

SVEUČILIŠTE U ZAGREBU
PREHRAMBENO-BIOTEHNOLOŠKI FAKULTET

DIPLOMSKI RAD

Zagreb, rujan 2013.

Hrvoje Rimac
285/MB

**ANOTACIJA GENSKIH NAKUPINA
PRIRODNIH SPOJEVA U GENOMU
BAKTERIJE *Streptomyces* sp. C**

Rad je izrađen u Kabinetu za bioinformatiku na Zavodu za biokemijsko inženjerstvo Prehrambeno-biotehnološkog fakulteta Sveučilišta u Zagrebu pod mentorstvom dr. sc. Jurice Žučka, docenta te uz pomoć dr. sc. Antonija Starčevića, docenta.

Zahvaljujem se mentoru, doc. dr. sc. Jurici Žučku što mi je omogućio izradu diplomskog rada, na stručnom vodstvu, konstruktivnim savjetima i pomoći tijekom cijele izrade i pisanja ovog rada.

Također se zahvaljujem i doc. dr. sc. Antoniju Starčeviću i prof. dr. sc. Daslavu Hranueliju na pomoći i vrijednim savjetima koji su mi uistinu olakšali ovaj cijeli proces.

Hvala i mojoj šefici, prof. dr. sc. Marici Medić-Šarić i doc. dr. sc. Željku Debeljaku, docentu na razumijevanju i strpljivosti u protekle dvije godine, kao i obitelji i prijateljima, uz koje je sve bilo puno lakše.

Jos jedan je pao! ☺

TEMELJNA DOKUMENTACIJSKA KARTICA

Diplomski rad

Sveučilište u Zagrebu
Prehrambeno-biotehnoški fakultet
Zavod za biokemijsko inženjerstvo
Kabinet za bioinformatiku

Znanstveno područje: Biotehničke znanosti
Znanstveno polje: Biotehnologija

ANOTACIJA GENSKIH NAKUPINA PRIRODNIH SPOJEVA U GENOMU BAKTERIJE *Streptomyces sp.* C

Hrvoje Rimac, 285/MB

Sažetak: Zbog sve veće pojave rezistencije bakterija na postojeće lijekove, do koje je došlo zbog pretjerane i neutemeljene upotrebe trenutno dostupnih antibiotika, u zadnjih se desetak godina pojavila potreba za pronašlaskom novih antibiotika. Antibiotici spadaju u skupinu spojeva koji se nazivaju sekundarni metaboliti, a čiju veliku većinu čine poliketidi i neribosomalno sintetizirani polipeptidi. Takve sekundarne metabolite uglavnom proizvode bakterije roda *Streptomyces*, a služe im kao kompetitivna prednost u odnosu na druge bakterije. Međutim, problem u pronašljenju novih antibiotika je što ih većina nije eksprimirana u mjeri da bi mogla biti detektirana i okarakterizirana standardnim znanstvenim metodama. Cilj ovog diplomskog rada je bilo anotiranje genoma slabo poznate bakterijske vrste *Streptomyces sp.* C bioinformatičkim programskim paketom *ClustScan* kako bi se pronašli geni koji kodiraju za sintezu potencijalno novih antibiotika nepoznatih struktura.

Ključne riječi: PKS, NRPS i hibridne genske nakupine, genom bakterije *Streptomyces sp.* C, računalni programski paket *ClustScan*

Rad sadržava: 59 stranica, 33 slike, 6 tablica, 65 literaturna navoda, 1 prilog

Jezik izvornika: hrvatski

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

Mentor: doc. dr.sc. Jurica Žučko

Pomoć pri izradi: doc. dr.sc. Antonio Starčević

Stručno povjerenstvo za ocjenu i obranu:

1. Prof. dr. sc. Želimir Kurtanjek
2. Doc. dr. sc. Jurica Žučko
3. Doc. dr. sc. Antonio Starčević
4. Izv. prof. dr. sc. Ana Vukelić (zamjena)

Datum obrane: 20. rujna 2013.

BASIC DOCUMENTATION CARD

Graduate Thesis

University of Zagreb

Faculty of Food Technology and Biotechnology

Department of Biochemical Engineering

Section for Bioinformatics

Scientific area: Biotechnical Sciences

Scientific field: Biotechnology

ANNOTATION OF NATURAL PRODUCTS GENE-CLUSTERS IN *Streptomyces* sp. C GENOME

Hrvoje Rimac, 285/MB

Abstract: Due to increased bacterial resistance to existing drugs, which was caused by an excessive and unjustified use of currently available antibiotics, in the last several years a need for new novel antibiotics arose. Antibiotics are a group of compounds which, among others, are called secondary metabolites and mostly consist of polyketides and nonribosomally synthesized peptides. These secondary metabolites are mostly produced by bacteria of the *Streptomyces* genus and serve to their producers as a competitive advantage relative to other bacteria. The problem in finding new antibiotics is that most of them are not expressed to the extent of being detected and characterized by standard scientific methods. The goal of this thesis is to annotate the genome of a not well known species of bacteria, *Streptomyces* sp. C using *ClustScan*, the bioinformatics computer program, in order to find genes which encode for enzymes which produce potentially new antibiotics of unknown structures.

Keywords: PKS, NRPS and hybrid gene-clusters, *Streptomyces* sp. C genome, *ClustScan* program package

Thesis contains: 59 pages, 33 figures, 6 tables, 65 references, 1 supplement

Original in: Croatian

Graduate Thesis in printed and electronic (pdf format) version is deposited in: Library of the Faculty of Food Technology and Biotechnology, Kačićeva 23, Zagreb

Mentor: Ph. D. Jurica Žučko, Assistant Professor

Technical support and assistance: Ph. D. Antonio Starčević, Assistant Professor

Reviewers:

1. Ph. D. Želimir Kurtanjek, Full Professor
2. Ph. D. Jurica Žučko, Assistant Professor
3. Ph. D. Antonio Starčević, Assistant Professor
4. Ph. D. Ana Vukelić, Associate Professor (substitute)

Thesis defended: 20 September 2013

SADRŽAJ

1. UVOD.....	1
2. TEORIJSKI DIO.....	2
2.1. PRIRODNI SPOJEVI MIKROORGANIZAMA	2
2.1.1. Poliketidi.....	4
2.1.1.1. Poliketid sintaze tipa I.....	9
2.1.1.2. Poliketid sintaze tipa II	10
2.1.1.3. Poliketid sintaze tipa III	10
2.1.2. Neribosomalno sintetizirani peptidi.....	11
2.1.3. Poliketidno/peptidni hibridi	13
2.1.3.1. Sustavi koji ne uključuju neposrednu hibridizaciju između enzima NRPS i PKS....	14
2.1.3.2. Sustavi koji uključuju neposrednu funkcionalnu hibridizaciju između enzima PKS i NRPS	16
2.2. SEKVENCIRANJE GENOMA RAZLIČITIH VRSTA AKTINOBAKTERIJA	16
2.2.1. Važnost sekvenciranja genoma različitih vrsta aktinobakterija	19
2.2.2. Sekvenciranje genoma bakterije <i>Streptomyces sp. C</i>	19
2.3. BIOINFORMATIČKI ALATI ZA ANOTACIJU GENSKIH NAKUPINA.....	20
2.3.1. Baze podataka <i>CSDB</i> , <i>SEARCHPKS</i> , <i>MAPSI</i> , <i>ClusterMine360</i>	21
2.3.2. Bioinformatički alati za anotaciju genskih nakupina prirodnih spojeva	23
2.3.3. Generički programski paket <i>ClustScan</i>	24
3. EKSPERIMENTALNI DIO	26
3.1. MATERIJAL.....	26
3.1.1. Računalna podrška i operativni sustav	26
3.1.2. Baze podataka	26
3.1.2.1. Baza podataka GenBank	26
3.1.2.2. Baza podataka NRPS-PKS.....	26
3.1.2.3. Baza podataka Pfam	27
3.1.3. Bioinformatički računalni paketi i programi.....	28
3.1.3.1. Računalni program Glimmer	28
3.1.3.2. Računalni program GeneMark-PS.....	28
3.1.3.3. Programski paket HMMER.....	28
3.1.3.4. Računalni program BLAST.....	29
3.1.3.5. Programski paket ClustScan.....	29

3.2. METODE.....	30
3.2.1. Prikupljanje sekvenca DNA.....	30
3.2.2. Pronalaženje strukturnih gena	31
3.2.3. Analiza sekvenca proteina programskim paketom <i>HMMER</i>	34
3.2.4. Preuzimanje gotovih profila proteina iz baze podataka Pfam.....	34
3.2.5. Izrada vlastitih profila proteina pomoću programa <i>ClustScan</i>	37
4. REZULTATI.....	41
4.1. REZULTATI ANOTACIJE GENOMA BAKTERIJE <i>Streptomyces sp.</i> C PROGRAMSKIM PAKETOM <i>ClustScan</i>	41
5. RASPRAVA	46
5.1. ANOTACIJA GENOMA BAKTERIJE <i>Streptomyces sp.</i> C PROGRAMSKIM PAKETOM <i>ClustScan</i>	46
6. ZAKLJUČCI.....	52
7. LITERATURA.....	53
8. PRILOZI	58
8.1. POPIS U RADU UPOTRIJEBLJENIH KRATICA	58
8.2. SADRŽAJ KOMPAKTNOG DISKA.....	59
8.2.1. Rezultat anotacije genoma bakterije <i>Streptomyces sp.</i> C programskim paketom <i>ClustScan</i>	59

1. UVOD

Mikroorganizmi su sposobni sintetizirati prirodne spojeve koji nisu uključeni u njihov vegetativni rast, odnosno koji im nisu nužni za preživljavanje. Ti im prirodni spojevi daju kompetitivnu prednost pred ostalim organizmima u njihovoj okolini, bilo da su to antibiotici koji ubijaju njihove neprijatelje, spojevi koji ih štite od vlastitih antibiotika, spojevi koji povećavaju njihovu patogenost, spojevi koji omogućuju veću apsorpciju potrebnih tvari iz okoline, npr. željeza i sl. Takvi prirodni spojevi se nazivaju sekundarnim metabolitima, a obično se sintetiziraju nakon eksponencijalne faze rasta. Najveći proizvođači sekundarnih metabolita su vrste roda *Streptomyces* i srodni rodovi aktinomiceta. Njihovo važno svojstvo je što su oni karakteristični za pojedinu vrstu bakterija te je tako pronađeno već više od 8000 antibiotički aktivnih supstancija, od čega ih 70% sintetiziraju vrste iz roda *Streptomyces* (Broad Institute, 2013).

S obzirom na njihovu kemijsku strukturu, među sekundarnim metabolitima najveću skupinu spojeva čine poliketidi (PKS), a drugu veliku skupinu čine neribosomalno sintetizirani polipeptidi (NRPS). Biosintetski put obje skupine spojeva obuhvaća spajanje jednostavnih građevnih jedinica u složene kemijske strukture. S obzirom da se može upotrijebljavati velik broj različitih građevnih jedinica s različitim redoslijedom ugradnje u produkt i s različitim načinima modifikacija tih građevnih jedinica, postoji praktički neograničen broj mogućih produkata, od kojih bi neki moguće bili primjereni i za ljudsku upotrebu.

U posljednjih desetak godina u znanstvenoj javnosti postoji zanimanje za sintezu, odnosno pronađak novih farmaceutskih supstancija, posebice antibiotika, zbog sve učestalije rezistencije mikroorganizama na postojeće lijekove, do koje je došlo zbog pretjerane i neutemeljene upotrebe postojećih antibiotika. Jedno od predloženih rješenja tog problema (Broad Institute, 2013) je sekvenciranje i anotiranje genoma, kako već opisanih bakterijskih vrsta, tako i metagenoma, kako bi se pronašli genske nakupine koje kodiraju za enzime koji sudjeluju u sintezi još neotkrivenih antibiotika. Ti, još neotkriveni antibiotici u većini slučajeva nisu eksprimirani u tolikoj mjeri od strane svojih proizvođača da bi se mogli otkriti uobičajenim znanstvenim metodama. Druga predložena mogućnost je korištenje *in silico* metoda i sintetske biologije kako bi se kombinirali dijelovi genoma različitih vrsta s ciljem dobivanja posve novih genskih produkata i posljedično antibiotika posve nove strukture.

Svrha ovog diplomskog rada je anotacija i klasifikacija genskih nakupina PKS, NRPS i PKS/NRPS hibridnih sustava vrste *Streptomyces* sp. C s ciljem pronađaka novih ili već opisanih genskih nakupina odgovornih za biosintezu poliketida, neribosomalno sintetiziranih polipeptida i njihovih hibrida.

2. TEORIJSKI DIO

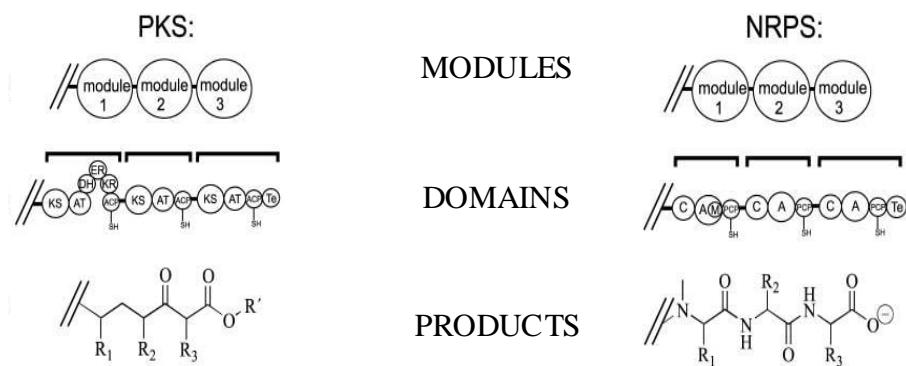
2.1. PRIRODNI SPOJEVI MIKROORGANIZAMA

Mikroorganizmi su sposobni proizvoditi prirodne spojeve vrlo raznolikih kemijskih struktura. Zanimljivi prirodni spojevi najčešće se sintetiziraju nakon eksponencijalne faze rasta, tijekom sekundarnog metabolizma. Zbog toga se nazivaju i sekundarnim metabolitima. Sekundarni metaboliti nisu nužni za vegetativni rast stanica već su uključeni u razne mehanizme preživljavanja u prirodi, dajući svojim proizvođačima određenu selektivnu prednost. Od takvih spojeva, za ljudsku su se upotrebu do sada najkorisnijima pokazali oni s antibiotskim svojstvima te se njihovom istraživanju posvećuje i najviše pozornosti. Pored toga, mnogi od njih pokazuju i antivirusna, citostatska, imunosupresivna, antihipertenzivna i druga svojstva. Važno je također naglasiti i da su farmakološki aktivni prirodni spojevi često prethodno nepoznate osnovne strukture te se mogu koristiti u osnovnim istraživanjima jer mogu pružiti uvid u mehanizam djelovanja pojedinih lijekova. Također se mogu izravno koristiti u medicinske i industrijske svrhe (Woodruff, 1980; Grabley i Thiericke, 1999; Hranueli i Cullum, 2001).

Obično su geni čiji produkti sudjeluju u biosintezi prirodnih spojeva kod bakterija i nekih pljesni (npr. *Penicillium* i *Aspergillus*) organizirani u jednu genetsku nakupinu smještenu na kromosomskoj DNA, dok se kod nekih pljesni, kao što su *Acremonium chrysogenum* i *Cephalosporium acremonium*, geni čiji produkti sudjeluju u biosintezi cefalosporina, nalaze na različitim kromosomima (Marahiel i sur., 2002). U biosintetski put sekundarnog metabolita može biti uključeno i više od 40-ak gena, čineći sekvencu duljine i do 100 kb. Cjeloviti skup proteina uključenih u biosintezu sekundarnog metabolita može biti i dvostruko veći od ribosoma (Fishbach i sur., 2008). Iz razloga velike potrošnje energije u sintezi proteina koji sudjeluju u sintezi sekundarnih metabolita, geni koji kodiraju za te proteine su pod selektivnim pritiskom te predstavljaju najbrže evoluirajuće genetičke elemente i posljedično dovode do velike raznolikosti između pojedinih vrsta bakterija. Takva brza evolucija je posljedica kratkog replikacijskog vremena mikroorganizama i mogućnosti horizontalnog prijenosa gena između mikroorganizama (Fischbach i sur., 2008).

S obzirom na kemijsku strukturu spojeva, među prirodnim spojevima proizvedenim od strane mikroorganizama, najveću skupinu čine poliketidi (npr. eritromicin, nistatin ili borelidin), koji se sintetiziraju djelovanjem enzimskog sustava poliketid sintaze (engl. „Polyketide Synthase“,

PKS). Druga velika skupina prirodnih spojeva su neribosomalno sintetizirani peptidi u čijoj sintezi sudjeluje enzimski sustav sintetaza neribosomalnih peptida (engl. „Non Ribosomal Peptide Synthetase“, NRPS). Iako sustavi NRPS upotrijebljavaju različite građevne jedinice, enzimi sustava PKS i NRPS imaju sličnu modularnu građu i slične mehanizme sinteze produkata (Slika 1). Biosintetski put obje skupine obuhvaća spajanje jednostavnih građevnih jedinica u složene kemijske strukture, a cijela se biosinteza odvija kao na proizvodnoj traci. Prema tome, svakom ugradnjom jednostavne građevne jedinice u nastajući produkt upravlja jedan modul kojeg čini međusobno povezani skup domena, a konačni produkt ovisi o broju domena i prisutnosti odnosno izostanku pojedinih domena unutar modula (Hranueli i Cullum, 2001; Hranueli i sur., 2008).

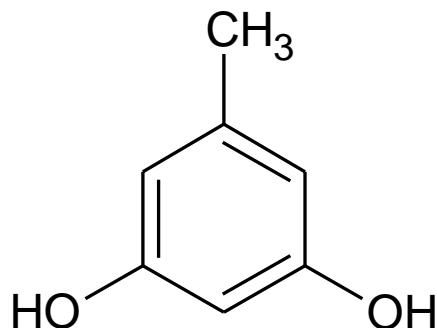


Slika 1. Od modula do produkta: moduli sustava PKS i NRPS podijeljeni su u domene koje kataliziraju pojedine enzimske reakcije. Sastav produkata određen je nizom aktivnih domena u pripadajućim modulima (Schwarzer i Marahiel, 2001).

U posljednje se vrijeme javio interes za oblikovanjem novih farmaceutskih supstancija inovativnih struktura. Metodama kombinatorne sinteze, odnosno manipulacijom genskih nakupina u uvjetima *in vitro* pokušava se doći do novog načina proizvodnje lijekova. Međutim, pri tome se javlja problem da tako konstruirani sojevi uopće ne proizvode željeni produkt ili su prinosi vrlo niski. Jedno od mogućih rješenja tog problema bi moglo biti oblikovanje novih genskih nakupina homolognom rekombinacijom u uvjetima *in silico* primjenom bioinformatičkih alata (Hranueli i Cullum, 2003). Za to je neophodno sekvenciranje novih genoma kako bi se pronašle nove genske nakupine te provele analize metagenoma mikroorganizama koji žive u tlu, moru ili zraku ne bi li se pronašli novi biosintetski putovi sekundarnih metabolita (Hranueli i sur., 2008).

2.1.1. Poliketidi

Poliketidi su velika i raznovrsna skupina kemijskih spojeva s vrlo širokim spektrom biološke aktivnosti. Proizvode ih različiti organizmi poput bakterija, pljesni, biljaka, insekata, dinoflagelata, mekušaca i spužvi. Poliketidi posjeduju antimikrobnu, antifungalnu, antiparazitsku, antitumorsku i mnoga druga svojstva (Hranueli i Cullum, 2001). Prvi poliketid su izolirali 1893. godine Collie i Myers (Collie, 1893) i nazvali su ga orcinol (Slika 2). Collie je također ustvrdio kako se orcinol sintetizira ponavljačim kondenzacijskim ili polimerizacijskim reakcijama. Kako orcinol i slični spojevi ne spadaju ni u jedan prethodno opisan kemijski razred, smješteni su u poseban kemijski razred nazvan poliketidi. Članovi tog razreda su okarakterizirani ponavljajućom pojmom $-\text{CH}_2\text{-CO-}$ motiva, kojeg je Collie nazvao „ketid“ (Bentley i Bennett, 1999).

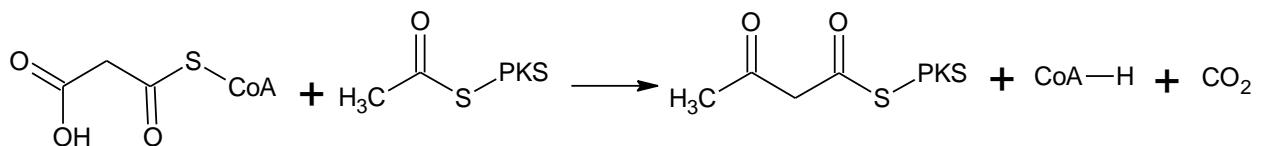


Slika 2. Struktura orcinola

Tijekom idućih godina, brojni poliketidi su bili izolirani iz filamentoznih biljaka i eubakterija (*Streptomyces* sp.), a 1953. godine su Birch i Donovan predložili novi mehanizam biosinteze poliketida, koji u velikoj mjeri nalikuje predloženom mehanizmu sinteze masnih kiselina. Ta je hipoteza postala poznata pod imenom poliacetatna hipoteza i glasi: „Poliketidi se sintetiziraju u smjeru glava-rep od poliacetatnih jedinica, nakon koje slijedi ciklizacija aldolnom reakcijom ili acilacijom na fenole“ (Birch i Donovan, 1953).

