S V E U Č I L I Š T E U Z A G R E B U
PREHRAMBENO-BIOTEHNOLOŠKI FAKULTET

Diplomski rad

Zagreb, studeni 2009. Ida Trninić, 5081/BM

ANOTACIJA GENSKIH NAKUPINA POLIKETIDA I HIBRIDNIH SUSTAVA U GENOMU BAKTERIJE *Streptomyces scabies*

Ovaj je rad izrađen u Kabinetu za bioinformatiku, Zavoda za biokemijsko inženjerstvo, Prehrambeno-biotehnološkog fakulteta, Sveučilišta u Zagrebu pod vodstvom

dr. sc. DASLAVA HRANUELI, redovitog profesora

uz svesrdnu pomoć

dr. sc. ANTONIA STARČEVIĆA, višeg asistenta

Zahvaljujem se mentoru, prof. dr. sc. Daslavu Hranueli, što mi je omogućio izradu ovog diplomskog rada, na stručnom vodstvu, konstruktivnim savjetima, strpljenju i iscrpnoj pomoći tijekom pisanja.

Također se zahvaljujem dr. sc. Antoniu Starčeviću na pomoći, podršci i uvijek dobrodošlim savjetima.

Hvala obitelji i prijateljima, posebice roditeljima, na ljubavi, bezuvjetnoj potpori i razumijevanju.
Saša, hvala što si uz mene.

TEMELJNA DOKUMENTACIJSKA KARTICA

Diplomski rad

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

Znanstveno područje: Biotehničke znanosti
Znanstveno polje: Biotehnologija

ANOTACIJA GENSKIH NAKUPINA POLIKETIDA I HIBRIDNIH SUSTAVA U GENOMU BAKTERIJE *Streptomyces scabies*

Ida Trninić, 5081/BM

Sažetak: Poliketidi i neribosomalno sintetizirani peptidi su najznačajniji predstavnici sekundarnih metabolita. U zadnje vrijeme sve veću pozornost privlače i hibridni sustavi. Mehanizmi sinteze navedenih sekundarnih metabolita zasnivaju se na enzimskim kompleksima nazvanim poliketid sintaze (PKS) i sintetaze neribosomalnih peptida (NRPS). U biosintezi hibridnih sustava uključeni su enzimski kompleksi obje skupine. Geni čiji enzimi sudjeluju u biosintezi sekundarnih metabolita obično su organizirani u genske nakupine. Cilj ovog diplomskog rada bio je sistematizirati znanje o dosada istraženim i detaljno opisanim biosintetskim mehanizmima sustava PKS i hibrida. Takav pristup trebao bi dovesti do bolje klasifikacije navedenih sekundarnih metabolita. Taj je pristup testiran na novosekvencioniranom i do sada neopisanom genomu bakterije *Streptomyces scabies*. Osnovni alat koji je upotrijebljen u ovom radu jest bioinformatički programski paket *ClustScan* razvijen na Prehrambeno-biotehnološkom fakultetu Sveučilišta u Zagrebu.

Ključne riječi: PKS genske nakupine, hibridne genske nakupine, genom bakterije *Streptomyces scabies*, računalni programski paket *ClustScan*

Rad sadržava: 57 stranica, 30 slika, 6 tablica, 52 literaturna navoda, 4 priloga

Jezik izvornika: hrvatski

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

Mentor: dr. sc. Daslav Hranueli, red. prof.
Pomoć pri izradi: dr.sc. Antonio Starčević, viši asistent

Stručno povjerenstvo za ocjenu i obranu:
1. Dr. sc. Vladimir Mrša, red. prof.
2. Dr. sc. Daslav Hranueli, red. prof.
3. Dr. sc. Ivan-Krešimir Svetec, doc.
4. Dr. sc. Ana Vukelić, doc. (zamjena)

Datum obrane: 10. studeni, 2009.

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

ANOTATION OF POLYKETIDE AND HYBRID GENE-CLUSTERS IN *Streptomyces scabies* GENOME

Ida Trninić, 5081/BM

Abstract: Polyketides and nonribosomal peptides represent most valued bioactive compounds among secondary metabolites. Recently their hybrid systems also attract a lot of attention. The biosynthetic mechanisms involved in the production of these compounds are based on large multifunctional enzymes called polyketide synthases (PKS) and nonribosomal peptide synthetases (NRPS). Genes coding for these systems in bacteria or fungi are usually arranged in the gene clusters. The goal of this work was to try to systematize existing knowledge on these enzymes and produce a better, more detailed classification which could help distinguish them on the level of DNA sequence. This classification was tested on a newly sequenced un-annotated genome of *Streptomyces scabies*. The main bioinformatic tool used in this work has been the *ClustScan* program package, a computer program developed at the Faculty of Food Technology and Biotechnology.

Keywords: PKS gene-clusters, hybrid gene-clusters, *Streptomyces scabies* genome, *ClustScan* program package

Thesis contains: 57 pages, 30 figures, 6 tables, 52 references, 4 supplements

Original in: Croatian

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

Mentor: Ph. D. Daslav Hranueli, Full Professor
Technical support and assistance: Ph. D. Antonio Starčević, Senior Assistant

Reviewers:
1. Ph. D. Vladimir Mrša, Full Professor
2. Ph. D. Daslav Hranueli, Full Professor
3. Ph. D. Ivan-Krešimir Svetec, Assistant Professor
4. Ph. D. Ana Vukelić, Assistant Professor (substitute)

Thesis defended: 10th November, 2009

SADRŽAJ

stranica

1. U V O D

Bakterije i plijesni proizvode veliki broj molekula koje nisu uključene u vegetativni rast stanica koje ih proizvode. Takove se supstancije nazivaju sekundarnim metabolitima. Geni čiji produkti sudjeluju u biosintezi sekundarnih metabolita, kao i geni za otpornost prema vlastitom antibiotiku, obično su u bakterija i nekih gljiva organizirani u jednoj genskoj nakupini. Mnogi se od tih sekundarnih metabolita primjenjuju u medicini, veterini i agroindustriji. Gotovo 50 % najvažnijih lijekova, koji se danas klinički primjenjuju, kao ljekovite supstancije sadržavaju sekundarne metabolite kao prirodne spojeve.

S obzirom na kemijsku strukturu spojeva, među sekundarnim metabolitima najveću skupinu čine poliketidi. Druga velika skupina sekundarnih metabolita jesu neribosomalno sintetizirani peptidi. Biosintetski put obje skupine obuhvaća spajanje jednostavnih građevnih jedinica u složene kemijske strukture katalitičkim djelovanjem enzimskih kompleksa poliketid sintaza (engl. "Polyketide Synthases", PKS) ili sintetaza neribosomalnih peptida (engl. "Non Ribosomal Peptide Synthetases", NRPS). Iako upotrebljavaju različite građevne jedinice, enzimi PKS i NRPS imaju sličnu modularnu građu katalitičkih domena i slične mehanizme sinteze produkta. Sve je veći broj opisanih hibridnih sustava koje karakteriziraju moduli i domene svojstvene za navedene enzimske komplekse upravo zbog slične organizacije. Zbog iznimne važnosti sekundarnih metabolita, neophodno je istražiti i detaljnije opisati biosintetske mehanizme kojima nastaju spojevi poliketidnog, neribosomalno peptidnog i hibridnog porijekla.

