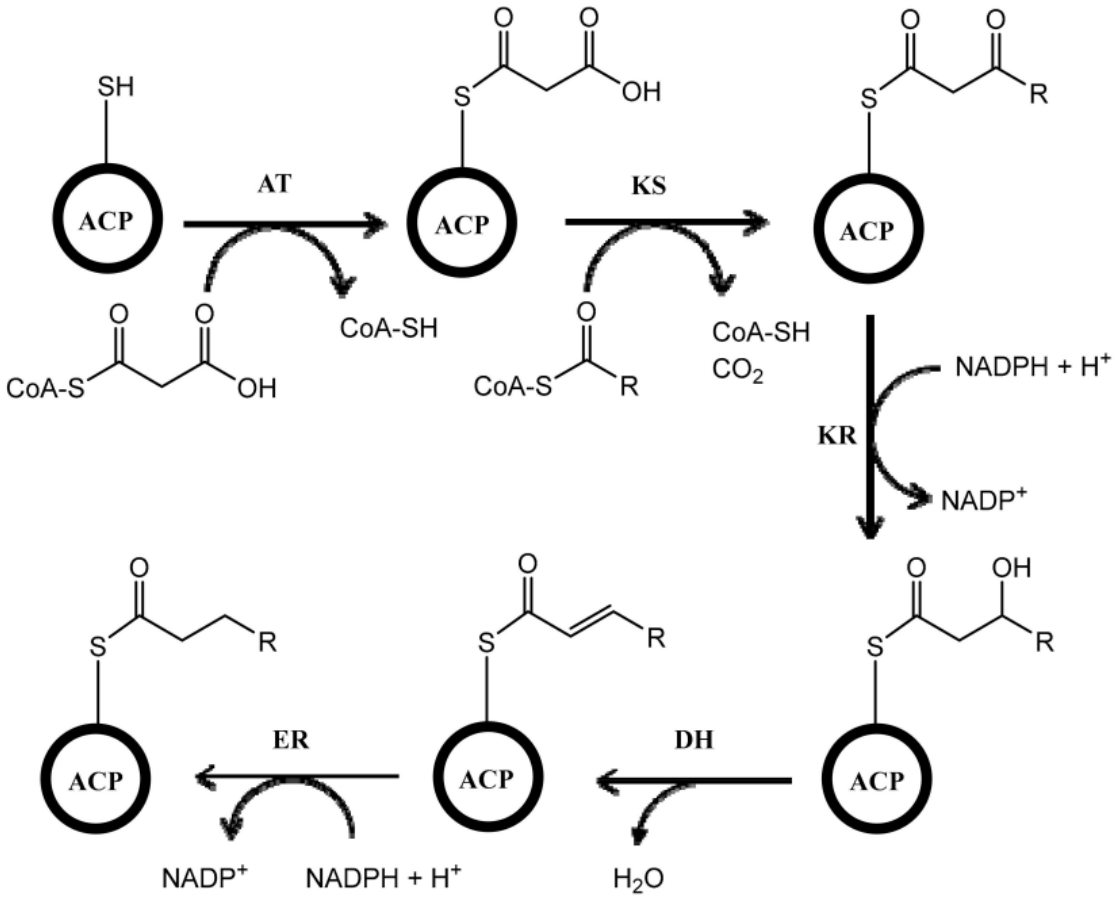
Slika 1. Kondenzacija acetil koenzima A (McNaught, Wilkinson i Union, 1997)

Poliketidi pokazuju široki raspon bioaktivnosti poput antibakterijske (npr. Tetraciklin), antifungalne (npr. Amfotericin B), antikancerogene (npr. Doxorubicin), antivirusne (npr.

Baltikolid), imunosupresivne (npr. Rapamicin), protiv kolesterola (npr. Lovastatin) i protuupalne aktivnosti (npr. Flavonoidi). Organizmi koji proizvode poliketide su bakterije, gljive, biljke, protisti, insekti i mekušci. Ovi organizmi proizvode poliketide u svrhu obrane, a insekti i za komunikaciju feromonima.

Od početka 1940-tih, povijest antibiotika uvelike je povezana s mikroorganizmima. Jedna od skupina bakterija koja proizvodi mnogo važnih antibiotika je *Actinobacteria*. *Actinobacteria* su gram-pozitivne, imaju velik udio GC veza i uključuju različite rodove koji su poznati po proizvodnji sekundarnih metabolita poput *Streptomyces, Micromonospora, Kitasatospora, Nocardiopsis, Pseudonocardia, Nocardia, Actinoplanes, Saccharopolyspora* i *Amycolatopsis*. Njihov najvažniji rod je *Streptomyces*. Pripadnici ovog roda imaju filamentozni oblik poput gljiva i izvor su oko dvije trećine svih poznatih prirodnih antibiotika. Među antibioticima koje proizvode bakterije roda *Streptomyces*, poliketidi su grupa jako važnih spojeva. Neki primjeri poliketida koje proizvode bakterije roda *Streptomyces* su rapamycin, oleandomycin, aktinorodin, daunorubicin i kaprazamicin.

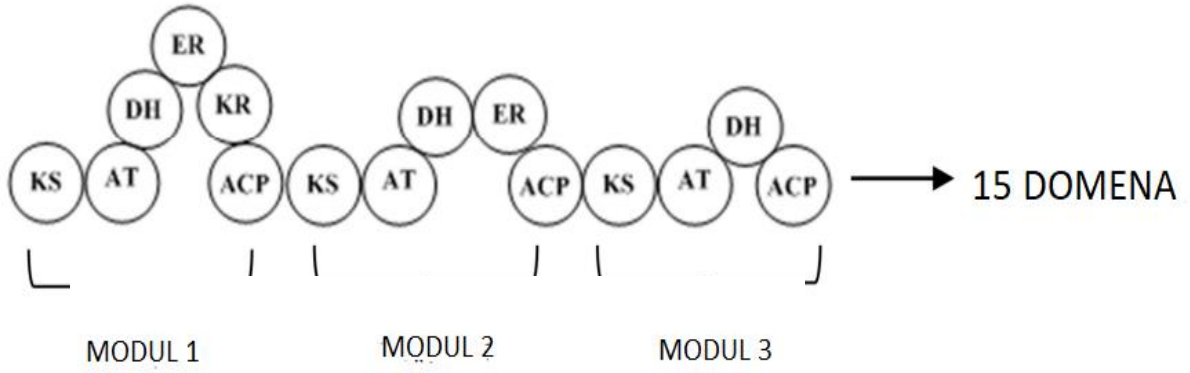
Biosinteza poliketida je vrlo kompleksna jer proces uključuje multifunkcionalne enzime koji se zovu poliketid sintaze (PKS). Mehanizam djelovanja PKS je sličan sintazi masnih kiselina (engl. „fatty acid synthase“, FAS). Proces uključuje mnoge enzimske reakcije sa različitim enzimima poput aciltransferaze (AT), kojoj je uloga kataliza vezanja supstrata (npr. Acetil ili malonil) na protein nosač acila (engl. „acyl carrier protein“, ACP) i ketosintaze (KS) koja katalizira kondenzaciju supstrata vezanih na ACP. Slijedi kondenzacije supstrata, a reakcija se nastavlja uključivanjem ketoreduktaze (KR), koja reducira keto ester, dehidrataze (DH) koja dehidrira spoj i enoilreduktaze (ER) koja reducira dvostruku ugljik-ugljik vezu u molekuli (Slika 2). Za razliku od FAS, proces kataliziran sa KR, DH i ER je opcionalan kod PKS, što može dati različite strukture poliketida sa keto grupama, hidroksilnim grupama i/ili dvostrukim vezama na različitim mjestima u molekuli. U bakterijama roda *Streptomyces* postoje tri vrste PKS (tip 1, tip 2 i tip 3) (Risdian i sur., 2019).



Slika 2. Shema reakcije koju kataliziraju poliketid sintaze (PKS) (Risdian i sur., 2019)

2.1.1.1. Poliketid sintaze tipa I

Poliketid sintaze tipa I uključuju velike multifunkcionalne proteine koji imaju puno modula koji sadrže domene u kojima se događa određena enzimska reakcija (Slika 3). Svaki modul ima zadatak da provede jedan ciklus kondenzacije na ne-iterativan način. Zbog toga što ovaj sustav radi pomoću modula, još se naziva i modularni PKS sustav. Esencijalne domene koje postoje u svakom modulu su AT, KS i ACP koje rade zajedno kako bi proizvele β-keto esterski međuprodukt. Ostale domene koje mogu biti prisutne su KR, DH i ER, koje su odgovorne za modifikaciju keto grupe. U procesu proizvodnje poliketida rastući poliketidni lanac se premješta s jednog modula na drugi dok se gotova molekula ne odcijepi sa zadnjeg modula pomoću posebnog enzima (Risdian i sur., 2019).



Slika 3. Struktura PKS tipa I kojega čine 3 modula i 15 domena (*prema* Risdian i sur., 2019)

PKS tipa I su uglavnom odgovorni za proizvodnju makrocikličkih poliketida (makrolida), a postoji istraživanje koje kaže da su tip I PKS također uključeni u biosintezu linearnog poliketida tautomicetina. Makrolidi pripadaju u poliketidne spojeve koji sadrže karakteristični makrociklički laktonski prsten koji ima različite aktivnosti poput antibakterijske, antifungalne, imunosupresivne i antikancerogene. Kao antibakterijski agens, makrolid funkcionira tako što inhibira sintezu proteina vezanjem za 50S podjedinicu ribosoma i blokira translokaciju kod sinteze proteina. Neki primjeri makrolida koje proizvode bakterije roda *Streptomyces* su rapamicin, FK506, spiramicin, avermektin, metimicin, narbomicin i pikromicin. Ove komponente su proizvedene pomoću multifunkcionalnih polipeptida za koje kodira biosintetski genski klaster (Risdian i sur., 2019).

2.1.1.2. Poliketid sintaze tipa II

Poliketid sintaze tipa II odgovorne su za proizvodnju aromatskih poliketida. Na temelju sustava polifenolnih prstenova i njihovog biosintetskog puta, aromatski poliketidi koje proizvode PKS tipa II obično se klasificiraju u sedam skupina, tj. antraciklini, anguciklini, aureolne kiseline, tetraciklini, tetracenomicini, polifenoli tipa pradimicina i benzoizokromankinoni.

Antraciklini se sastoje od linearnog tetracikličkog sustava prstenova sa kinon-hidrokinon grupama u prstenovima B i C. Anguciklini imaju kutni tetraciklički prstenasti sustav. Aureolne kiseline imaju triciklički kromofor. Tetraciklini sadrže linearni tetraciklički sustav prstenova bez kinon-hidrokinon grupa u prstenovima B i C. Tetracenomicini imaju linearni tetraciklički sustav prstenova sa kinonskom skupinom u prstenu B. Polifenoli tipa pradimicina smatraju se kao prošireni anguciklini. Benzoizokromankinoni sadrže kinonski derivat iz izokromske

strukture. Neki primjeri aromatskih poliketida koje proizvode bakterije roda Streptomyces su aktinorodin (benzoizokromankinoni), doksorubicin (antraciklini), jadomicin B (anguciklini), oksitetraciklin (tetraciklini), mitramicin (aureolne kiseline), tetracenomicin C (tetracenomicini i benstatin A (polifenoli tipa pradimicina).

Za razliku od PKS tipa I koji uključuju velike multifunkcionalne proteine koji imaju mnogo modula koji sadrže domene i kataliziraju enzimske reakcije na ne-iterativni način, PKS tipa II sadrže monofunkcionalne polipeptide i rade iterativno kako bi proizveli aromatske poliketide. Međutim, poput PKS tipa I, PKS tipa II također uključuju ACP koji služi kao sidro za nascentni poliketidni lanac. Uz to što sadrže ACP, PKS tipa II se također sastoje od jedinica ketosintaza (KS_α i KS_β) koje rade kooperativno kako bi proizvele poli-β-keto lanac. KS_α jedinica katalizira kondenzaciju prekursora, dok KS_β u PKS tipa II služi kao faktor određivanja duljine lanca. Ova tri velika sustava (ACP, KS_α i KS_β) zovu se „minimalni PKS“ i oni rade iterativno da bi proizveli aromatski poliketid. Ostali dodatni enzimi poput ketoreduktaza, ciklaza i aromataza surađuju kako bi transformirali poli-β-keto lanac u jezgru aromatskog spoja. *Post-tailoring* proces provodi se pomoću oksigenaza te glikozil i metil transferaza (Risdian i sur., 2019).

2.1.1.3. Poliketid sintaze tipa III

Za razliku od PKS tipa I i PKS tipa II, PKS tipa III ne koriste ACP kao sidro za proizvodnju poliketidnih metabolita. U ovom slučaju, acil-CoA se koriste direktno kao supstrati za proizvodnju poliketidnih spojeva. Kako bi se proizveli poliketidi, ovaj sustav koristi enzime koji konstruiraju homodimere i katalizira mnoge reakcije poput vezanja primera, produljivanja lanca i ciklizacije na iterativni način. Drugim riječima, PKS tipa III su najjednostavnije strukture među ostalim tipovima PKS. PKS tipa III su po prvi puta otkrivene u bakterijama 1999. godine, a prije toga su bile detektirane samo u biljkama.

Neka prethodna istraživanja su otkrila da bi PKS tipa III također mogle biti identificirane u bakterijama roda *Streptomyces*, poput RppA koji je pronađen u bakteriji *Streptomyces griseus*, a odgovoran je za sintezu 1,3,6,8-tetrahidroksinaftalena (THN). Gcs koji je identificiran u bakteriji *Streptomyces coelicolor* A3(2) igra važnu ulogu u biosintezi germicidina. SrsA, za kojeg kodira gen *srsA*, izoliran je iz *Streptomyces griseus* i poznato je da igra važnu ulogu u biosintezi fenolnih lipida, tj. alkilrezocinola i alkilpirona. Predloženo je da je Ken2 koji je izoliran iz Streptomyces violaceoruber uključen u proizvodnju 3,5-dihidroksifenilglicin (3,5-DHPG). Ovaj spoj je neproteinogena aminokiselina koja je potrebna za formiranje kendomicina i nekoliko ostalih glikopeptidnih antibiotika poput balhimicina,

kloroeremomicina i vankomicina. Cpz6, za kojeg kodira gen *cpz6*, izoliran je iz *Streptomyces* sp. MK730–62F2 i uključen je u biosintezu kaprazamicina tako što proizvodi skupinu novih triketidepirena (presulficidina). Štoviše, još jedno otkriće je također predložilo da DpyA katalizira formiranje alkildihidropirona u bakteriji *Streptomyces reveromyceticus* (Risdian i sur., 2019).

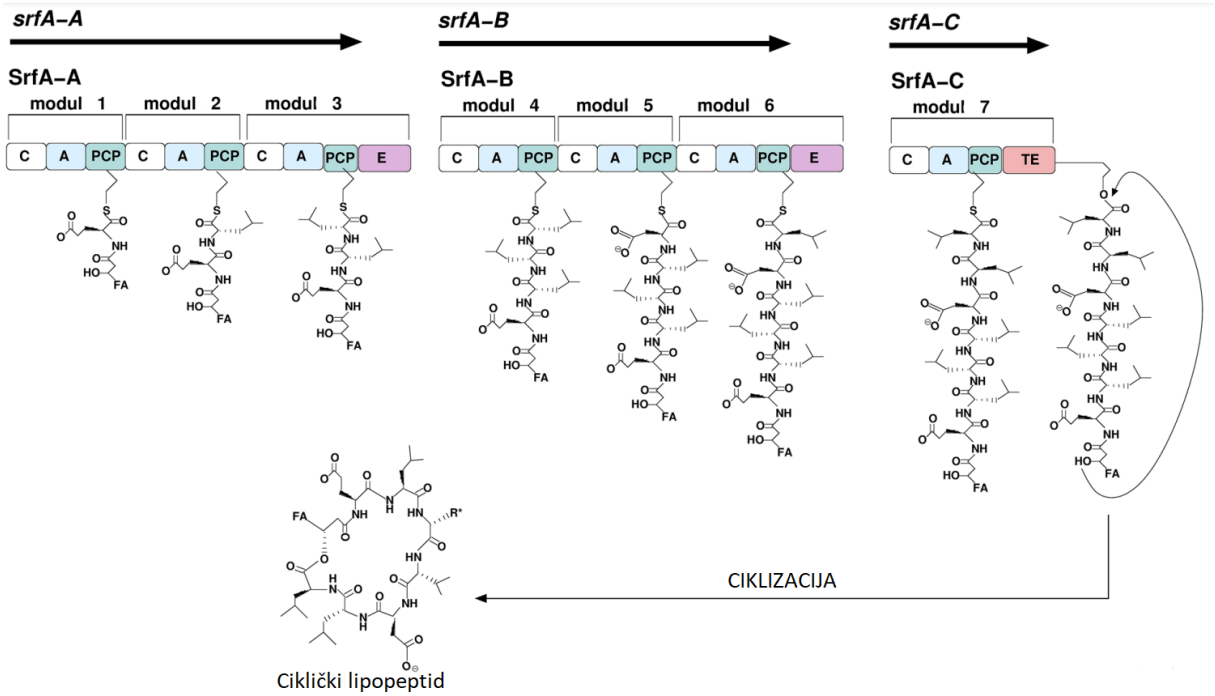
2.1.2. Neribosomalno sintetizirani polipeptidi

Neribosomalno sintetizirani polipeptidi su raznolika obitelj prirodnih produkata i pripadaju skupini sekundarnih metabolita sa različitim svojstvima poput toksina, siderofora, pigmenata, antibiotika, citostatika, imunosupresiva i antikancerogenih agensa. Oni su posebni po tome što je njihova sinteza neovisna o ribosomskoj mašineriji. Mikroorganizmi koji nastanjuju zemlju poput bakterija redova *Actinomycetes* i *Bacilli* te eukariotski filamentozni fungi najčešći su proizvođači neribosomalno sintetiziranih polipeptida, ali i morski mikroorganizmi su se također pojavili kao izvor takvih peptida. Ovi peptidi sadrže strukturne značajke poput aminokiselina kao što je ornitin ili imino kiselina i njihove strukture su makrociklički razgranate, makrociklički dimeri ili trimeri identičnih strukturnih elemenata. Obično se neribosomalno sintetizirani peptidi sintetiziraju na velikim enzimskim kompleksima neribosomalnih peptid sintetaza (NRPS), koji se definiraju kao modularni multidomenski enzimi. NRPS enzimi su prisutni u sve tri domene stabla života, a najzastupljeniji su u bakterijama, nešto manje u eukariotima i rijetki u arhejama. U domeni bakterija, *Proteobacteria*, *Actinobacteria*, *Firmicutes* i *Cyanobacteria* rodovi su koji sadrže obilje ovih enzima te se promatra korelacija između veličine genoma i broja NRPS klastera (Martínez-Núñez i López, 2016).