S implementacijom modernih rekombinantnih tehnika (ligaza 1967. god., restriktični enzimi 1970. god., rekombinantna DNA 1972. god., sekvenciranje 1983. god. i PCR 1984. god), sredinom 1980-ih godina je postalo moguće analizirati genetske osnove proizvodnje poliketida. Prvi poliketid koji je bio istražen genetskim i molekularnim pristupom je bio plavi pigment aktinorodin, koji proizvodi vrsta *Streptomyces coelicolor* (Hopwood, 1997).

Danas je poznato da se sinteza poliketida odvija velikim više-funkcionalnim enzimima ili više-enzimskim kompleksima nazvanim poliketid sintaze (PKS). Sama se sinteza odvija kao na proizvodnoj traci, s time da je redoslijed kemijskih reakcija uvjetovan redoslijedom enzimskih domena koje ih kataliziraju i, kao što je već rečeno, takav mehanizam biosinteze je vrlo sličan onome ravnolančanih masnih kiselina sintetiziranih pomoću sintaze masnih kiselina (engl. „Fatty Acid Synthase“, FAS). S obzirom da oba sustava enzima kataliziraju kondenzaciju aktiviranog primarnog metabolita (acetil-CoA, malonil-CoA i butiril-CoA), nastaju β -ketoacil polimeri povezani s enzimom tioesterskom vezom (Slika 3).



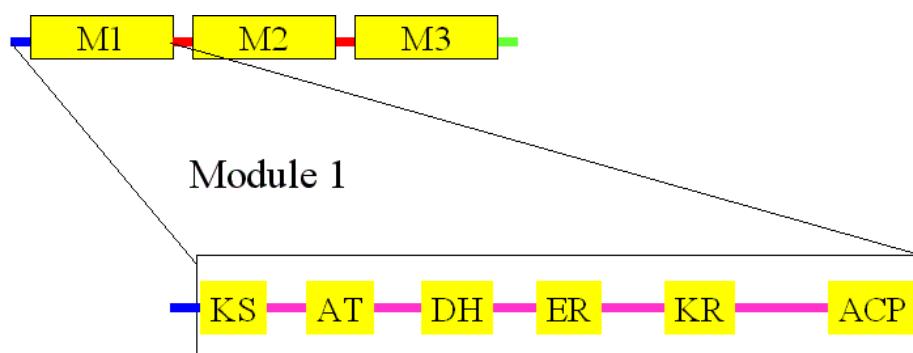
Slika 3. Shematski prikaz nastajanja β -ketoacilnog polimera na enzimu PKS.

Pri sintezi masnih kiselina nakon kondenzacije slijedi β -ketoredukcija, dehidratacija i redukcija enoilne skupine kako bi se sintetizirala potpuno reducirana (zasićena) masna kiselina. Nasuprot tome, u sintezi poliketida ovi reduksijski koraci mogu biti djelomično ili potpuno izostavljeni te rezultiraju značajno različitim poliketidnim lancima s prisutnim β -keto, β -hidroksi i alkilnim skupinama (Fujii i sur., 2001; Chan i sur., 2009).

Enzimi poliketid sintaze se obično kategoriziraju na temelju broja njihovih podjedinica (jedna ili više) te načinu sinteze (modularne i iterativne) (Tablica 1). Također je pronađeno 11 različitih katalitičkih domena koje se mogu nalaziti u modulima (Tablica 2.). Najjednostavnija PKS se sastoji od domena KS, AT, ACP i TE. Standardni redoslijed domena je prikazan na Slici 4, dok su na Slici 5 prikazani primjeri nekih poliketidnih spojeva. Najbolje opisana vrsta PKS enzima je modularna poliketid sintaza tip I, ali se većina tako dobivenih informacija može primjeniti i na druge vrste enzima PKS (Fujii i sur., 2001; Cheng i sur., 2003). Na Slici 6 se mogu vidjeti razlike u mehanizmu biosinteze između PKS sustava I, II i III.

Tablica 1. Razlike između pojedinih sustava PKS na s obzirom na strukturu, mehanizam sinteze, evolucijske odnose i učestalost u prirodi (preuzeto od Watanabe i Ebizuka, 2004).

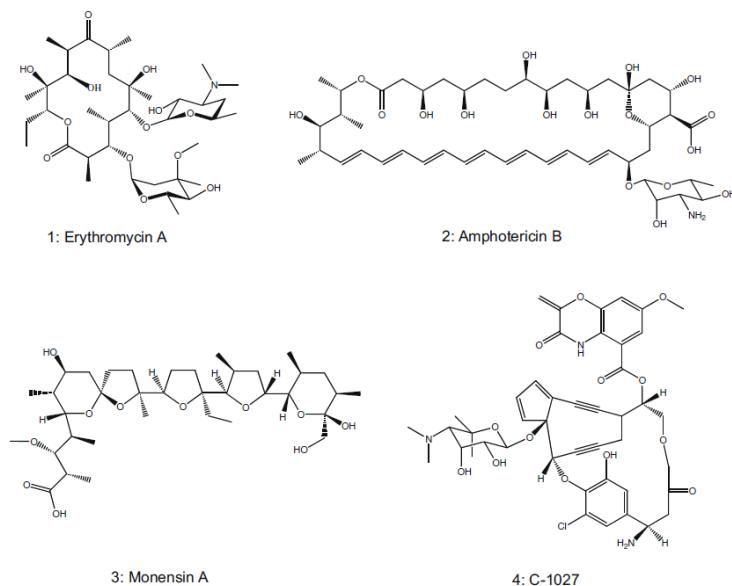
Vrsta	Struktura enzima	Mehanizam sinteze	Predviđljivost strukture	Sličnost s	Pronađen kod
Tip I (modularan)	Jedan protein s više modula	Linearan („proizvodna traka“), gdje je svako aktivno mjesto upotrijebljeno samo jednom	Da, do određene mjeru	/	Bakterije
Tip I (iterativan)	Jedan protein s jednim modulom	Iterativan, aktivna mjesta su opetovano korištena	Ne	FAS kralježnjaka	Pljesni
Tip II	Više proteina, svaki s jednim monofunkcionalnim aktivnim mjestom	Iterativan, aktivna mjesta mogu biti korištena ili jednom ili više puta	Ne	FAS bakterija	Bakterije
Tip III	Jedan protein (homodimer) s jednim aktivnim mjestom	Iterativan, aktivna mjesta su opetovano korištena	Ne	/	Biljke i bakterije



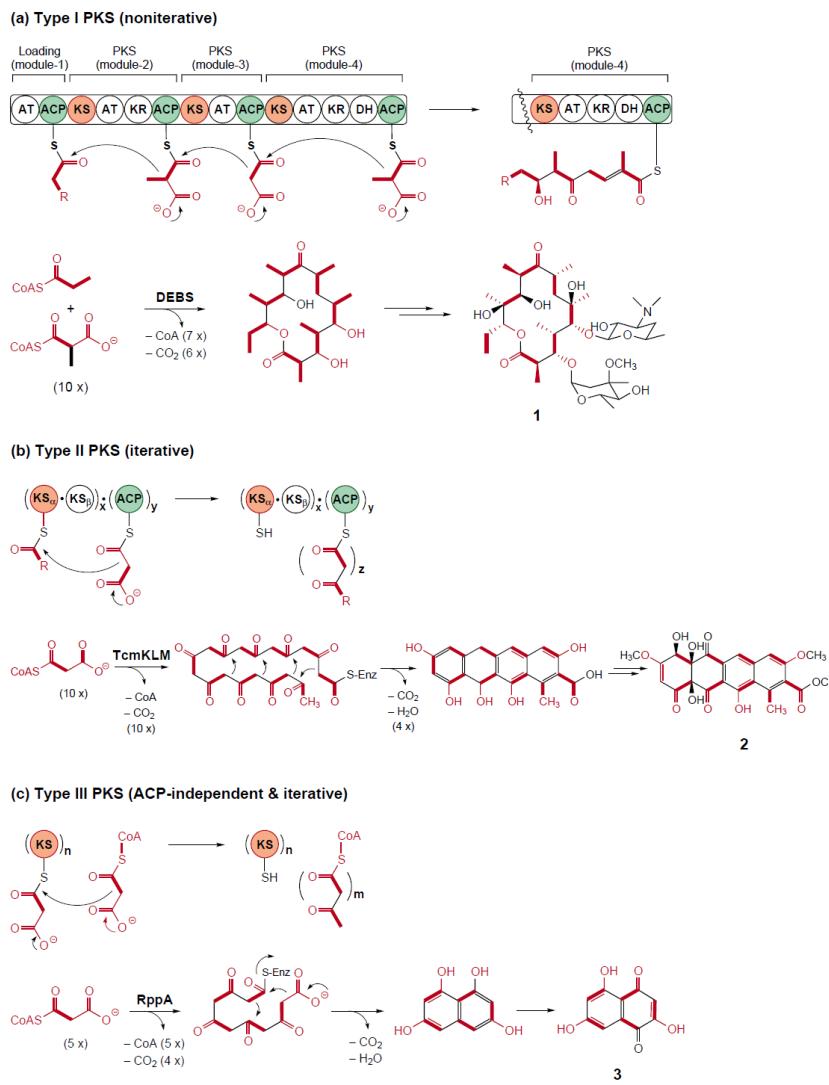
Slika 4. Prikaz redoslijeda standardnih domena koje se mogu pojaviti u modulu PKS-a.

Tablica 2. Različite vrste domena pronađenih u poliketid sintazama. Jedanaest različitih domena može biti podijeljeno u 3 grupe na temelju toga u kojem dijelu sinteze sudjeluju (C = reakcije kondenzacije, R = redukcija β -ketona i M = druge post-kondenzacijske modifikacije). S obzirom da su kratice na engleskom jeziku, radi lakšeg sporazumijevanja su i puna imena domena napisana na istome jeziku.

Aktivno mjesto		Funkcija
Starter Acyltransferase (SAT)	C	Uvođenje početnog acilnog monomera
Acyltransferaze (AT)	C	Uvođenje početnog, produžnog ili središnjeg acilnog monomera
Acyl Carrier Protein (ACP)	C	Zadržavanje rastućeg poliketidnog lanca u obliku tiolnog estera (KS-S-poliketid)
β -ketoacyl synthaze (KS)	C	Kondenzacija između početnog/središnjeg i produžnog monomera
β -keto reductase (KR)	R	Redukcija β -keto skupine u hidroksilnu skupinu
Dehydratase (DH)	R	Redukcija hidroksilne skupine u enoilnu skupinu (nezasićenu)
Enoyl Reductase (ER)	R	Redukcija enoilne skupine u alkilnu skupinu (zasićenu)
Thioesterase (TE)	C	Potpomaganje otpuštanja konačnog produkta s enzima
Methyltransferase (MT)	M	Prijenos metilne skupine na rastući poliketid
Product Template domain	C	Određivanje načina smatanja poliketida u nereducirajućim PKS-ovima
Claisen Cyclase (CYC)	M	Potpomaganje stvaranja prstena Claisenovom vrstom ciklizacije
Condensation (CON)	M	Potpomaganje kondenzacije sintetiziranog poliketida s drugim poliketidima



Slika 5. Primjeri nekih poliketidnih kemijskih spojeva. Proizvođači navedenih poliketida: 1. *Saccharopolyspora erythraea*, 2. *Streptomyces nodosus*, 3. *Streptomyces cinnamonensis*, 4. *Streptomyces globisporus*.



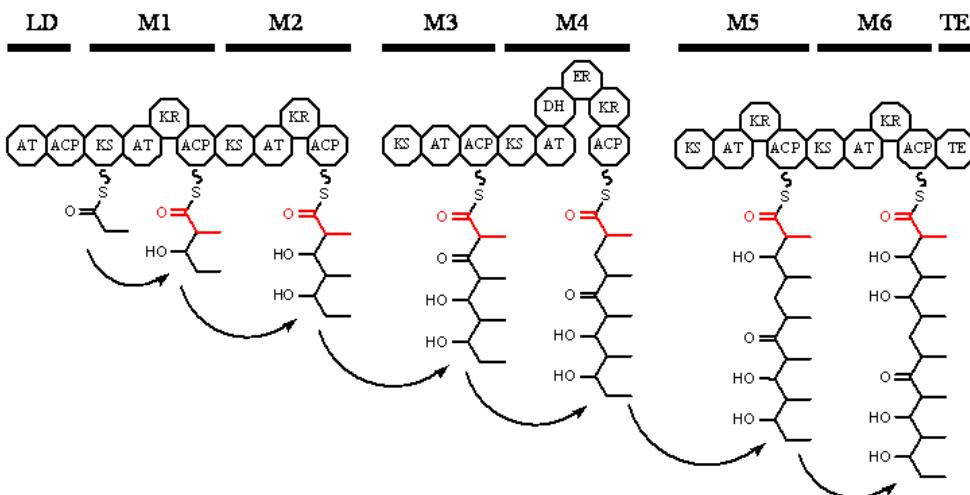
Slika 6. Enzimski sustavi: PKS tip I (a), PKS tip II (b), PKS tip III (c) (Shen, 2003).

Enzimski sustav PKS tip I (Slika 6 (a)) je multifunkcionalni enzim organiziran u module, od kojih se svaki sastoji od više domena koje provode po jednu reakciju u sintezi. Svaki modul upravlja sintezom jednog monomera u oligomeru, kao što je prikazano na primjeru deoksieritromicin B sintetaze (DEBS) u sintezi reduciranih poliketida (npr. makrolida, polietera i poliena), poput eritromicina A (iterativna podvrsta PKS sustava tipa I se sastoji od jednog modula). Enzimski sustavi PKS tip II (Slika 6 (b)) su multienzimski kompleksi koji izvršavaju iterativne reakcije, kao što je pojašnjeno na primjeru tetracenomicin PKS sustava u sintezi aromatskih poliketida (često policikličkih), poput tetracenomicina C. I na kraju, enzimski sustav PKS tip III (Slika 6 (c)) su homodimerni iterativni kondenzacijski enzimi, kao što je prikazano na primjeru sinteze RppA. Ta se sintaza upotrijebljava za sintezu aromatskih poliketida (često mono- i bicikličkih), kao što je flavolin (Shen, 2003).

2.1.1.1. Poliketid sintaze tipa I

Kao što je već rečeno, poliketid sintaze tipa I karakteriziraju veliki više-modularni i više-funkcionalni polipeptidi na kojima se nalaze sve aktivne domene za svaku enzimom kataliziranu reakciju. Ovi se sustavi dalje dijele na dvije različite vrste podsustava: modularne podsustave (sustav PKS se sastoji od sekvence odvojenih modula i sinteza teče linearno), koji su zaslužni za sintezu npr. eritromicina, rapamicina i rifamicina te na iterativne podsustave (gdje se domene koriste ciklički) koji sintetiziraju poliketide plijesni, npr. aflatoksine. Iterativni podsustavi se nadalje mogu podijeliti na nereducirajuće PKS (NR-PKS), čiji produkti su pravi poliketidi, djelomično reducirajuće PKS (PR-PKS) i potpuno reducirajuće PKS (FR-PKS), čiji su produkti derivati masnih kiselina (Nicholson i sur., 2001; Hutchinson, 2003).

U modularnim sustavima PKS tip I broj ketidnih podjedinica je jednak broju modula od kojih se taj PKS sastoji, dok kod iterativnog tipa to nije tako, s obzirom da se moduli koriste više puta (Cheng i sur., 2003; Jenke-Kodama i Dittmann, 2009). Prikaz modularnog PKS sustava tipa I je na Slici broj 7.



Slika 7. Prikaz linearnog načina sinteze eritromicina pomoću 3 modularna proteina PKS tipa I. Rastući međuprodot ide s jednog na drugi modul. Domena TE potpomaže otpuštanje konačnog produkta s enzima. Najnovije dodana acetatna skupina je označena crvenom bojom. Debele crte na vrhu prikazuju duljinu pojedinih modula. Eritromicin je sintetiziran pomoću 3 različita enzima, od kojih se svaki sastoji od 2 modula (Frandsen, 2010).

2.1.1.2. Poliketid sintaze tipa II

Za razliku od sustava PKS tip I, za enzime sustava PKS tip II su svojstveni više-enzimski kompleksi koji se sastoje od zasebnih, uglavnom mono-funkcionalnih proteina koji djeluju iterativno. Enzimsku osnovu čine dvije podjedinice β -ketoacil sintaze (KS α i KS β) i jedan mali proteinski nosač acila (ACP), koji zajedno tvore tzv. "minimalni PKS". Podjedinica KS α zajedno s podjedinicom KS β katalizira kondenzaciju acil-tioestera prilikom sinteze ugljikovog lanca. Podjedinica KS β ima i dodatnu ulogu faktora koji određuje duljinu ugljikova lanca (engl. "Chain Length Factor", CLF). Nasuprot tome, domena ACP djeluje kao sidro za rastući poliketidni lanac tijekom kondenzacijskih i modifikacijskih stupnjeva. Sustavi PKS tip II sudjeluju u biosintezi pigmenata spora te aromatskih antibiotika poput tetraciklina i antrakinona (Komaki i Harayama, 2006; Li i Müller, 2009).

Tako, primjerice, biosintezu poliketidnog antibiotika aktinorodina u vrste *S. coelicolor* sintetizira jedan tipičan sustav PKS tip II, koji upotrijebjava šest genskih produkata kako bi se oblikovalo osam acetatnih jedinica u tri spojena supstituirana šesteročlana prstena (dva aromatska prstena i jedan laktinski). Tri su podjedinice specifične za reakciju: ketoreduktaza, aromataza i ciklaza (engl. "Keto Reductase", KR; "Aromatase", ARO; "Cyclase", CYC), dok su preostale tri (KS α , KS β i ACP) ponavljujući "minimalni PKS" koji je zajednički svim sustavima PKS tipa II (Taguchi i sur., 2006). Važno je također spomenuti kako je prije nekoliko godina pronađen i PKS tipa II koji se ne ponaša iterativno – nedostaje mu ACP domena. Kao supstrat se izravno koristi acil-CoA i katalizira se nastajanje i C-C i C-O veza u sintezi makrotetrolida (Shen, 2003).

2.1.1.3. Poliketid sintaze tipa III

Enzimski sustavi PKS tip III, također poznati i pod imenom šalkon sintazi slični sustavi (engl. „Chalcone Synthase Like, CHS like“) su homodimerni iterativni kondenzacijski enzimi koje pronalazimo kod biljaka i bakterija. Oni sintetiziraju aromatske poliketide, često mono- ili bicikličke (uključujući i flavonoide) (Shen, 2003; Gokhale i sur., 2007).

Za razliku od PKS sustava tip I i II, sustav PKS tip III koristi jedno aktivno mjesto da iterativno izvede kompleksne serije reakcija: dovodenje acil-CoA tioestera do aktivnog mjesta

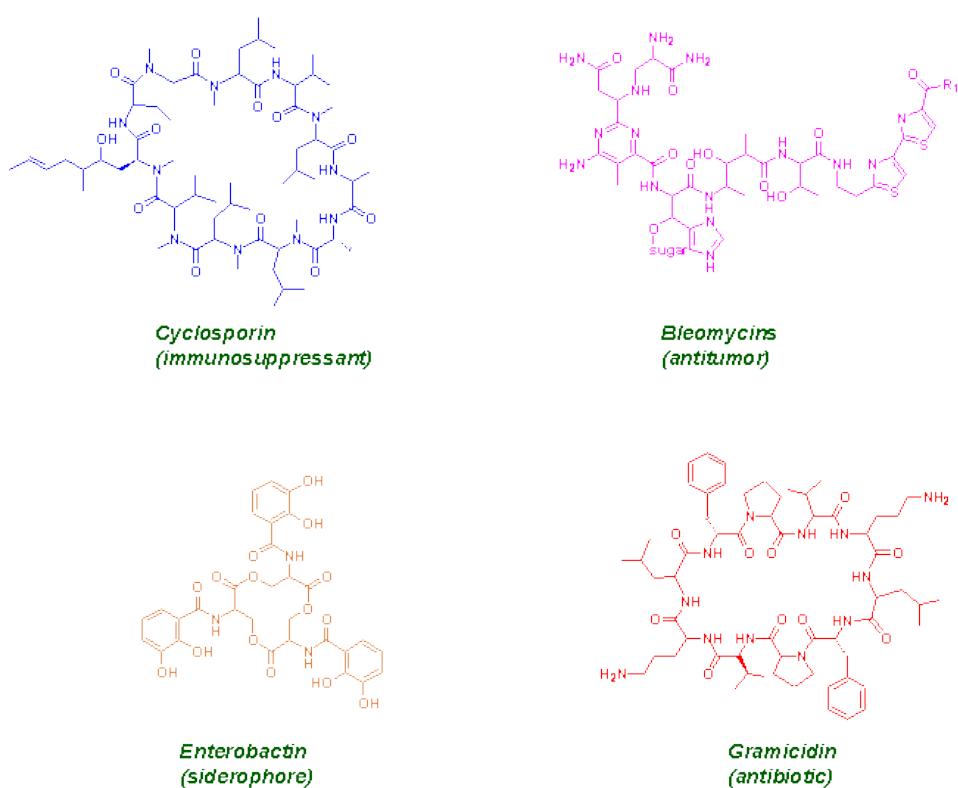
cisteina, produživanje lanca dekarboksilativnom kondenzacijom produžnog monomera (obično malonil-CoA) i intramolekularna ciklizacija za otpuštanje konačnog produkta. Variranjem početnog monomera, broja kondenzacijskih ciklusa i mehanizma ciklizacije (aldolna kondenzacija, Claisenova kondenzacija ili laktonizacija), sustav PKS tip III može proizvesti širok raspon konačnih produkata te se na temelju svojih produkata dijele na još 5 podvrsta (Li i Müller, 2009).