U izradi ovog diplomskoga rada pozornost će se usmjeriti na prikupljanje sekvencija DNA i proteina PKS i hibridnih sustava iz baza podataka u svrhu izrade klasifikacijske tablice sustava PKS. Na temelju klasifikacijske tablice, izradit će se vlastiti profili proteina za proteine sustava PKS, kao i za proteine hibridnih sustava. Dobiveni rezultati upotrijebit će se za anotaciju novosekvencioniranog i dosada neopisanog genoma bakterije *Streptomyces scabies,* koja je ujedno i glavni cilj ovog diplomskoga rada. Na temelju sličnosti genoma vrste *S. scabies* i dosad anotiranih genoma iz streptomiceta (*S. coelicolor* i *S. avermitilis*) pretpostavlja se da će anotiranje ovog genoma također dovesti do pronalaska novih ili već opisanih genskih nakupina odgovornih za biosintezu poliketida i polipeptid/poliketid hibridnih produkata.

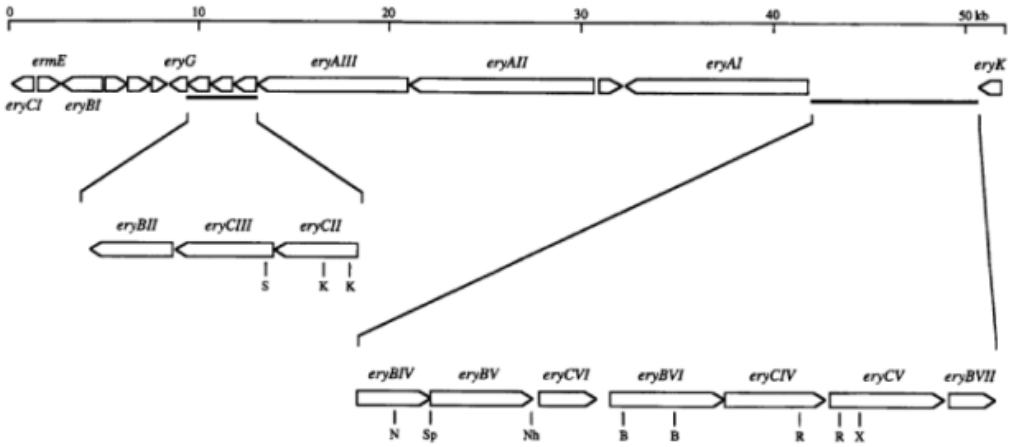
2. TEORIJSKI DIO

2.1. SEKUNDARNI METABOLITI MIKROORGANIZAMA

Bakterije i plijesni proizvode veliki broj molekula koje nisu neophodno potrebne za vegetativni rast stanica, već su uključene u različite mehanizame preživljavanja u prirodi, te kao čimbenici u ekološkim natjecanjima i simbiozi. Takove se supstancije nazivaju sekundarnim metabolitima. Sekundarni metaboliti se najčešće sintetiziraju nakon eksponencijalne faze rasta, odnosno pri brzinama rasta manjim od maksimalne, dakle tijekom idiofaze. Mogu se smatrati kemijskim oblikom diferencijacije. Međutim, nasuprot morfološkoj diferencijaciji koja je tipična za rod, ekspresija sekundarnog metabolizma je specifična za vrstu (to jest različite vrste istog roda mogu proizvoditi različite sekundarne metabolite) (Demain, 1999; Grabley i Thiericke, 1999; Hranueli i Cullum, 2001).

Geni čiji produkti sudjeluju u biosintezi sekundarnih metabolita (npr. antibiotika), kao i geni za otpornost prema vlastitom antibiotiku, obično su u bakterija i nekih gljiva (npr. *Penicillium* i *Aspergillus*) organizirani u jednoj genskoj nakupini. U drugih se gljiva, kao npr. *Acremonium chrysogenum* i *Cephalosporium acremonium*, geni čiji produkti sudjeluju u biosintezi cefalosporina, nalaze na različitim kromosomima (Marahiel i sur., 2002). U biosintetskom putu sekundarnog metabolita može biti uključeno i više od 40-ak gena koji čine DNA sekvenciju dužine od preko 100 kb. To jest, cijeli je skup proteina uključen u biosintezu sekundarnog metabolita, i može biti dvostruko veći od ribosoma (Fischbach i sur., 2008). Naime, ribosomi sintetiziraju tisuće različitih proteina, dok sekundarnim biosintetskim mehanizmom nastaje tek jedna mala molekula. Na Slici 1 shematski je prikazana genska nakupina odgovorna za biosintezu poliketidnog antibiotika eritromicina A u kojoj se vide 22 gena.

Zbog velike potrošnje energije, koju stanica mora podnijeti prilikom sinteze proteina uključenih u sekundarni metabolizam, veliki su geni (engl. "giant genes") u arhebakterija i eubakterija rijetki i neprestano pod selektivnim pritiskom (Reva i Tümmler, 2008). Pored toga, genske nakupine sekundarnih metabolita predstavljaju i vrlo različite te najbrže evoluirajuće genetičke elemente. Brza evolucija posljedica je: kratkog replikacijskog vremena mikroorganizama, mogućnosti horizontalnog prijenosa gena između mikroorganizama što dovodi do duplikacija i/ili fuzije gena, odnosno do sinteze strukturalno vrlo raznolike grupe sekundarnih metabolita (Fischbach i sur., 2008).



Slika 1. Shematski prikaz genske nakupine za biosintezu poliketidnog antibiotika eritromicina A. Produkti gena *ery*AI, *ery*AII i *ery*AIII sudjeluju u biosintezi poliketidnog kostura (tzv. aglikona), dok su produkti ostalih gena uglavnom uključeni u poslije-poliketidne modifikacije.

Mnogi se od tih sekundarnih metabolita primjenjuju u medicini, veterini i agroindustriji. Gotovo 50 % najvažnijih lijekova, koji se danas klinički primjenjuju, kao ljekovite supstancije sadržavaju sekundarne metabolite kao prirodne spojeve (Demain, 1999). Osim antibiotika, fungicida, antivirusnih sredstava i citostatika, u kliničkoj se primjeni nalaze i imunosupresori, antihipertenzivna sredstva, antidiabetici, antimalarici, te antiholesterolemici (Grabley i Thiericke, 1999; Hranueli i Cullum, 2001). Komercijalno su važni i proizvodi za agroindustriju, kao što su to antiparazitici, kokcidiostatici, životinjski promotori rasta i prirodni insekticidi. S obzirom na kemijsku strukturu spojeva, među sekundarnim metabolitima najveću skupinu čine poliketidi. Druga velika skupina sekundarnih metabolita jesu neribosomalno sintetizirani peptidi. Biosintetski put obje skupine obuhvaća spajanje jednostavnih građevnih jedinica u složene kemijske strukture katalitičkim djelovanjem enzimskih kompleksa poliketid sintaza (engl. "Polyketide Synthases", PKS) ili sintetaza neribosomalnih peptida (engl. "Non Ribosomal Peptide Synthetases", NRPS). Iako upotrebljavaju različite građevne jedinice, enzimi PKS i NRPS imaju sličnu modularnu građu katalitičkih domena i slične mehanizme sinteze produkta (Hranueli i sur., 2008).