NRPS su modularno organizirani multi enzimski kompleksi koji služe kao kalupi i biosintetska mašinerija koja funkcionira preko mehanizma tio-kalupa koji je neovisan o ribosomima. Modul se definira kao sekcija NRPS enzima koja na specifičan način pridružuje jednu aminokiselinu na peptidnu okosnicu. Moduli se mogu podijeliti u domene, koje kataliziraju individualne korake neribosomalne sinteze peptida. Svaki modul sastoji se od minimalno tri domene: adenilacijske (A), protein nosač peptidila (engl. peptidyl carrier protein, PCP) ili tiolacijske (T) i kondenzacijske (C) domene koje provode sintezu neribosomalno sintetiziranih polipeptida (Slika 4). Redoslijed modula obično je kolinearan produktu. Sinteza se odvija u smjeru od N- do C-terminalnog kraja, a nastali peptidi su obično dugački oko 3-15 aminokiselina i mogu biti linearni, ciklički ili razgranati ciklički. Postoji konzervirani motiv u DNA sekvenci koja kodira za domenu koji omogućuje identifikaciju njihovih sekvenci pomoću

alata kao što je BLAST.

Prvi korak biosinteze katalizira A domena. Ona prepoznaje i izvršava aktivaciju aminokiselinskog supstrata adenilacijom pomoću Mg-ATP što rezultira nastajanjem aminoacil adeniliranog međuprodukta. A domena sastoji se od oko 550 aminokiselina i ima 10 aminokiselinskih ostataka koji se mogu smatrati „kodonima“ NRPS enzima i važni su za specifičnost supstrata. Supstrati koje prepoznaje A domena mogu sadržavati D- i L- oblike 20 aminokiselina koje se koriste u ribosomalnoj sintezi proteina kao i neproteinogenske aminokiseline poput ornitina, iminokiselina i hidroksi kiselina poput α-aminoadipinske i β-butirične kiseline. Reakcija koju provodi A domena dijeli istu kemiju kao i ona koju provode aminoacil-tRNA sintetaze koje sudjeluju u prvom koraku ribosomalne sinteze proteina. Drugi korak provodi PCP domena koja se sastoji od oko 80 aminokiselina. Ona kovalentno veže aktiviranu aminokiselinu na njezin kofaktor, 4'-fosfopanteteinsku ruku, a zatim tioesterskom vezom prenosi aktivirani supstrat i elongacijske međuprodukte na C- domenu. Posljednji korak provodi C domena koje je dugačka oko cca. 450 aminokiselina. Ona katalizira formiranje peptidne veze između karboksilne grupe nascentnog peptida i aminokiseline koju prenosi bočni modul, što rezultira translokacijom rastućeg lanca na sljedeći modul. Nakon kondenzacijskog koraka, linearni međuprodukt se oslobađa pomoću tioesterazne (TE) domene hidrolizom ili unutarnjom ciklizacijom kod bakterija (Martínez-Núñez i López, 2016).



Slika 4. Biosinteza neribosomalno sintetiziranih polipeptida pomoću NRPS sintetaze (*prema* Martínez-Núñez i López, 2016)

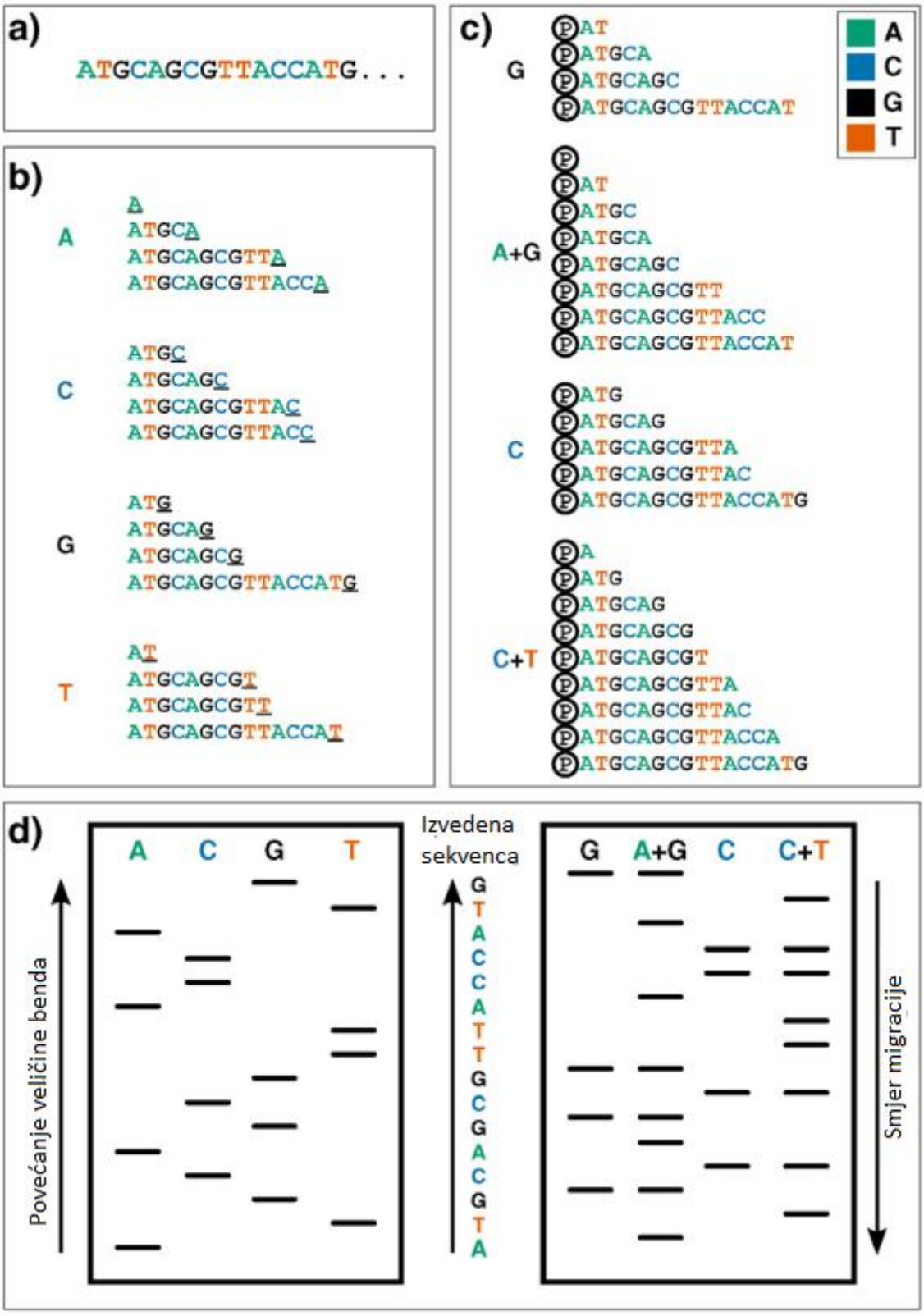
2.2. SEKVENCIONIRANJE DNA

Redoslijed nukleinskih kiselina u DNA i RNA molekulama sadrži važnu informaciju o nasljednim i biokemijskim svojstvima zemaljskog života. Stoga je mogućnost čitanja tih sekvenci nužna za biološka istraživanja (Heather i Chain, 2015).

2.2.1. Sekvencioniranje prve generacije

Prva tehnika sekvencioniranja koja je široko prihvaćena bila je Maxam-Gilbert metoda, tj. tehnika kemijskog cijepanja. Ova metoda smatra se početkom metoda sekvencioniranja prve generacije. Ova metoda se ne oslanja na DNA polimerazu za dobivanje fragmenata DNA već se radioaktivno obilježena DNA tretira kemikalijama koje lome lanac kod specifičnih baza. Nakon nanošenja na poliakrilamidni gel moguće je odrediti duljinu fragmenata i time poziciju specifičnih nukleotida (Slika 5, desno).

Važno otkriće za sekvencioniranje dogodilo se 1977. kada je Sanger razvio dideoksi metodu sekvencioniranja. Ova metoda koristi kemijske analoge deoksiribonukleotida (dNTP) koji su monomeri DNA lanca. Dideoksinukleotidi (ddNTP) nemaju 3' hidroksilnu grupu koja je potrebna za produljivanje DNA lanaca i time ne mogu formirati vezu sa 5' fosfatom slijedećeg dNTP-a. Dodatkom radioaktivno označenih ddNTP-ova u reakciju produljivanja DNA lanca u maloj koncentraciji u odnosu na standardne dNTP-ove dolazi do nastajanju DNA lanaca svih mogućih duljina jer se ddNTP-ovi nasumično ugrađuju u rastući lanac što zaustavlja daljnji rast lanca. Izvođenjem četiri paralelne reakcije od kojih svaka sadrži po jednu vrstu ddNTP baze i dobivanjem rezultata na četiri stupca poliakrilamidnog gela, moguće je korištenjem autoradiografije zaključiti kako izgleda nukleotidna sekvenca u originalnom lancu - kalupu, jer će u odgovarajućem stupcu gela biti vidljiva radioaktivna vrpca (Slika 5, lijevo). Tijekom godina napravljena su brojna poboljšanja ove metode, uključujući i zamjenu fosfo- ili tricijevog radioobilježavanja sa florimetrijskom detekcijom što je omogućilo odvijanje reakcije u jednoj reakcijskoj posudi umjesto u četiri te poboljšanu detekciju kapilarnom elektroforezom. Ova dva poboljšanja doprinijela su razvoju automatiziranog sekvencioniranja DNA i prvih komercijalnih strojeva za sekvencioniranje DNA koji su korišteni za sekvencioniranje genoma sve kompleksnijih vrsta (Heather i Chain, 2015).



Slika 5. Tehnologije sekvencioniranja prve generacije, Sangerova i Maxam-Gilbertova metoda (*prema* Heather i Chain, 2015)

2.2.2. Druga generacija sekvencioniranja

Pirosekvencioniranje je luminiscentna metoda koja se temelji na sintezi pirofosfata. To je proces koji se odvija pomoću dva enzima. ATP sulfurilaza koristi se za konverziju pirofosfata u ATP koji se tada koristi kao supstrat za luciferazu. Time se proizvodi svjetlost koja je proporcionalna količini pirofosfata. Ovaj pristup koristi se za dobivanje sekvence mjerenjem nastajanja pirofosfata dok se nukleotidi kreću kroz sustav koji sadrži kalup DNA pričvršćen za čvrstu fazu. Ova metoda ima nekoliko prednosti: može se provesti korištenjem prirodnih nukleotida i može se promatrati u stvarnom vremenu (real-time). Glavni nedostatak su poteškoće u identifikaciji više istih nukleotida za redom. Pirosekvencioniranje je kasnije preuzela biotehnološka kompanija 454 Life Sciences te je kasnije evoluiralo u prvu veliku uspješnu komercijalnu tehnologiju za sekvenciranje sljedeće generacije (engl. next-generation sequencing, NGS).

Nekoliko paralelnih tehnika sekvencioniranja pojavilo se nakon uspjeha 454. Najvažnija od njih je metoda Solexa koju je kasnije preuzela Illumina. DNA molekule omeđene adapterima prelaze preko velikog broja komplementarnih oligonukleotida koji su povezani na protočnu ćeliju. Tada slijedi PCR čvrste faze koji proizvodi susjedne nakupine klonskih populacija od svakog lanca DNA vezanog na protočnu ćeliju. Ovaj proces je dobio ime „bridge amplification“ zbog DNA lanaca koji se savijaju kako bi započeli slijedeću polimerizaciju sa susjednih oligonukleotida povezanih na površinu.

Uz 454 i Solexa/Illumina sekvencioniranje, ističe se još i sekvenciranje ligacijom i detekcijom oligonukleotida (engl. „sequencing by oligonucleotide ligation and detection“, SOLiD) koje je razvila tvrtka Applied Biosystems. Kao što samo ime govori, SOLiD ne radi na principu sinteze, već ligacije pomoću DNA ligaze. Nedostatak ove metode je nemogućnost proizvodnje dugačkih slijedova DNA kao što to uspijeva Illumina metodi i metodi na bazi pirosekvencioniranja, što otežava sastavljanje genome (assembly). Bez obzira na to, ova metoda je prisutna je je njezina prednost vrlo niska cijena (Heather i Chain, 2015).

2.2.3. Sekvencioniranje treće generacije

Treću generaciju sekvencioniranja karakterizira mogućnost sekvencioniranja jedne molekule DNA, bez potrebe za umnažanjem DNA koje karakterizira sve prethodne tehnologije sekvencioniranja. Od treće generacije najviše se koristi “single molecule real time” (SMRT) tehnologija koju je razvila tvrtka Pacific Biosciences (Heather i Chain, 2015). Od izoliranog genetičkog materijala konstruira se SMRTbell® knjižnica ligiranjem adaptera na krajeve

dsDNA. U knjižnicu se također dodaju početnica i polimeraza, koje se vežu na adapter. DNA s početnicom i polimerazom nanosi se na SMRT® ćeliju koja sadrži puno mikro jažica koji se zovu Zero-Mode Waveguides (ZMW). U te jažice smješta se po jedna molekula DNA i u njima se događa reakcija polimerizacije. Pri vezanju nukleotida na rastući lanac DNA emitira se svjetlost (drukčije boje za svaku od četiri dušične baze koje su vezane na nukleotid trifosfat) i tako se reakcija prati u stvarnom vremenu (Real-Time). Ova tehnologija omogućuje vrlo precizno sekvenciranje i veliku duljinu DNA sljedova (reads) (PacBio, 2020).

2.3. SINTENIJA

Blokovi sintenije definiraju se kao kromosomske regije genoma koje dijele zajednički redoslijed homolognih gena koji potječu od zajedničkog pretka. Naizmjenično se koriste i alternativni nazivi poput konzervirana sintenija ili kolinearnost. Usporedbe sintenije genoma između vrsta i unutar iste vrste otvorile su mogućnost za proučavanje evolucijskih procesa koji vode ka raznolikosti broja i strukture kromosoma u mnogim izolatima. Analiza sintenije u blisko povezanim vrstama sada je norma za svaki novi objavljeni genom. Međutim, kvaliteta assembly-ja dolazi u pitanje, pošto se pokazalo da utječe na analize koje slijede, poput anotacije. Općenito, identifikacija sintenije proces je filtriranja i organiziranja svih lokalnih sličnosti između sekvenci genoma u koherentnu globalnu sliku. Najintuitivniji način za identifikaciju sintenije bio bi njezino rekonstruiranje iz selektivnih poravnanja genoma, međutim razine nukleotidne raznolikosti između vrsta mogle bi učiniti takve metode zahtjevnima. Mnogi alati umjesto toga koriste ortologne povezanosti između protein-kodirajućih gena kao sidra za pozicioniranje statistički bitnih lokalnih poravnanja. Pristupi uključuju korištenje usmjerenog acikličkog grafa, matricu homologije gena (engl. „gene homology matrix“, GHM) te algoritam koji koristi recipročne najbolje hitove (engl. „reciprocal best hits“). Sve ove metode slažu se u dugačkim blokovima sintenije, ali razlikuju se u rukovanju lokalnim preraspodjelama kao i u rezoluciji identifikacije sintenije. Stoga analiza sintenije jako ovisi o kvaliteti assembly-ja. Na primjer, sekvence koje nedostaju u assembly-ju mogu dovesti do toga da izostane anotacija gena, a time i ortologna povezanost. Još uvijek je nejasno utječe li fragmentacija assembly-ja na zadatak homologije da identificira sidra, rasporede sekvenci za provjeru reda i praznina ili druge faktore u analizi sintenije (Liu i sur., 2018).

2.4. ONTOLOGIJA GENA

Cilj ontologije gena (GO) je stvaranje strukturiranog, precizno definiranog, standardiziranog i kontroliranog vokabulara kojime se može opisati uloga gena i genskih produkata u bilo kojem organizmu. Postoje tri kategorije ontologije gena: biološki procesi, molekularna funkcija i stanična komponenta.