2.1.2. Neribosomalno sintetizirani peptidi

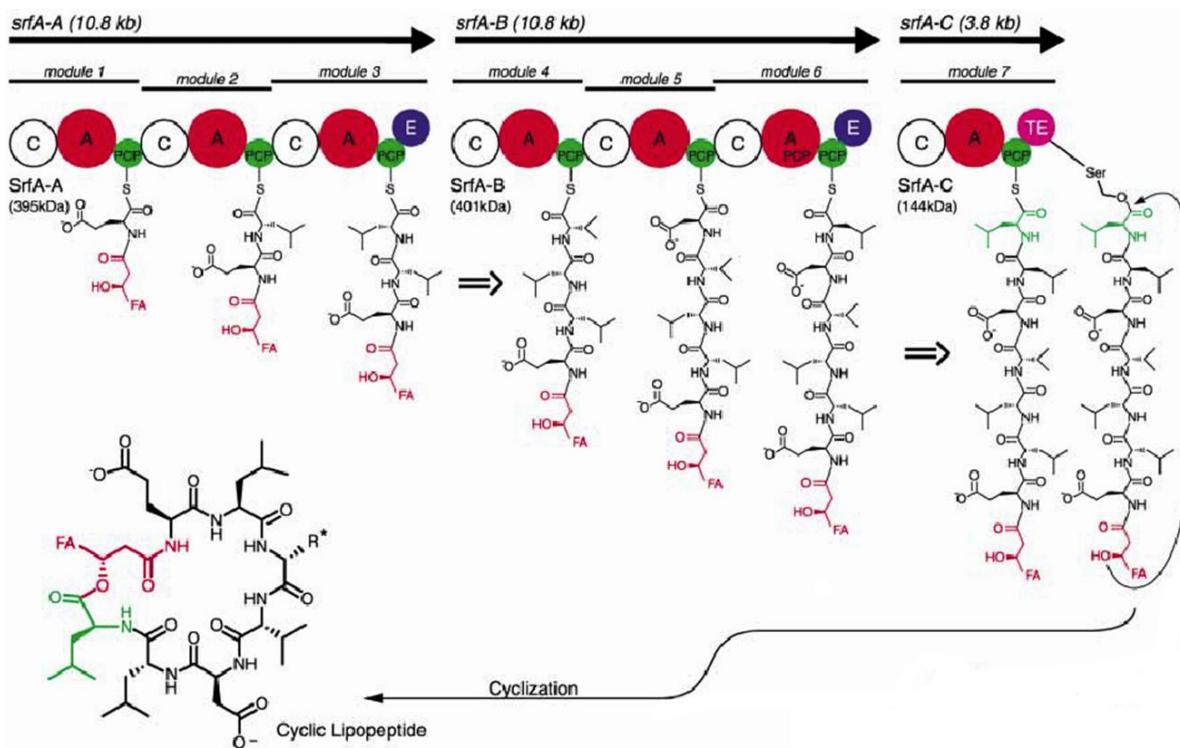
Neribosomalno sintetizirani peptidi su vrsta sekundarnih peptidnih metabolita. Obično ih proizvode mikroorganizmi poput bakterija i pljesni (Schwarzer i Marahiel, 2001). Sintetiziraju se uz djelovanje neribosomalne peptid sintetaze (NRPS) koja je, za razliku od ribosoma, neovisna o glasničkoj RNA (mRNA). Svaka neribosomalna peptid sintetaza može sintetizirati samo jednu vrstu peptida. Neribosomalno sintetizirani peptidi često imaju cikličku i/ili razgranatu strukturu. Mogu sadržavati i neproteinske aminokiseline uključujući i D-aminokiseline, mogu biti modificirani, npr. N-metilnim ili N-formilnim skupinama, a mogu biti i glikozilirani, acilirani, halogenirani ili hidroksilirani (sveukupno postoji više od 50 vrsta aminokiselina koje mogu biti upotrijebljene (Ansari i sur., 2004)). Ciklizacija sintetiziranih peptida može rezultirati i nastankom oksazolina i tiazolina, koji se nakon toga mogu oksidirati ili reducirati. Također, neribosomalni polipeptidi su često dimeri ili trimeri identičnih sekvenci povezanih u prsten ili razgranati. To su vrlo raznoliki prirodni produkti mikroorganizama sa širokim spektrom bioloških i farmakoloških svojstava. Često mogu biti i toksini, siderofori ili pigimenti, a neki od neribosomalno sintetiziranih peptida s antibiotskim, citostatskim i imunosupresivnim svojstvima se također nalaze u komercijalnoj upotrebi, npr. bacitracin, gramicidin i vankomicin od antibiotika, bleomicin od citostatika i ciklosporin od imunosupresiva (Schwarzer i sur., 2003).

Do danas je izolirano oko 1000 strukturno različitih neribosomalno sintetiziranih proteina. Međutim, za vrlo je malen broj od njih poznat mehanizam biosinteze (Caboche i sur., 2008). Poznato je da su enzimski sustavi NRPS organizirani u module koji se sastoje od domena (slično kao i enzimski sustavi PKS) u kojem svaki modul katalizira jedan ciklus elongacije i modifikacije funkcionalne skupine. Broj i redoslijed modula i vrsta domena unutar modula određuju strukturne varijacije konačnog peptidnog produkta tako što određuju broj, redoslijed

i izbor aminokiseline koja će biti uključena u sintezu i modifikaciju nastajućeg peptida. Osnovni modul enzima NRPS sadržava domenu za kondenzaciju (engl. "Condensation domain", C), domenu za adenilaciju aminokiselina (engl. "Adenylation domain", A) i mali proteinski nosač peptidila (engl. "Peptidyl Carrier Protein", PCP). Domena A određuje koji će se supstrat koristiti i veže ga za fosfopanteteinsku ruku domene PCP u obliku tioestera uz utrošak jednog ATP. Domena C katalizira stvaranje peptidne veze između aminoacil- ili peptidil-S-PCP-a i sljedeće aminokiseline vezane na domenu PCP. Rezultat je produženje lanca za jednu aminokiselinu, nakon čega može slijediti epimerizacija, N-metilacija ili heterociklizacija supstrata. Na kraju se produkt odvaja s enzima uz pomoć C-terminalne tioesteraze (Te) (Marahiel i sur., 2002; Challis i Naismith, 2004). Primjeri nekih neribosomalno sintetiziranih peptida su na Slici 8, dok je sinteza cikličkog lipopeptida je prikazana na Slici 9.



Slika 8. Primjeri nekih neribosomalno sintetiziranih peptida.



Slika 9. Stupnjevit prikaz sinteze cikličkog lipopeptida (Bruner i sur., 2002).

Strukturna raznolikost produkata enzimskih sustava NRPS je posljedica, uz različite organizacije domena unutar modula, i različitih mehanizama sinteze produkata. Po tome se enzimski sustavi NRPS dijele u tri grupe: 1) linearne (tip A) – gdje se svaki modul upotrijebjava samo jednom, odnosno broj modula odgovara broju aminokiselina u polipeptidu (analogno modularnom enzimskom sustavu PKS tip I), 2) ponavljajuće (tip B) – gdje se jedan modul upotrijebjava više puta tijekom sinteze jednog polipeptida (analogno iterativnom enzimskom sustavu PKS tip II) i 3) nelinearnom (tip C) – gdje se neki moduli ponavljaju više puta i može se odstupati od uobičajene organizacije modula C-A-PCP zbog prisutstva barem jednog neuobičajenog modula (Marahiel i sur., 2002; Challis i Naismith, 2004).

2.1.3. Poliketidno/peptidni hibridi

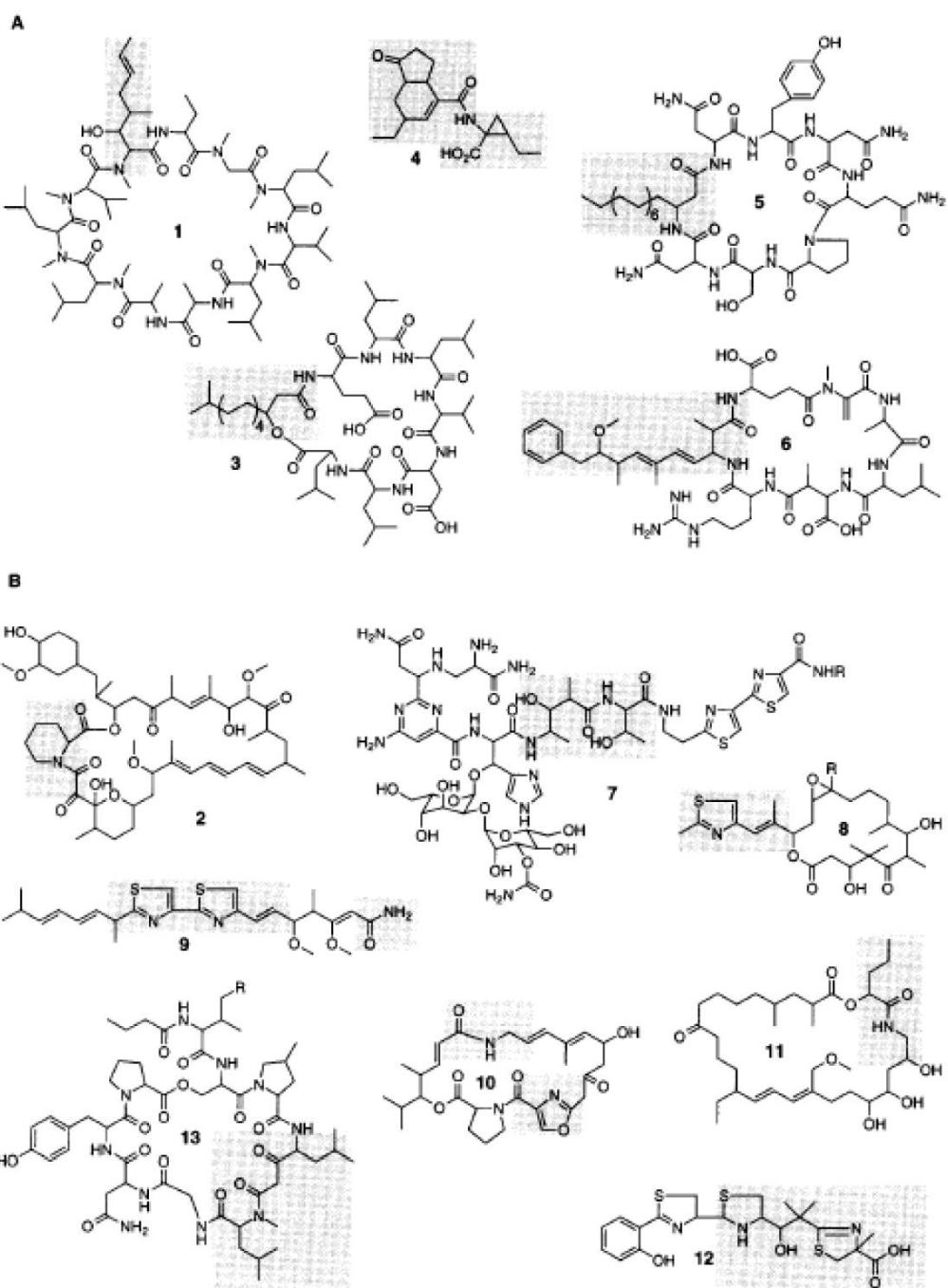
Poliketidno/peptidni sustavi upotrijebjavaju i PKS i NRPS module u sintezi konačnog produkta, a ovdje spadaju i sekundarni metaboliti koji nastaju biosintezaom iz aminokiselina i kratkih karboksilnih kiselina, kao i cjeloviti PKS ili NRPS sustavi uz pokoju domenu iz drugog tipa enzima (Slika 10). To je moguće zbog slične modularne strukture, ali i sličnih

reakcija koje kataliziraju pojedine domene. Oba sustava također koriste i proteinske nosače, PCP kod enzima NRPS i ACP kod enzima PKS, koji pridržavaju rastuće lance. Oni post-translacijskom modifikacijom dobivaju 4'-fosfopanteteinsku prostetsku skupinu na koju je vezan rastući intermedijer tijekom cijelog procesa produljivanja. Kada produkt dostigne punu dužinu, otpušta se pomoću domene tioesteraze koja se obično nalazi na C-terminalnom kraju enzima PKS, odnosno NRPS (Du i sur., 2001). Međutim, takav tip organizacije otežava mogućnost predviđanja strukture konačnog produkta. Dodatan problem predstavlja i izostavljanje (engl. „module skipping“) cijelih modula, koji se smatraju inaktivnima tijekom biosinteze. Problem također predstavlja i otkrivanje osnovnih katalitičkih i molekularnih svojstava i povezivanje strukturno-funkcionalnih odnosa ovih neobičnih sustava. Međutim, takvi hibridni sustavi svejedno predstavljaju veliki potencijal u razvoju „neprirodnih“ prirodnih lijekova i predstavljaju najpogodnije komplekse za kombinatorijalnu sintezu. (Wenzel i Müller, 2005).

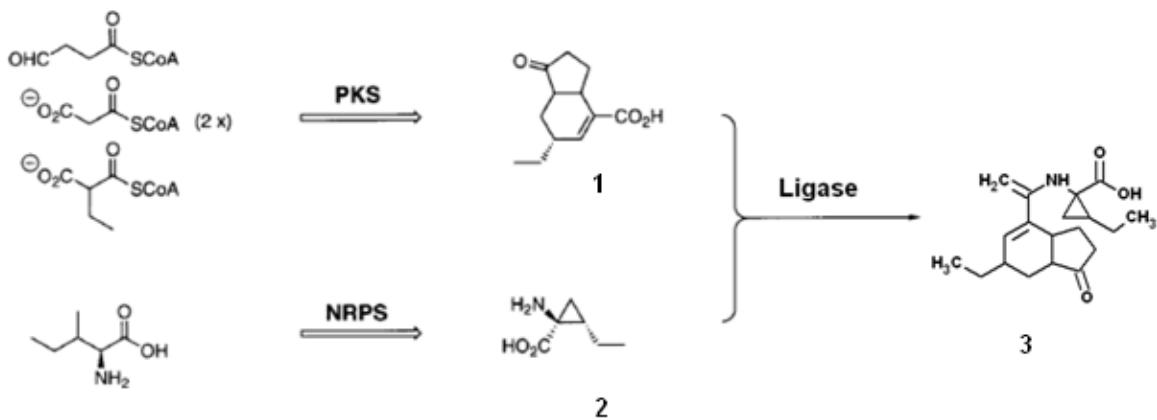
Na temelju biosintetskih mehanizama kojima se karboksilni/poliketidni i aminokiselinski/peptidni dijelovi ugrađuju u konačne proekte, poliketidno/peptidni prirodni produkti mogu se podijeliti u dvije skupine: 1. Sustavi koji ne uključuju neposrednu hibridizaciju između enzima NRPS i PKS i 2. Sustavi koji uključuju neposrednu funkcionalnu hibridizaciju između enzima NRPS i PKS (Du i sur., 2001).

2.1.3.1. Sustavi koji ne uključuju neposrednu hibridizaciju između enzima NRPS i PKS

Kod ovih sustava okosnica hibridnog produkta nastaje spajanjem poliketidnih i aminokiselinskih dijelova mehanizmima koji ne zahtijevaju izravnu funkcionalnu hibridizaciju između proteina NRPS i PKS. Naime, poliketidni i aminokiselinski dijelovi se zasebno sintetiziraju pomoću enzima PKS i NRPS, a potom se spajaju u poliketidno/peptidni hibrid pomoću enzima ligaze. Primjer je fitotoksin koronatin, kojeg proizvode mnogi sojevi bakterije *Pseudomonas syringae*, a sastoji se od dvije različite komponente, poliketidne koju čini koronafaksna kiselina (Slika 11, struktura 1) i aminokiselinske koju čini koronaminska kiselina (Slika 11, struktura 2). Genska nakupina za koronatin nalazi se na dva lokusa, koji kodiraju za enzime NRPS i PKS, odijeljena regulatornom regijom. Oba lokusa sadrže vlastitu tioesterazu što potvrđuje hipotezu da se koronafaksna kiselina i koronaminska kiselina otpuštaju s NRPS i PKS enzima prije nego se spoje i formiraju koronatin (Slika 11, struktura 3) (Du i sur., 2001).



Slika 10. Primjeri poliketidno/peptidnih hibridih prirodnih produkata: ciklosporin A (1), rapamicin (2), surfaktin (3), koronatin (4), mikosubtilin (5), mikrocistin (6), bleomicin (7), epotiloni (8), miksofiazol (9), pristinamicin II_B(10), TA (11), jersiniabaktin (12) i nostopeptolidi (13). Spojevi između peptidnih i poliketidnih dijelova su osjenčani. (A) Biosinteza ovih spojeva ne zahtijeva izravnu funkcionalnu hibridizaciju između enzima NRPS i PKS. (B) Biosinteza ovih spojeva uključuje hibridne sustave NRPS/PKS.



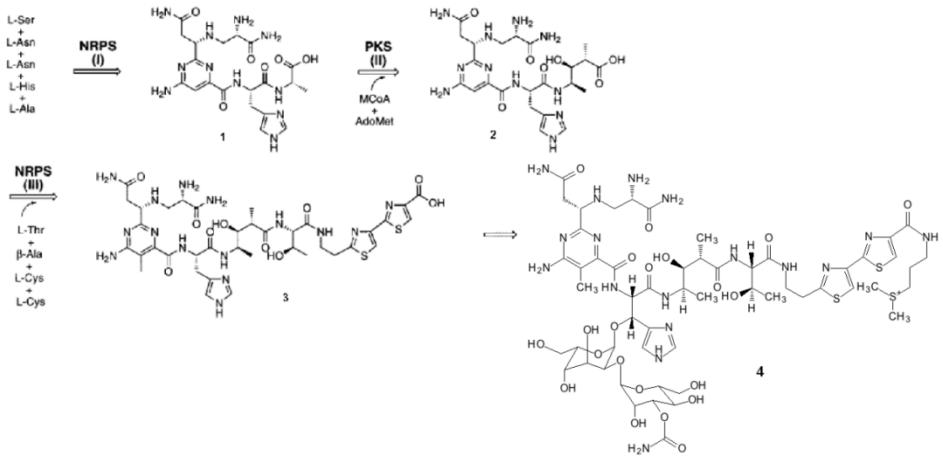
Slika 11. Biosintetski put koronatina u bakteriji *P. syringe*. Poliketidni dio koronafaksne kiseline i aminokiselinski dio koronaminske kiseline se spajaju u koronatin pomoću enzima ligaze (Parry i sur., 1991).

2.1.3.2. Sustavi koji uključuju neposrednu funkcionalnu hibridizaciju između enzima PKS i NRPS

Ovim putem nastaje većina poliketidno/peptidnih metabolita čiji su biosintetski putovi opisani. Okosnica se hibridnog produkta sastavlja djelovanjem hibridnog PKS/NRPS sustava, koji posreduje u izravnom produljivanju peptidnog međuproducta vezanog za enzim NRPS pomoću PKS modula i obratno. Sinteza aglikona bleomicina se odvija prema hibridnom modelu NRPS/PKS/NRPS u tri stupnja. U prvom stupnju iz aminokiselina Ser, Asn, Asn i His nastaje P-3A (Slika 12, struktura 1) pomoću enzima NRPS. U drugom stupnju enzim PKS katalizira elongaciju dodatkom malonil-CoA i adenozilmethionina čime nastaje P-4 (Slika 12, struktura 2). Trećim se stupnjem nastavlja elongacija P-4 (Slika 12, struktura 3), koju katalizira enzim NRPS, dodatkom aminokiselina β-Ala, Cys i Cys, čime nastaje P-6m, iz kojeg nastaje bleomicin (Slika 12, struktura 4) (Du i sur., 2001).