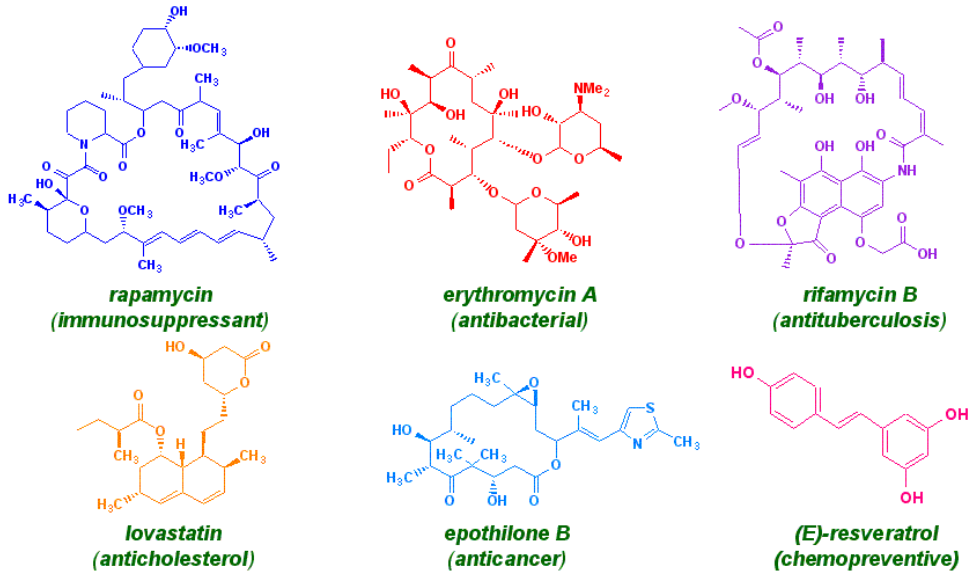
U posljednjih nekoliko godina u znanstvenoj javnosti postoji osobito zanimanje za oblikovanje novih supstancija u proizvodnji lijekova manipulacijom genskih nakupina tih enzimskih kompleksa u uvjetima *in vitro*, postupcima kombinatorne biosinteze. Međutim, značajna je prepreka napretku na tom području što većina promjena u uvjetima *in vitro* ne

dovodi do sinteze produkta ili su prinosi vrlo mali. Jedno od mogućih rješenja toga problema bilo bi oblikovanje novih genskih nakupina homolognom rekombinacijom u uvjetima *in vivo* jer bi se tako omogućilo spajanje identičnih sekvencija i smanjile poteškoće zbog pojave nefunkcionalnih čvorišta. Tome, međutim, prethodi sekvencioniranje novih genoma kako bi se pronašle nove genske nakupine (tj. geni koji nemaju svoje homologe u bazama podataka), odnosno anotacija genskih nakupina u metagenomima mikroorganizama što žive u tlu ili u simbiozi s morskim organizmima ne bi li se pronašli novi biosintetski putovi sekundarnih metabolita (Hranueli i sur., 2008).

2.1.1. Poliketidi

Poliketidi su velika i raznolika skupina kemijskih spojeva s vrlo širokim spektrom biološke aktivnosti, te velikim područjem primjene. U farmakološki važne aktivnosti poliketidnih prirodnih produkata spadaju antimikrobna, antifungalna, antiparazitska, antitumorska i agrokemijska svojstva. Ovi su sekundarni metaboliti vrlo rasprostranjeni, a proizvode ih različiti organizmi, poput bakterija, gljiva, biljaka, insekata, dinoflagelata, mekušaca te spužvi. Široki spektar aktivnosti poliketida čini ih ekonomski, klinički i industrijski najvažnijim kemijskim spojevima. Mnogi su poliketidi dobro poznati spojevi poput eritromicina A (makrolidni antibiotik širokog spektra) ili rapamicina (imunosupresor) (Slika 2) (Hranueli i Cullum, 2001).

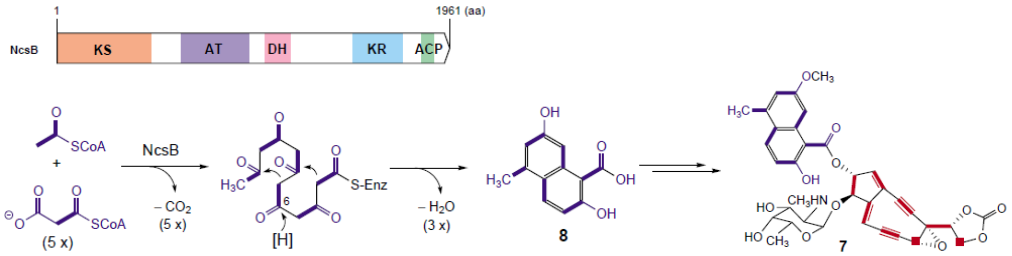
Biosinteza poliketida odvija se na velikim više-funkcionalnim enzimima ili više-enzimskim kompleksima nazvanim poliketid sintaze (PKS). Na tim se enzimskim kompleksima biosinteza odvija kao na "proizvodnoj traci", s time da je redoslijed kemijskih reakcija uvjetovan redoslijedom enzimskih domena koje ih kataliziraju. Takav je mehanizam biosinteze vrlo sličan onome ravnolančanih masnih kiselina pomoću sintaza masnih kiselina (engl. "Faty Acid Synthase", FAS). Za biosintezu se upotrebljavaju jednostavne građevne jedinice od 2, 3 ili 4 ugljikova atoma, poput acetil-CoA, propionil-CoA i butiril-CoA (Chan i sur., 2009). Takve se građevne jedinice kondenziraju sve dok se ne sintetizira ugljikov lanac potrebne dužine. U svakom se ciklusu kondenzacije lanac produljuje za dva ugljikova atoma, s time da je na ß-ugljikovom atomu uvijek keto skupina, koja najčešće ostaje nepromijenjena. Poznata su tri tipa poliketid sintaza koje se međusobno znatno razlikuju po strukturi i funkciji (Shen, 2003).



Slika 2. Primjeri nekih poliketidnih kemijskih spojeva.