Biološki procesi uključuju biološke ciljeve čijem izvršavanju doprinose gen ili genski produkt. Proces se ostvaruje putem jednog ili više uređenih skupova molekularnih funkcija. Procesi često uključuju kemijsku ili fizičku transformaciju u smislu da nešto ulazi u proces i nešto drugo izlazi iz procesa. Primjeri termina za opće biološke procese su „stanični rast i održavanje“ ili „prijenos signala“. Primjeri specifičnih termina za biološke procese su „translacija“, „metabolizam pirimidina“ ili „biosinteza cAMP“.

Molekularna funkcija definirana je kao biokemijska aktivnost genskog produkta (uključujući specifično vezanje za ligande ili strukture). Ova definicija također se odnosi na sposobnost genskog produkta (ili kompleksa genskih produkata) da vrši specifičnu aktivnost. Molekularna funkcija opisuje samo ono što se dogodilo bez da specificira gdje ili kada se događaj odvio. Primjeri općih funkcionalnih termina su „enzim“, „transporter“ ili „ligand“. Primjeri specifičnih funkcionalnih termina su „adenilat ciklaza“ ili „ligand za toll receptor“.

Stanična komponenta odnosi se na mjesto u stanici gdje je genski produkt aktivan. Ovi termini odražavaju naše razumijevanje strukture stanice. Ona uključuje termine poput „ribosom“ ili „proteosom“ koji specificiraju gdje se genski produkti mogu nalaziti.

Biološki proces, molekularna funkcija i stanična komponenta atributi su gena, genskih proizvoda ili skupina genskih proizvoda. Odnosi između genskog produkta (ili skupine genskih produkata) i biološkog procesa, molekularne funkcije te stanične komponente ukazuju na to da neki protein može funkcionirati u nekoliko procesa, sadržavati domene koje nose različite molekularne funkcije i sudjelovati u više alternativnih interakcija sa drugim proteinima, organelama ili lokacijama u stanici (Ashburner i sur., 2000).

3. EKSPERIMENTALNI DIO

3.1. MATERIJAL

3.1.1. Računalna podrška i operativni sustav

Ovaj je diplomski rad izrađen na računalu sljedećeg sklopovlja: prijenosno računalo *Hewlett Packard*, procesor Intel®Core™ i5-4300M CPU @ 2,60GHz, 8GB radne memorije. Na čvrstom disku upotrijebljenog računala instaliran je operativni sustav „Microsoft Windows 10 Pro“.

3.1.2. Canu assembler

Canu assembler nasljednik je Celera assemblera i dizajniran je za sastavljanje sekvenci (engl. „assembly“) dobivenih visokošumnim sekvenciranjem jedne molekule (engl. „high-noise single-molecule sequencing“) kao što je PacBio SMRT tehnologija sekvenciranja. MinHash proces poravnanja (engl. „MinHash Alignment Process“, MHAP) dizajniran je kako bi se prevladalo usko grlo preklapanja šumnih readova dobivenih sekvencioniranjem jedne molekule. Jedna od prednosti Canu assemblera (Koren i sur., 2021) je što je uvelike skratio vrijeme trajanja procesa sastavljanja genoma. Canu također ima poboljšani algoritam za konstrukciju grafa koji razdvaja blisko povezana ponavljanja i alele, a temelji se na statističkom modelu greške čitanja. Protočna obrada podataka (engl. „pipeline“) Canu assemblera sastoji se od tri etape: korekcija, skraćivanje i assembly, a svaka od njih može raditi samostalno ili u seriji (npr. Samo korekcija readova ili assembly bez korekcije) (Koren i sur., 2017).

3.1.3. Aplikacije unutar KBase

Kbase (URL: https://www.kbase.us/) je softver otvorenog koda (engl. „open-source software“) i podatkovna platforma koja omogućuje dijeljenje podataka, integraciju i analizu mikroba, biljaka i njihovih zajednica. KBase sadrži internu referentnu bazu podataka koja objedinjuje informacije iz široko korištenih repozitorija javno dostupnih podataka. Ona uključuje preko 90,000 mikrobnih genoma iz RefSeq baze podataka, preko 50 biljnih genoma iz Phytozome-a te više od 30,000 reakcija i spojeva iz KEGG, BIGG i MetaCyc baza podataka. Ovi podaci dostupni su za integraciju sa podacima korisnika gdje je to prikladno (npr. usporedba genoma ili izrada filogenetičkog stabla). KBase povezuje ove različite tipove

podataka sa rasponom analitičkih funkcija unutar web-baziranog korisničkog sučelja (Arkin i sur., 2018).

Narativ, što je naziv za primarno sučelje Kbase-a, korisniku pruža drugačije iskustvo od ostalih platformi za analizu, iako dijeli neke zajedničke karakteristike sa nekoliko njih. Na ovom sučelju, koje je napravljeno na platformi Jupyter (URL: https://jupyter.org/), korisnik može učitati svoje podatke, pretražiti i koristiti opsežne javno dostupne referentne podatke, pristupiti podacima koje dijele drugi korisnici, dijeliti svoje podatke sa ostalima , odabrati i pokrenuti aplikacije unosom svojih podataka, vidjeti i analizirati rezultate tih aplikacija te zabilježiti svoja razmišljanja i interpretacije skupa sa koracima analize (Arkin i sur., 2018).

KBase sadrži brojne aplikacije koje korisniku pružaju različite funkcije kao što su assembly (meta)genoma, generaliziranje contiga, anotacija genoma, analiza homologije sekvenci, izrada filogenetskog stabla, komparativna genomika, metaboličko modeliranje, RNA-seq procesiranje i analiza ekspresije. Korisnici također mogu kreirati i pokrenuti blokove koda unutar narativa korištenjem „kodnih ćelija“. KBase sadrži sučelje za programiranje aplikacija (engl. „Application programming interface“, API) koje dozvoljava korisniku da pozove bilo koju KBase aplikaciju programski, unutar tih ćelija koda (Arkin i sur, 2018). Aplikacije koje su korištene u ovom radu navedene su u sljedećim poglavljima.

3.1.3.1. Aplikacija QUAST

QUAST (QUality ASsessment Tool) (Gurevich i sur., 2018) je aplikacija za procjenu assemblyja na temelju različitih podataka poput broja contiga, duljine contiga, N50/75, L50/75 i udjela GC veza. Aplikacija koristi jedan ili više assemblyja kao input i generira output sa statistikama za sve assemblyje. Aplikacija može evaluirati kvalitetu assemblyja čak i bez referentnog genoma, pa ga korisnici mogu koristiti za procjenu genoma novih vrsta koje još nemaju referentni genom. Dodatna prednost je i njegova brzina, a koraci koji oduzimaju najviše vremena su zaustavljeni, što omogućuje njegovu efektivnu upotrebu na višejezgrenim procesorima (Gurevich i sur., 2013).

3.1.3.2. Aplikacija CheckM

CheckM (Parks i sur., 2015) je aplikacija za procjenu kvalitete genoma izolata, pojedinačnih stanica ili metagenoma usporedbom s postojećom bazom podataka genoma. Ova aplikacija dio je M-suite kolekcije bioinformatičkih alata iz Ecogenomics Group na Sveučilištu

Queensland u Australiji. CheckM pruža detaljne procjene o kompletnosti genoma i kontaminaciji korištenjem setova sveprisutnih gena koji se nalaze u jednoj kopiji unutar filogenetske loze. Procjena kvalitete genoma također se može ispitati korištenjem dijagrama koji prikazuju ključne genomske karakteristike (npr. broj GC veza) koje ističu sekvence izvan očekivanih raspodjela unutar referentnog genoma. Kvaliteta genome je najbolje ocjenjena u genomu koji ima maksimalnu kompletnost i minimalnu kontaminaciju (Parks i sur., 2014).

3.1.3.3. Aplikacija za anotaciju više mikrobnih assembly-ja pomoću RASTtk

Ova aplikacija koristi RAST (Rapid Annotations using Subsystems Technology) (Brettin i sur., 2015) alat za anotaciju prokariotskih assembly-ja ili ažuriranje anotacija assembly-ja. Genomske datoteke koje aplikacija generira imaju isto ime kao i uneseni assembly-ji sa ".RAST" nastavkom. Ovime se genomi pripremaju za daljnju analizu pomoću drugih Kbase aplikacija.

RAST server razvijen je 2008. godine u svrhu anotacije mikrobnih genoma. On radi na način da projecira ručno odabrane anotacije gena iz SEED baze podataka (Overbeek i sur., 2014) na genome. Ključ konzistencije i točnosti RAST algoritma su precizno strukturirani podaci za anotaciju unutar SEED baze podataka koji su organizirani u podsustave (setove logički povezanih funkcionalnih uloga). Prednosti ovog pristupa su pružanje brzine, pogodnosti i konzistentnosti korisniku. Kako bi napravio anotaciju pomoću RAST-a, korisnik šalje svoje contige na server na kojem se obavlja posao. Ovo oslobađa korisnika od potrebe za preuzimanjem i instaliranjem više programa ili izvođenjem kompleksnih radnji. Međutim, bez obzira na prednosti, ovaj pristup ima i neka ograničenja. Na primjer, zadana protočna obrada podataka možda nije najbolji izbor za zadani genom. Također je kompliciranо prilagoditi protočnu obradu anotacije odabirom različitih alata te dodavanjem vlastitih dijelova i anotacija (Brettin i sur., 2015).

3.1.3.4. Aplikacija za izradu feature setova od genoma

Feature setovi korisni su za grupiranje skupine elemenata i izvođenje nekih operacija na toj skupini elemenata. Na primjer, umjesto specificiranja svih elemenata na listi za svaku metodu, ova metoda se može iskoristiti za izgradnju jednog tzv. "Feature Set" koji se može koristiti kao input za ostale metode.

3.1.3.5. Aplikacija Batch Create Assembly Set

Ova aplikacija omogućuje korisniku da napravi set assembly-ja koji se može koristiti kao „input“ u ostalim aplikacijama koje podržavaju takav format. Svi assembly-ji koji se koriste za izradu seta moraju biti dio narativa. Svaki assembly čije se ime poklapa sa odabranim uzorkom biti će uključen u set. Ukoliko uzorak nije odabran tada će svi assembly-ji u narativu biti uključeni.

3.1.4. Alat za izradu Venn-ovih dijagrama

Ovaj alat radi na principu usporedbe elemenata dvaju ili više skupova (URL: http://bioinformatics.psb.ugent.be/webtools/Venn/). On detektira elemente koji su tim skupovima zajednički i različiti. Ako je broj skupova manji od sedam, ovaj alat može nacrtati Venn-ov dijagram na temelju zajedničkih i jedinstvenih elemenata tih skupova. Korisnik može birati između simetričnog i asimetričnog Venn-ovog dijagrama. Grafički output moguće je preuzeti u SVG i PNG formatu. Skupove koje se koriste za input moguće je učitati u obliku datoteke ili kopirati čitav skup i zalijepiti u prazna polja. Svakom skupu i moguće je dodijeliti vlastito ime. Skupovi moraju sadržavati samo jedan element u svakoj liniji, ali nema ograničenja po pitanju broja linija. Elementi se obrađuju na način koji razlikuje velika i mala slova. Duplicirani elementi u skupu procesuiraju se na taj način da se redundantne kopije uklanjaju.

3.1.5. PANTHER

PANTHER (Protein Analysis Through Evolutionary Relationships, http://pantherdb.org) je resurs koji služi za evolucijsku i funkcionalnu klasifikaciju gena iz različitih organizama. Evolucijska klasifikacija podrazumijeva grupiranje prema klasi proteina, porodici proteina ili potporodici, a funkcionalna klasifikacija uključuje ontologiju gena i puteve (Mi i sur., 2020). U ovom radu korištena je funkcionalna klasifikacija, tj. ontologija gena (engl. „gene onthology“, GO).

Funkcionalna klasifikacija grupira proteine prema njihovim funkcijama, a ne porodicama. Funkcije se klasificiraju na više načina, a tri su definirana ontologijom gena (molekularna funkcija, stanična komponenta i biološki proces). Funkcije se anotiraju korištenjem dvije različite metode. Prva metoda anotira tzv. klade (definirane korištenjem unutarnjih čvorova stabla) povezanih proteina unutar filogenetskog stabla što rezultira anotacijama svih proteina u

anotiranoj kladi. Ove anotacije uključuju termine za ontologiju gena iz PANTHER GO-slim, pojednostavljenog podskupa kompletne GO, kao i PANTHER Pathway ontologiju. Druga metoda anotira individualne proteine u funkcionalne klase, što podrazumjeva kompletnu ontologiju gena ili Reactome puteve (Mi i sur., 2020). U ovome radu korištena je druga metoda.

3.1.6. AmiGO 2

AmiGO (URL: http://amigo.geneontology.org/amigo) je web aplikacija koja dopušta korisnicima da pretražuju i vizualiziraju ontologije i anotacije povezanih genskih produkata. AmiGO se može koristiti online na Gene Ontology (GO) stranici (URL: http://www.geneontology.org) kako bi se pristupilo podacima koje pruža konzorcij za ontologiju gena. Također se može preuzeti i instalirati za pretraživanje lokalnih ontologija i anotacija. AmiGO dopušta korisnicima da pretražuju, sortiraju, analiziraju, vizualiziraju i preuzimaju podatke od interesa, a također sadrži i funkcije kao što su BLAST, obogaćivanje termina (engl. „Term Enrichment“), GO Slimmer alati, GOOSE („GO Online SQL Environment“) i vodič za pomoć korisnicima (Carbon i sur., 2008).

Alat za obogaćivanje termina određuje je li promatrana razina anotacije za grupu gena značajna u kontekstu pozadinskog skupa, obično svih gena u genomu. Ovo može biti korisno za otkrivanje povezanosti između gena. Alat za obogaćivanje termina koristi GO-TermFinder Perl modul i nudi brojne opcije za unos, filtriranje i izlazne podatke, uključujući vizualizaciju komponenti koja prikazuje značajno prezastupljene termine u kontekstu GO stabla (Carbon i sur., 2008).

3.1.7. AntiSMASH

AntiSMASH (Antibiotics & Secondary Metabolite Analysis Shell) je alat koji omogućava brzu identifikaciju, anotaciju i analizu genskih klastera koji su uključeni u sintezu sekundarnih metabolita u bakterijskim i fungalnim sekvencama. Ovaj alat može brzo detektirati sve poznate vrste genskih klastera koji sudjeluju u biosintezi sekundarnih metabolita, pružiti detaljnu funkcionalnu anotaciju NRPS/PKS kao i predvidjeti kemijsku strukturu NRPS/PKS produkata sa većom preciznosti nego ostale poznate metode. Također sadrži i komparativni modul analize klastera gena koji omogućuje detekciju i vizualizaciju evolucijskih sličnosti između klastera gena iz unesenog genoma i drugih genskih klastera kako bismo mogli brzo donijeti zaključke o funkcijama gena i operona na temelju homologije (Medema i sur., 2011).

Unos podataka moguć je učitavanjem datoteke nekog od podržanih formata (FASTA, GBK ili EMBL). Dobiveni rezultati mogu se vidjeti na internetskoj stranici na kojoj je moguće pregledavati različite klastere gena. Detektirani geni na kojima je temeljena identifikacija genskog klastera prikazani su u vlastitim bojama. Rezultati ClusterBlasta prikazani su na sličan način, kao mape poravnatih genskih klastera u kojima su geni sa dijeljenim BLAST rezultatima označeni identičnim bojama (Medema i sur., 2011).

3.2. METODE

3.2.1. Sekvencioniranje i sastavljanje genoma

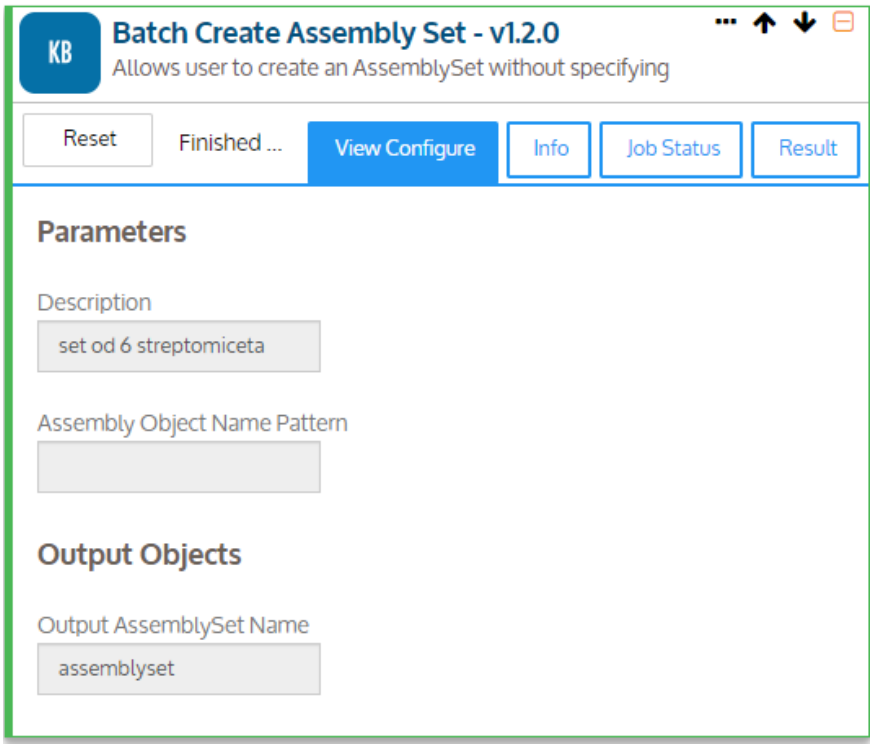
Genomi šest organizama roda Streptomyces izoliranih iz Jadranskog mora sekvencionirani su pomoću PacBio SMRT tehnologije sekvencioniranja. Dobiveni readovi sastavljeni su u contige pomoću Canu assemblera.