2.2. SEKVENCIRANJE GENOMA RAZLIČITIH VRSTA AKTINOBakterija

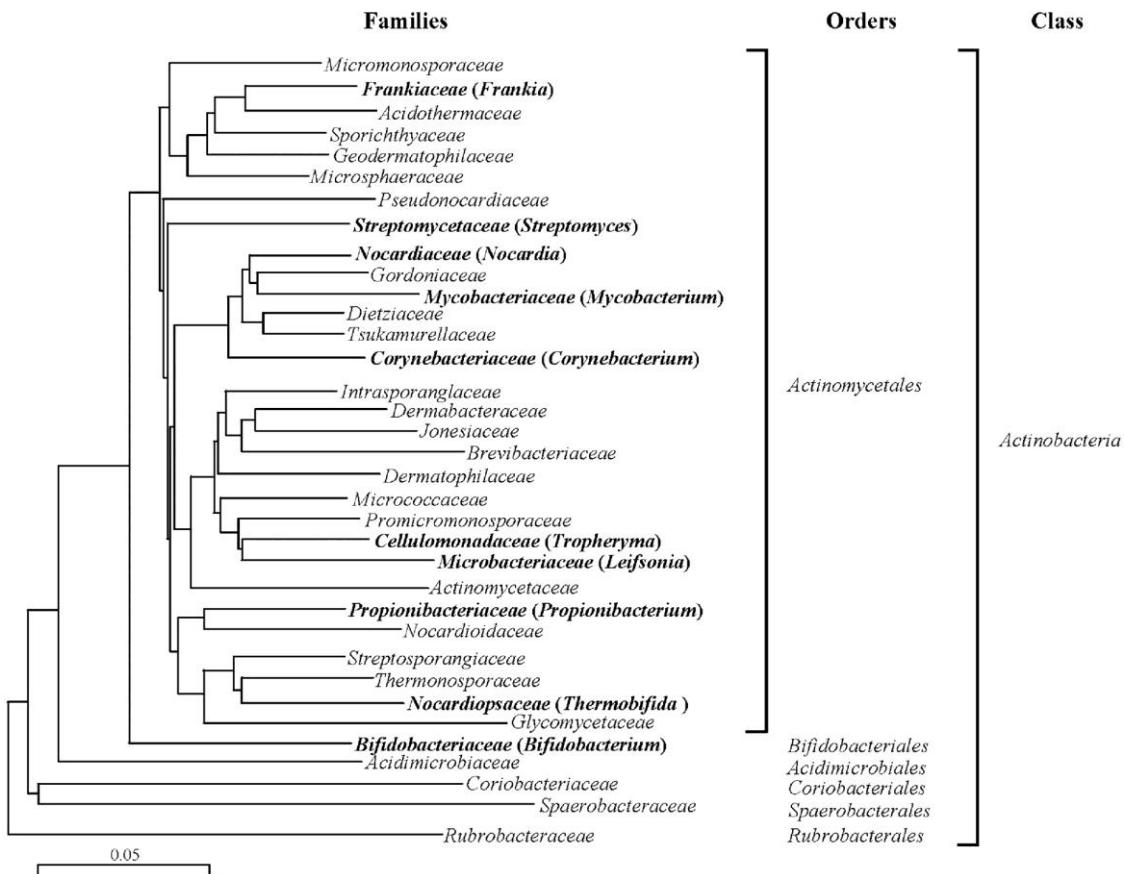
Aktinobakterije su jedna od najvećih taksonomskih skupina među 18 glavnih odjeljaka koji su poznati unutar carstva *Bacteria*. Ta skupina uključuje 5 pod-razreda i 14 pod-redova. To su Gram-pozitivne bakterije uglavnom visokog G+C sastava; do više od 70% u vrsta rodova *Streptomyces* i *Frankia* (Ventura i sur., 2007), iako su se nedavno pronašle i slatkovodne aktinobakterije s niskim G+C sastavom, koji može iznositi i do 42% (Ghai i sur., 2012).



Slika 12. Biosintetski put bleomicina (4) u bakterije *S. verticillus* uključuje hibridni sustav NRPS/PKS/NRPS. Rastući hibridni peptidni-poliketidni biosintetski međuproducti P-3A (1), P-4 (2) i P-6m (3) izolirani su iz divljeg tipa bakterije *S. verticillus* i određene su im strukture.

Aktinobakterije se mogu pronaći u zemlji i u vodi (Servin i sur., 2008) te obuhvaćaju vrlo širok spektar morfološki različitih oblika stanica: okrugle stanice (*Micrococcus*), štapićasto-okrugle stanice (*Arthrobacter*), stanice u obliku fragmentiranih hifa (*Nocardia* spp.) ili trajno visoko diferencirane razgranate stanice u obliku micelija (*Streptomyces* spp.). Među njima su također očite i fiziološke razlike te imaju i različita svojstva metabolizma. Ovom su skupinom obuhvaćene patogene bakterije (primjerice, *Mycobacterium* spp., *Corynebacterium* spp., *Propionibacterium* spp.), bakterije koje žive u tlu (*Streptomyces* spp.), biljni simbionti (*Leifsonia* spp.), bakterije koje mogu, ali i ne moraju fiksirati dušik (*Frankia* spp.) te stanovnici animalnog probavnog trakta (*Bifidobacterium* spp.) (Ventura i sur., 2007).

Do sada sekvencirani genomi aktinobakterija uglavnom pripadaju mikroorganizmima koji su bitni za humanu i veterinarsku medicinu, biotehnologiju i ekologiju, s obzirom da su proizvođači brojnih spojeva koji se upotrijebljavaju u praksi kao antibiotici, citostatici, imunosupresivi i dr. Važnost sekvenciranja genoma aktinomiceta je povećanje znanja njihovog biosintetskog potencijala, otkrivanje načina stimulacije proizvodnje „kriptičnih“ prirodnih produkata – onih koji se ne proizvode pod standardnim laboratorijskim uvjetima. Sekvenciranjem i analizom gena za 16S rRNA utvrđeno je 39 porodica i 130 rodova (Slika 13) (Ventura i sur., 2007), a prema posljednjim podacima, aktivna su 746 genomska projekta za aktinobakterije (NCBI, 2013).



Slika 13. Filogenetsko stablo aktinobakterija temeljeno na redoslijedu 1500 nukleotida 16S rRNA. Genomi sekvencirani do 2007. godine su istaknuti (Ventura i sur., 2007).

Za sada su samo su tri micelijska roda aktinobakterija (*Streptomyces*, *Thermobifida* i *Frankia*) zastupljena u javno dostupnim bazama genoma. Od ta tri roda, rod *Streptomyces* je dobio posebnu pozornost zbog tri glavna razloga:

1. prisutnost u tlu - streptomikete su prisutne u izobilju i od velike su važnosti jer sadrže hidrolitičke enzime pomoću kojih oslobađaju netopljivi ugljik iz organskih ostanaka biljaka i gljiva;
2. vrsta *S. coelicolor* A3(2) predstavlja izuzetno važan modelni organizam zbog svoje složene diferencijacije;
3. streptomikete su bogat prirodni izvor antibiotika i drugih biološki aktivnih sekundarnih metabolita te su zbog toga od velikog značenja za medicinu i veterinu.

Kromosomi bakterija roda *Streptomyces* su vrlo veliki (8 – 10 Mb). Kromosom vrste *S. coelicolor* je prvi slučaj u kojem je potvrđeno da bakterijski genom sadržava više gena od jednostavnog eukariotskog organizma kvasca *Saccharomyces cerevisiae* (genom kvasca sadržava 6 203 gena, dok genom vrste *S. coelicolor* sadržava 7 825 gena). Iako je većina bakterijskih kromosoma kružna, kromosomi roda *Streptomyces* su linearni. Pretpostavlja se da je linearnost kromosoma streptomiceta posljedica jednostrukih recipročnih izmjena (engl. "single-crossover recombination") između prvotnog kružnog kromosoma i linearne plazmide (Hranueli i Cullum, 2001).

2.2.1. Važnost sekvenciranja genoma različitih vrsta aktinobakterija

Rezistencija mikroorganizama na postojeće lijekove predstavlja svjetski problem. Zbog pretjerane i neutemeljene upotrebe antibiotika došlo je do globalnog povećanja otpornosti patogenih mikroorganizama na postojeće antibiotike, tako da su se već pojavile i bakterijske infekcije koje se ne mogu više uspješno liječiti antibioticima ili su doktori primorani koristiti tzv. rezervne antibiotike, koji mogu imati ozbiljne nuspojave (WHO, 2013).

Jedno od predloženih rješenja je upotreba sekvenciranja genoma kako bi se pronašli novi antibiotici, koji nisu u tolikoj mjeri eksprimirani od strane svojih proizvođača, a prirodno im služe za stjecanje kompetitivne prednosti nad ostalim mikroorganizmima. Najnovije pretpostavke temeljene na genomskim sekvencama genoma bakterija *Streptomyces avermitilis*, *Streptomyces coelicolor* i *Saccharopolyspora erythraea* pokazuju kako se uobičajenim probirom pronađi samo 10-20% antibiotika koje proizvode ispitivani sojevi, tako da bi ti slabo eksprimirani, „kriptični“, antibiotici mogli postati novi lijekovi u borbi protiv rezistentnih mikroorganizama. Kao najizgledniji proizvođači novih antibiotika se nameće red *Actinomycetales*, u koji spadaju i vrste roda *Streptomyces*, koji su proizvođači oko 70% svih do sada poznatih antibiotika (Broad Institute, 2013).

2.2.2. Sekvenciranje genoma bakterije *Streptomyces sp. C*

Vrsta *Streptomyces sp. C* malo je poznata vrsta roda *Streptomyces* s visokim udjelom G+C, izolirana u Kini iz uzorka tla. Za njezino sekvenciranje je bio zadužen „Broad Institute“ te su

rezultati sekvenciranja objavljeni i javno dostupni (mogu se preuzeti s njihove web stranice). Opisana je i anotacija dijela genoma, do čega je došlo tijekom pisanja ovog diplomskog rada. Genom vrste *Streptomyces sp.* C se nalazi u 4 dijela (što je posljedica visokog G+C sastava, 72,61% te posljedičnog otežanog slaganja dijelova genoma), sveukupne veličine 8,46 Mb. Predviđa se da se u genomu nalazi 7 540 gena. Sekvenciranje je izvršeno ABI metodom, s prosječnim pokrivanjem genoma od 8 puta, a sastavljanje genoma je izvršeno Arachne metodom (Broad Institute, 2013).

2.3. BIOINFORMATIČKI ALATI ZA ANOTACIJU GENSKIH NAKUPINA

Uz napredak informatike i smanjenje troškova pohranjivanja podataka se vezuje i razvoj bioinformatike, odnosno primjene informatike u biološkim istraživanjima. Kako je cijena pohrane informacija pala, tako se povećao broj pohranjenih sekvenci gena i proteina, odnosno genoma i proteoma te se javila i potreba za smislenim organiziranjem te velike količine podataka. Također je pala i cijena sekvenciranja genoma te su se razvili i računalni programi za polu-automatsku i automatsku anotaciju genoma (Yandell i Ence, 2012).

Anotacija genoma je proces koji objedinjuje više različitih pristupa analizi nukleinskih kiselina ili proteinskih sekvenci sa svrhom da se pridodaju strukturalne i funkcionalne karakteristike modelima gena. Kada se opiše više tisuća modela gena u nekom organizmu, konstrukcija i vizualizacija mreže gena predstavlja nove izazove u razumijevanju i kompleksnim uzorcima ekspresije tih gena i generiranju novog znanja u istraživanju genoma. Da bi se uopće mogle iskoristiti informacije dobivene sekvenciranjem gena, potrebni su organizacijski i vizualizacijski alati (Castillo i sur., 2012). Anotacija genoma uključuje predviđanje genetičkih elemenata: strukturnih gena, pseudogena, promotora, tj. regulatornih regija, dijelova DNA koji ne sadržavaju genetičku uputu i ponovljenih sekvenca. Anotacija se može izvršiti na različite načine: ručno, automatski ili *ab initio* (Hranueli i sur., 2008; Yandell i Ence, 2012).

Ručnom anotacijom se dobivaju rezultati najveće točnosti jer metoda uključuje prethodno pretraživanje literature, međutim, ta je metoda spora i zbog toga obuhvaća vrlo malen dio anotiranih podataka. Informacije dobivene ručnom anotacijom jedne sekvene mogu se upotrijebiti za automatsku anotaciju (pomoću računalnih programa) drugih sekvenca koje

pokazuju određeni stupanj sličnosti. Točnost ove metode ovisi o evolucijskoj udaljenosti organizama čije genome analiziramo (što je veća udaljenost, manja je vjerojatnost točnosti predviđanja genetičkih elemenata). Iako je uspješnost automatske metode niža, zbog svoje se brzine češće upotrijebljava za anotaciju velikih sekvenci DNA. Neke se anotacije mogu izvesti upotrebom *ab initio* modela primjenom podatka prikupljenih analizom fizičko-kemijskih svojstava molekule produkta za koju genetičku uputu sadržava promatrana sekvenca DNA. *Ab initio* modeli predviđaju položaje gena u genomu na temelju matematičkih modela, a ne vanjskih dokaza (npr. usporedbom s već anotiranim genomom) i to im je velika prednost. Međutim, njihova je primjena ograničena jer ne mogu predvidjeti netranslatirane regije i transkripte koji podliježu alternativnoj post-transkripcijskoj obradi. Također, problem je i to što se matematički model mora izraditi za svaki organizam posebno, npr. učestalost kodona, distribucija duljina introna i eksona itd. Naime, rezultati mogu jako varirati, čak i između srodnih organizama, tako da se prilikom odabira metode bira između brzine anotacije i njezine točnosti (Reeves i sur., 2009; Yandell i Ence, 2012).

Za analizu genskih nakupina poliketida i neribosomalno sintetiziranih peptida razvijeno je nekoliko bioinformatičkih alata koji upotrijebljavaju donekle različite strategije i pristupe. Takvi bioinformatički alati i programski paketi omogućuju analizu i manipulaciju genskim nakupinama čija funkcija i konačni kemijski produkt još nisu poznati. Neki od alata navedeni su u tekstu koji slijedi.

2.3.1. Baze podataka *CSDB*, *SEARCHPKS*, *MAPSI*, *ClusterMine360*

Baza podataka *CSDB* (engl. „*ClustScan DataBase*“) je bioinformatička baza podataka koja sadrži genetičke i biokemijske informacije o prirodnim produktima sintetiziranim od strane sustava TMS (engl. „*Thiotemplate Modular Systems*“). Ovi sustavi uključuju sintaze PKS, sintetaze NRPS, sintetaze NRPS-neovisnih siderofora (NIS) i hibridne sintaze/sintetaze anotirane upotrebom programskog paketa *ClustScan*. Baza podataka *CSDB* sadrži sve podatke počevši s genomskim DNA sekvencama aktinobakterija zajedno sa sekvencama DNA i proteina koje anotiraju gene, module, domene i odgovarajuće poveznice sustava TMS. Ona također sadrži sve poznate poliketidne i peptidne građevne elemente u obliku izomernih zapisa SMILES-a (engl. „*Simplified Molecular Input Line Entry System*“), zajedno s programiranim slijedom koji dozvoljava predikciju linearnih i cikličkih poliketidnih i

peptidnih lanaca i aglikona u 2D i 3D oblicima prikladnim za daljnju računalnu obradu. Ovu bazu podataka je moguće pretraživati koristeći anotacije genskih nakupina sustava TMS, kao i strukture spojeva sustava TMS. S podacima baze *CSDB* se također može manipulirati koristeći različite uobičajene bioinformatičke alate (Thiotemplate Modular Systems Studies, 2013).

Bioinformatički alat *SEARCHPKS* (Yadav i sur., 2003) je program za otkrivanje i analizu domena poliketid sintaze. On omogućuje jednostavnu identifikaciju različitih domena PKS i modula iz zadane sekvene polipeptida. Uz to, može predvidjeti specifičnost potencijalnih domena aciltransferaza za različite početne ili produžne građevne jedinice. Bioinformatički alat *SEARCHPKS* služi za koreliranje kemijske strukture poliketida s organizacijom domena i modula s odgovarajućom modularnom poliketid sintazom, kao i za ekstenzivnu analizu i procjenu homologije sekvenci različitih poliketid sintazna, pružajući informacije za izvođenje pokusa izmjene domena i modula. Program *SEARCHPKS* se temelji na analizi različitih okarakteriziranih nakupina modularnih poliketid sintaza koje se nalaze u bazi podataka *PKSDB* (bazi podataka modularnih poliketid sintaza) (Yadav i sur., 2003; Ansari i sur., 2004).

Još jedan popularan program za analizu modularnih poliketid sintaza je sustav *MAPSI* (engl. „Management and Analysis for Polyketide Synthase Type I“) (Tae i sur., 2009), nasljednik sustava *ASMPKS* (Tae i sur., 2007). On pretražuje genome mikroorganizama u potrazi za modularnim nakupinama PKS i otkriva ih na temelju sličnosti s već anotiranim poliketidima. Također može predvidjeti specifičnost domena AT te time omogućuje predikciju sintetiziranog poliketidnog lanca. Sustav *MAPSI* je Web aplikacija s kompleksnijom strukturom od bioinformatičkog alata *SEARCHPKS*, s obzirom da koristi bazu podataka svih anotiranih poliketidnih nakupina, koji se koriste za identifikaciju homolognih nakupina, a koristi i informacije trenutno analiziranih genoma. Još jedna od odlika sustava *MAPSI* je i mogućnost sastavljanja umjetne poliketid sintaze tako što se jednostavno odaberu moduli s predefiniranim sastavom domena i specifičnosti za supstrate (Tae i sur., 2009). Nedavno je opisan bioinformatički programski paket *antiSMASH* (Medema i sur., 2011.) koji je danas najpopularniji jer predstavlja kompilaciju svi do sada opisanih programa.

Baza podataka *ClusterMine360* (Conway i Boddy, 2013) je novija baza podataka mikrobnih poliketidnih i neribosomalno peptidnih genskih nakupina. Ta baza podataka omogućava

članovima zajednice da svojim učešćem doprinose bazi podataka. Istovremeno, automatizacija procesa (preko veze s bazom podataka NCBI) omogućuje visoku dosljednost i kvalitetu podataka. U bazi se trenutno nalazi više od 200 genskih nakupina koje sintetiziraju više od 185 tipova spojeva. Također sadrži i preko 10 000 sekvenci domena PKS/NRPS. Te se sekvene mogu filtrirati i pohraniti kao individualne ili višestruke datoteke u FASTA obliku zapisa. Cijela je baza podataka organizirana oko tipova spojeva. Svaki tip spojeva je povezana s jednim ili više nakupina, kao i sa sinonimom, slikom strukture spoja, biosintetskim putevima koji su uključeni u sintezu tog spoja. Povezana je i sa spojevima koji imaju sličnu kemijsku strukturu ili su im slične genske nakupine. Baza podataka *ClusterMine360* također može odrediti za koje je metabolite određena domena aciltransferaze ili domena za adenilaciju domena specifična. Također može predvidjeti i je li neka ketoreduktaza aktivna ili inaktivna te može predvidjeti i stereokemiju produkta (Conway i Boddy, 2013).

2.3.2. Bioinformatički alati za anotaciju genskih nakupina prirodnih spojeva

Jedan od novijih programa za analizu genskih nakupina sustava NRPS je program *CLUSEAN* (engl. „*CLuster SEquence ANalyzer*“) – računalno bazirana okosnica za automatsku analizu genskih nakupina bakterijskih sekundarnih metabolita (Weber i sur., 2009). On integrira standardne metode analize poput bioinformatičkog programa *BLAST* (Altschul i sur., 1997) i bioinformatičkog programa *HMMER* (Eddy, 1998) s alatima specifičnima za identifikaciju funkcionalnih domena i motiva u sustavima NRPS/PKS tip I i za predikciju specifičnosti sustava NRPS. Kombinacija bioinformatičkih programa *BLAST* i *HMMER* je upotrijebljena za identifikaciju homolognih proteina i za identifikaciju domena, dok je predikcija specifičnosti domena za adenilaciju temeljena na pristupu razvijenom za računalni program *NRPSpredictor* (Rausch i sur., 2005).

Računalni program *NP.searcher* traži genske nakupine prirodnih produkata kao što su poliketidi, neribosomalni peptidi i hibridne molekule u sekvenciranim genomima. Omogućuje i predviđanje moguće strukture konačnog produkta primjenjujući 20 standardnih aminokiselina i 22 druga intermedijera u biosintezi neribosomalno sintetiziranih peptida, te četiri acil-CoA građevne jedinice poliketida (Li i sur., 2009).

Od ostalih alata, važno je spomenuti i alat *DecipherIT™* tvrtke Ecopia BioSciences Inc. koji automatski anotira genske nakupine mikroorganizama iz tla (Zazopoulos i sur., 2003). Sličnu primjenu ima i alat pod nazivom *Biogenerator*, norveške tvrtke Biosergen AS, koji omogućuje i predviđanje biološke aktivnosti kemijskih supstancija (Zotchev i sur., 2006) te internetski poslužitelj *SBSPKS* (engl. „Structure Based Sequence Analysis of PolyKetide Synthases“) koji omogućuje analizu genskih nakupina sustava PKS i strukturni prikaz više-enzimskih sustava, a sastoji se od tri dijela: NRPS_PKS (baza podataka NRPS-PKS koja omogućuje korisnicima povezivanje kemijskih struktura prirodnih spojeva s domenama i modulima u odgovarajućim NRPS i PKS sustavima), Model_3D_PKS (koji omogućuje predviđanje tercijarne i kvartarne strukture više-enzimskog modularnog sustava PKS) i Dock_Dom_Anal (koja predviđa sve moguće kombinacije interakcija između polipeptidnih lanaca, odnosno njihovih poveznica i domena smještenih na terminalnim krajevima) (Anand i sur., 2010).

2.3.3. Generički programski paket *ClustScan*

Programski paket *ClustScan* pruža mogućnost brze, poluautomatske anotacije modularnih biosintetskih genskih nakupina sa na znanju utemeljenim predviđanjima aktivnosti i specifičnosti njihovih modula i katalitički aktivnih domena. Nakon učitavanja, sekvenca DNA se automatski prevodi u šest otvorenih okvira čitanja sekvence proteina pomoću programa *Transeq* (Rice i sur., 2000). Programska paket omogućuje traženje gena na temelju sekvene translatiranoga proteina pomoću programa *GeneMark-PS* (Besemer i Borodovsky, 2005) ili *Glimmer* (Delcher i sur., 2007). Katalitički aktivne domene mogu se prepoznati unutar pronađenih enzima pomoću programskog paketa *HMMER* (Eddy, 1998) upotrebom proteinских profila preuzetih iz baze podataka Pfam (Finn i sur., 2008), ili pomoću vlastitih proteinских profila, strogo definiranim ili relaksiranim parametrima. Nakon obavljenе anotacije genske nakupine i biosintetskog puta, konačnu je anotaciju moguće "eksportirati" u obliku zapisa *GenBank*, *EMBL* ili *XML* i upotrijebiti je u drugim aplikacijama ili za upis u baze podataka *GenBank* (NCBI, 2013) odnosno *EMBL* (EBI, 2013) (Starčević i sur., 2008).