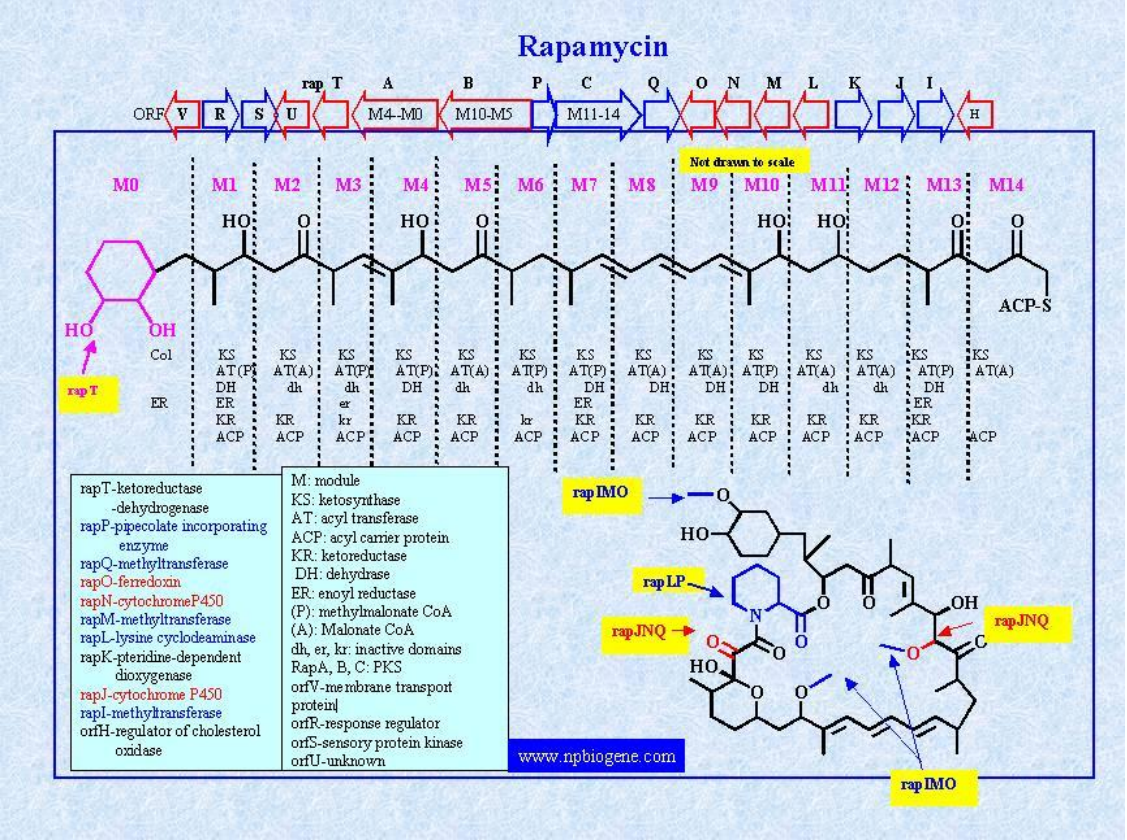
2.1.1.1. Poliketid sintaze tipa I

Sustave PKS 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 dva različita podsustava: ponavljajući podsustav (tzv. iterativni, primjerice PKS neokarzinostatina, Slika 3) i modularni podsustav (npr. sustav odgovoran za biosintezu rapamicina, Slika 4).



Slika 3. Organizacija domena PKS neokarzinostatina. Gen *Ncs*B sadržava genetičku uputu za PKS tipa I sastavljenu od karakterističnih domena KS, AT, DH, KR i ACP. Hipotetski prikaz biosinteze naftalinske kiseline (8) iz prekursora acil-CoA ponavljajućim procesom. Aromatski dio poliketida i enedinska jezgra prikazani su plavom i crvenom bojom. Atomi ugrađeni neposredno u odgovarajuće poliketidne dijelove iz prekusora acil-CoA su istaknuti.

Za sustave PKS tipa I svojstvena je modularna organizacija. Svaki se modul sastoji od minimalno tri domene, ketosintaze, acil transferaze i malog proteinskog nosača acila (engl. "Keto Synthase", KS; "Acyl Transferase", AT; "Acyl Carrier Protein", ACP) koje odabiru, aktiviraju i kataliziraju dekarboksilativnu Claisenovu kondenzaciju između produžne građevne jedinice i rastućeg poliketidnog lanca stvarajući međuprodukt β-ketoacil-S-ACP. Dodatne domene nalaze se između domena AT i ACP, a provode promjenjiv niz redukcijskih modifikacija β-keto skupine prije slijedećeg stupnja produžetka lanca. Redoslijed modula u enzima PKS određuje slijed biosintetskih događaja, a varijacije domena unutar modula pružaju strukturnu raznolikost koja se očituje u krajnjim poliketidnim produktima.



Slika 4. Prikaz organizacije gena i proteina modularnog sustava PKS tipa I odgovorne za biosintezu imunosupresora rapamicina.

Sustavi PKS tipa I uobičajeni su u bakterija. Primjerice, sekvencioniranjem cjelovitih genoma bakterijskih vrsta *Streptomyces avermitilis* i *S. coelicolor*, utvrđeno je da one posjeduju osam, odnosno tri genske nakupine PKS tipa I (Hutchinson, 2003).

2.1.1.2. Poliketid sintaze tipa II

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

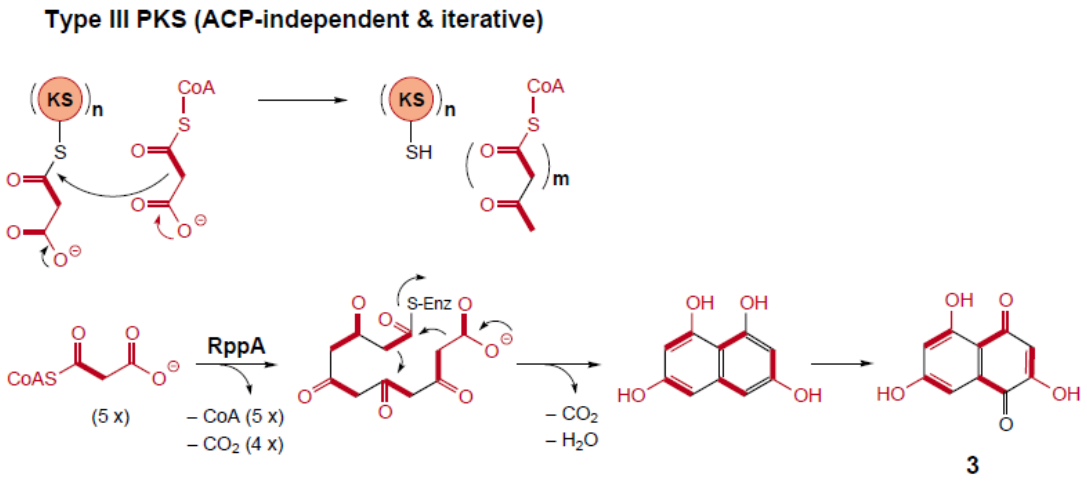
Tako, primjerice, biosintezu poliketidnog antibiotika aktinorodina u vrste *S. coelicolor* sintetizira jedan tipičan sustav PKS tipa II, koji upotrebljava šest genskih produkata kako bi se oblikovalo osam acetatnih jedinica u tri spojena supstituirana šesteročlana prstena (dva aromatska prstena i jedan laktonski). Tri su podjedinice specifične za reakciju: ketoreduktaza, aromataza i ciklaza (engl. "Keto Reductase", KR; "Aromatase", ARO; "Cyclase", CYC), dok su preostale tri (KSα, KSβ i ACP) ponavljajući "minimalni PKS" zajednički svim sustavima PKS tipa II (Taguchi i sur., 2006).