3.2.2. Procjena kvalitete assembly-ja

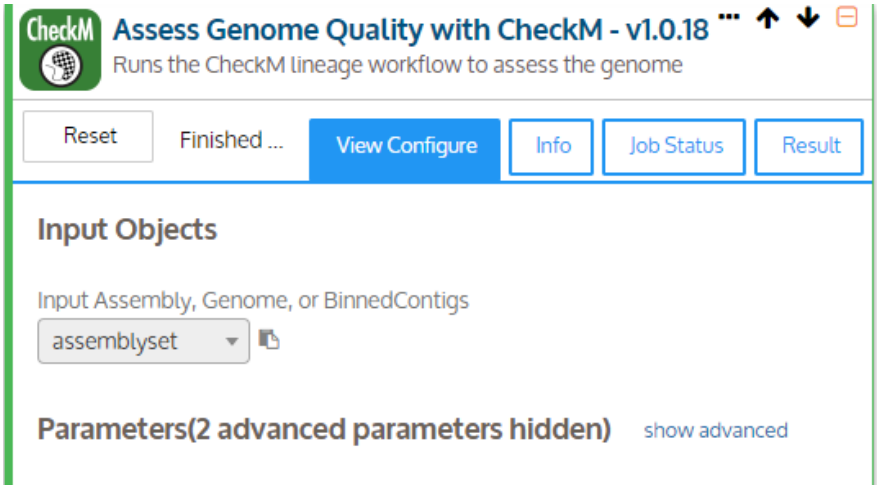
Šest assembly-ja (BC134, BC138, BC152, BC156, BC157 i BC164) učitano je u KBase narativ odabirom opcije „Add Data“. Tada je odabrana opcija „import“ i odabrani su assembly-ji koji su spremljeni na računalu u fasta formatu. U stupcu „Import As...“ potrebno je odabrati opciju „FASTA assembly“ za svaki učitani assembly te pritisnuti na „import selected“. Svaka aplikacija odabire se iz širokog popisa aplikacija koje su raspoređene po kategorijama te postoji mogućnost pretrage pomoću ključnih riječi. Procjena kvalitete assembly-ja napravljena je pomoću aplikacija QUAST i CheckM.

Iz popisa aplikacija odabrana je aplikacija QUAST. Ulazni podaci za aplikaciju QUAST su assembly-ji. Ovih šest assembly-ja učitano je u aplikaciju nakon čega je pokrenuta analiza istih pritiskom na tipku „Run“. Nakon određenog vremena dobiveni su rezultati analize u obliku tablice.

Kao ulazni parametar za aplikaciju CheckM, korišten je assembly set koji je prethodno napravljen pomoću aplikacije „Batch Create Assembly Set“ (Slika 6). Ovo je potrebno napraviti jer aplikacija CheckM dozvoljava unos samo jednog assembly-ja ili genoma. Kako bi se napravio assembly set potrebno je samo imenovati ga i pokrenuti aplikaciju. Aplikacija će tada šest samostalnih assembly-ja posložiti u jedan set koji se može koristiti kao ulazni parametar za ostale aplikacije. Za procjenu kvalitete assembly-ja aplikacijom CheckM potrebno je postaviti assembly set kao ulazni parametar i pokrenuti aplikaciju (Slika 7).



Slika 6. Ulazni parametri za kreiranje assembly seta

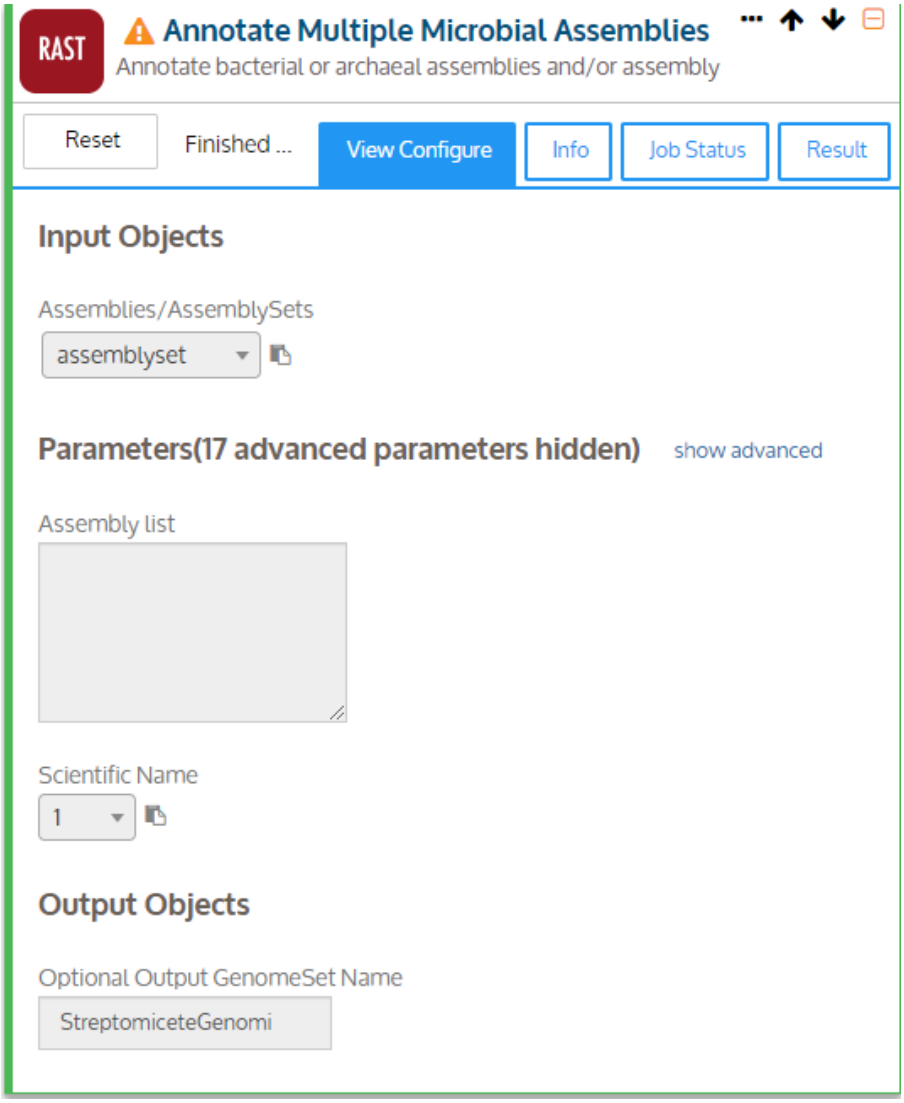


Slika 7. Ulazni parametri za aplikaciju CheckM

3.2.3. Anotacija assembly-ja

Za anotaciju assembly-ja korištena je aplikacija „Annotate Multiple Microbial Assemblies with RASTtk“ (Slika 8). Kao što joj samo ime govori, ova aplikacija omogućuje anotaciju više genoma odjednom i potpuno je automatizirana. Kao ulazni objekt korišten je prethodno napravljeni assembly set. Time je dodatno pojednostavljena anotacija assembly-ja te je manje vremena potrebno za anotaciju. Pokretanjem ove aplikacije generira se šest anotiranih genoma,

kao i set genoma u kojemu su svi oni na istome mjestu.

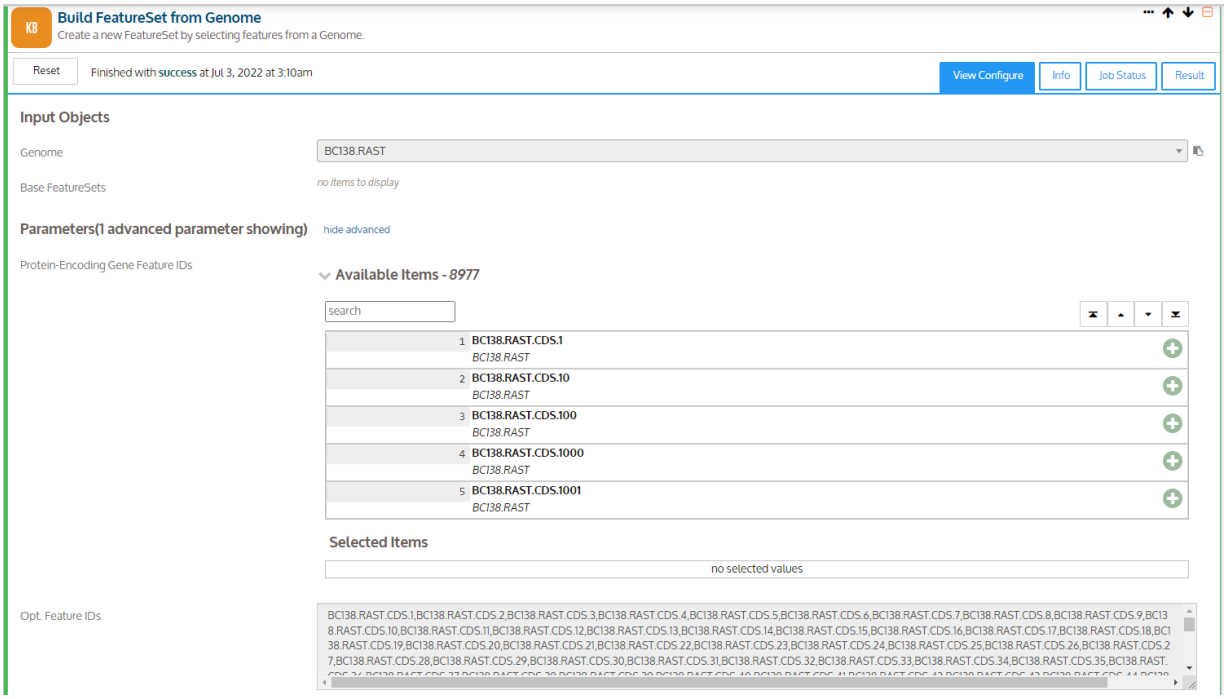


Slika 8. Ulazni parametri za aplikaciju Annotate Multiple Microbial Assemblies with RASTtk

3.2.4. Izrada Venn-ovih dijagrama

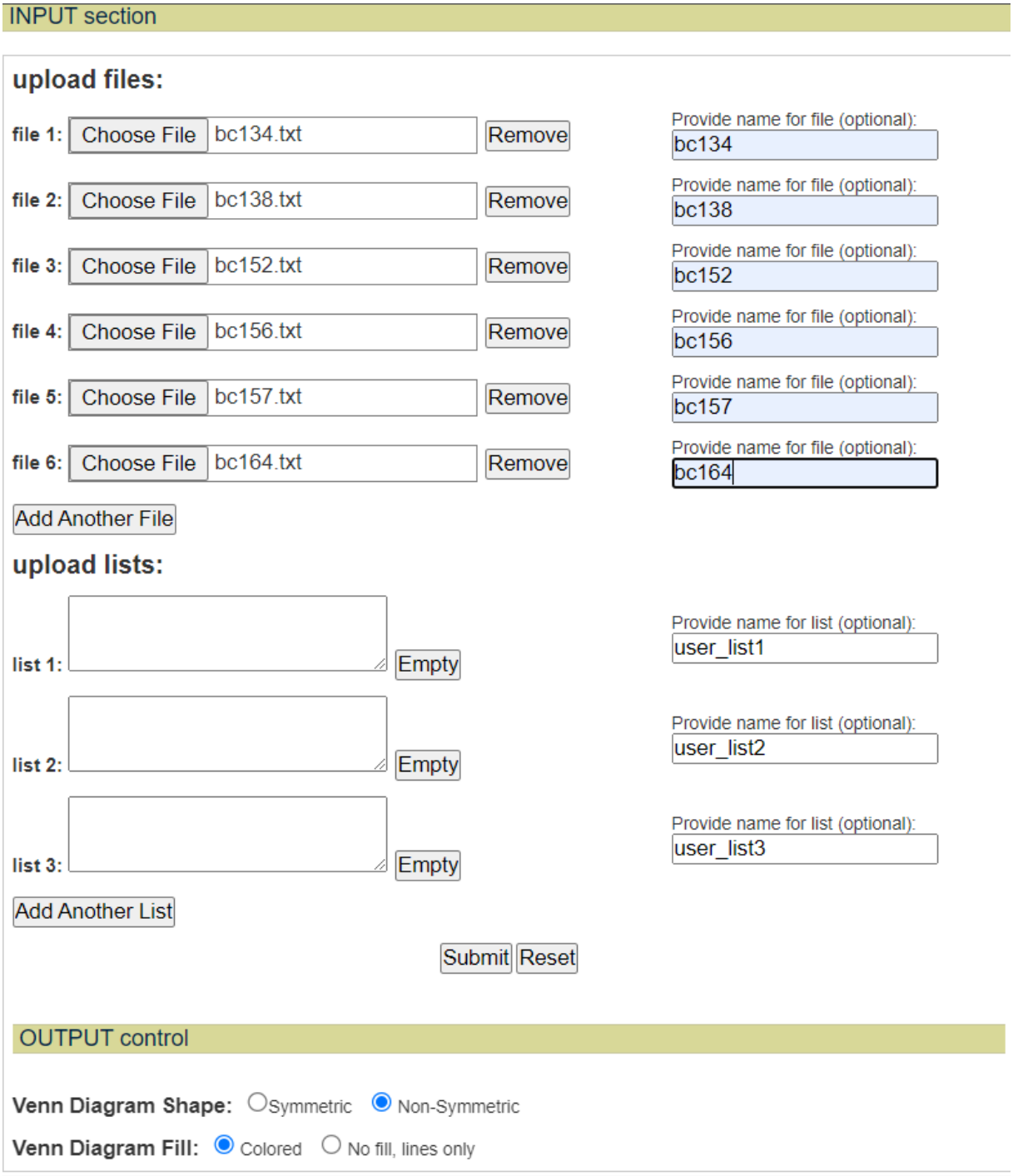
Prije izrade samih Venn-ovih dijagrama potrebno je od genoma napraviti tzv. Feature setove, odnosno setove s genima iz anotiranih genoma. Ovi setovi sadrže identifikacijske brojeve tih gena, kao i funkciju proteina za kojeg oni kodiraju. Za to je korištena aplikacija Build FeatureSet from Genome (Slika 9). Odabirom jednog anotiranog genoma aplikacija će izlistati sve gene koji postoje u tom genomu. Iz tog popisa gena odabiru se oni koji će činiti Feature set. Pošto je za izradu ovog diplomskog rada bilo potrebno uzeti sve gene iz svakog pojedinačnog organizma, korištena je opcija „Opt. Feature IDs“ koja postaje vidljiva pritiskom na opciju „show advanced“. Pod tom opcijom potrebno je navesti identifikacijske brojeve svih gena, odvojene zarezom i bez razmaka. Kako bi ovaj postupak bio lakši i brži, u Python-u je napravljen kod koji pravi listu svih identifikacijskih brojeva iz odabranog genoma te ih odvaja

zarezom. Unutar koda je potrebno mijenjati neke parametre poput količine identifikacijskih brojeva koji će činiti listu i imena organizma koje je sadržano unutar svakog identifikacijskog broja. Aplikacija je pokrenuta i ovaj postupak je ponovljen za svih šest genoma.



Slika 9. Ulazni parametri za aplikaciju Build FeatureSet from Genome

Ovi feature setovi preuzeti su na računalo u obliku .txt datoteka nakon čega je iz svake izbrisano sve osim funkcije proteina za koje kodiraju navedeni geni. Ovo je također učinjeno pomoću koda u Python-u koji omogućuje zamjenu identifikacijskih brojeva sa nekom drugom riječi i brisanje te riječi. Budući da se svaki identifikacijski broj sastojao od riječi i rastućih brojeva, bilo je potrebno prvo zamijeniti te brojke sa nekom riječi, a tek onda izbrisati tu riječ iz datoteke. Nakon što je to učinjeno, ove datoteke učitane su u web alat za izradu Venn-ovih dijagrama (Slika 10). Pošto je učitano šest datoteka, oblik Venn-ovog dijagrama mora biti asimetričan. Pritiskom na tipku „Submit“ dobiven je asimetričan Venn-ov dijagram koji prikazuje broj dijeljenih i individualnih gena u ovim genomima, kao i listu s imenima dijeljenih - zajedničkih proteina kao i onih koje sadrže pojedinačni organizmi - individualnih.