Starčević i suradnici (2008) su usporedili uspješnost generičkog programskog paketa *ClustScan* (Tablica 3) u odnosu na sustave *SEARCHPKS* i *MAPSI* (Yadav i sur., 2003; Tae i sur., 2007). Za procjenu uspješnosti generičkog programskog paketa *ClustScan* važna su dva

kriterija. To su funkcionalnost i točnost predviđanja te brzina i prikladnost programa za anotaciju velikih skupina sekvenca DNA. Točnost predviđanja je demonstrirana na primjeru jedne dobro opisane genske nakupine koja sadržava genetičku uputu za biosintezu poliketidnog antibiotika eritromicina (Starčević i sur., 2008).

Tablica 3. Usporedba generičkog računalnog programskog paketa *ClustScan* sa sustavima *SEARCHPKS* i *MAPSI*.

SVOJSTVA	<i>CLUSTSCAN</i>	<i>SEARCHPKS</i>	<i>MAPSI</i>
Učitavanje sekvence DNA	Da	Samo protein	Da ¹
Prepoznavanje specifičnosti domena AT	Da	Da	Da
Prepoznavanje stereokemije domena KR	Da	Ne	Ne
Prepoznavanje neaktivne domene KR	Da	Ne	Ne
Prepoznavanje neaktivne domene DH	Da	Ne	Ne
Prepoznavanje neaktivne domene ER	Da	Ne	Ne
Uređivanje predviđanja	Da	Ne	Ne
Eksport zapisa anotacije	Da	Ne	Ne
Predviđanje kemijskih struktura	Da	Ne	Da ²
Eksport kemijske strukture u standardnom obliku	Da	Ne	Ne

¹ Zahtijeva dugačke sekvence DNA (> 200 kb za bakterije visokog G+C-sastava) za preciznu identifikaciju gena.

² Ograničeno predviđanje kemijskih struktura bez mogućnosti njihova uređivanja.

3. EKSPERIMENTALNI DIO

3.1. MATERIJAL

3.1.1. Računalna podrška i operativni sustav

Ovaj je diplomski rad izrađen na računalu sljedećeg sklopolja: prijenosno računalo *Helwett Packard*, procesor Intel®Core™2 Duo CPU T5870 @ 2,00 GHz, 3GB radne memorije. Na čvrstom disku upotrijebljenog računala instaliran je operativni sustav „Microsoft Windows XP Professional, Service Pack 3“.

3.1.2. Baze podataka

3.1.2.1. Baza podataka GenBank

Baza podataka *GenBank* osnovana je u listopadu 1982. godine i održava je američki Nacionalni centar za biotehnološke informacije (engl. "National Center for Biotechnology Information", NCBI). Sadržava dobro opisane, tj. anotirane sekvene DNA. U trenutno aktualnom izdanju baze – 197. Izdanje, 15. kolovoza 2013 nalazi se 154 192 921 011 baza iz 167 295 840 analiziranih sekvena DNA. Cjeloviti se pregled trenutnog izdanja baze podataka *GenBank* nalazi na poslužitelju centra NCBI (NCBI, 2013). Baza podataka *GenBank* dio je međunarodne suradnje (engl. "International Nucleotide Sequence Database Collaboration") koja pruža mogućnost neprestane razmjene i dostupnosti sekvenca DNA s bazama podataka EMBL (engl. "European Molecular Biology Laboratory") i DDBJ (engl. "DNA Data Bank of Japan"). Baza se svakodnevno nadopunjava, a svaka dva mjeseca izlazi novo izdanje.

3.1.2.2. Baza podataka NRPS-PKS

Baza podataka *NRPS-PKS*, utemeljena pri indijskom Nacionalnom centru za imunologiju (engl. "National Institute of Immunology", NII), New Delhi, Indija, sadržava podatke o enzimima uključenim u biosintezu farmaceutski važnih prirodnih produkata, kao i računalne programe za njihovo analiziranje. Na internetskoj stranici baze podataka *NRPS-PKS* nudi se mogućnost jednostavne ekstrakcije različitih domena iz sekvence enzima te određivanja katalitičkih aktivnosti domena, aktivnih mjesta i specifičnosti za supstrat. Baza također omogućuje analizu

domena neopisanih NRPS/PKS genskih nakupina na temelju njihove sličnosti s podacima o enzimima koji se nalaze u bazi te određivanje njihove specifičnosti za supstrat i motiva aktivnog mesta. Rezultati analiziranih NRPS/PKS genskih nakupina organizirani su u četiri baze podataka za pretraživanje i određivanje organizacije domena i specifičnosti za supstrate neribosomalnih peptid sintetaza i tri tipa poliketid sintaza. Te su baze podataka nazvane *NRPSDB* (engl. "A Database of Non-Ribosomal Peptide Synthetases"), *PKSDB* (engl. "A Database of Modular Polyketide Synthases"), *ITERDB* (engl. "A Database of Type I Iterative PKS Gene Clusters") i *CHSDB* (engl. "A Database of Chalcone Synthases"), prema tipu prirodnih produkata i mehanizmu biosinteze. Baze su međusobno povezane kako bi se moglo provoditi analize genskih nakupina odgovornih za biosintezu poliketidno/peptidnih hibridnih produkata. Obrađivanjem velikog broja biosintetskih genskih nakupina dokazano je da, uz točno određivanje domena NRPS i PKS sustava, baza podataka *NRPS-PKS* može također predvidjeti i specifičnosti za supstrat adenilacijskih i acetiltransferaznih domena s relativno visokom točnošću. Ova svojstva čine bazu podataka *NRPS-PKS* vrijednim izvorom za identifikaciju prirodnih produkata koje sintetiziraju NRPS/PKS genske nakupine pronađenih u novosekveniranim genomima. Baza *NRPS-PKS* putem grafičkog sučelja omogućuje korisnicima povezivanje kemijskih struktura prirodnih spojeva s domenama i modulima u odgovarajućim neribosomalnim peptid sintetazama ili poliketid sintazama. Također nudi smjernice za zamjenu domena/modula kao i za eksperimente za mjesno specifičnu mutagenezu koji omogućuju promjenu postojećih genskih nakupina s ciljem dobivanja novih prirodnih produkata. Bazi podataka NRPS-PKS moguće je pristupiti preko internetske stranice <http://www.nii.res.in/nrps-pks.html> (Anonymous 1, 2013; Ansari i sur., 2004).

3.1.2.3. Baza podataka Pfam

Baza podataka Pfam je velika zbirka skrivenih Markovljevih modela proteinskih domena organiziranih u proteinske obitelji. Za svaku proteinsku obitelj unutar baze podataka Pfam moguće je pregledati: višestruko poravnanje, strukturu proteinskih domena, filogenetska stabla, poveznice na druge baze podataka i poznate strukture proteina. Jedan protein može pripadati u nekoliko Pfam obitelji, a 74% sekvenca proteina ima barem jednu sličnu proteinsku sekvencu unutar baze podataka Pfam. Taj se broj zove pokrivenost baze. Baza podataka Pfam-A je ručno izrađeni dio baze koji sadržava više od 9 000 unosa. Za svaki unos pohranjena su poravnajanja proteinskih sekvenca i skriveni Markovljevi modeli. Postoji još i baza podataka Pfam-B koja

sadržava veliki broj malih obitelji slabije kvalitete podataka. Ona se koristi u slučajevima kada nisu pronađene sekvene sa sličnošću unutar baze podataka Pfam-A (Finn i sur., 2008).

3.1.3. Bioinformatički računalni paketi i programi

3.1.3.1. Računalni program Glimmer

Računalni program *Glimmer* (engl. "Gene Locator and Interpolated Markov ModelER") služi za pronalaženje gena u DNA mikroorganizama, posebno u genomima bakterija, arheja i virusa. Razvijen je i prvotno upotrijebljen u instituciji "The Institute for Genomic Research" (TIGR) za anotaciju više od 100 različitih bakterijskih vrsta. Programom *Glimmer* može se konstruirati model za dijelove DNA koji kodiraju za proteine, koristeći druge otvorene okvire čitanja u učitanoj (analiziranoj) sekvenci DNA kao podatak za uvježbavanje (engl. "training data"). Manje je učinkovit prilikom analize kratkih sekvenca DNA, kao i u sekvenca DNA s visokim G+C sastavom baza. Takvu DNA imaju upravo genomi streptomiceta (Delcher i sur., 2007).

3.1.3.2. Računalni program GeneMark-PS

Računalni program *GeneMark* je razvijen 1993. godine kao prvi bioinformatički program za prepoznavanje gena. Učinkovit je i precizan alat za analizu genoma. Sadržava knjižnicu modela gena različitih vrsta bakterija. Prilikom pretraživanja željene sekvene DNA odabire se model koji je najbliži vrsti čija se DNA analizira (Besemer i Borodovsky, 2005).

3.1.3.3. Programska paket HMMER

Programski paket *HMMER* pripada skupini alata za pretrage na osnovi sličnosti sekvenca. Njegova je specifičnost da upotrijebjava profile proteina u obliku skrivenih Markovljevih modela (engl. "Hidden Markov Models", HMM) prilikom pretrage. Zbog toga ovaj programski paket dobro razlikuje ključne konzervirane aminokiseline (odgovorne za katalitičku aktivnost) od onih manje važnih. Pored toga, programski paket *HMMER* uzima u obzir insercije i delekcije te im pridodaje stupanj vjerojatnosti. Program *hmmpfam* (sastavni dio paketa *HMMER*) čita sekvencu po sekvencu upita i traži sličnost s nekim od proteinskih profila u bazi podataka. Profili

proteina koji se nalaze u bazi podataka koju program *hmmpfam* pretražuje organizirani su u obitelji proteina (npr. obitelj proteina globina). Za sve sekvene unutar jedne obitelji proteina smatramo da su nastale od zajedničkog pretka te pokazuju značajan stupanj sličnosti. Uspješnost određivanja sličnosti točnija je prilikom usporedbe dviju sekvenca proteina nego u usporedbi dviju sekvenca DNA (Eddy, 1998).

3.1.3.4. Računalni program *BLAST*

Računalni program *BLAST* (engl. „**B**asic **L**ocal **A**lignment **S**earch **T**ool“) je zapravo algoritam za usporedbu primarnih sekvenca DNA ili proteina. Pretraživanjem pomoću programa *BLAST* se mogu uspoređivati istraživane sekvene s bazom podataka sekvenci kako bi se u bazi podataka pronašle sekvene koje su slične istraživanoj sekvenci. Različite vrste računalnog programa *BLAST* se mogu koristiti s obzirom na ispitivanu sekvencu, npr. usporedba nukleotidne sekvene s nukleotidnom bazom podataka, proteinske sekvene s proteinskom bazom podataka, translatirane genske sekvene s proteinskom bazom podataka, proteinske sekvene s bazom podataka translatiranih nukleotida i translatirane nukleotidne sekvene s bazom podataka translatiranih nukleotida. Računalni program *BLAST* je dizajniran od strane Stephena Altschula, Warrena Gisha, Webba Millera, Eugenea Myersa i Davida J. Lipmana pri National Institutes of Health (NIH), Bethesda, Maryland, a objavljen je 1990. godine (Altschul i sur., 1990).

3.1.3.5. Programske pakete *ClustScan*

Računalni generički programski paket *ClustScan* je razvijen 2008. godine u Kabinetu za bioinformatiku Prehrambeno-biotehnološkog fakulteta, Sveučilišta u Zagrebu. Generički programski paket *ClustScan* se upotrijebjava za anotaciju modularnih genskih nakupina s naznanju utemeljenim predviđanjima aktivnosti i specifičnosti njihovih modula i katalitički aktivnih domena (Starčević i sur., 2008). Tim se programskim paketom mogu anotirati genske nakupine u novosekvenciranim genomima i metagenomima kao i predvidjeti kemijske strukture produkata biosinteze iz sekvenca DNA genskih nakupina. Računalni generički programski paket ClustScan se može se preuzeti sa Web stranice <http://bioserv.pbf.hr/cms/index.php?page=clustscan> (Thiotemplate Modular Systems Studies, 2013). Za detaljniji opis vidi potpoglavlje 2.3.3.

3.2. METODE

3.2.1. Prikupljanje sekvenca DNA

Kod prikupljanja primarnih sekvenca DNA vrste *Streptomyces sp. C* pristupljeno je mrežnim stranicama „Broad Institute“-a, otkuda su se preuzele primarne sekvence genoma navedene vrste te svi statistički podaci vezani uz sam genom (Slika 14) i njegovo sastavljanje iz podataka dobivenih sekvenciranjem (engl. „assembly“) (Slika 15).

Potrebno je naglasiti kako sastavljanje genoma *Streptomyces sp. C* nije dovršeno te se zbog trenutne nesavršenosti tehnologije sekvenciranja i metoda za sastavljene očitane dijelove DNA u genom još uvijek nalazi u 4 dijela.

Genome Statistics

Summary

	Size	Chrs	%GC	Genes	tRNAs	rRNAs
<i>S. sp. Mg1</i>	7.64 Mb	N/A	71.60	7,647	65	5
<i>S. sp. SPB74</i>	6.97 Mb	N/A	73.00	5,879	60	6
<i>S. clavuligerus ATCC 27064</i>	9.07 Mb	N/A	71.94	7,645	60	7
<i>S. pristinaespiralis ATCC 25486</i>	8.13 Mb	N/A	71.16	7,009	64	4
<i>S. sviceus ATCC 29083</i>	9.31 Mb	N/A	70.46	8,394	65	5
<i>S. ghanaensis ATCC 14672</i>	8.51 Mb	N/A	72.22	7,891	65	4
<i>S. roseosporus NRRL 15998</i>	7.82 Mb	N/A	71.36	6,925	65	8
<i>S. albus J1074</i>	6.82 Mb	N/A	73.15	5,902	66	7
<i>S. sp. AA4</i>	9.18 Mb	N/A	69.78	8,419	56	4
<i>S. sp. C</i>	8.46 Mb	N/A	72.61	7,540	71	7
<i>S. sp. SPB78</i>	7.56 Mb	N/A	72.90	6,387	63	5
<i>S. griseoflavus Tu4000</i>	8.05 Mb	N/A	71.67	6,338	62	6
<i>S. hygroscopicus ATCC 53653</i>	11.03 Mb	N/A	71.22	9,196	62	4
<i>S. lividans TK24</i>	8.32 Mb	N/A	72.18	7,551	63	6
<i>S. viridochromogenes DSM 40736</i>	8.65 Mb	N/A	71.12	7,714	65	6
<i>S. sp. E14</i>	7.93 Mb	--	72.67	6,198	68	7
<i>M. carbonacea ATCC 39149</i>	6.82 Mb	--	72.33	5,633	51	4
<i>Micromonospora sp. M42</i>	6.78 Mb	--	72.60	6,093	54	9
<i>Kutzneria sp. 744</i>	11.65 Mb	--	69.81	10,200	54	9
<i>S. roseosporus NRRL 11379 (v4)</i>	7.85 Mb	--	71.42	6,931	66	8

Table header definitions

- ▶ **Size** length of complete genome sequence, calculated by adding lengths of all scaffolds together
- ▶ **Chrs** number of chromosomes
- ▶ **%GC** GC content of scaffolds
- ▶ **Genes** number of predicted protein-coding genes in genome
- ▶ **tRNAs** number of predicted tRNA genes in genome
- ▶ **rRNAs** number of predicted rRNA genes in genome

Slika 14. Statistički podaci genoma vrste *Streptomyces sp. C* (Broad Institute, 2013).

Assembly Statistics

Assembly Stats

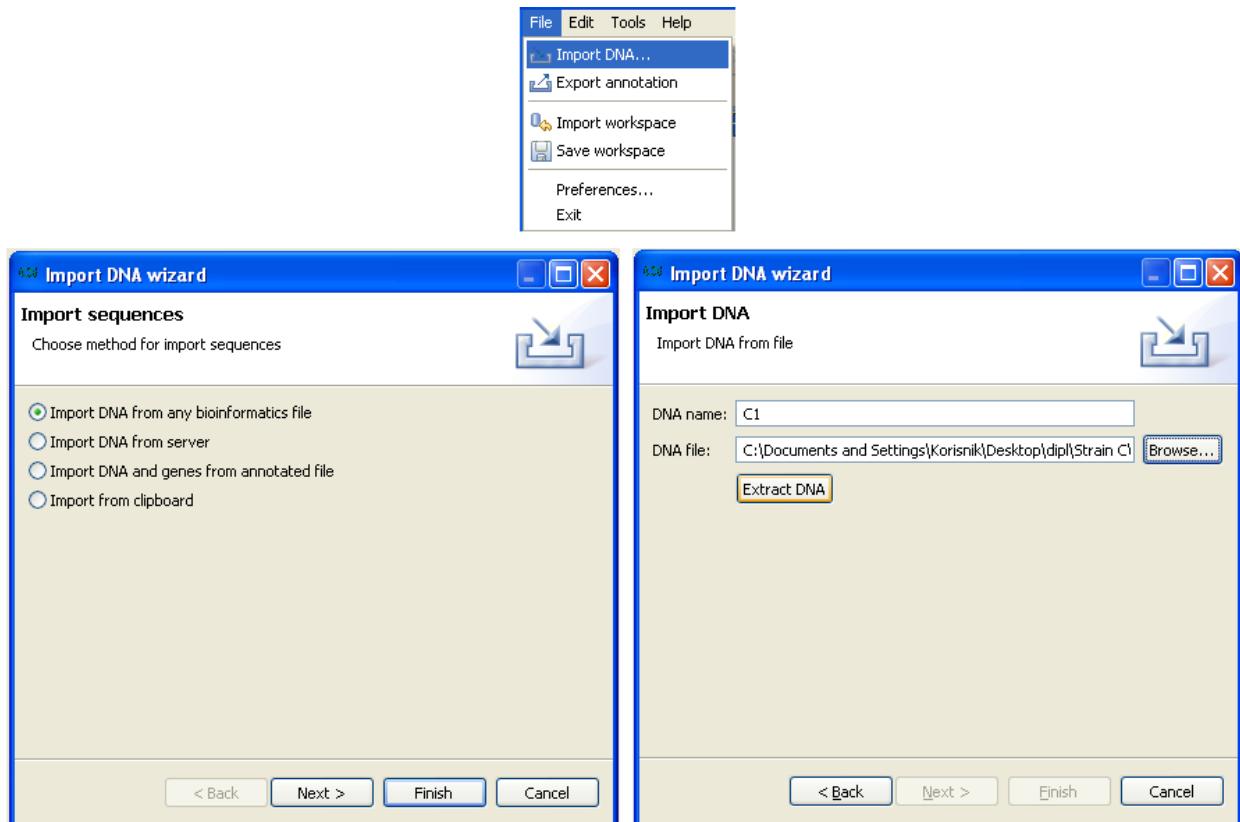
	Coverage	Assembly Size	Total		Scaffold N50	Contigs	Contig N50	
			Contig Length	Scaffolds			%Q40	
S. sp. Mg1	9.3X (8.9X Q>20)	7.64 Mb	7.4 Mb	87	149.65 Kb	590	20.55 Kb	98.90
S. sp. SPB74	12.3X (11.65X Q>20)	6.97 Mb	6.51 Mb	2	6.81 Mb	845	13.13 Kb	98.47
S. clavuligerus ATCC 27064	11.6X (9.2X Q>20)	9.07 Mb	8.36 Mb	10	6.76 Mb	1,036	15.17 Kb	98.21
S. pristinaespiralis ATCC 25486	9.54X (8.96X Q>20)	8.13 Mb	7.63 Mb	1	8.13 Mb	844	17.56 Kb	98.45
S. sviceus ATCC 29083	8.37X (7.9X Q>20)	9.31 Mb	9.06 Mb	1	9.31 Mb	552	33.52 Kb	98.84
S. ghanaensis ATCC 14672	11.6X (10.8X Q>20)	8.51 Mb	8.22 Mb	3	8.27 Mb	616	24.17 Kb	99.06
S. roseosporus NRRL 15998	11.9X (11.4X Q>20)	7.82 Mb	7.56 Mb	1	7.82 Mb	371	38.58 Kb	99.17
S. albus J1074	14.3X (13.5X Q>20)	6.82 Mb	6.62 Mb	2	6.81 Mb	501	23.68 Kb	99.01
S. sp. AA4	10.9X (10.4X Q>20)	9.18 Mb	9.15 Mb	4	8.48 Mb	98	218.5 Kb	99.73
S. sp. C	8.17X (7.64X Q>20)	8.46 Mb	7.92 Mb	4	7.45 Mb	652	25.61 Kb	97.68

Slika 15. Statistički podaci vezi uz sastavljanje genoma vrste *Streptomyces* sp. C iz podataka dobivenih sekvenciranjem (Broad Institute, 2013).