2.1.1.3. Poliketid sintaze tipa III

Za razliku od sustava PKS tipa I i II, sustavi PKS tipa III, nazvani i šalkon sintazi slični sustavi (engl. "Chalcone Synthase Like", CHS like), imaju jednostavnu građu (oblik homodimera identičnih KS monomernih domena). Zbog toga su prikladniji za istraživanja *in vitro* i manipulacije kao i za detaljnije strukturne analize (Austin i Noel, 2002). Ovi proteini djeluju kao ponavljajući enzimi koji provode kondenzaciju. Dobar primjer je sintaza RppA odgovorna za biosintezu aromatskih poliketida (često mono-cikličnih ili bi-cikličnih), poput flavolina (Slika 5) (Shen, 2003). Sustavi PKS tipa I i II upotrebljavaju domenu ACP za aktivaciju supstrata acil-CoA i prijenos poliketidnih međuprodukata, dok enzimi PKS tipa III ne ovise o domeni ACP i djeluju neposredno na supstrate acil-CoA.

Od prvih otkrića bakterijskih poliketid sintaza tipa I (1990.), tipa II (1984.) i tipa III (1999.), ovi su sustavi PKS poslužili znanstvenicima kao molekularne osnove s ciljem da se

objasni značajna strukturna raznolikost poliketidnih prirodnih produkata, te kao biotehnološka platforma za proizvodnju "neprirodnih" prirodnih spojeva uz pomoć metoda kombinatorne biosinteze. Razvoj metodologije za kloniranje biosintetskih genskih nakupina, napredak u tehnologijama sekvencioniranja DNA te bioinformatici, otvaraju nove mogućnosti za pronalazak potpuno novih mehanizama biosinteze (Shen, 2003).



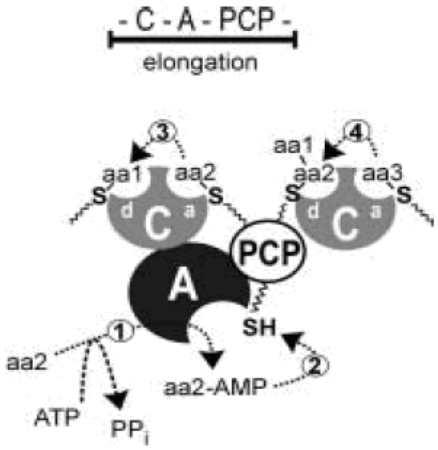
Slika 5. Sustav PKS Tipa III sastoji se od ponavljajuće jednostruke podjedinice kao što je prikazano na primjeru sintaze RppA za biosintezu flavolina (3). Atomi ugrađeni neposredno iz prekursora acil-CoA u nastalom poliketidu su podebljani.

2.1.2. Neribosomalno sintetizirani peptidi

Neribosomalno sintetizirani peptidi su druga velika skupina sekundarnih metabolita. U prirodi ih sintetiziraju mikroorganizmi, koji većinom prebivaju u zemlji, kao što su to Gram pozitivne bakterije rodova *Actinomyces* i *Bacillus*, te eukariotski mikroorganizmi iz skupine plijesni (Marahiel i sur., 2002). Do danas je opisano preko 1000 strukturalno različitih neribosomalno sintetiziranih peptida. Za vrlo je malen broj od njih, međutim, poznat mehanizam biosinteze (Caboche i sur., 2008). Neribosomalno sintetizirani peptidi (obično dugački od 3 do 15 aminokiselina) ne sintetiziraju se translacijom na ribosomima, već uzastopnom kondenzacijom aminokiselina katalitičkim djelovanjem više-funkcionalnih neribosomalnih peptid sintetaza (NRPS).

Enzim NRPS može biti izgrađen od jednog polipeptidnog lanca ili od nekoliko podjedinica. Enzimima NRPS se sintetiziraju gotovo svi peptidni antibiotici i dio siderofora, koje mikroorganizmi izlučuju radi vezivanja iona željeza iz okoline. Genske nakupine koje

sadržavaju genetičku uputu za enzime NRPS imaju sličnu organizaciju kao i genske nakupine za sustave PKS (vidi: podpoglavlje 2.1.1.). Osnovni modul enzima NRPS (Slika 6) 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 ili T) (Marahiel i sur., 2002; Challis i Naismith, 2004).



Slika 6. Primjer osnovnog modula enzima NRPS.

Domena A bira određenu aminokiselinu kao početnu ili produžnu građevnu jedinicu i aktivira je adenilacijom. Za to je potrebna jedna molekula ATP. Adenilat te rastući peptidni lanac prenose se pomoću fosfopanteteinske ruke na domenu PCP za koji se vežu preko tiolne skupine serina, formirajući aktiviran tioesterski vez. Domena C katalizira stvaranje peptidne veze između aminoacil tioestera na domeni PCP istog modula i prethodnog modula (Challis i Naismith, 2004). Početni modul enzima NRPS obično ne sadržava domenu C, dok se na kraju posljednjeg modula, kao i u enzima PKS (vidi: podpoglavlje 2.1.1.), nalazi domena tioesteraza (engl. "Thioesterase domain", Te) odgovorna za odvajanje linearnoga peptidnog lanca od enzima i ciklizaciju. Prošireni osnovni modul može sadržavati i domene: epimeraza (engl. "Epimerisation domain", E), metiltransferaza (engl. "Methyltransferase domain", MT) (Marahiel i sur., 2002). Neribosomalno sintetizirane peptide karakterizira velika strukturalna raznolikost. Oni mogu biti linearni, ciklički ili razgranato-ciklički, te makrociklički laktami ili laktoni što je posljedica ugradnje uobičajenih (proteinogenih) i neuobičajenih aminokiselina kao što su to alfa-hidroksi- i karboksi-kiseline povezane vezama koje nisu samo peptidne ili disulfidne (Wenzel i Müller, 2005).

Strukturalna raznolikost posljedica je i različite organizacije domena unutar modula, odnosno različitih mehanizama sinteze produkata, zbog čega se enzimi NRPS dijele u tri grupe: linearne (tip A), ponavljajuće (tip B) i nelinearne (tip C) (Marahiel i sur., 2002).