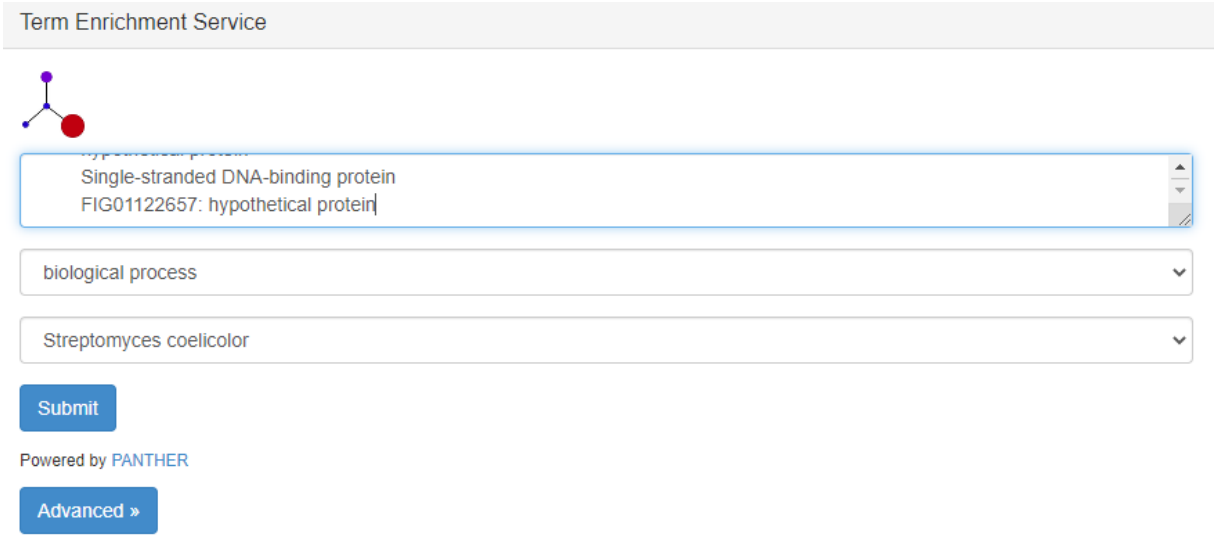


Slika 10. Ulazni podaci za izradu Venn-ovih dijagrama

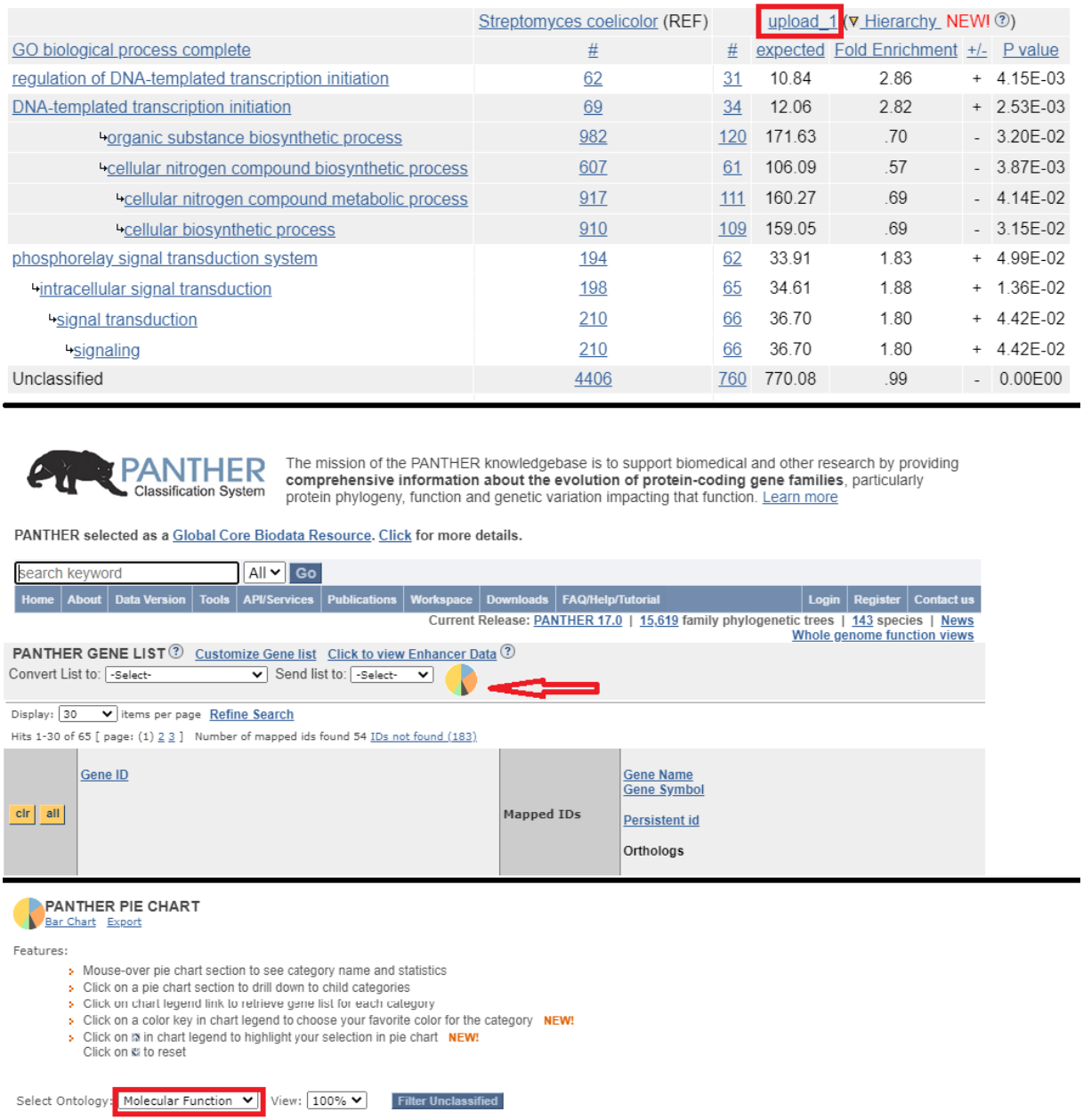
3.2.5. Ontologija gena

Ontologija gena napravljena je preko PANTHER-a kojem je pristupljeno preko korisničkog sučelja AmiGO 2. AmiGO 2 pruža brojne usluge vezane za ontologiju gena pa tako i uslugu za „term enrichment“ (Slika 11). Kao ulazni podaci korištena su imena proteina ovih šest pojedinačnih organizama, a kao referentni organizam naveden je *Streptomyces coelicolor*.

Taj organizam je odabran jer je to jedini organizam roda *Streptomyces* koji postoji unutar PANTHER-a, a ujedno je i modelni organizam te pripada rodu *Streptomyces*. Pritiskom na tipku „Submit“ ovaj će alat usporediti navedene proteine sa odabranim referentnim organizmom te će se učitati tablica u kojoj se nalazi kompletan pregled ontoloških termina za odabranu kategoriju ontologije gena, p-vrijednost, broj identificiranih gena u svakoj skupini za referentni genom i učitanu listu proteina te stupac u kojemu je naznačeno jesu li ti geni prezastupljeni (engl. „overrepresented“) ili podzastupljeni (engl. „underrepresented“). Kako bi se došlo do vizualnog prikaza u obliku tortnog dijagrama (engl. „pie chart“) potrebno je kliknuti na „upload_1“ gdje će se otvoriti tablica sa proteinima iz odabranog organizma koje je PANTHER identificirao. Iznad te tablice nalazi se ikona tortnog dijagrama. Pomicanjem kursora miša preko te ikone nude se opcije za vizualizaciju pojedinačnih kategorija ontologije gena. Odabirom jedne od njih otvara se grafički prikaz za tu kategoriju (Slika 12). Pod opcijom „Select Ontology“ moguće je mijenjati kategoriju ontologije i time dobiti grafički prikaz za ostale kategorije ontologije gena. Ovaj proces potrebno je ponoviti za svih šest organizama.



Slika 11. Ulazni podaci na AmiGO 2 sučelju



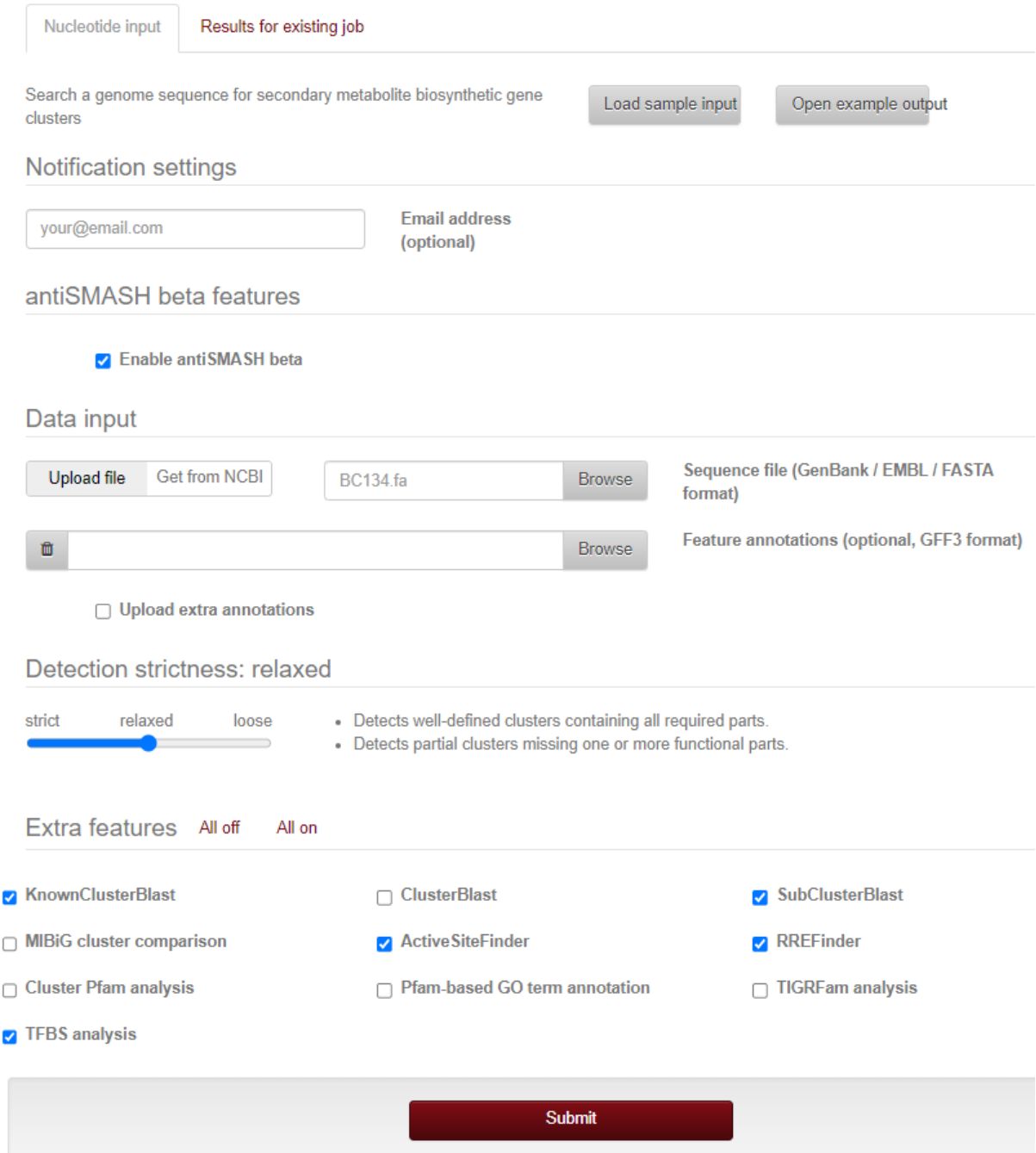
	Streptomyces coelicolor (REF)		upload_1 (▿ Hierarchy NEW!)			
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
regulation of DNA-templated transcription initiation	62	31	10.84	2.86	+	4.15E-03
DNA-templated transcription initiation	69	34	12.06	2.82	+	2.53E-03
↳organic substance biosynthetic process	982	120	171.63	.70	-	3.20E-02
↳cellular nitrogen compound biosynthetic process	607	61	106.09	.57	-	3.87E-03
↳cellular nitrogen compound metabolic process	917	111	160.27	.69	-	4.14E-02
↳cellular biosynthetic process	910	109	159.05	.69	-	3.15E-02
phosphorelay signal transduction system	194	62	33.91	1.83	+	4.99E-02
↳intracellular signal transduction	198	65	34.61	1.88	+	1.36E-02
↳signal transduction	210	66	36.70	1.80	+	4.42E-02
↳signaling	210	66	36.70	1.80	+	4.42E-02
Unclassified	4406	760	770.08	.99	-	0.00E00

Slika 12. Redoslijed operacija za vizualizaciju ontoloških termina korištenjem PANTHER-a

3.2.6. Identifikacija genskih klastera koji su uključeni u sintezu sekundarnih metabolita

Za identifikaciju genskih nakupina koje su uključene u sintezu sekundarnih metabolita korišten je web alat po imenu AntiSMASH. Kao ulazni podaci mogu se koristiti NCBI pristupni broj (engl. „accession number“) ili datoteka s računala u podržanom formatu. Za ovaj rad korištena je opcija „Upload file“ te su pojedinačno učitani assembly-ji u fasta formatu (Slika 13). Također postoji mogućnost davanja e-mail adrese prije pokretanja aplikacije kako bi rezultati analize stigli korisniku na mail jednom kada je proces završen. Moguće je mijenjati i strogost detekcije, a ponuđene su tri opcije „strict“, „relaxed“ i „loose“. Odabrana je zadana opcija „relaxed“ koja detektira dobro definirane genske nakupine, kao i neke djelomične

nakupine kojima nedostaju neki funkcionalni dijelovi. Odabrane su i neke od ponuđenih dodatnih opcija kao što je to prikazano na slici 13. Pritiskom na tipku „Submit“ ovaj alat počinje sa anotacijom genskih klastera i kada završi daje tablični prikaz rezultata. Ovaj proces potrebno je ponoviti za svih 6 assembly-ja pojedinačno.



Slika 13. Ulazni podaci za AntiSMASH

4. REZULTATI I RASPRAVA

Za procjenu kvalitete assembly-ja koriste se brojni alati poput QUAST i CheckM. Oni se razlikuju prema rezultatima koje daju. QUAST je alat koji pruža informacije o tehničkim parametrima poput N50, dok CheckM daje informaciju o kompletnosti genoma i kontaminacijama. CheckM se često uspoređuje s alatom po imenu BUSCO jer oba alata izvršavaju funkcionalnu procjenu kvalitete assembly-ja. Na prokariotskom genomu, BUSCO ima rezultate usporedive s CheckM-om, a velika prednost BUSCO-a leži u činjenici da može detektirati i procijeniti eukariotske genome. Također, CheckM zahtijeva puno više memorije za procijenu bakterijskih genoma pa je za laptope sa ograničenom memorijom i CPU resursima preporučljivo korištenje BUSCO alata. Međutim, procijenom duplikacija, odnosno kontaminacija u genomu pomoću ova dva alata dobivaju se isti rezultati tako da odabir alata ne utječe značajno na rezultate istraživanja (Manni i sur., 2021). Venn-ovi dijagrami u bioinformatici koriste se za analiziranje povezanosti između bioloških podataka i njihovu vizualizaciju (Jia, Xu i Wang, 2021). Na Venn-ove dijagrame i ontologiju gena utječe anotacija jer su upravo rezultati anotacije korišteni kao ulazni podaci. Dvije najpoznatije aplikacije koje omogućuju anotaciju assembly-ja unutar KBase-a su RAST i Prokka. Karimi i sur. (2021) navode kako Prokka u prosjeku proizvodi više EC brojeva i veće metaboličke mreže, a RAST proizvodi manje lažno pozitivnih reakcija i tzv. „dead-end“ metabolita. Za pronalaženje genskih nakupina koje kodiraju za proteine sekundarnog metabolizma koristi se AntiSMASH koji daje tablični i grafički prikaz tih nakupina. U ovome radu, rezultati AntiSMASH analize biti će prikazani tablično.

4.1. REZULTATI PROCJENE KVALITETE ASSEMBLY-JA POMOĆU APLIKACIJA QUAST I CHECKM UNUTAR KBASE-A

Tablica 1 prikazuje rezultate procjene kvalitete assembly-ja za 6 organizama. Prvi dio tablice prikazuje parametre dobivene pomoću aplikacije QUAST. Idealno je da se genom sastoji od što manjeg broja contiga. Za navedene uzorke ovaj parametar varira od idealnog do jako velikog broja. Veliki dio duljine svih uzoraka čini jedan (najdulji) contig, čija je duljina također navedena u tablici 1. BC138 ima najveću ukupnu duljinu, a BC157 najmanju. BC138 također ima najveću N50 vrijednost (zbroj duljina svih contiga, počevši od najvećeg, koji je veći od 50 % duljine genoma), a BC152 najmanju. L50 prikazuje broj contiga koji čine N50. Svi ovi

genomi zapravo sadrže po jedan toliko dugačak contig da čini više od 90 % genoma. Prema broju contiga, najbolji uzorak je BC157, a prema N50 najbolji je BC138. Međutim, N50 nije dobar pokazatelj kvalitete assembly-ja, a to se najbolje vidi na primjeru BC156 koji se sastoji od jednog jako velikog contiga i puno manjih, a ima veću N50 vrijednost nego BC157 koji se sastoji od samo jednog contiga. Zbog toga je potrebno napraviti i CheckM analizu kako bi se dobila pouzdanija usporedba.

Drugi dio tablice 1 prikazuje rezultate dobivene pomoću aplikacije CheckM. Za ovih šest organizama iz roda *Streptomyces* generirano je 563 markera. Markeri koji nedostaju daju informaciju o kompletnosti genoma, a markeri koji postoje u više kopija ukazuju na kontaminacije. Markeri koji postoje u samo jednoj kopiji (engl. „single-copy marker“) su poželjni. Iz ovih rezultata vidi se da je kompletnost genoma poprilično visoka i da ima malo kontaminacija. Prema tome najbolje sastavljeni genom je BC164 jer ima najveću kompletnost i najmanje kontaminacija.