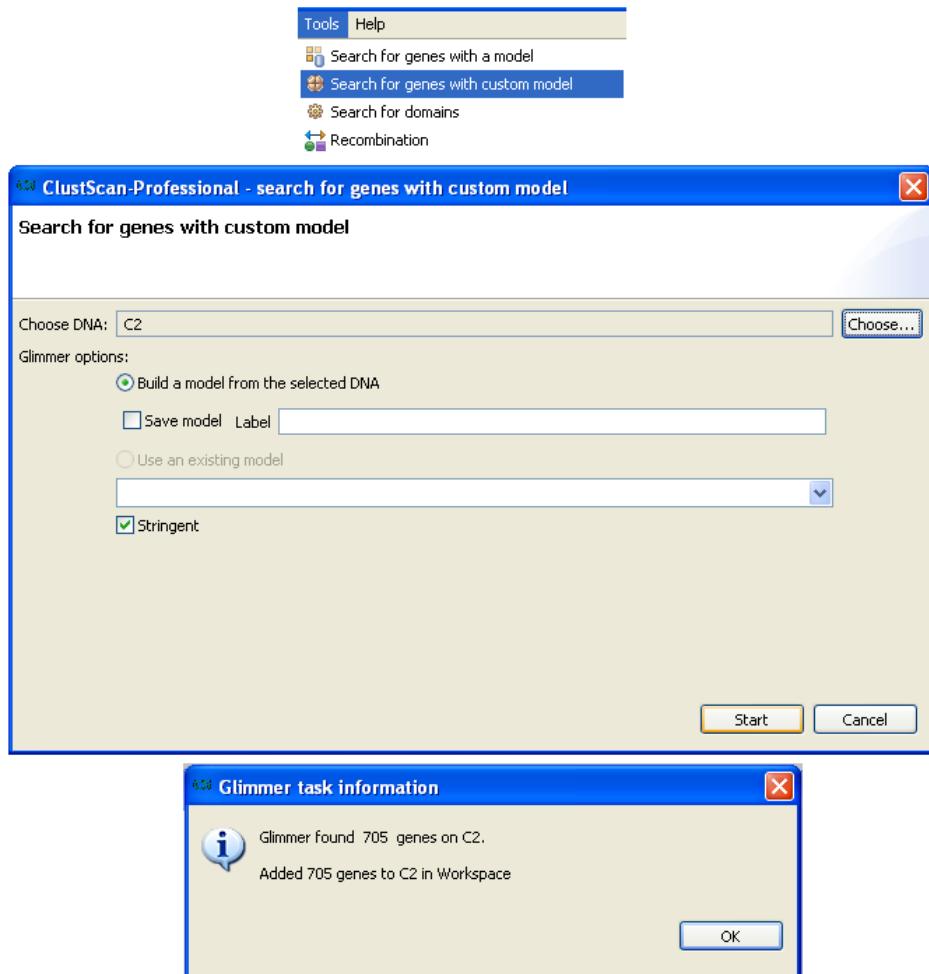
3.2.2. Pronalaženje strukturnih gena

Nedovršena sekvenca vrste *Streptomyces* sp. C je preuzeta s mrežnih stranica Broad Instituta

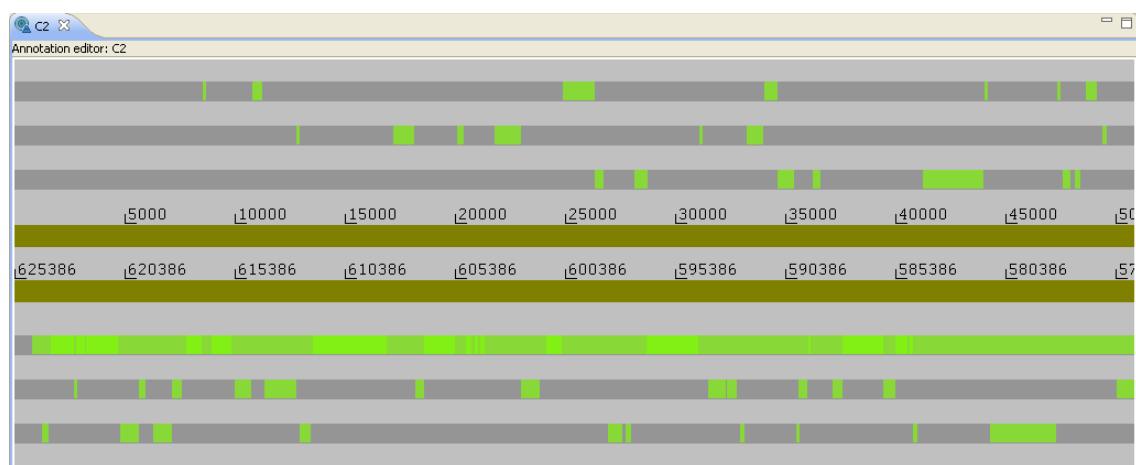
(Broad Institute, 2013). Za pronalaženje gena korišteni su alati dostupni u programskom paketu *ClustScan* (Thiotemplate Modular Systems Studies, 2013). Sama analiza sekvence DNA odvija se na Linux serveru (svaki korisnik ima svoju lozinku (engl. "password")). Lozinka korisniku omogućuje: pristup vlastitoj radnoj površini (engl. "workspace"), učitavanje sekvence DNA i provođenje analize, dok se sami rezultati pohranjuju na serveru i ili na čvrstom disku te je stoga je svaku analizu dovoljno provesti samo jednom (to je važno jer pretraživanje jednog genoma može trajati i nekoliko sati). Na početku analize potrebno je učitati DNA genoma vrste *Streptomyces sp.* C u programske paket *ClustScan* odabirom opcije "File/Import DNA" (Slika 16), pri čemu se DNA automatski prevodi u šest otvorenih okvira čitanja pomoću programa *Transeq* (Rice i sur., 2000). Genom se zatim obrađuje bioinformatičkim programima *Glimmer* ili *GeneMark*, kako bi se identificirali kodirajući dijelovi DNA u genomu. U slučaju ovog diplomskog rada korišten je program *Glimmer* odabirom opcije "Tools/Search for genes with custom model" i odabirom naše DNA sekvence (Slika 17). Na Slici 18 je prikazan rezultat obrade ispitivane DNA sekvence računalnim programom *Glimmer*.



Slika 16. Redoslijed operacija za učitavanje DNA sekvence koja će se analizirati.



Slika 17. Redoslijed operacija za pronaalaenje gena pomoću bioinformatičkog programa *Glimmer*.



Slika 18. Prikaz rezultata dobivenih nakon obrade ispitivane DNA sekvene programom *Glimmer*. Vidljiv je prikaz gena, u obliku zelenih pravokutnika, na svih 6 okvira čitanja.

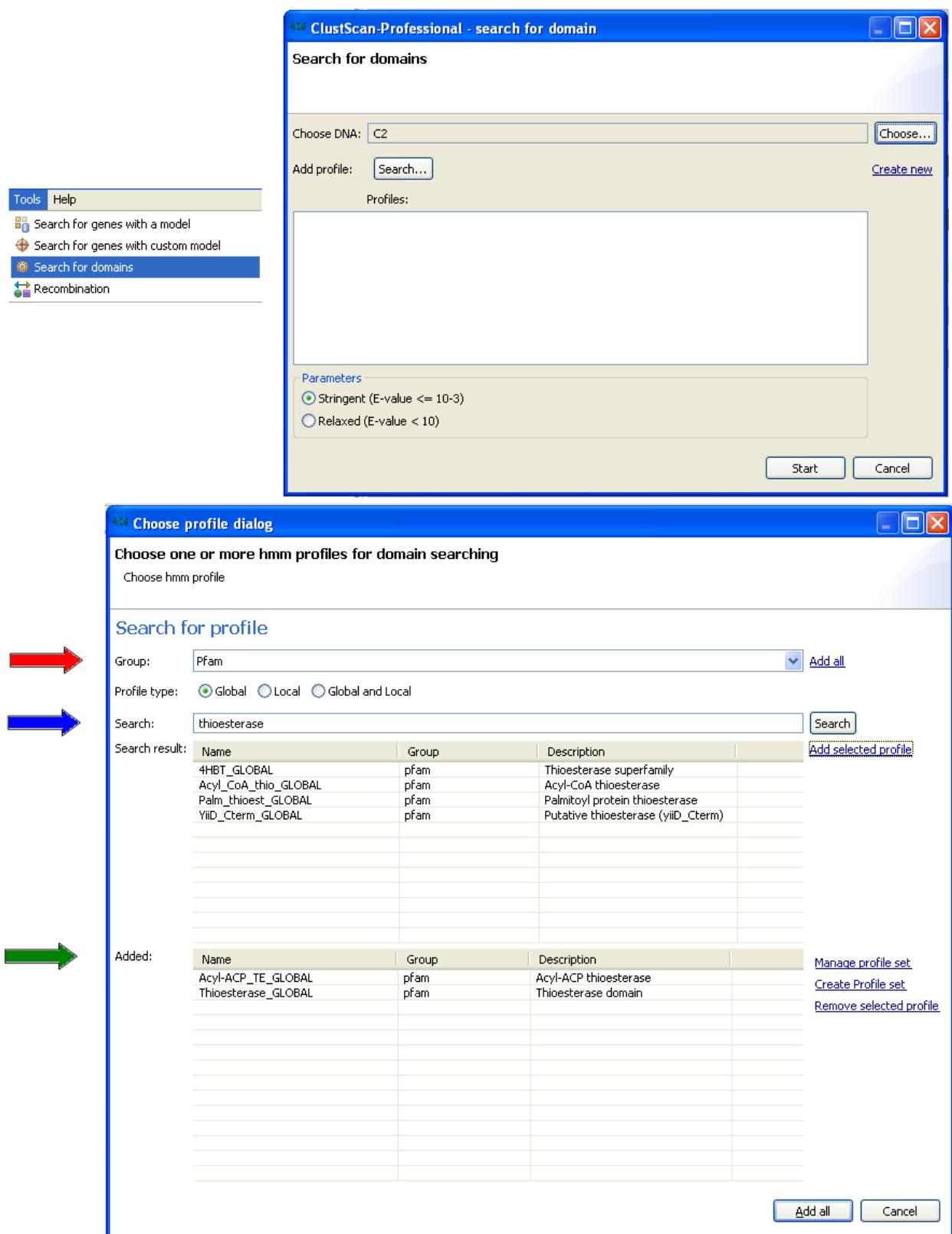
3.2.3. Analiza sekvenca proteina programskim paketom *HMMER*

Programskim paketom *HMMER* se nakon toga na ispitivanoj DNA sekvenci traže geni za određene proteine koristeći se HMM profilima za već postojeće obitelji proteina unutar Pfam baze podataka ili bazom podataka prilagođenom od strane korisnika.

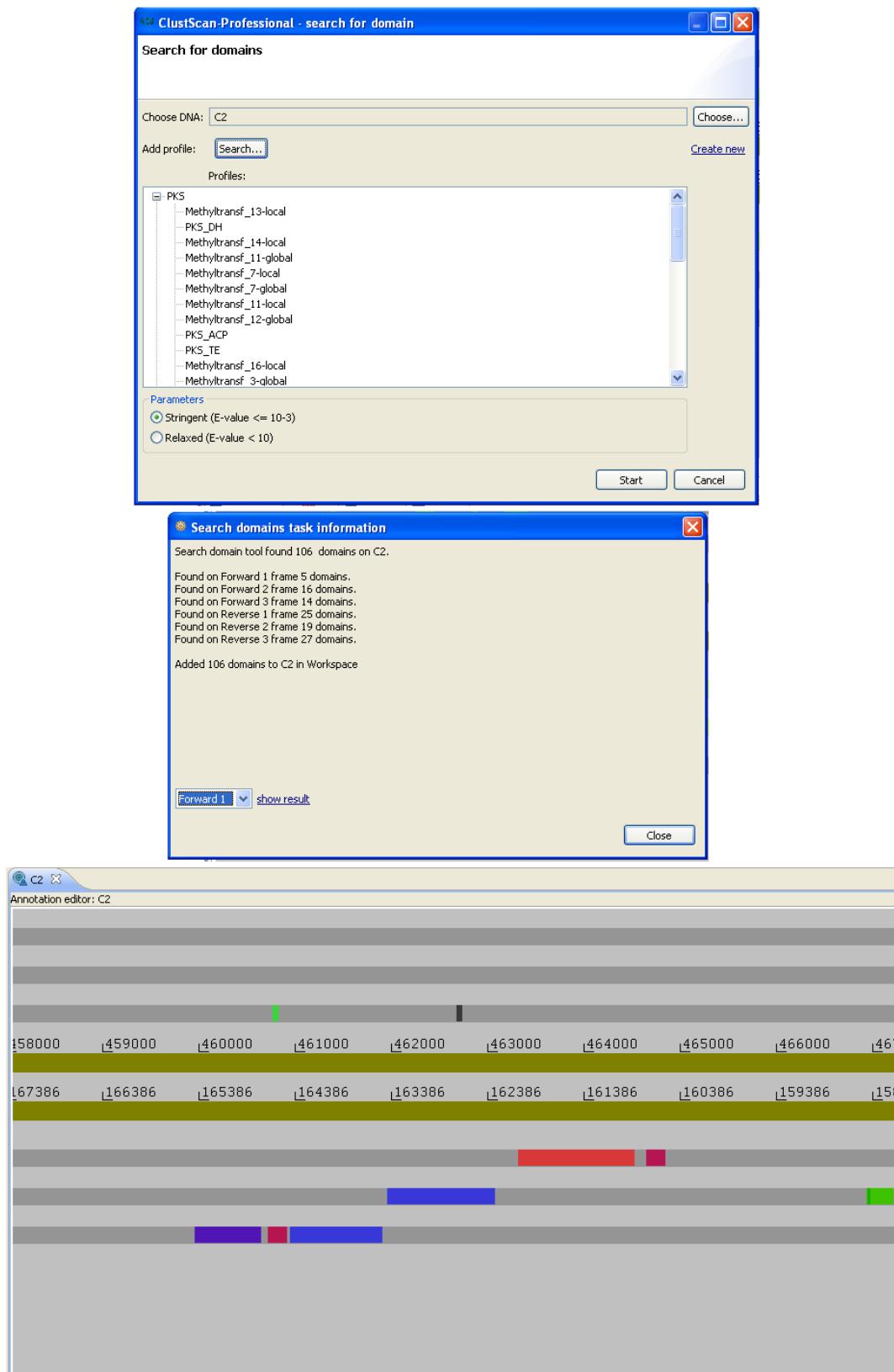
3.2.4. Preuzimanje gotovih profila proteina iz baze podataka Pfam

Baza profila proteina za pretraživanje genoma vrste *Streptomyces sp.* C izrađena je upotrebom baze podataka Pfam pomoću programskog paketa *ClustScan* odabirom opcije "Tools/Search for domains". Na učitanom prozoru izabrana je DNA i opcija "Add profile: Search" (Slika 19).

Za pretraživanje je odabrana baza podataka Pfam (globalni profili proteina, Slika 19, označeno crveno), a pod "Search" su upisana imena traženih domena (Slika 19, označeno plavo). Nakon pretraživanja baze podataka Pfam rezultati upita prikazani su pod "Search result". Ako se prikazana domena prihvata, označava se klikom miša i odabire se opcija "Add selected profile" te se odgovarajući profil automatski prikazuje pod izbornikom "Added" (Slika 19, označeno zeleno). Nakon učitavanja svih domena, odabrana je opcija kreiranja kompleta profila (engl. "create profile set"). Komplet profila zatim je učitan odabirom opcije "Manage profile set". Podešeni su statistički parametri programskog paketa *HMMER* (opcija "stringent") i pokrenuto je pretraživanje. Podešeni parametri i rezultati pretraživanja su prikazani na Slici 20.



Slika 19. Prozor programskog paketa *ClustScan* za odabir domena i izradu proteinskih profila.



Slika 20. Prikaz podešenih parametara i rezultata pretraživanja pomoću programskog paketa *HMMER* korištenjem Pfam baze podataka

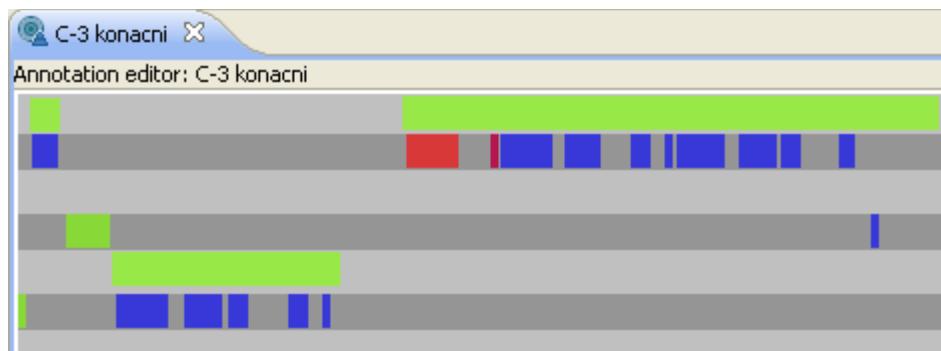
3.2.5. Izrada vlastitih profila proteina pomoću programa *ClustScan*

Odabirom opcije "Tools/Search for domains/Create new/I want to create an alignment" moguće je izraditi bazu profila proteina i bez upotrebe baze podataka Pfam. Kako bi se to napravilo, najprije je potrebno prikupiti po nekoliko sekvenci DNA za svaku od domena iz tablica profila proteina koristeći se bazom podataka *NRPS-PKS* te ih spojiti u jednu datoteku. Tako je za svaku vrstu domena izrađena zasebna datoteka koja sadržava nekoliko primjeraka sličnih sekvenci. Sekvence koje čine jednu datoteku zlijepljene su u za to predviđeni prozor i odabrana je opcija provjere zapisa (sekvence moraju biti u točnom zapisu FASTA). Temeljni zadatak pri izradi vlastitih profila proteina HMM je odabir opcije "Start Align" čime se pokreće bioinformatički program za višestruko poravnavanje sekvenca DNA ili proteina, *ClustalW* (Thompson i sur., 1994). Nakon poravnjanja program sam izrađuje profil proteina i daje mogućnost njegovog imenovanja.

3.2.6. Anotacija genoma bakterije *Streptomyces sp.* C programskim paketom *ClustScan*

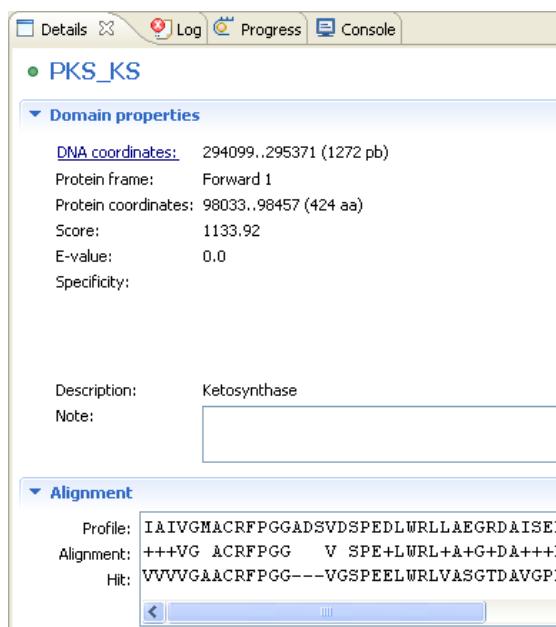
Nakon što su učitani profili proteina i podešeni parametri programskog paketa *HMMER* pokrenuto je pretraživanje genoma. Programska paket *ClustScan* prikazuje rezultate u obliku:

- liste ili stabla, tj. prozor radne površine (engl. "Workspace window") prikazuje sve pronađene proteine, strukturirane u tri okvira čitanja s lijeva na desno (engl. "forward") i tri okvira čitanja s desna na lijevo (engl. "reverse"). Za svaki otvoreni okvir čitanja prikazan je ukupan broj pronađenih proteinskih domena;
- grafičkom obliku, tj. prozor za uređivanje rezultata anotacije (engl. "Annotation editor") prikazuje sve pronađene gene (Slika 21, zeleni pravokutnici) i proteinske domene (Slika 21, plavi pravokutnici predstavljaju domene PKS, a crveni NRPS).



Slika 21. Prozor programskog paketa *ClustScan* za uređivanje rezultata anotacije

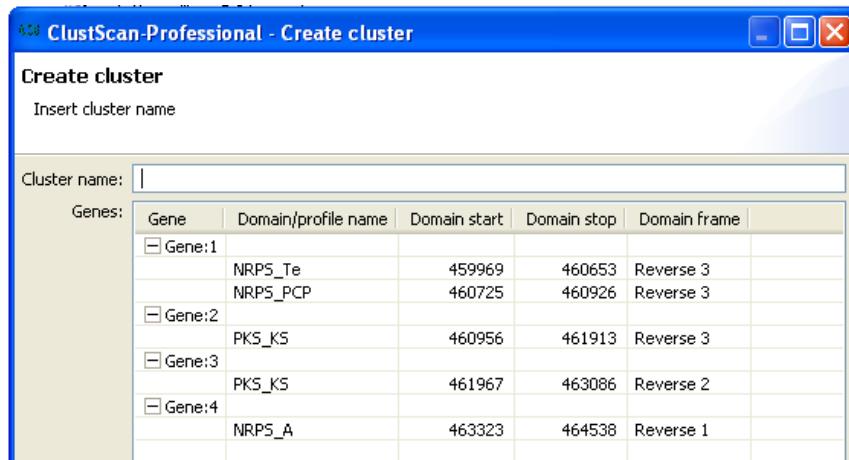
Detaljne informacije (poput koordinata proteinske domene u DNA i proteinu, vrijednosti parametara dobivenih programom *HMMER*, kao i samo poravnanje sekvenca) o svakoj domeni možemo saznati klikom lijeve tipke miša na domenu, čime pokrećemo učitavanje prozora detalja (Slika 22).