2.1.3. Poliketidno/peptidni hibridi

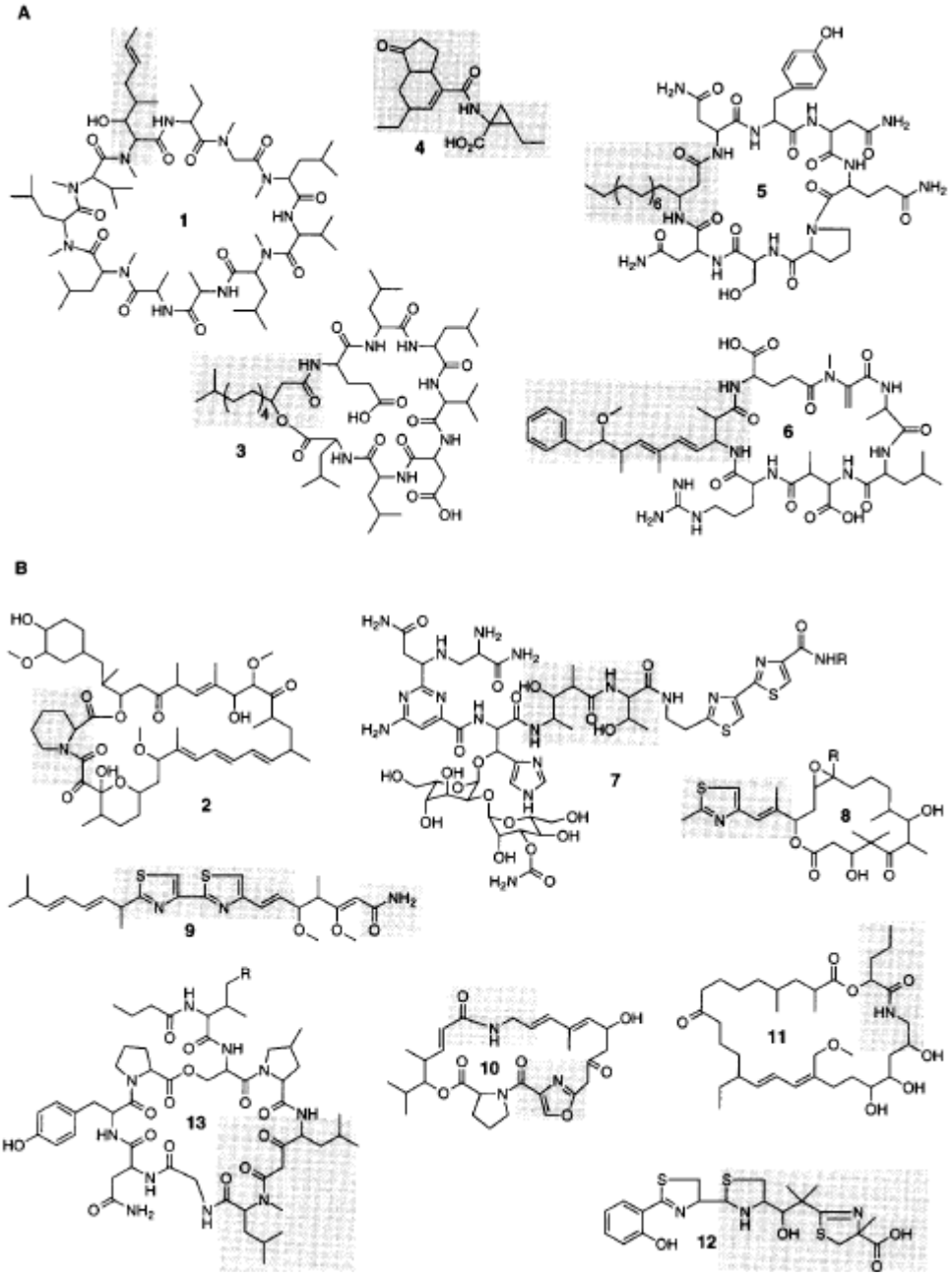
Poliketidno/peptidni hibridni prirodni produkti sastavljeni su od miješanih poliketidnih i peptidnih struktura (Slika 7). Ovi sekundarni metaboliti nastaju biosintezom iz aminokiselina i kratkih karboksilnih kiselina djelovanjem poliketid sintaza (PKS) i neribosomalnih peptid sintetaza (NRPS).

Poliketidno/peptidni hibridi upotrebljavaju vrlo slične mehanizme prilikom biosinteze prirodnih produkata koji spadaju u dvije različite skupine. Osim zajedničke modularne organizacije, oba sustava koriste proteinske nosače, PCP kod enzima NRPS i ACP kod enzima PKS, koji pridržavaju rastuće lance. Obje domene, i PCP i ACP, posttranslacijskom modifikacijom dobivaju 4´-fosfopanteteinsku prostetsku skupinu, a modificiranje katalizira enzim iz obitelji 4´-fosfopantetein transferaza (PPTaze). Tijekom cijelog procesa produljivanja, rastući intermedijer ostaje kovalentno vezan tioesterkom vezom preko tiolne skupine 4´-fosfopanteteinske grupe na proteinski nosač. Kada dostigne potpunu duljinu, peptidni ili poliketidni produkt se otpušta sa proteina pomoću tioesterazne domene koja se obično nalazi na krajnjem C-terminalnom kraju enzima PKS, odnosno NRPS (Du i sur., 2001).

Na temelju biosintetskih mehanizama kojima se karboksilni/poliketidni i aminokiselinski/peptidni dijelovi ugrađuju u konačne produkte, poliketidno/peptidni hibridini prirodni produkti mogu se podijeliti u dvije skupine:

A. skupina u kojoj se okosnica hibridnog produkta sastavlja djelovanjem hibridnog NRPS/PKS sustava, a koji posreduje u izravnom produljivanju peptidnog međuprodukta vezanog na enzim NRPS pomoću PKS modula i obrnuto;

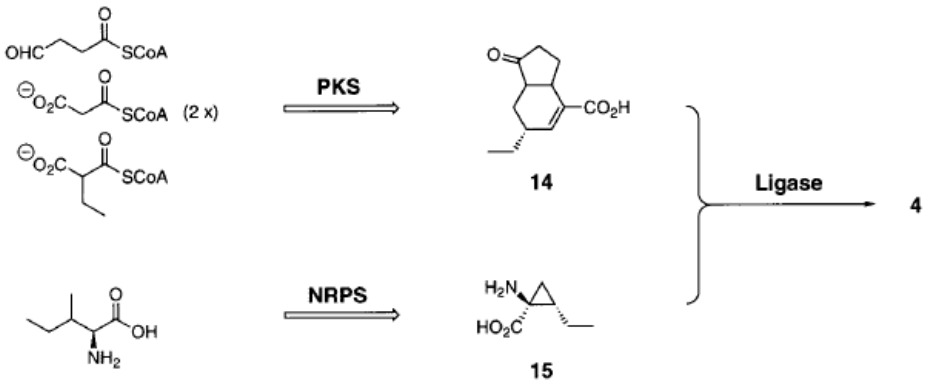
B. skupina u kojoj okosnica hibridnog produkta nastaje spajanjem pomoću drugih mehanizama koji ne zahtijevanju izravnu funkcionalnu hibridizaciju između NRPS i PKS proteina.



Slika 7. Primjeri poliketidno/peptidnih hibridih prirodnih produkata: ciklosporin A (1), rapamicin (2), surfaktin (3), koronatin (4), mikosubtilin (5), mikrocistin (6), bleomicin (7), epotiloni (8), miksotiazol (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.

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

Biosinteza koronatina je primjer u kojemu se aminokiselinski i poliketidni dijelovi zasebno sintetiziraju pomoću enzima NRPS i PKS, a potom se spajaju u poliketidno/peptidni hibridni metabolit pomoću enzima ligaze. Fitotoksin koronatin (Slika 7, struktura 4), kojeg proizvode mnogi sojevi bakterije *Pseudomonas syringae*, sastoji se od dvije različite komponente: poliketidni dio čini koronafaksna kiselina (Slika 8, struktura 14), a aminokiselinski dio čini koronaminska kiselina (Slika 8, struktura 15). Genska nakupina za koronatin nalazi se u 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.