Tablica 1. Rezultati procjene kvalitete assemblyja dobiveni pomoću aplikacija QUAST i CheckM

		BC164	**BC157**	**BC156**	**BC152**	**BC138**	**BC134**
Broj contiga	ukupno	4	1	11	4	2	6
	>=1000pb	4	1	11	4	2	6
	>= 10000 bp	3	1	10	3	2	5
	>= 100000 bp	2	1	3	3	1	3
	>= 1000000 bp	1	1	1	1	1	1
Duljina najduljeg contiga		8450404	8241966	8349386	8221791	9650943	8445163
Ukupna duljina	Svih contiga	8639416	8241966	8902767	8620176	9690616	9135422
	>= 1000 bp	8639416	8241966	8902767	8620176	9690616	9135422

Tablica 1. Rezultati procjene kvalitete assemblyja dobiveni pomoću aplikacija QUAST i

CheckM - *nastavak*

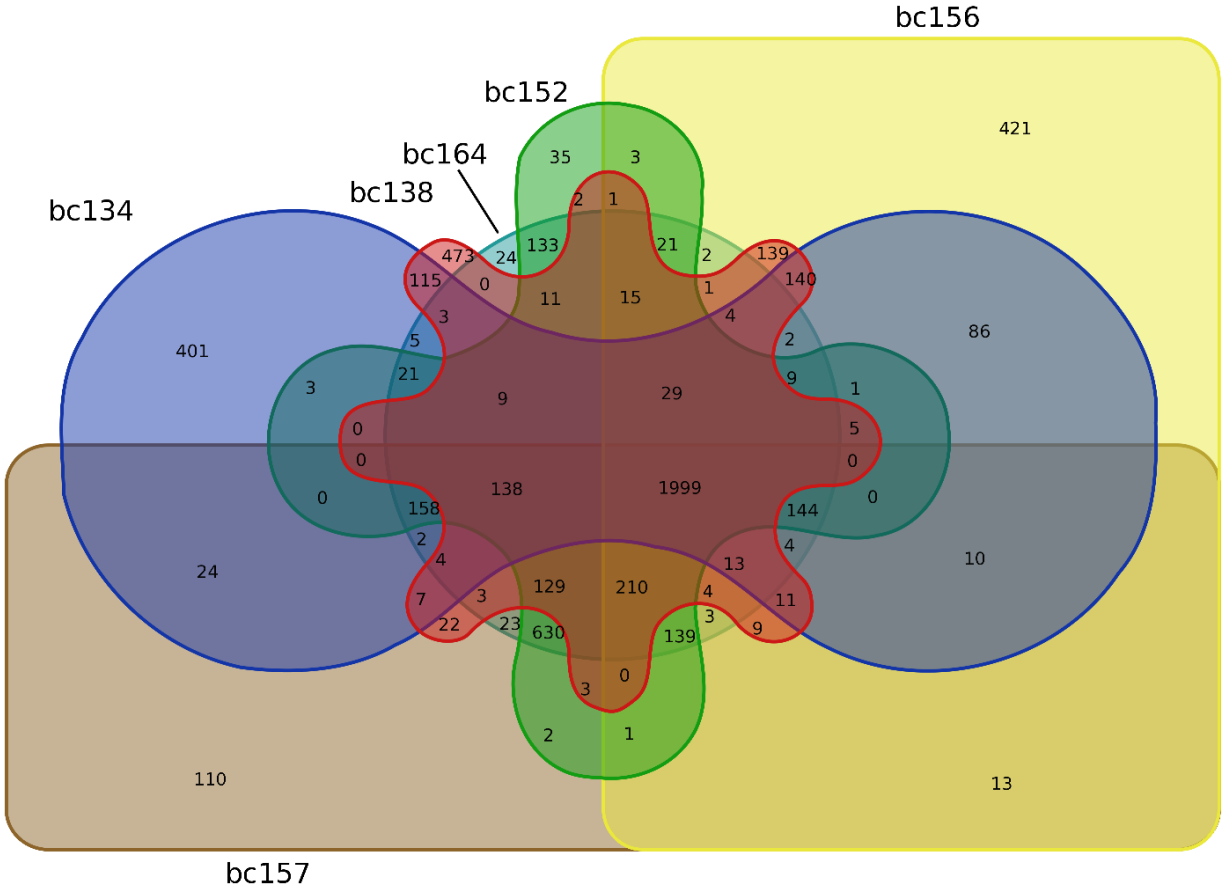
	>= 10000 bp	8638121	8241966	8900793	8618202	9690616	9133448
	>= 100000 bp	8624879	8241966	8617753	8618202	9650943	9033849
	>= 1000000 bp	8450404	8241966	8349386	8221791	9650943	8445163
N50 (i N75)		8450404	8241966	8349386	8221791	9650943	8445163
L50 (i L75)		1	1	1	1	1	1
GC (%)		71.48	71.48	71.74	71.46	70.8	68.99
Kompletnost genoma (%)		99,7	97,39	97,81	97,64	99,44	97,5
Kontaminacija (%)		0,0	0,07	2,48	0,0	1,42	2,47
Broj markera		563	563	563	563	563	563
Broj markera koji nedtostaju	**0**	3	18	12	15	3	14
Broj single-copy markera	**1**	560	544	538	548	553	537
Broj markera sa više kopija	**2**	0	1	13	0	7	12
	3	0	0	0	0	0	0
	4	0	0	0	0	0	0
	5+	0	0	0	0	0	0

Ljubičasto – najveća duljina, bijelo – medijan, crveno – najmanja duljina

4.2. VENN-OV DIJAGRAM

Slika 14 prikazuje asimetrični venn-ov dijagram za šest uzoraka. Svaka od obojenih zona prikazuje koliko proteina se nalazi u navedenom organizmu, a područja preklapanja prikazuju brojnost proteina koji su zajednički tim organizmima. Iz tih podataka može se iščitati koliko je proteina zajedničko ovim organizmima, a koliko njih je prisutno u samo jednom od organizama. Moguće je da su ove brojke veće nego što je to prikazano na slici zbog toga što je anotacijom pomoću RAST alata dobiveno puno tzv. hipotetskih proteina, a alat koji je korišten za izradu

Venn-ovih dijagrama ima svojstvo da sve linije u dokumentu koje su iste tretira kao samo jednu liniju, a svaku istu zanemaruje. RAST je također detektirao puno gena kojima nije određena nikakva funkcija pa dokument sadrži i puno praznih linija koje su također zanemarene prilikom izrade Venn-ovih dijagrama. Međutim, pošto nije riječ o proteinima koji su od značajne važnosti, već o pretpostavljenim proteinima za koje nije sigurno da se eksprimiraju ili im se ne može odrediti neka funkcija, ovi gubici i nisu previše značajni.

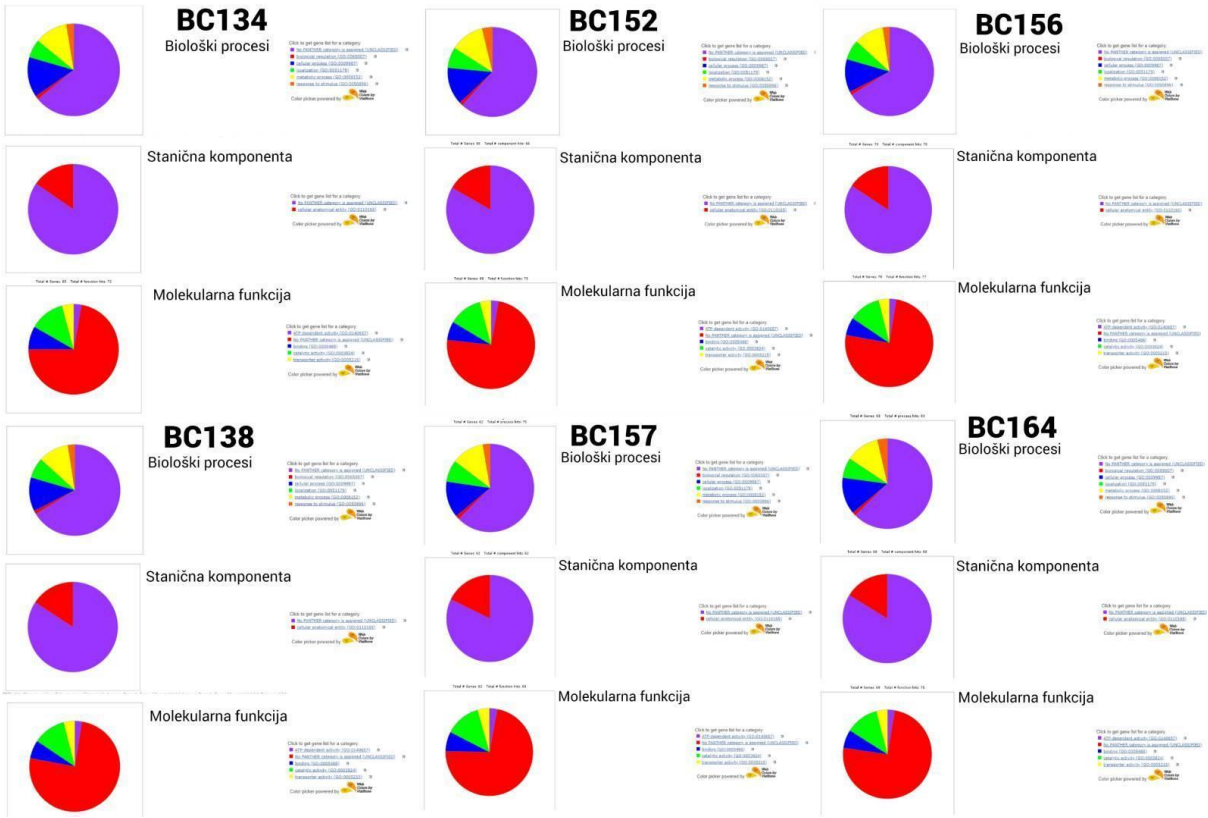


Slika 14. Venn-ov dijagram za genome šest organizama roda *Streptomyces*

4.3. ONTOLOGIJA GENA

Na slici 15 prikazani su tortni dijagrami koji prikazuju koji udio detektiranih proteina pripada u navedene ontološke termine. Iz ovih rezultata može se vidjeti kako za sve organizme i sve kategorije ontologije gena većina proteina pripada u nerazvrstanu skupinu za koju nije dodijeljena PANTHER kategorija. Razlog tome može biti činjenica da u PANTHER-u još nisu opisane mnoge kategorije ontoloških termina u koje bi ovi proteini bili svrstani. Nadalje, brojno stanje detektiranih proteina manje je od očekivanog, a razlog tome može biti činjenica da je kao referentni organizam odabran *Streptomyces coelicolor* jer je to jedini organizam roda *Streptomyces* koji je prisutan u PANTHER-u. Stoga je svim detektiranim proteinima dodijeljen

UniProt pristupni broj (engl. „accession number“) koji pripada tim proteinima iz bakterije vrste *Streptomyces coelicolor*. Moguće je da postoje proteini koje ove vrste proizvode, a *Streptomyces coelicolor* ne proizvodi te da su se tako izgubili u procesu. Također, moguće je da neki od proteina iz ovih organizama ne postoje unutar UniProt baze podataka jer se možda radi o novim ili slabo okarakteriziranim organizmima. Još jedan od razloga može biti i nesavršenost RAST alata za anotaciju assembly-ja. Anotacijom je dobiveno puno hipotetskih proteina i proteina kojima nije dodijeljen nikakav opis. Takve proteine PANTHER nije mogao detektirati te su zato izostavljeni.



Slika 15. Grafički prikaz ontologije gena za šest organizama roda Streptomyces

Legenda: Biološki procesi: ljubičasto-nesvrstano; crveno-biološka regulacija; plavo-stanični proces; zeleno-lokalizacija; žuto-metabolički proces; narančasto-odgovor na stimulus; Stanična komponenta: ljubičasto-nesvrstano; crveno-stanična anatomska cjelina; Molekularna funkcija: ljubičasto-ATP ovisna aktivnost; crveno-nesvrstano; plavo-vezanje; zeleno-katalitička aktivnost; žuto-transportna aktivnost

4.4. REZULTATI ANTISMASH ANALIZE

AntiSMASH analizom dobiveno je šest tablica u kojima su prikazane sve genske nakupine koje sadrže gene koji kodiraju za sekundarne metabolite iz navedenih organizama. Oni su

razvrstani po contizima na kojima se nalaze, a neki od contiga ne sadrže takve genske klastere.

Tablica 2. Rezultati antiSMASH analize za organizam BC134

Genska nakupina	**BC134**	**BC138**	**BC152**	**BC156**	**BC157**	**BC164**
1.1	+	-	-	-	-	-
1.2	+	-	-	-	-	-
3.1	+	-	-	-	-	-
3.2	+	-	+	+	+	+
3.3	+	+	-	+	-	-
3.4	+	-	+	-	-	+
3.5	+	-	-	-	-	-
3.6	+	-	-	-	-	-
3.7	+	+	+	+	+	+
3.8	+	+	-	-	-	-
3.9	+	-	-	-	-	-
3.10	+	-	-	-	+	-
3.11	+	-	+	+	+	+
3.12	+	-	-	-	-	+
3.13	+	-	-	-	-	-
3.14	+	+	-	-	-	-
3.15	+	-	-	-	-	-
3.16	+	+	+	+	+	+
3.17	+	-	-	-	-	-
3.18	+	-	-	+	-	-
3.19	+	-	-	-	-	-
3.20	+	-	-	-	-	-

Tablica 3. Rezultati antiSMASH analize za organizam BC156

Genska nakupina	**BC134**	**BC138**	**BC152**	**BC156**	**BC157**	**BC164**
1.1	-	+	-	+	-	-
1.2	-	-	-	+	+	+
1.3	-	-	-	+	-	-
1.4	+	-	+	+	+	+
1.5	-	-	-	+	-	-
1.6	-	+	-	+	-	-
1.7	-	-	+	+	+	+
1.8	-	-	-	+	-	-
1.9	-	-	-	+	-	-
1.10	-	-	-	+	-	-
1.11	+	+	+	+	+	+

Tablica 3. Rezultati antiSMASH analize za organizam BC156 – *nastavak*

Genska nakupina	**BC134**	**BC138**	**BC152**	**BC156**	**BC157**	**BC164**
1.12	+	-	-	+	-	-

1.13	-	-	-	+	-	-
1.14	-	+	+	+	-	-
1.15	-	-	-	+	-	-
1.16	-	-	-	+	-	-
1.17	-	-	-	+	-	-
1.18	-	-	-	+	-	-
1.19	-	-	-	+	-	-
1.20	-	-	-	+	-	-
1.21	-	-	-	+	-	-
1.22	-	-	-	+	-	-
1.23	-	-	-	+	-	-
1.24	-	-	-	+	-	-
1.25	-	-	-	+	-	-
1.26	-	-	-	+	-	-
1.27	+	+	+	+	+	+
1.28	-	-	-	+	-	-
4.1	-	-	-	+	-	-
6.1	-	-	-	+	-	-

Tablica 4. Rezultati antiSMASH analize za organizam BC157

Genska nakupina	**BC134**	**BC138**	**BC152**	**BC156**	**BC157**	**BC164**
1	-	-	-	-	+	-
2	-	+	+	-	+	+
3	-	-	-	-	+	+
4	-	-	-	-	+	-
5	-	-	-	-	+	-
6	-	-	-	-	+	-
7	-	-	+	-	+	+
8	-	-	-	-	+	-
9	+	+	+	+	+	+
10	+	-	+	-	+	+
11	-	-	-	-	+	-
12	-	-	-	-	+	-
13	-	-	-	-	+	-
14	-	-	-	-	+	-
15	-	-	-	-	+	-
16	-	-	-	-	+	-
17	-	-	-	-	+	-
18	-	-	+	-	+	+
19	-	-	-	-	+	-
20	-	-	+	+	+	+
21	-	-	-	-	+	-

Tablica 4. Rezultati antiSMASH analize za organizam BC157 – *nastavak*

Genska nakupina	**BC134**	**BC138**	**BC152**	**BC156**	**BC157**	**BC164**
22	-	-	-	-	+	-

23	-	-	+	-	+	+
24	-	-	-	-	+	-
25	-	-	-	-	+	-
26	+	+	+	+	+	+
27	-	-	-	-	+	+
28	-	-	-	-	+	-
29	-	-	+	-	+	+
30	+	-	+	+	+	+
31	-	-	-	-	+	-
32	-	-	-	+	+	+
33	-	-	-	-	+	-

Tablica 5. Rezultati antiSMASH analize za organizam BC152

Genska nakupina	**BC134**	**BC138**	**BC152**	**BC156**	**BC157**	**BC164**
1.1	-	-	+	-	-	+
1.2	-	-	+	-	-	-
1.3	+	-	+	+	+	+
1.4	-	-	+	-	+	+
1.5	-	-	+	-	-	+
1.6	-	-	+	-	-	-
1.7	+	-	+	+	+	+
1.8	-	-	+	-	-	-
1.9	+	-	+	-	-	+
1.10	-	-	+	-	+	+
1.11	-	-	+	-	-	+
1.12	-	+	+	+	-	-
1.13	-	-	+	+	+	+
1.14	-	-	+	-	+	+
1.15	-	-	+	-	-	+
1.16	-	-	+	-	-	+
1.17	-	-	+	-	-	+
1.18	-	-	+	-	-	+
1.19	-	-	+	-	-	+
1.20	-	-	+	-	-	+
1.21	-	-	+	-	-	+
1.22	-	-	+	-	-	+
1.23	-	-	+	-	-	-
1.24	-	-	+	-	-	+
1.25	+	+	+	+	+	+
1.26	-	-	+	-	+	+
1.27	-	-	+	-	-	-
1.28	-	-	+	-	-	+