Slika 22. Prozor programa *ClustScan* s detaljnim informacijama o svojstvima pojedine domene

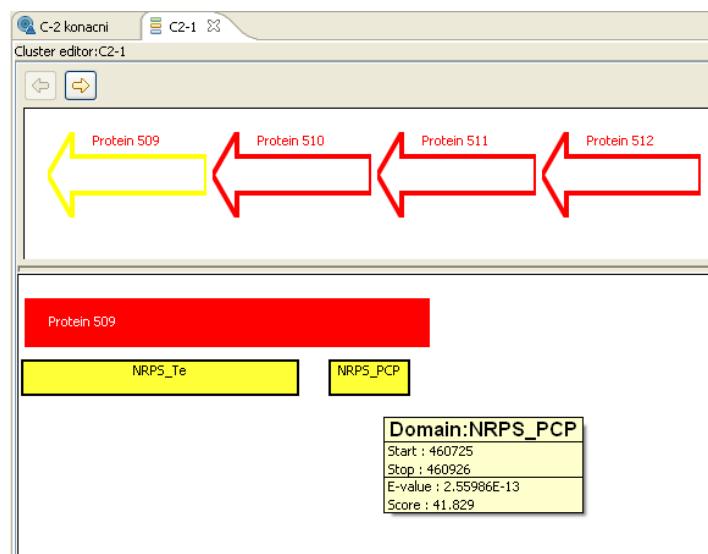
Sve domene čiji parametri programskog paketa *HMMER* ne zadovoljavaju postavljane uvjete (vrijednost E, uspjeh pogotka) se obrišu, a zatim se definiraju genske nakupine. Nakon toga se usporede proteinske sekvene pronađenih proteina s bazom podataka sekvenca pomoću programa *BLAST* te se na temelju toga odredi pripadaju li geni istoj genskoj nakupini. Postupak

definiranja genske nakupine je vrlo jednostavan. Najprije se lijevom tipkom miša označi prvi gen, a zatim pritisne i drži tipku "Ctrl" i označi zadnji gen u genskoj nakupini. Klikom desne tipke miša na zadnji gen pojavljuje se opcija "Create cluster". U prozoru (Slika 23) su ispisani svi geni obuhvaćeni u gensku nakupinu, proteinske domene unutar njih te njihov položaj unutar sekvene DNA. Kreiranje genske nakupine za vršava njezinim imenovanje (upisom imena u polje „Cluster name“).



Slika 23. Prozor programskog paketa *ClustScan* s detaljnim prikazom gena koji čine jednu gensku nakupinu.

Nakon kreiranja genske nakupine učitava se prozor za uređivanje genskih nakupina "Cluster editor" (Slika 24).



Slika 24. *ClustScan* prozor za uređivanje genskih nakupina.

Svaki gen (Slika 24) je prikazan crvenom strelicom koja ujedno i pokazuje njegov smjer čitanja, odabirom pojedinog gena učitava se dodatni prozor koji prikazuje domene unutar označenog gena. Nakon završene anotacije genoma bakterije *Streptomyces sp. C*, "Workspace" s pronađenim genskim nakupinama sačuvan je na čvrstom disku (vidi: potpoglavlje 8.2.1.).

4. REZULTATI

4.1. REZULTATI ANOTACIJE GENOMA BAKTERIJE *Streptomyces sp. C* PROGRAMSKIM PAKETOM *ClustScan*

Analizom nedovršenog genoma bakterije *Streptomyces sp. C* bioinformatičkim alatom za predviđanje gena *Glimmer* pronađeno je sveukupno 994 gena. Korištenjem gotovih profila PKS i NRPS domena dobivenih pretragom baze podataka Pfam, koristeći program *hmmpfam* sa sljedećim kriterijima pretrage: vrijednost $E < 10^{-5}$ i uspjeh pogotka (engl. „score“) > 0 , izdvojena su 166 proteina kao potencijalni kandidati za promatrane sustave PKS, NRPS i njihove hibride (Tablica 4).

Tablica 4. Statistika anotacije genoma *Streptomyces sp. C*

	C-1	C-2	C-3	C-4	Ukupno
Broj gena	15	704	248	27	994
Broj potencijalnih protein-a PKS ili NRPS sustava					
Otvoreni okvir čitanja s lijeva na desno 1	0	1	50	0	51
Otvoreni okvir čitanja s lijeva na desno 2	0	2	39	0	41
Otvoreni okvir čitanja s lijeva na desno 3	0	2	32	1	35
Otvoreni okvir čitanja s desna na lijevo 1	0	9	2	0	11
Otvoreni okvir čitanja s desna na lijevo 2	0	5	11	0	16
Otvoreni okvir čitanja s desna na lijevo 3	0	8	4	0	12

Na osnovi usporedbe *BLAST* računalnim programom, sveukupno je definirano 7 sustava, a od toga su:

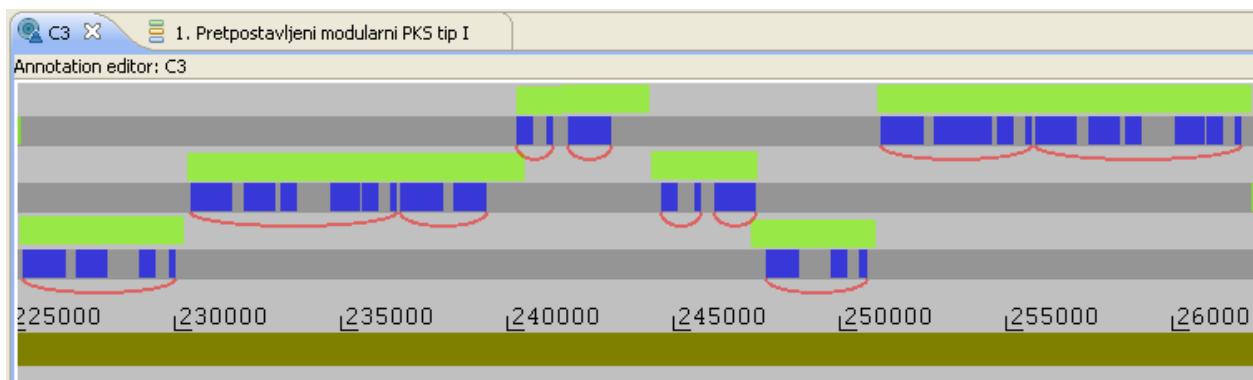
- 2 PKS sustava,
- 2 NRPS sustava i
- 3 hibridna sustava.

Detaljan prikaz se nalazi u Tablici 5.

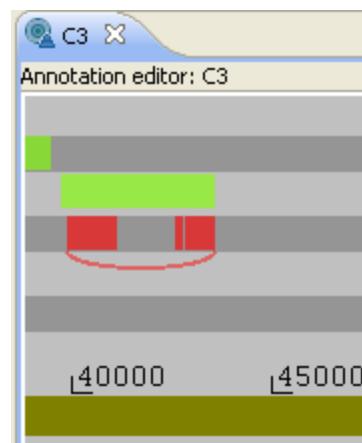
Tablica 5. Raspodjela genskih nakupina sekundarnih metabolita u genomu bakterije *Streptomyces sp. C.*

Genska nakupina	Odsječak genoma	Početak (bp)	Kraj (bp)	Broj gena	Dužina genske nakupine (bp)
Poliketid sintaze					
1. Prepostavljeni modularni tip I	C-3	192157	270739	13	78582
2. Prepostavljeni iterativni tip I	C-3	66345	71660	4	5315
NRPS sintetaze					
3. Prepostavljeni tip B	C-3	39797	43648	1	3851
4. Prepostavljeni tip B	C-3	83070	85225	2	2155
Hibridni sustavi					
5. Prepostavljeni PKS tip II/NRPS	C-2	459960	464991	4	5031
6. Prepostavljeni modularni PKS tip I/NRPS	C-3	284382	345795	17	61413
7. Prepostavljeni modularni PKS tip I/NRPS	C-3	146688	153193	3	6505

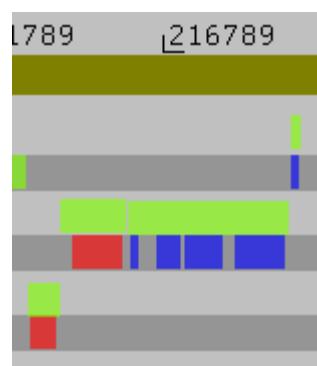
U nastavku je prikazan po jedan primjer od svake skupine anotiranih sustava (slike 25, 26 i 27), dok se potpuni rezultati anotacije genoma bakterije *Streptomyces sp. C* pomoću programskog paketa *ClustScan* nalaze u prilozima (vidi potpoglavlje 8.2.1). Na slikama 25, 26 i 27 zeleni pravokutnici predstavljaju gene unutar kojih su smještene domene proteina označene plavom (PKS) ili crvenom (NRPS) bojom, a crvenim su linijama označeni moduli. U shematskom prikazu (Tablica 6) su navedeni geni koji sadržavaju module i domene (domene unutar istog modula su označene istom bojom, a moduli koji se nalaze na jednom genu se nalaze u istom retku pod zajedničkim brojem gena), dok „prazni“ geni, tj. geni koji ne sadržavaju uputu za neku od PKS ili NRPS domena, a koji su također u nekim slučajevima sastavni dio genskih nakupina, nisu posebno navedeni. Takvi „prazni“ geni su vjerojatno uključeni u postsintetske modifikacije ili sadržavaju neke nove i do sada nepoznate funkcionalnosti. Moduli označeni zvjezdicom (*) su dodatno pojašnjeni u poglavlju 5.1., kao i domene označene žutom bojom. Kratice oznaka domena su objašnjene u poglavljima 2.1.1. i 2.1.2.



Slika 25. Prikaz dijela anotirane genske nakupine br. 1. Prepostavljeni modularni PKS tip I anotiran programskim paketom *ClustScan*.



Slika 26. Prikaz anotirane genske nakupine br. 3. Prepostavljeni NRPS tip B anotiran programskim paketom *ClustScan*.



Slika 27. Prikaz anotirane genske nakupine br. 7. Prepostavljeni modularni PKS tip I/NRPS hibrid anotiran programskim paketom *ClustScan*.

Tablica 6. Shematski prikaz pronađenih genskih nakupina.

Genska nakupina br. 1: prepostavljeni modularni PKS tip I

Gen 186: -AT-ACP--KS-AT-DH-KR-ACP--KS-AT-DH-ER-KR-ACP-
Gen 187: -KS-AT-ACP--KS-AT-*
Gen 188: -KS-AT-KR-ACP-
Gen 189: -KS-AT-DH-ER-KR-ACP--KS-AT-*
Gen 190: -KR-ACP-*
Gen 191: -KS-*
Gen 192: -DH-ACP-* -KS-*
Gen 193: -AT-KR-ACP-*
Gen 195: -KS-AT-(DH)-KR-ACP--KS-AT-DH-ER-KR-ACP-
Gen 196: -KS-AT-KR-ACP--KS-*
Gen 198: -TE-

Genska nakupina br. 2: prepostavljen iterativni PKS tip II

Gen 76: -KS-*
Gen 75: -AT-*
Gen 74: -TE-*

Genska nakupina br. 3: prepostavljeni NRPS tip B

Gen 52: -A-PCP-Te-

Genska nakupina br. 4: prepostavljeni NRPS tip B

Gen 88: -A-
Gen 89: -Te-

Genska nakupina br. 5: prepostavljeni PKS tip II/NRPS hibrid

Gen 512: -PCP-A-
Gen 511: -KS-
Gen 510: -KS-
Gen 509: -ACP-TE-

Nastavak Tablice 6. Shematski prikaz pronađenih genskih nakupina.

Genska nakupina br. 6: prepostavljeni modularni PKS tip I/NRPS hibrid

Gen 210: -KS-AT-DH-KR-ACP-

Gen 212: -A-PCP--KS-AT-KR-ACP--KS-AT-DH-KR-*

Gen 217: -KS-AT-*

Gen 218: -KR-ACP-* -KS-AT-ACP--KS-AT-KR-ACP-

Gen 219: -KS-AT-KR-ACP-

Gen 220: -KS-*

Gen 221: AT-KR-ACP-*

Gen 222: -KS-AT-DH-*

Gen 223: -KR-*

Gen 224: -ACP-* -KS-*

Gen 225: -AT-*

Gen 226: -KR-ACP-* -TE-

Genska nakupina br. 7: prepostavljeni modularni PKS tip I/NRPS hibrid

Gen 154: -KS-AT-DH-ACP-

Gen 152: -A-

Gen 153: -Te-

5. RASPRAVA

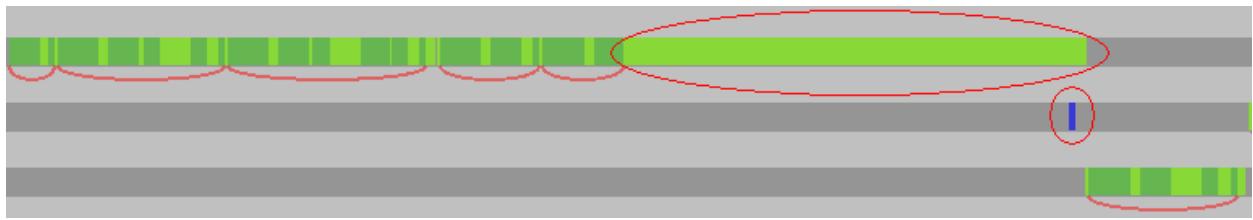
5.1. ANOTACIJA GENOMA BAKTERIJE *Streptomyces* sp. C PROGRAMSKIM PAKETOM *ClustScan*

Prilikom analize rezultata uočeni su slučajevi u kojima su domena i gen bili u različitim okvirima čitanja. Uzrok toj grešci je najvjerojatnije krivo prepoznavanje otvorenog okvira čitanja od strane programa korištenog za prepoznavanje kodirajućih regija DNA. Budući da programski paket *ClustScan* omogućuje prilagođavanje koordinata, u nekoliko su slučajeva geni pomaknuti iz jednog okvira čitanja u drugi kako bi se domene smjestile unutar gena. Ovaj je postupak korišten zbog nemogućnosti programskog paketa *ClustScan* da domenu koja nije unutar granica gena obuhvati u gensku nakupinu.

Bilo je i slučajeva da se neke domene, koje se djelomično nalaze unutar gena, ne prikazuju kao njihovi sastavni dijelovi, odnosno ne prikazuju se u prozoru za uređivanje genskih nakupina. Do toga dolazi jer programski paket *ClustScan* zahtijeva da se domena cijelom svojom duljinom nalazi unutar granica gena da bi se mogla smatrati cijelom genske nakupine, odnosno kako bi se prikazala u prostoru za uređivanje genskih nakupina. Taj je problem riješen prilagođavanjem koordinata gena na način da nove koordinate obuhvaćaju domenu u njenoj cijelosti. Također, u dva se slučaja domena ACP nalazila izvan gena i u različitom okviru čitanja od ostatka gena. Pretpostavka je da je do takvog pomaka okvira čitanja došlo zbog greške u sekvenciranju genoma jer se u oba slučaja ta domena nalazi u blizini dijela genoma koji se nije mogao kvalitetno sekvencirati.

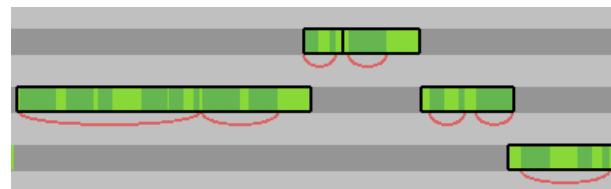
1. Genska nakupina br. 1: prepostavljeni modularni PKS tip I

- U genskoj nakupini br. 1 je prvi takav primjer gen 187 (Slika 28), gdje se ACP domena nalazi u različitim okviru čitanja od ostatka gena, ispred čega je veliki dio DNA koji nije mogao biti kvalitetno sekvenciran (veliki dio DNA u kojem se nalaze „N“ nukleotidi), vjerojatno zbog velikog udjela G+C, tako da postoji i mogućnost postojanja i još dodatnih domena u tom području.



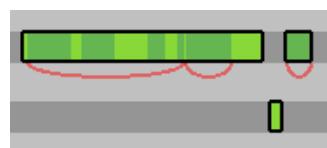
Slika 28. Prikaz gena 187 i ACP domene koja se nalazi u drugom okviru čitanja.

- Također, u genskoj nakupini br. 1 se nalazi nekoliko gena koji nisu organizirani na uobičajeni način (geni 189-193), Slika 29. U prikazanom slučaju modul je podijeljen između dva okvira čitanja, a redoslijed domena u potpunosti odgovara uobičajenom redoslijedu domena PKS sustava. Grešku je najvjerojatnije izazvala insercija/delecija jednog nukleotida zbog greške u sekvenciranju ili sastavljanju genoma.



Slika 29. Prikaz gena 189-193 genske nakupine 1.

- U genu 195 genske nakupine 1 dolazi i do preklapanja domena AT (vjerojatnost pogotka 507,682 i $E=1,48764 \times 10^{-153}$) i DH (uspjeh pogotka 0.162998 i $E=3.41642 \times 10^{-9}$), koja je također i inaktivna te je uklonjena.
- Gen 196 završava KS domenom, nakon čega slijedi dio gena u kojem nije prepoznata ni jedna domena. Nakon gena 196 slijedi „prazan“ gen – gen bez i jedne PKS/NRPS domene te ga slijedi gen s domenom ACP. Prepostavlja se da je ovdje došlo do delecije AT domene (Slika 30).



Slika 30. Prikaz gena 196-198 genske nakupine 1.

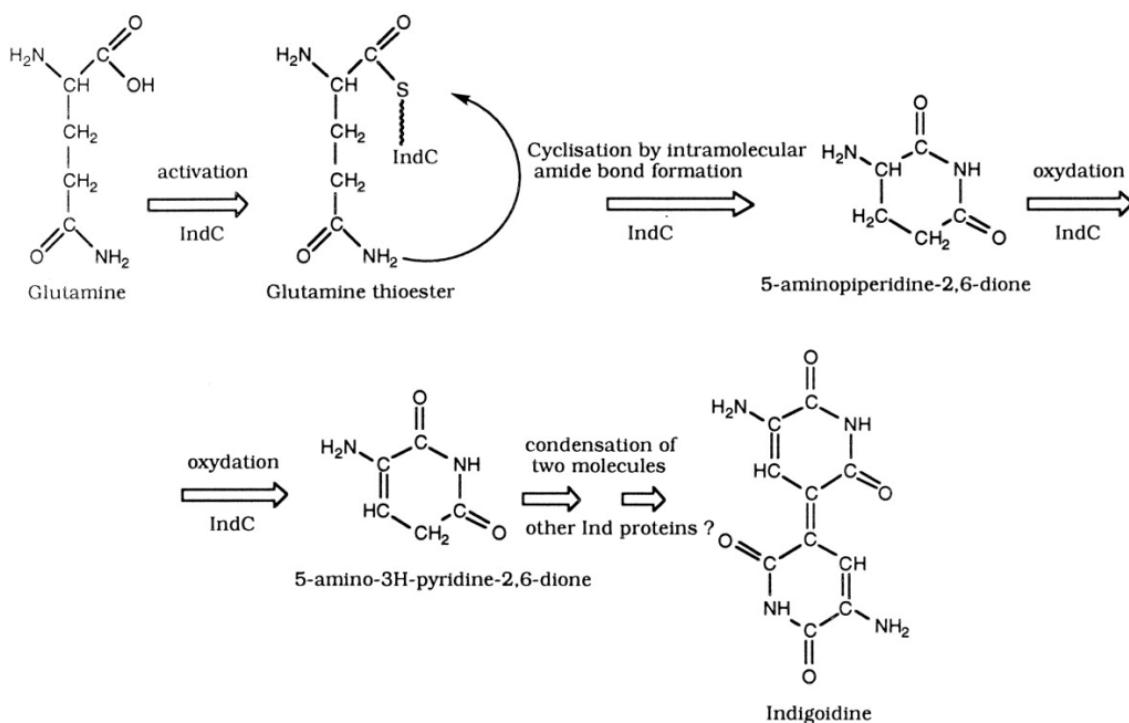
- Za gensku nakupinu 1 se pretpostavlja da pripada modularnom PKS sustavu tipa I. Gen 187 sliči genu koji kodira za protein koji sudjeluje u sintezi rifamicina, a geni 186 i 193 genima vrste *Streptomyces lasaliensis* koji kodiraju za proteine koji sudjeluju u sintezi lasalocida. Gen 193 djelomično nalikuje i genu vrste *Streptomyces himastatinicus* koji kodira za protein koji sudjeluje u sintezi oleandomicina. Sva tri navedena spoja imaju antibiotska svojstva.

2. Genska nakupina br. 2: pretpostavljeni iterativni PKS tip II

- Genska nakupina br 2 ne nalikuje ni na jedan do sada anotirani protein te je moguće da je ova genska nakupina lažno pozitivni rezultat. U prilog toj hipotezi ide i organizacija same genske nakupine – tri domene koje se sve nalaze na različitim genima, što ne odgovara ni jednom dosad opisanom PKS sustavu.

3. Genska nakupina br. 3: pretpostavljeni NRPS tip B

- Ovaj NRPS sustav nalikuje sintetazi plavog pigmenta indigoidina vrste *Erwinia chrysanthemi*, biljnog patogena. Sama sintetaza indigoidina nije NRPS sustav, ali s njom dijeli mnoge sličnosti. Navedena sintetaza se sastoji od 3 enzima, IndA, IndB i IndC. Uloga enzima IndA je još uvijek nepoznata, dok IndB pokazuje sličnost s fosfatazama koje su uključene u sintezi spojeva s antibiotskim svojstvima, a IndC pokazuje značajnu sličnost s mnogim NRPS sustavima. IndC enzim sadržava adenilacijsku domenu sa sekvencom DAWCFG LI za prepoznavanje glutamina i oksidacijsku sekvencu sličnu onima nađenima u različitim NRPS sustavima odgovornima za sintezu tiazolnih spojeva. Indigoidin se sintetizira iz dvije molekule glutamina (slika 31), međutim točan mehanizam biosinteze još nije otkriven. Također, otkriveno je da je ovaj spoj odgovoran za patogenost vrste *E. chrysanthemi* i da povećava njezinu otpornost na oksidativni stres prouzročen reaktivnim kisikovim vrstama proizvedenim od strane biljke kao dio njenog obrambenog odgovora (Reverchon i sur., 2002).



Slika 31. Prepostavljeni mehanizam biosinteze plavog pigmenta indigoindina (Reverchon i sur., 2002).