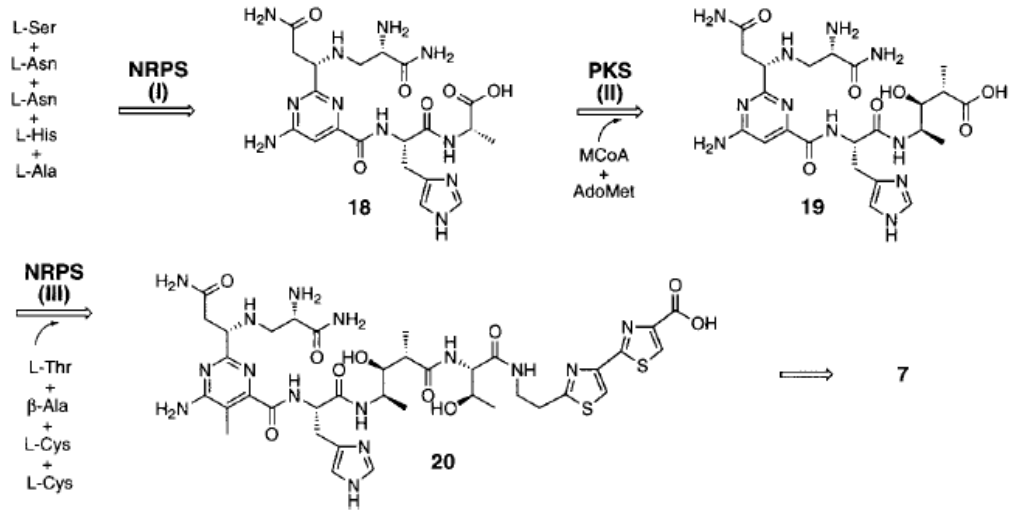
Slika 8. Biosintetski put koronatina (4) u bakteriji *P. syringe*. Poliketidni dio koronafaksne kiseline (14) i aminokiselinski dio koronaminske kiseline (15) se spajaju u koronatin (4) pomoću enzima ligaze.

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

Većina poliketidno/peptidnih hibridnih metabolita, čiji su biosintetski putevi opisani, nastaje pomoću hibridnih sustava NRPS/PKS koji posreduju u izravnom prijenosu peptidnog međuprodukta vezanog na enzim NRPS na modul enzima PKS i obratno.

Prema hibridnom modelu NRPS/PKS/NRPS može se predvidjeti da se biosinteza aglikona bleomicina (Slika 7, struktura 7) odvija u tri stupnja. U prvom stupnju iz aminokiselina Ser, Asn, Asn i His nastaje P-3A (Slika 9, struktura 18) pomoću enzima NRPS. U drugom stupnju enzim PKS katalizira elongaciju dodatkom malonil-CoA i

adenozilmetionina čime nastaje P-4 (Slika 9, struktura 19). Trećim se stupnjem nastavlja elongacija P-4, koju katalizira enzim NRPS, dodatkom aminokiselina β-Ala, Cys i Cys, čime nastaje P-6m (Slika 9, struktura 20).



Slika 9. Biosintetski put bleomicina (7) u *S. verticillus* uključuje hibridni sustav NRPS/PKS/NRPS. Rastući hibridni peptidni-poliketidni biosintetski međuprodukti P-3A (18), P-4 (19) i P-6m (20) izolirani su iz divljeg tipa *S. verticillus* i određene su im strukture.

Dok priroda kroz stvaranje poliketidno/peptidnih hibridnih metabolita pokazuje svu svoju raznolikost, hibridni sustavi NRPS/PKS odgovorni za nastajanje ovih metabolita predstavljaju najpogodnije komplekse za kombinatornu biosintezu. Najveći izazov u istraživanju biosinteze poliketidno/peptidnih hibridnih prirodnih produkata jest otkrivanje osnovnih katalitičkih i molekularnih svojstava i povezivanje struktuno-funkcionalnih odnosa ovih neobičnih sustava, bez kojih potencijalna kombinatorna biosinteza novih poliketidno/peptidnih metabolita ne bi bila ostvariva (Du i sur., 2001).

2.2. SEKVENCIRANJE GENOMA RAZLIČITIH VRSTA AKTINOBAKTERIJA

Aktinobakterije su jedna od najvećih taksonomskih skupina između 18 glavnih linija koje su u ovom trenutku poznate unutar carstva *Bacteria*. Ta skupina uključuje 5 podrazreda i 14 podreda. Uglavnom se sastoji od Gram-pozitivnih bakterija visokog G+C sastava DNA (od 51% u nekih vrsta roda *Corynebacterium* do više od 70% u vrsta rodova *Streptomyces* i *Frankia*). Aktinobakterije 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.). Također su, među njima, očite i fiziološke razlike, te 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*), te stanovnici animalnog probavnog trakta (*Bifidobacterium* spp.) (Ventura i sur., 2007).

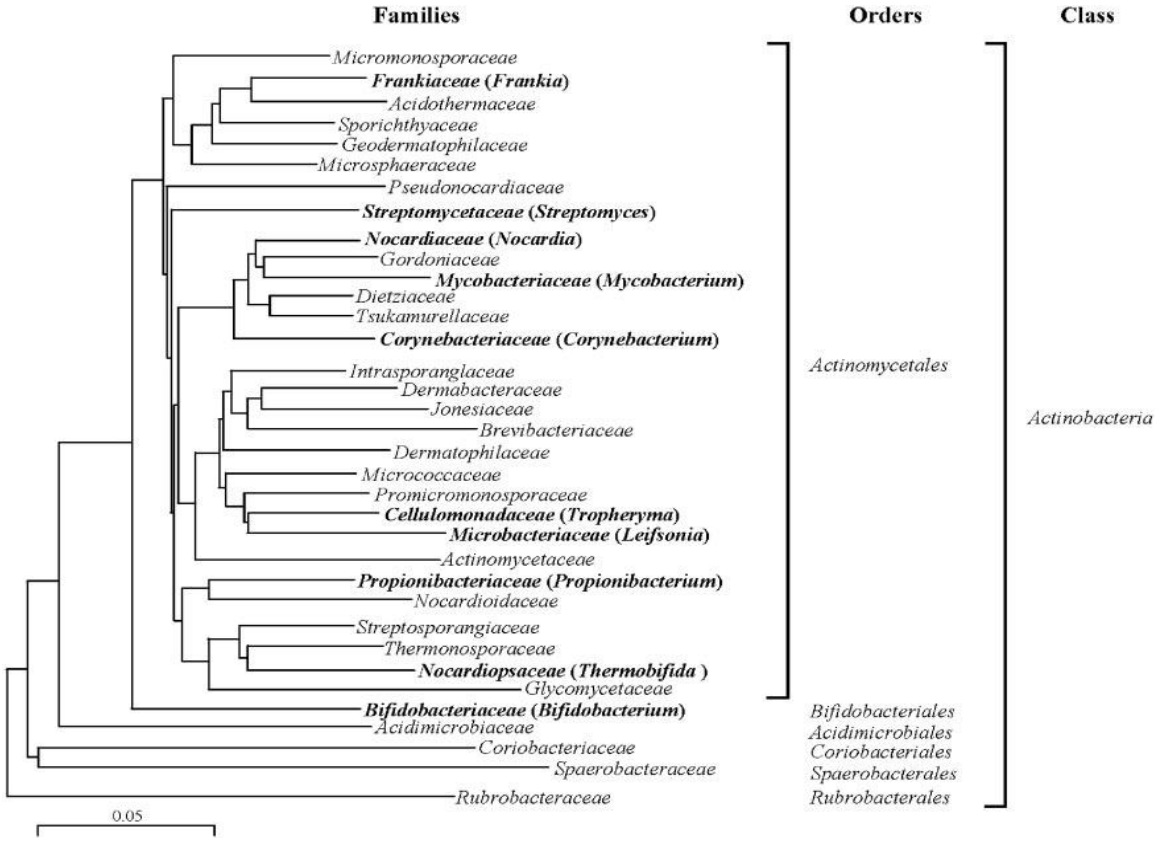
Do sada sekvencionirani genomi aktinobakterija uglavnom pripadaju mikroorganizmima koji su bitni za humanu i veterinarsku medicinu, biotehnologiju i ekologiju. Sekvencioniranjem i analizom gena za 16S rRNA utvrđeno je 39 porodica i 130 rodova (Slika 10) (Ventura i sur., 2007). Prvi sekvencionirani genom aktinobakterija jest humani patogen *Mycobacterium tuberculosis* H37Rv (Cole i sur., 1998). U posljednjih nekoliko godina sekvencionirano je još 20-ak genoma, dok je sekvencioniranje genoma još 82 predstavnika s visokim G+C sastavom u tijeku (NCBI, 2009).