Tablica 5. Rezultati antiSMASH analize za organizam BC152 - *nastavak*

Genska nakupina	**BC134**	**BC138**	**BC152**	**BC156**	**BC157**	**BC164**
1.29	-	-	+	-	-	+

1.30	-	-	+	-	-	+
2.1	-	-	+	-	+	+
2.2	-	-	+	-	-	-
2.3	-	-	+	-	-	+

Tablica 6. Rezultati antiSMASH analize za organizam BC164

Genska nakupina	**BC134**	**BC138**	**BC152**	**BC156**	**BC157**	**BC164**
4.1	-	-	+	-	-	+
4.2	-	-	-	+	+	+
4.3	+	-	+	+	+	+
4.4	-	-	+	-	+	+
4.5	-	-	+	-	-	+
4.6	-	-	+	-	+	+
4.7	+	-	+	+	+	+
4.8	+	-	+	-	-	+
4.9	-	-	+	-	+	+
4.10	-	-	+	-	+	+
4.11	+	-	-	-	-	+
4.12	-	-	+	+	+	+
4.13	-	-	+	-	+	+
4.14	-	-	+	-	-	+
4.15	-	-	-	-	-	+
4.16	-	-	+	-	-	+
4.17	-	-	+	-	-	+
4.18	-	-	+	-	-	+
4.19	-	-	+	-	-	+
4.20	-	-	+	-	-	+
4.21	-	-	+	-	-	+
4.22	-	-	+	-	-	+
4.23	-	-	-	-	-	+
4.24	-	-	+	-	-	+
4.25	+	+	+	+	+	+
4.26	-	-	+	-	+	+
4.27	-	-	-	-	-	+
4.28	-	-	+	-	-	+
4.29	-	-	+	-	-	+
4.30	-	-	+	-	-	+
4.31	-	-	-	-	+	+
4.32	-	-	+	-	+	+
4.33	-	-	-	-	-	+

Tablica 7. Rezultati antiSMASH analize za organizam BC138

Genska nakupina	**BC134**	**BC138**	**BC152**	**BC156**	**BC157**	**BC164**
2.1	-	+	-	-	-	-
2.2	-	+	-	-	-	-

2.3	-	+	-	-	-	-
2.4	-	+	-	-	-	-
2.5	-	+	-	-	-	-
2.6	-	+	-	-	-	-
2.7	-	+	-	-	-	-
2.8	-	+	-	-	-	-
2.9	-	+	-	+	-	-
2.10	-	+	-	-	-	-
2.11	-	+	-	-	-	-
2.12	+	+	-	-	-	-
2.13	-	+	-	-	-	-
2.14	+	+	-	-	-	-
2.15	-	+	-	-	-	-
2.16	-	+	-	-	-	-
2.17	-	+	-	-	-	-
2.18	+	+	+	+	+	+
2.19	-	+	-	-	-	-
2.20	-	+	-	-	-	-
2.21	-	+	-	-	-	-
2.22	-	+	-	-	-	-
2.23	-	+	-	-	-	-
2.24	-	+	-	-	-	-
2.25	-	+	+	+	-	-
2.26	-	+	-	-	-	-
2.27	+	+	+	+	+	+
2.28	-	+	-	-	-	-
2.29	-	+	-	-	-	-
2.30	-	+	-	-	-	-
2.31	-	+	-	-	-	-
2.32	-	+	-	-	-	-
2.33	+	+	-	-	-	-
2.34	-	+	-	-	-	-
2.35	-	+	-	-	-	-
2.36	-	+	-	-	-	-

Ove tablice prikazuju sažete rezultate analize, a detaljna analiza prikazana je u prilozima.

Iz rezultata dobivenih antiSMASH analizom može se utvrditi pripada li pronađena genska nakupina u NRPS, PKS ili neki drugi sustav. Ovaj alat pruža detaljnu klasifikaciju koja uključuje mnogo tipova genskih nakupina, kao i informaciju o tome koja poznata genska nakupina je najsličnija pronađenoj i u kojem postotku. Zbog navedenoga moguće je s lakoćom usporediti rezultate analiza različitih organizama. Iz priloženih tablica vidi se da su brojne genske nakupine, poput melanina, zajedničke svim organizmima, ali također postoje one jedinstvene, koje se pojavljuju u samo jednom od organizama. BC138 je organizam sa najviše jedinstvenih genskih nakupina koje ne dijeli sa ostalim analiziranim bakterijama, a BC152 i

BC164 su međusobno najsličniji organizmi prema svom biosintetskom potencijalu. Ovime se pokazalo da su ovih šest organizama izoliranih iz Jadranskog mora potencijalno vrlo dobar izvor različitih sekundarnih metabolita, a antiSMASH se pokazao kao pouzdan i veoma detaljan alat za pronalaženje genskih nakupina koje kodiraju za sekundarne metabolite. Ovo istraživanje je važno zato što su streptomicete provjeren i bogat izvor ljekovitih supstanci, a bioinformatički alati poput KBase-a i antiSMASH-a su omogućili da se složena analiza genoma ovih organizama približi velikom broju istraživača, što je donedavno bilo teško zamislivo.

5. ZAKLJUČCI

1. Korištenjem aplikacija QUAST i CheckM utvrđeno je da se radi o kvalitetnim assembly-jima. Svaki od njih sadrži jedan veoma dugačak contig koji osigurava savršenu vrijednost L50 koja iznosi 1 za sve assembly-je. Nadalje, kompletnost svakog od genoma je iznad 97 %, a postotak kontaminacija ne premašuje 2,5 %. To su jako dobri rezultati s obzirom da je poželjna što veća kompletnost genoma i što manje kontaminacija.
2. Web alat koji je korišten za izradu Venn-ovih dijagrama pronašao je 1999 istih genskih produkata u svih šest analiziranih vrsta. Geni koji kodiraju za njih vjerojatno se nalaze u blokovima sintenije i ortologni su, uz iznimku hipotetskih proteina. Svaki od navedenih genoma također sadrži i određen broj genskih produkata koji su specifični za tu vrstu, a neki od njih mogli bi biti od značajne važnosti.
3. Ontologija gena pokazala je da je, ako se zanemare nesvrstani proteini, većina njih pripala u metaboličke i stanične procese za kategoriju biološki procesi, staničnu anatomsku cjelinu za kategoriju stanična komponenta te katalitičku aktivnost za kategoriju molekularna funkcija. Neki od tih proteina mogli bi biti enzimi sekundarnog metabolizma koji sudjeluju u procesu sinteze metabolita korisnih za čovječanstvo.
4. AntiSMASH analiza ovih šest organizama roda *Streptomyces* pokazala je da oni imaju biosintetski potencijal za proizvodnju raznih korisnih spojeva kao što su bafilomicin B1 koji se koristi kao antibiotik, pigment melanin koji ima različite primjene i ektoin koji se koristi u kozmetici.

6. LITERATURA

Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, i sur. (2018) KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol* **36** 566-569. https://doi.org/10.1038/nbt.4163

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM i sur. (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29. https://doi.org/10.1038/75556.

Barbuto Ferraiuolo S, Cammarota M, Schiraldi C, Restaino OF (2021) Streptomycetes as platform for biotechnological production processes of drugs. *Appl Microbiol Biot* **105**, 551–568. https://doi.org/10.1007/s00253-020-11064-2

bioinformatics.psb.ugent.be. (n.d.) *Draw Venn Diagram*. [online] Dostupno na: http://bioinformatics.psb.ugent.be/webtools/Venn/.

Brettin T, Davis J, Disz T, Edwards RA, Gerdes S, Olsen GJ (2015) RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep* **5**. https://doi.org/10.1038/srep08365

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S (2008) AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288–289. https://doi.org/10.1093/bioinformatics/btn615.

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

Heather JM, Chain B (2016) The Sequence of sequencers: the History of Sequencing DNA. *Genomics* **107**, 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003.

Jia A, Xu L, Wang Y (2021) Venn diagrams in bioinformatics. *Brief Bioinform* [online] **22**, bbab108. https://doi.org/10.1093/bib/bbab108.

Jupyter (2019) Project Jupyter. [online] Jupyter.org. Dostupno na: https://jupyter.org/.

Karimi E, Geslain E, Belcour A, Frioux C, Aïte M, Siegel A i sur. (2021) Robustness analysis of metabolic predictions in algal microbial communities based on different annotation pipelines. *PeerJ* **9**, e11344. https://doi.org/10.7717/peerj.11344

Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* https://doi.org/10.1101/gr.215087.116

Liu D, Hunt M, Tsai IJ (2018) Inferring synteny between genome assemblies: a systematic evaluation. *BMC Bioinformatics* **19**, 26. https://doi.org/10.1186/s12859-018-2026-4

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021) BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**, 4647–4654. https://doi.org/10.1093/molbev/msab199.

Martínez-Núñez MA, López VELy (2016) Nonribosomal peptides synthetases and their applications in industry. *Sustain Chem Process* **4**, 13. https://doi.org/10.1186/s40508-016-0057-6

McNaught AD, Wilkinson A, Union I (1997) Compendium of chemical terminology: IUPAC recommendations, Oxford Blackwell Science, str. 1150-1151.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA i sur. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* **39**, W339–W346. https://doi.org/10.1093/nar/gkr466.

Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, Thomas PD (2020) PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res* **49**, D394–D403. https://doi.org/10.1093/nar/gkaa1106.

Overbeek R, Olson R, Pusch G., Olsen GJ, Davis JJ, Disz T i sur. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* **42**, D206–D214. https://doi.org/10.1093/nar/gkt1226.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2014) Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055. https://doi.org/10.1101/gr.186072.114

Quail MA, Smith M, Coupland P i sur. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341. https://doi.org/10.1186/1471-2164-13-341

Rimac H (2013) Anotacija genskih nakupina prirodnih spojeva u genomu bakterije *Streptomyces* sp. C (diplomski rad), Prehrambeno-biotehnološki fakultet, Sveučilište u Zagrebu, Zagreb.

Risdian C, Mozef T, Wink J (2019) Biosynthesis of Polyketides in Streptomyces. *Microorganisms* **7**, 124. https://doi.org/10.3390/microorganisms7050124.

Ruiz B, Chávez A, Forero A, García-Huante Y, Romero A, Sánchez M i sur. (2010). Production of microbial secondary metabolites: Regulation by the carbon source. *Crit Rev Microbiol* **36**, 146–167. https://doi.org/10.3109/10408410903489576

www.youtube.com. (2020). PacBio Sequencing – How it Works. [online] Dostupno na: https://www.youtube.com/watch?v=_lD8JyAbwEo&ab_channel=PacBio [Pristupljeno 19. svibnja 2023].

7. PRILOZI

Prilog 1. Rezultati antiSMASH analize za organizam BC134

Regija	**Tip**	**Pozicija od**	**Pozicija do**	**Najsličniji poznati klaster**		**Sličnost (%)**
tig00000002						
1.1	butirolakton	24,590	39,950	laktonamicin	Poliketid	10
1.2	Fragment sličan NRPS, fragment sličan PKS, NRPS	57,932	168,471	Virginiamicin S1	NRP+poliketid	77
tig00000005						
3.1	NRPS	105,872	154,424	frankobaktin A1/frankobaktin A2/frankobaktin A3/frankobaktin B1/frankobaktin B2/frankobaktin B3/frankobaktin C1	NRP	25
3.2	Terpen, melanin	321,802	342,886	melanin	ostalo	100
3.3	Siderofor neovisan o NRPS	382,278	395,172	peucechelin	NRP	25

3.4	transAT-PKS, PKS tip 1	651,579	745,479	Bafilomicin B1	Poliketid: Modularni poliketid tipa I	100
3.5	PKS tip 1, butirolakton, fragment sličan PKS, Linearni peptidi koji sadrže azol (LAP)	1,099,460	1,181,952	4-heksadekanoil-3-hidroksi-2-(hidroksimetil)-2H-furan-5-on	poliketid	54
3.6	Oligosaharid, PKS tip 2, ostalo, fragment sličan PKS	1,243,594	1,335,022	Komodokinon B	Poliketid: tip 2 Poliketid + saharid: hibrid/tailoring saharid	73
3.7	Ektoin	1,880,861	1,891,259	ektoin	Ostalo	100
3.8	Terpen	2,820,635	2,844,610	Izorenieraten	Terpen	85
3.9	Terpen	3,752,818	3,772,499	kremimicin	Poliketid	5
3.10	Fragment sličan NRPS	4,759,376	4,801,471	kolibrimicin	NRP+ostalo	17
3.11	Melanin	5,230,754	5,241,230	melanin	Ostalo	100
3.12	Siderofor neovisan o NRPS	6,056,327	6,068,329			
3.13	Fragment sličan RiPP (ribosomalno sintetiziran i posttranslacijski modificiran peptidni produkt)	6,212,490	6,222,822			

3.14	PKS tip 2, NRP-metalofor, NRPS, lantipeptid klase III	6,411,840	6,542,954	Pigment spore	Poliketid	83
3.15	PKS tip 1, terpen, tiopeptid, LAP	6,817,921	6,882,937	Ajudazol A	NRP+poliketid: modularni poliketid tipa 1	76
3.16	Terpen	7,071,616	7,097,109	hopen	Terpen	61
3.17	Indol, lasso peptid	7,193,554	7,218,135	a201a	Ostalo: nukleozid	8
3.18	Terpen, NRPS	7,434,801	7,504,658	omnipeptin	NRP: ciklički depsipeptid	21
3.19	Fragment sličan NRPS, NRPS	7,542,430	7,588,100	nukleocidin	Ostalo	39
3.20	transAT-PKS, fragment sličan RiPP, NRPS, NRP-metalofor, fragment sličan NRPS, PKS tip 1	7,860,680	8,164,318	weishanmicin	NRP+poliketid	90

tig00000004, tig00000006, tig00000007 i tig00000008 ne sadrže sekundarne metabolite

Prilog 2. Rezultati antiSMASH analize za organizam BC156

Regija	**Tip**	**Pozicija od**	**Pozicija do**	**Najsličniji poznati klaster**		**Sličnost (%)**
tig00000001						
1.1	PKS sličan heterocistično	128,538	169,893	heksakosalaktoza A	Ostalo	9

	j glikolipid sintazi (hglE-KS)					
1.2	PKS tip 3	344,641	385,648	alkilrezorcinol	Poliketid	100
1.3	Siderofora neovisna o NRPS	433,378	446,037	lankamicin	Poliketid	6
1.4	melanin	505,777	533,232	melanin	Ostalo	100
1.5	NRPS	576,408	616,638	Foksicin A/ foksicin B/ foksicin C/ foksicin	NRP + poliketid	4
1.6	Lantipeptid 3. klase, terpen	676, 103	704,221	SapB	RiPP: lantipeptid	100
1.7	Terpen	1,042,036	1,062,478			
1.8	terpen	1,119,755	1,140,489	Dudomicin A	NRP	13
1.9	Fragment sličan NRPS, terpen	1,217,337	1,257,413	Desertomicin B/ desertomicin A/ desertomicin G	NRP	5
1.10	PKS tipa 1	1,333,452	1,417,655	ionostatin	Poliketid	43
1.11	Terpen	1,439,227	1,465,549	hopen	Terpen	61
1.12	Fragment sličan NRPS, NRPS, terpen	1,797,831	1,868,790	omnipeptin	NRP: ciklički depsipeptid	21
1.13	Fragment sličan NRPS	2,147,094	2,187,120	streptotricin	NRP	83
1.14	Siderofora neovisna o NRPS	2,248,193	2,263,164	shizokinen	Ostalo	25
1.15	CDPS	2,599,545	2,620,366			
1.16	NRPS, tioamiditi	2,723,905	2,764,569	arginomicin	Ostalo	20

1.17	Tioamid-NRP	3,008,270	3,066,250	Avilamicin A/ avilamicin C	Saharid: oligosaharid	7
1.18	Linaridin	4,367,616	4,388,194	Pentostatin/ vidarabin	Ostalo	9
1.19	NRP-metalofora, NRPS	4,473,615	4,530,573	peucechelin	NRP	55
1.20	Lantipeptid 2. klase	4,602,961	4,625,786			
1.21	NRPS, PKS tipa 1, fragment sličan NRPS	4,858,729	4,911,748	Oksalomicin B	NRP + poliketid	9
1.22	Lantipeptid 5. klase, tioamiditi	5,064,752	5,107,653	neotioviridamid	RiPP	100
1.23	Ostalo, fragment sličan NRPS, NRPS	5,480,828	5,537,065	WS9326	NRP	7
1.24	fragment sličan NRPS	6,883,107	6,924,595	Orizanaftopiran A/ orizanaftopiran B/ orizanaftopiran C/ orizantron A/ orizantron B/ klorizantron A klorizantron B	Poliketid	20
1.25	Ne-alfa poli aminokiseline (NAPAA)	7,138,895	7,172,803	ε-Poli-L-lizin	NRP	100