4. Genska nakupina br. 4: prepostavljeni NRPS tip B

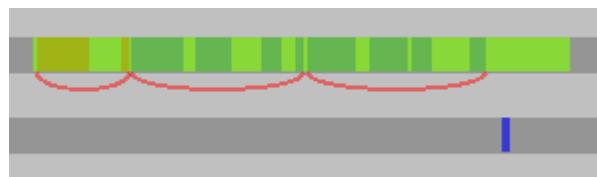
- Ovaj NRPS sustav se sastoji od adenilacijske domene koja veže *O*-sukcinilbenzoat i tioesterazne domene te je moguće da je konačni produkt oligomer *O*-sukcinilbenzoata, iako je također moguće da je i ova genska nakupina lažno pozitivni rezultat, odnosno ostatak NRPS sustava ostao nakon delecije dijela genoma.

5. Genska nakupina br. 5: prepostavljeni PKS tip II/NRPS hibrid

- Za gensku nakupinu br. 5 se prepostavlja da je hibridni sustav PKS tip II i NRPS sustava. Međutim, adenilacijska domena NRPS sustava nalikuje poliribonukleotid nukleotidil transferazi, a KS domene PKS sustava, iako imaju nisku E vrijednost (2.86209×10^{-24} i 6.99091×10^{-11}), imaju i negativan uspjeh pogotka (-117.394 i -308.231, redom). To nam govori da je ovo ili lažno pozitivni rezultat ili je ovaj sustav jako divergirao i predstavlja dosad neopisanu vrstu hibridnog PKS/NRPS sustava.

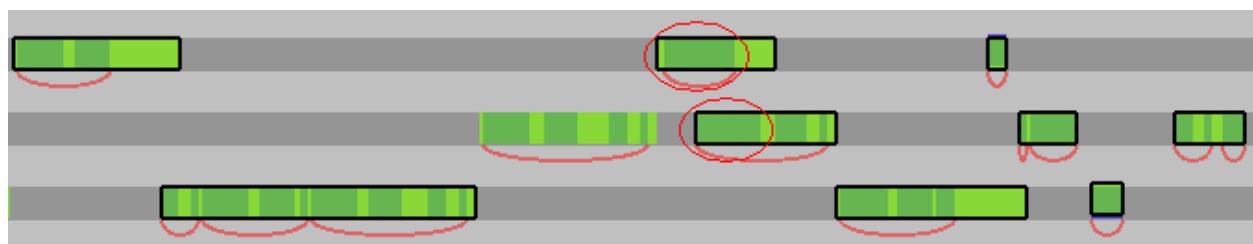
6. Genska nakupina br. 6: pretpostavljeni modularni PKS tip I/NRPS hibrid

- Kao i kod gena 187 genske nakupine 1, i u genu 212 genske nakupine 6 domena ACP se nalazi u drugom okviru čitanja te nije obuhvaćena navedenim genom. Razlog ovoj situaciji je moguće identičan, odnosno nemogućnost kvalitetnijeg sekvenciranja navedenog dijela genoma (slika 32).



Slika 32. Prikaz gena 212 genske nakupine 6.

- Još jedna sličnost s genskom nakupinom 1 je neuobičajena organizacija domena po genima (geni 217, 218, 220-226) – pri sekvenciranju ili sastavljanju genoma je zbog insercije ili delecije jednog nukleotida došlo do pomaka okvira čitanja pojedinih domena, međutim njihov je sveukupni raspored uobičajen za PKS sustave (slika 33). Gen 222, i u manjoj mjeri gen 210, na osnovi rezultata pretrage računalnim programom *BLAST*, također pokazuju sličnost s genima koji kodiraju za protein koji sudjeluju u sintezi lasalocida vrste *Streptomyces lasaliensis*, a gen 218 pokazuje najveću sličnost s genom vrste *Streptomyces filamentosus* koji kodira za protein koji sudjeluje u sintezi rifamicina.



Slika 33. Prikaz gena 217, 218, 220-226 genske nakupine 6.

- Kod genske nakupine 6 je još važno napomenuti kako kod gena 220 i 221 (slika 33) domene KS i AT djelomično pokrivaju isto područje genoma, iako se nalaze u različitim

okvirima čitanja. Do toga je došlo zbog pogreške u sekvenciranju. Na 3' strani domene KS (E vrijednost iznosi 531,192, a uspjeh pogotka je $1,245 \times 10^{-160}$), odnosno na 5' strani domene AT (E vrijednost iznosi 221,32, a uspjeh pogotka je $2,377 \times 10^{-67}$) se genom nije mogao kvalitetno sekvencirati, odnosno to područje obiluje „N“ nukleotidima.

7. Genska nakupina br. 7: Pretpostavljeni modularni PKS tip I/NRPS hibrid

- U ovoj genskoj nakupini PKS dio hibrida pokazuje sličnost modularnim PKS tip I sustavima, dok NRPS dio hibrida ne pokazuje sličnost s dosad anotiranim proteinima.

Iz rezultata pretraživanja genoma bakterije *Streptomyces sp. C* (Broad Institute, 2013), pomoću profila proteina koji su učitani u programski paket *ClustScan*, mogu se s donekle visokom sigurnošću klasificirati određene genske nakupine, odnosno spada li određena genska nakupina u PKS, NRPS ili hibridne sustave, međutim teže je odrediti kojoj klasifikacijskoj grupi unutar tih sustava pronađene genske nakupine pripadaju. Također, s obzirom na standardne bioinformatičke metode, ovim programskim paketom je omogućen brži pregled sekvence i lakše pronalaženje genskih nakupina. Manipulacija genima, modulima i domenama je jednostavna i donekle se mogu ukloniti greške nastale nekvalitetnim sekvenciranjem i sastavljanjem genoma.

6. ZAKLJUČCI

Na temelju dobivenih rezultata i provedene rasprave mogu se izvesti sljedeći zaključci:

- Korištenjem programskog paketa *ClustScan* u genomu bakterije *Streptomyces sp.* C identificirano je 7 genskih nakupina:
 - dvije genske nakupine kodiraju za produkte koji su uključeni u sintezu poliketida,
 - dvije genske nakupine kodiraju za produkte koji su uključeni u sintezu neribosomalno sintetiziranih peptida i
 - četiri genske nakupine kodiraju za produkte koji su uključeni u sintezu hibridnih poliketidno/peptidnih produkata.
- Navedene genske nakupine, budući da su analizirane koristeći se samo *in silico* metodom, opisane su s određenom vjerojatnošću, pri čemu se za neke sustave može sumnjati da su lažno pozitivni rezultati.
- Genske nakupine 1 i 6, zbog svoje sličnosti u organizaciji domena s genskim nakupinama koje kodiraju za antibiotike rifamicin i lasalocid (i oleandomicin), predstavljaju genske nakupine koje imaju najviše vjerojatnosti da kodiraju za enzime koji sudjeluju u sintezi prirodnih spojeva antibiotskog djelovanja.
- Daljnja istraživanja genskih nakupina koji sadržavaju genetičku uputu za sustave PKS i NRPS bi osigurala precizniju anotaciju i predviđanje strukture produkata takvih nakupina, kao i njihove karakterizacije i potencijalnih učinaka. To bi sve omogućilo i bolje definiranje navedenih genskih nakupina i na razini podgrupa, a ne samo glavnih tipova te bi olakšalo i pronađazak novih prirodnih spojeva od interesa.

7. LITERATURA

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
2. Altschul, S. F, Madden, T. L, Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
3. Anand, S., Prasad, M. V., Yadav, G., Kumar, N., Shehara, J., Ansari, Z., Mahanty, D. (2010) SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.* **38**, Suppl:W487-496.
4. Anonymous 1 (2013) A knowledge based resource for analysis of Non-ribosomal Peptide Synthetases and Polyketide Synthetases, <<http://www.nii.res.in/nrps-pks.html>>. Pristupljeno 10.8.2013.
5. Ansari, M. Z., Yadav, G., Gokhale, R. S., Mohanty, D. (2004) NRPS-PKS: a Knowledge-based Resource of NRPS/PKS Megasynthases. *Nucleic Acids Res.* **32**, 405-413.
6. Bentley, R., Bennet, J. W. (1999) Construction polyketides: from Collie to combinatorial biosynthesis. *Annu. Rev. Microbiol.* **53**, 411-446.
7. Besemer, J., Borodovsky, M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33**, 451-454.
8. Birch, A. J., Donovan, F.W. (1953) Studies in relation to biosynthesis. I. Some possible routes to derivatives of orcinol and phloroglucinol. *Aust. J. Chem.* **6**, 360-368.
9. Broad Institute (2013) Antibiotic Discovery: New Strategies to Combat Antibiotic Resistance initiative, Broad Institute,<<http://www.broadinstitute.org/>>. Pristupljeno 1.5.2013.
10. Bruner, S. D., Weber T., Kohli, R. M., Schwarzer, D., Marahiel M. A., Walsh C. T., Stubbs, M. T. (2002) Structural Basis for the Cyclization of the Lipopeptide Antibiotic Surfactin by the Thioesterase Domain SrfTE. *Structure* **10**, 301-310.
11. Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., Kucherov, G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.* **36**, 326-331.
12. Castillo, L. F., Galeano, N., Isazal, G. A., Gaitan, A. (2012) Construction of coffee transcriptome networks based on gene annotation semantics. *J. Integr. Bioinform.* **9**, 205-212.
13. Challis, L. G., Naismith, J. H., (2004) Structural aspects of non-ribosomal peptide biosynthesis. *Curr. Opin. Chem. Biol.* **14**, 748-756.
14. Chan, Y. A., Podevels, A. M., Kevanya, B. M., Thomas, M. G. (2009) Biosynthesis of polyketide synthase extender units. *Nat. Prod. Rep.* **26**, 90–114.

15. Collie, J. N. (1893) The production of naphthalene derivatives from dehydracetic acid. *J. Chem. Soc.* **63**, 329-337.
16. Conway, K. R., Boddy, C. N. (2013) ClusterMine360: a database of microbial PKS/NRPS biosynthesis. *Nucleic Acids Res.* **41**, D402-407.
17. Delcher, A. L., Bratke, K. A., Powers, E. C., Salzberg, S. L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673-679.
18. Du, L., Sánchez, C., Shen, B. (2001) Hybrid peptide-polyketide natural products: biosynthesis and prospects toward engineering novel molecules. *Metab. Eng.* **3**, 78-95.
19. EBI (2013) European Bioinformatics Institute - European Molecular Biology Laboratory EMBL, <<http://www.ebi.ac.uk/>>. Pristupljeno 15. 07. 2013.
20. Eddy, S. R. (1998) Profile Hidden Markov models. *Bioinformatics* **14**, 755-763.
21. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K. (2008) The Pfam protein families database. *Nucleic Acids Res.* **36**, 281-288.
22. Fischbach, M. A., Walsh, C. T., Clardy, J. (2008) The evolution of gene collectives: How natural selection drives chemical innovation. *Proc. Natl. Acad. Sci. USA* **105**, 4601-4608.
23. Fujii, I., Watanabe, A., Sankawa, U., Ebizuka, Y. (2001) Identification of Claisen cyclase domain in fungal polyketide synthase WA, a naphthopyrone synthase of *Aspergillus nidulans*. *Chem. Biol.* **8**, 189-197.
24. Ghai, R., McMahon, K. D., Rodriguez-Valera, F. (2012) Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. *Environ. Microbiol. Rep.* **4**, 29–35.
25. Gokhale, R. S., Sankaranarayanan, R., Mohanty, D. (2007) Versatility of polyketide synthases in generating metabolic diversity. *Curr. Opin. Struc. Biol.* **17**, 736-743.
26. Grabley, S., Thiericke, R. (1999) The impact of natural products on drug discovery. U: Drug Discovery from Nature (Grabley, S., Thiericke, R., ured.), Springer, Berlin, str. 3-37.
27. Hopwood, D.A. (1997) Genetic contributions to understanding polyketide synthases. *Chem. Rev.* **97**, 2465-2497.
28. Hranueli, D., Cullum, J. (2001) Novi hibridni poliketidi dobiveni kombinatornom biosintezom. *Kem. Ind.* **50**, 381-411.
29. Hranueli, D., Cullum, J. (2003) Bioinformatics of *Streptomyces* species and food production. U: Current Studies of Biotechnology - Vol. III. Food. Croatian Society of Biotechnology (Kniewald, Z. i sur. ured.), Zagreb, Hrvatska, str. 333-340.

30. Hranueli, D., Starčević, A., Žučko, J., Diminić, J., Škunca, N., Željeznak, V., Kovaček, D., Pavlinušić, D., Šimunković, J., Long, P. F., Cullum, J. (2008) Oblikovanje novih prirodnih spojeva u uvjetima *in silico*. *Kem. Ind.* **57**, 245–256.
31. Hutchinson, R. C. (2003) Polyketide and non-ribosomal peptide synthases: Falling together by coming apart. *Proc. Natl. Acad. Sci. USA* **100**, 3010-3012.
32. Jenke-Kodama, H., Dittmann, E. (2009) Evolution of metabolic diversity: Insights from microbial polyketide synthases. *Phytochemistry* **70**, 1858-1866.
33. Komaki, H., Harayama., S. (2006) Sequence diversity of type-II polyketide synthase genes in *Streptomyces*. *Actinomycetologic.* **20**, 42–48.
34. Li, Y., Müller, R. (2009) Non-modular polyketide synthases in myxobacteria. *Phytochemistry* **70**, 1850-1857.
35. Li, M. H. T., Ung, P. M. U., Zajkowski, J., Garneau-Tsodikova, S., Sherman, D. H. (2009) Automated genome mining for natural products. *BMC Bioinformatics* **10**, 185.
36. Marahiel, M. A., Mootz, H. D., Schwarzer, D. (2002) Ways of Assembling Complex Natural Products on Modular Nonribosomal Peptide Synthetases. *ChemBioChem* **3**, 490-504.
37. Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E., Breitling, R. (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.
38. NCBI (2013) NCBI National Center for Biotechnology Information, <<ftp://ftp.ncbi.nih.gov/>>. Pristupljeno 10.8.2013.
39. Nicholson, T. P., Rudd, B. A. M., Dawson, M., Lazarus, C. M., Simpson, T. J., Cox, R. J. (2001) Design and utility of oligonucleotide gene probes for fungal polyketide synthases. *Chem. Biol.* **8**, 157-178.
40. Frandsen, R. (2010) Rasmus Frandsen, Ph.D. – Home Page, <<http://www.rasmusfrandsen.dk/>>. Pristupljeno 12.7.2013.
41. Parry, R. J., Lin, M. T., Walker, A. E., Mhaskar, S. (1991) Biosynthesis of coronatine: investigations of the biosynthesis of coronamic acid. *J. Am. Chem. Soc.* **113**, 1849-1850.
42. Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., Huson, D. H. (2005) Specificityprediction of adenylation domains in nonribosomal peptide synthetases (NRPS) usingtransductive support vector machines (TSVMs). *Nucleic Acids Res.* **33**, 5799-808.
43. Reeves, G.A., Talavera, D., Thornton, J.M. (2009) Genome and proteome annotation:organization, interpretation and integration. *J. R. Soc. Interface* **6**, 129-147.

44. Reverchon, S., Rouanet, C., Expert, D., Nasser, W. (2002) Characterization of indigoidine biosynthetic genes in *Erwinia chrysanthemi* and role of this blue pigment in pathogenicity. *J. Bacteriol.* **184**, 654-665.
45. Rice, P., Longden, I., Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277.
46. Schwarzer, D., Marahiel, M.A. (2001) Multimodular biocatalysts for natural product assembly. *Naturwissenschaften*, **88**, 93-101.
47. Servin, J. A., Herbold, C.W., Skophammer, R. G., Lake, J. A. (2008) Evidence excluding the root of the tree of life from the actinobacteria. *Mol. Biol. Evol.* **25**, 1-4.
48. Shen, B. (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr. Opin. Chem. Biol.* **7**, 285–295.
49. Starčević, A., Žučko, J., Šimunković, J., Long, P. F., Cullum, J., Hranueli, D. (2008) ClustScan: an integrated program package for semi-automatic annotation of modular biosynthetic gene cluster and *in silico* prediction of novel chemical structures. *Nucleic Acids Res.* **10**, 1-11.
50. Tae, H., Kong, E., Park, K. (2007) ASMPKS: an analysis system for modular polyketidesynthases. *BMC bioinformatics* **8**, 327.
51. Tae, H., Sohng, J. K., Park, K. (2009) Development of an analysis program of type Ipolyketide synthase gene clusters using homology search and profile hidden Markov model. *J. Microbiol. Biotechnol.* **19**, 140-146.
52. Taguchi, T., Okamoto, S., Lezhava, A., Ochi, K., Ebizuka, Y., Ichinose, K. (2006) Possible involvementof ActVI-ORFA in transcriptional regulation of act VI tailoring-step genes for actinorhodin biosynthesis. *FEMS Microbiol. Lett.* **269**, 234–239.
53. Thiotemplate Modular Systems Studies (2013) ClustScan Professional, <<http://bioserv.pbf.hr/cms/index.php?page=clustscan>>. Pristupljeno 5.5.2013.
54. Thompson, J. D., Higgins, D.G., Gibson T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.
55. Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G. F., Chater, K. F., van Sinderen, D. (2007) Genomics of *Actinobacteria*: tracing the evolutionary history of an ancient phylum. *Microbiol. Mol. Biol. Rev.* **71**, 495-548.
56. Watanabe, A., Ebizuka, Y. (2004) Unprecedented mechanism for chain length determination in fungal aromatic polyketide synthases. *Chem. Biol.* **11**, 1101-1106A.

57. Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Hudson, D. H., Wohlleben, W. (2009) CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.* **140**, 13-17.
58. Wenzel, S. C., Müller, R. (2005) Formation of novel secondary metabolites by bacterial multimodular assembly lines: deviations from textbook biosynthetic logic. *Curr. Opin. Chem. Biol.* **9**, 447-458.
59. WHO (2013) World Health Organization, <<http://www.who.int/en/index.html>>. Pristupljeno 25.8.2013.
60. Woodruff, H. B. (1980) Natural products from microorganisms. *Science* **208**, 1225-1229
61. Yadav, G., Gokhale, R. S., Mohanty, D. (2003) SEARCHPKS: a program for detection and analysis of polyketide synthase domain. *Nucleic Acids Res.* **31**, 3654-3658.
62. Yandell, M., Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329-342.
63. Cheng, Y. Q., Tang, G. L., Shen, B. (2003) Type I polyketide synthase requiring a discrete acyltransferase for polyketide biosynthesis. *P. Natl. Acad. Sci. USA* **100**, 3149-3154.
64. Zazopoulos, E., Huang, K., Staffa, A., Liu, W., Bachmann, B. O., Nonaka, K., Ahlert, J., Thorson, J. S., Shen, B., Farnat, C. M. (2003) A genomics – guided approach for discovering and expressing cryptic metabolic pathways. *Nat. Biotechnol.* **21**, 187-190.
65. Zotchev, S. B., Stepanichikova, A. V., Sergeyko, A. P., Sobolev, B. N., Filimonov, D. A., Poroikov, V. V. (2006) Rational design of macrolides by virtual screening of combinatorial libraries generated through *in silico* manipulation of polyketide synthases. *J. Med. Chem.* **49**, 2077-2087.

8. PRILOZI

8.1. POPIS U RADU UPOTRIJEBLJENIH KRATICA

A	- domena za adenilaciju aminokiselina (NRPS)
ACP	- mali peptidni nosač acila (PKS)
ARO	- domena aromataze (PKS)
AT	- domena aciltransferaze (PKS)
C	- domena za kondenzaciju (NRPS)
CHS like	- šalkon sintazi slični sustavi (PKS)
CLF	- faktor koji određuje duljinu ugljikova lanca (PKS)
CoA	- koenzim A
CYC	- domena ciklaze (PKS)
DEBS	- deoksieritromicin B sintetaza
DH	- domena dehidrataze (PKS)
DNA	- deoksiribonukleinska kiselina
ER	- domena enoilreduktaze (PKS)
FAS	- sintaza masnih kiselina
FR-PKS	- potpuno reducirajući PKS sustavi
kb	- kilobaza (tisuća baza)
KR	- domena ketoreduktaze (PKS)
KS	- domena ketosintaze (PKS)
Mb	- megalabaza (milijun baza)
mRNA	- „messenger“ RNA (glasnička RNA)
NIS	- sintetaza NRPS-neovisnih siderofora
NR-PKS	- nereducirajući PKS sustavi
NRPS	- sintetaza neribosomalno sintetiziranih peptida
pb	- parovi baza
PCP	- mali polipeptid nosač peptidila (NRPS)
PCR	- „polymerase chain reaction“ (lančana reakcija polimeraze)
PKS	- poliketid sintaza
PR-PKS	- djelomično reducirajući PKS sustavi
RNA	- ribonukleinska kiselina
rRNA	- ribosomska RNA
TE	- domena tioesteraze (PKS)
Te	- domena tioesteraze (NRPS)

8.2. SADRŽAJ KOMPAKTNOG DISKA

Na kompaktnom disku (CD-R) se nalazi tekst Diplomskog rada (Diplomski rad hrimac.pdf) i prilog pod poglavljem 8.2.1.

8.2.1. Rezultat anotacije genoma bakterije *Streptomyces sp.* C programskim paketom *ClustScan*

Priložena je radna površina programskog paketa *ClustScan* (S.sp.C_rezultati_anotacije_hrimac). Napomena: za njeno je pregledavanje potrebno instalirati programski paket *ClustScan* na osobnom računalu koji se može preuzeti s mrežnih stranica Kabineta za bioinformatiku Prehrambeno-biotehnološkog fakulteta.