1.26	Tiopeptid, LAP, PKS tipa 2, klaster koji sadrži RRE element	7,646,881	7,725,925	Pigment spore	Poliketid	66
1.27	Ektoin	7,874,370	7,883,851	ektoin	Ostalo: ektoin	100
1.28	Terpen	8,013,521	8,034,963	2-metilizoborneol	Terpen	100
1.29	ektoin	8,151,130	8,161,534	ektoin	Ostalo: ektoin	100
tig00000004						
4.1	tioamiditi	9,878	32,346	griselimicin	NRP	11
tig00000006						
6.1	Butirolakton, PKS tipa 3, terpen, ostalo	11,666	91,666	meroklorin A/ meroklorin B/ dekloro-meroklorin A/ dekloro-meroklorin B/ izokloro-meroklorin B/ diklor-meroklorin B/ meroklorin D/ meroklorin C	Terpen + poliketid: poliketid tipa 3	80

tig00000002, tig00000003, tig00000005, tig00000007, tig00000008, tig00000009, tig00000010 i tig00000011 ne sadrže sekundarne metabolite

Prilog 3. Rezultati antiSMASH analize za organizam BC157

Regija	**Tip**	**Pozicija od**	**Pozicija do**	**Najsličniji poznati klaster**		**Sličnost (%)**
1	Fragment sličan NRPS	11,966	52,739	A-500359 A/A-500359 B	NRP	5
2	Butirolakton	69,530	80,474	koelimicin P1	Poliketid: modularni	16

					poliketid tipa 1	
3	Terpen	111,451	132,423	geosmin	Terpen	100
4	NRP-metalofora, NRPS	167,009	275,337	griseobaktin	NRP	100
5	PKS tipa 3	345,842	385,505	lasalocid	Poliketid	7
6	NRPS, fragment sličan PKS, transAT-PKS, fragment sličan NRPS	920,521	1,028,900	Largimicin 38/ largimicin 39	NRP + poliketid: modularni poliketid tipa 1	62
7	Terpen	1,160,735	1,179,546	Stefimicin D	Poliketid: tip 2 Poliketid + saharid: hibrid/ tailoring saharid	19
8	Lantipeptid klase 1	1,388,727	1,414,286			
9	Ektoin	1,713,984	1,724,014	ektoin	Ostalo	100
10	Fragment sličan NRPS, NRPS	2,385,453	2,466,487	kolibrimicin	NRP + ostalo	53
11	Lantipeptid klase 2, lantipeptid klase 3	2,799,398	2,829,255			
12	Betalakton, NRPS, siderofora neovisna o NRPS	2,888,595	2,938,481	desferioksamin B	Ostalo	100
13	NRPS, PKS tipa 1	3,510,963	3,562,779	Variobaktin A/ variobaktin B	NRP + poliketid	14
14	NRPS	3,621,517	3,675,222	daptomicin	NRP	12

15	Fosfonat, lasso peptid	4,800,858	4,856,455	keywimysin	RiPP	100
16	Lantipeptid 3. klase	4,952,477	4,975,335	korbomicin	NRP	7
17	Fragment sličan NRPS, PKS tipa 1	5,631,860	5,684,729	Rotihibin A	NRP	21
18	Lantipeptid klase 3	5,775,706	5,797,643	AmfS	RiPP: lantipeptid	100
19	Ektoin	6,030,224	6,040,598	showdomicin	Ostalo	41
20	Terpen	6,167,059	6,186,487			
21	Siderofora neovisna o NRPS	6,590,567	6,605,324			
22	PKS tipa 1, oligosaharid	6,655,902	6,753,820	auroramicin	Poliketid	64
23	PKS tipa 1, NRPS	6,905,920	6,955,961	kolizmicin A	NRP + poliketid: modularni poliketid tipa 1	74
24	NRPS	7,054,030	7,096,423	holomicin	NRP	92
25	Terpen	7,280,334	7,300,292	formicamicini A-M	Poliketid	16
26	Terpen	7,680,670	7,706,832	hopen	Terpen	69
27	NRPS, fragment sličan RiPP	7,764,056	7,810,204	SGR PTM-ovi/ SGR PTM spoj b/ SGR PTM spoj c/ SGR PTM spoj d/	NRP + poliketid	100
28	PKS tipa 1, NRPS	7,914,482	7,961,918	A54145	NRP	5
29	fragment sličan RiPP	7,978,562	7,987,950	streptamidin	RiPP: ostalo	66
30	Melanin	7,990,365	8,000,847	melanin	Ostalo	100

31	Tiopeptid, LAP	8,011,062	8,039,229	laktazol	RiPP: tiopeptid	33
32	PKS tipa 3	8,041,916	8,082,968	alkilrezorcinol	poliketid	100
33	NRPS, fragment ovisan o NRPS	8,155,037	8,205,295	foksicin A/ foksicin B/ foksicin C/ foksicin	NRP + poliketid	7

Prilog 4. Rezultati antiSMASH analize za organizam BC152

Regija	**Tip**	**Pozicija od**	**Pozicija do**	**Najsličniji poznati klaster**		**Sličnost (%)**
tig00000001						
1.1	PKS tip 1, NRPS	26,397	72,516	paulomicin	Ostalo	7
1.2	PKS tip 3	148,398	187,497	alkilrezocinol	Poliketid	100
1.3	Melanin	222,261	232,662	melanin	Ostalo	100
1.4	Fragment sličan RiPP	235, 149	244,530	streptamidin	RiPP: ostalo	66
1.5	NRPS	252,674	296,599	Valinomicin/ montanastatin	NRP + saharid: hibrid/ tailoring saharid	30
1.6	Fragment sličan RiPP, fragment sličan NRPS	512,591	561,712	SGR PTM-ovi/ SGR PTM spoj b/ SGR PTM spoj c/ SGR PTM spoj d	NRP + poliketid	66
1.7	Terpen	619,742	645,860	hopen	Terpen	69
1.8	NRP-metalofor, NRPS, PKS tip 1	1,012,351	1,082,525	filipin	Poliketid	15
1.9	PKS tip 1, transAT-PKS, fragment sličan NRPS, terpen	1,088,042	1,212,708	Bafilomicin B1	Poliketid: modularni poliketid tipa 1	100

1.10	fragment sličan NRPS, PKS tip 1	1,516,826	1,566,522	kolizmicin A	NRP + poliketid: modularni poliketid tipa 1	74
1.11	Tioamid-NRP, NRPS, Ladderane, tiopeptid	1,676,030	1,770,977	kolibrimicin	NRP + ostalo	53
1.12	Siderofor neovisan o NRPS	1,841,821	1,854,646	shizokinen	Ostalo	30
1.13	Terpen	2,290,458	2,309,854			
1.14	Lantipeptid 3. klase	2,688,349	2,710,207	AmfS	RiPP: lantipeptid	100
1.15	PKS tip 1, fragment sličan NRPS	2,811,900	2,863,096	enduracidin	NRP	6
1.16	Lasso peptid	3,637,596	3,660,436	SRO15-2005	RiPP: lasso peptid	100
1.17	Ektoin, butirolakton	3,893,377	3,908,416	showdomicin	Ostalo	47
1.18	Lantipeptid 2. klase	4,018,579	4,042,019	kalkomicin A	Poliketid	9
1.19	Lantipeptid 1. klase	4,440,827	4,464,843			
1.20	Lantipeptid 1. klase	4,869,174	4,895,358			
1.21	NRPS	4,896,935	4,950,405	Azotobaktin D	NRP	16
1.22	Siderofor neovisan o NRPS	5,599,129	5,610,906	Deferoksamin B	Ostalo	100
1.23	Lantipeptid 3. klase	5,681,973	5,703,259			
1.24	PKS tip 2	5,735,046	5,807,519	ravidomicin	Poliketid	74

1.25	Ektoin	6,734,349	6,744,777	ektoin	Ostalo	100
1.26	Terpen	7,195,300	7,216,376	Stefimicin D	Poliketid: tip 2 Poliketid + saharid: hibrid/ tailoring saharid	19
1.27	NRPS	7,374,964	7,419,536	Koprisamid C/ koprisamid D	NRP	13
1.28	NRPS, fragment sličan NRPS, ostalo	7,570,073	7,662,488	Aktinomicin D	NRP	89
1.29	PKS tip 3	7,970,725	8,011,843	naringenin	Poliketid: poliketid tipa 3	100
1.30	NRP-metalofor, NRPS, fragment sličan NRPS, PKS tip 1	8,042,940	8,221,791	Kinolidomicin A	Poliketid	30
tig00000002						
2.1	Butirolakton	60,938	71,882	Koelimicin P1	poliketid: modularni poliketid tipa 1	16
2.2	Terpen	102,896	125,108	Vazabitid A	NRP	4
2.3	Fragment sličan NRPS, transAT-PKS, PKS tip 1	155,676	278,621	Kinolidomicin A	poliketid	30

tig00000003 i tig00000004 ne sadrže sekundarne metabolite

Prilog 5. Rezultati antiSMASH analize za organizam BC164

Regija	Tip	Pozicija od	Pozicija do	Najsličniji poznati klaster		Sličnost (%)
tig00000005						
4.1	Fragment sličan NRPS, NRPS	36,754	87,788	paulomicin	Ostalo	7
4.2	PKS tipa 3	161,694	201,328	alkilrezorcinol	Poliketid	100
4.3	Melanin	236,108	246,590	melanin	Ostalo	100
4.4	Fragment sličan RiPP	248,998	258,186	streptamidin	RiPP: ostalo	66
4.5	NRPS, PKS tipa 1	266,525	316,314	Valinomicin/ montanastatin	NRP + saharid: hibrid/ tailoring saharid	30
4.6	Fragment sličan RiPP, NRPS, PKS tipa 1	439,931	495,579	SGR PTM-ovi/ SGR PTM spoj b/ SGR PTM spoj c/ SGR PTM spoj d	NRP + poliketid	100
4.7	Terpen	547,131	573,264	Hopen	Terpen	69
4.8	NRP-metalofor, NRPS, PKS tipa 1, Fragment sličan NRPS, terpen	938,050	1,138,460	Bafilomicin B1	Poliketid: modularni poliketid tipa 1	100
4.9	NRPS, PKS tipa 1	1,442,382	1,492,084	kolizmicin A	NRP + Poliketid: modularni poliketid tipa 1	74
4.10	Tioamid-NRP, NRPS, Ladderane, tiopeptid	1,601,622	1,696,576	kolibrimicin	NRP + ostalo	53

4.11	Siderofor neovisan o NRPS	1,767,598	1,780,667			
4.12	Terpen	2,220,074	2,240,354			
4.13	Lantipeptid 3. klase	2,618,931	2,640,756	AmfS	RiPP: lantipeptid	100
4.14	PKS tipa 1, fragment sličan NRPS	2,743,478	2,794,704	enduracidin	NRP	6
4.15	Lantipeptid 4. klase	3,450,711	3,473,605	labirintopeptin A2 labirintopeptin A1 labirintopeptin A3	RiPP: lantipeptid	40
4.16	Lasso peptid	3,569,286	3,592,126	SRO15-2005	RiPP: lasso peptid	100
4.17	Ektoin, butirolakton	3,825,091	3,839,049	showdomicin	Ostalo	47
4.18	Lantipeptid 2. klase	3,946,915	3,972,700	Kalkomicin A	Poliketid	9
4.19	Lantipeptid 1. klase	4,372,615	4,396,652			
4.20	Lantipeptid 1. klase	4,801,043	4,827,227			
4.21	NRPS	4,828,806	4,882,271	Azotobaktin D	NRP	16
4.22	Siderofor neovisan o NRPS	5,539,668	5,551,446	Deferoksamin B	Ostalo	100
4.23	Lantipeptid 3. klase, lantipeptid 2. klase	5,622,529	5,653,704			
4.24	PKS tipa 2	5,675,629	5,748,102	ravidomicin	Poliketid	74
4.25	Ektoin	6.675.026	6,685,424	ektoin	Ostalo	100
4.26	Terpen	7,136,853	7,155,894	Stefimicin D	Poliketid: tip 2 Poliketid + saharid:	19

					hibrid/ tailoring saharid	
4.27	fragment sličan NRPS, NRPS	7,309,245	7,360,659	Aktinomicin D	NRP	10
4.28	NRPS, fragment sličan NRPS, ostalo	7,509,575	7,607,638	Aktinomicin D	NRP	89
4.29	PKS tipa 3	7,910,397	7,951,515	naringenin	Poliketid: poliketid tipa 3	100
4.30	NRP-metalofor, NRPS, PKS tipa 1, fragment sličan NRPS	7,981,639	8,285,362	Kinolidomicin A	Poliketid	40
4.31	Terpen	8.316.902	8,339,115	geosmin	Terpen	100
4.32	Butirolakton	8,370,153	8,381,097	Koelimicin P1	Poliketid: modularni poliketid tipa 1	16
4.33	fragment sličan NRPS	8.398.855	8,442,469	Higromicin A	saharid	9

tig00000002, tig00000003 i tig00000004 ne sadrže regije sa sekundarnim metabolitima

Prilog 6. Rezultati antiSMASH analize za organizam BC138

Regija	**Tip**	**Pozicija od**	**Pozicija do**	**Najsličniji poznati klaster**		**Sličnost (%)**
tig00000003						
2.1	Lasso peptid	35,218	57.324	Citrulasin D	RiPP	100
2.2	Butirolakton	180,433	189,505			
2.3	Lasso peptid	639,879	661.914			

2.4	fragment sličan NRPS, Aril polien	679,070	721,259	kitacinamicin A/ kitacinamicin B/ kitacinamicin C/ kitacinamicin D/ kitacinamicin E/ kitacinamicin F	NRP	4
2.5	Betalakton	803,205	826.373			
2.6	NRPS, tiopeptid, LAP, betalakton	950,718	1,023,687	Murajmicin C1	NRP + poliketid	23
2.7	Fragment sličan RiPP	1,088,509	1,098,102			
2.8	PKS tipa 1, fragment sličan NRPS, fragment sličan PKS, transAT-PKS	1,134,991	1,378,391	kaniferolid A/ kaniferolid B/ kaniferolid C/ kaniferolid D	Poliketid: modularni poliketid tipa 1	58
2.9	PKS tipa 3, lantipeptid 3. klase	1,421,681	1.471,660	SapB	RiPP: lantipeptid	100
2.10	NAPAA	1,576,538	1,612,642	stenotricin	NRP: ciklički depsipeptid	13
2.11	Fragment sličan RiPP	1,647,100	1.654.507	Heksakosalakton A	Ostalo	4
2.12	PKS tipa 2	1,688,026	1,760,541	Pigment spore	Poliketid	83
2.13	NRP-metalofor, NRPS, aminopolikarbok silna kiselina, melanin, PKS tipa 1	1,777,114	1,950,489	lidikamicin	NRP + poliketid: modularni poliketid tipa 1	100
2.14	Siderofor neovisan o NRPS	2,096,809	2,109,284	peucechelin	NRP	25

2.15	Beta laktam	2,193,513	2,217,014	Valclavam/(-)-2-(2-hidroksiletil)clavam	Ostalo: ne-NRP beta laktam	71
2.16	Lasso peptid	2,833,169	2,854,424	Anantin C	RiPP	50
2.17	Siderofor neovisan o NRPS	2,940,387	2,949,238	Deferoksamin E	Ostalo	100
2.18	Ektoin	3,032,644	3,043,060	ektoin	Ostalo: ektoin	100
2.19	PKS tipa 1	3,976,592	4,019,810			
2.20	Terpen	4,087,127	4,107,334	ebelakton	Poliketid	5
2.21	Terpen	5,102,870	5,119,144	klipibiciklen	Alkaloid	6
2.22	Terpen	5,536,642	5,558,498	salinomicin	poliketid: modularni poliketid tipa 1	6
2.23	LAP	5,611,023	5,632,800			
2.24	Tioamiditi, LAP, tiopeptid, NRPS, lasso peptid, fragment sličan NRPS	5,691,932	5,787,955	ulleungdin	RiPP: lasso peptid	100
2.25	Siderofor neovisan o NRPS	7,342,063	7,356,763	shizokinen	Ostalo	20
2.26	Butirolakton	7,512,472	7,522,591			
2.27	Terpen	8,290,538	8.316.298	hopen	Terpen	69
2.28	hglE-KS, PKS tipa 1	8,454,336	8,505,142	Heksakosalakton A	Ostalo	9
2.29	Lantipeptid 1. klase	8,643,855	8.668.394	scelifrolaktam	Poliketid	8
2.30	Fragment sličan RiPP	8,741,230	8,753,164			
2.31	Butirolakton	8,988,365	8,999,116	Koelimicin P1	poliketid: modularni	12

					poliketid tipa 1	
2.32	fragment sličan NRPS, NRPS, ostalo	9,091,050	9,148,211	antipain	NRP	100
2.33	Terpen	9,203,852	9,229,415	izorenieraten	Terpen	100
2.34	Aminoglikozid/aminociklitol	9,443,712	9,467,657	streptomicin	Saharid	55
2.35	Butirolakton	9,484,278	9,495,009	BE-14106	Poliketid: modularni poliketid tipa 1	7
2.36	fragment sličan NRPS, NRPS	9,559,604	9,607,029	daptomicin	NRP	25

tig00000002 ne sadrži regije sa sekundarnim metabolitima

IZJAVA O IZVORNOSTI

Ja (Andro Vuković) 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.

_____

Vlastoručni potpis