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:

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

Sekvencioniranjem vrste *S. coelicolor* (Challis i Hopwood, 2003) pronađena je 21 genska nakupina odgovorna za sintezu sekundarnih metabolita, dok je u vrste *S. avermitilis* identificirano čak 25 genskih nakupina (Ōmura i sur., 2001).

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 genoma kružna, genomi roda *Streptomyces* su linerni. Pretpostavlja se da je linearnost kromosoma streptomiceta posljedica jednostruke recipročne izmjene (engl. "single-crossover recombination") između prvotno kružnog kromosoma i linearnog plazmida (Hranueli i Cullum, 2001).



Slika 10. Filogenetsko stablo aktinobakterija temeljeno na redoslijedu 1500 nukleotida 16S rRNA. Do sad sekvencionirani genomi su istaknuti.

2.2.1. Sekvenciranje genoma bakterije *Streptomyces scabies*

Vrsta *S. scabies* je biljni patogen koji uzrokuje krastavost krumpira, ali i ostalog korjenastog povrća poput repe, pastrnjaka, mrkve i rotkve (Lambert i Loria, 1989). Ovu bakteriju karakterizira pseudomicelijski rast. Pseudohife su vegetativni oblici stanica koji se

lome ili fragmentiraju u spore. Vrsta *S. scabies* direktno inficira mlada tkiva (poput gomolja u razvoju), a u razvijena tkiva prodire kroz prirodne otvore ili oštećenja. Za infekciju je odgovoran toksin, kojeg ova bakterija sintetizira.

Institucija "The Sanger Institut" u suradnji s Odjelom za fitopatologiju sa Sveučilišta Cornell u Sjedinjenim Američkim Državama, bila je zadužena za sekvenciranje genoma bakterije *S. scabies* soja 87.22 (The Wellcome Trust Sanger Institute 4, 2009). Sekvenciranje genoma vrste *S. scabies* je završeno i javno dostupno, ali njegova anotacija još nije objavljena. Pošlo se od pretpostavke da je genom vrste *S. scabies* dugačak oko 8.1 Mb te da sadržava 72% G-C parova baza. Sekvencija DNA genoma vrste *S. scabies* je dostupna za pretraživanje na "Blast Serveru" Sangerova Instituta, te za preuzimanje sa njihove "FTP" stranice (The Wellcome Trust Sanger Institute 4, 2009). Kromosom bakterije sastoji se od 10.148.695 bp (engl. "base pairs", parovi baza) s G+C sadržajem od 71.45%. Genom je sekvencioniran nasumičnom metodom (engl. "shotgun"), a na stranici su dostupne sve njegove sekvencije. Ukupno ima 158.304 očitanja (engl. "reads"), što čini 89.897 Mb. Prema tome genom je teoretski pokriven 99,99%.

2.3. BIOINFORMATIČKI ALATI ZA ANOTACIJU GENSKIH NAKUPINA

Napredak na području bioloških i računalnih znanosti omogućio je pohranjivanje velikog broja genoma i proteoma, kao i njihovu organizaciju i anotaciju. Anotacija genoma uključuje predviđanje genetičkih elemenata: strukturnih gena, pseudogena, promotora, tj. regulatornih regija, dijelova DNA što ne sadržavaju genetičku uputu i ponovljenih sekvencija. Anotacija se može izvršiti na različite načine: ručno, automatski ili *ab initio*.

Ručnim anotiranjem dobivaju se rezultati najveće točnosti jer metoda uključuje prethodno pretraživanje literature, međutim, spora je i zbog toga obuhvaća vrlo malen dio anotiranih podataka. Informacije dobivene ručnim anotiranjem jedne sekvencije mogu se upotrijebiti za automatsku anotaciju drugih sekvencija 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 upotrebljava za anotaciju velikih sekvencija 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 sekvencija DNA. Prilikom odabira metode, bira se između brzine anotacije i njezine točnosti (Reeves i sur., 2009).

Za analizu genskih nakupina poliketida i neribosomalno sintetizitanih peptida razvijeno je nekoliko bioinformatičkih alata koji upotrebljavaju donekle različite strategije i pristupe. Programi za anotaciju, kao i već anotirane sekvencije, dostupni su javnosti, a baze podataka su visoko organizirane i specijalizirane.

2.3.1. *SEARCHPKS, MAPSI* i drugi alati

Bioinformatički alat *SEARCHPKS* (Yadav i sur., 2003) daje mogućnost korisniku da "učita" vlastitu sekvenciju proteina i upotrebom bioinformatičkog programa *BLAST* (Altschul i sur., 1990) identificira specifičnost domena u vlastitoj sekvenciji. Ovaj alat omogućuje jednostavnu identifikaciju različitih domena i modula enzima PKS unutar zadane polipeptidne sekvencije. Usto, predviđa specifičnost potencijalne acetiltransferazne domene za različite starter i produžujuće jedinice. Također omogućuje povezivanje kemijske strukture poliketidnih produkata sa organizacijom domena i modula u odgovarajućim modularnim poliketid sintazama. Računalni program *SEARCHPKS* može pomoći i u identifikaciji poliketidnih produkata nastalih djelovanjem PKS genskih nakupina pronađenih u novosekvencioniranim genomima. Računalni pristup upotrebljen u programu *SEARCHPKS* temelji se na sveobuhvatnim analizama različitih dobro opisanih genskih nakupina modularnih poliketid sintaza obuhvaćenih u bazi podataka PKSDB modularnih poliketid sintaza (Ansari i sur., 2004).

Drugi je koristan izvor podataka baza podataka ASMPKS (Tae i sur., 2007) s računalnim programom *MAPSI*, koji upotrebljava sličnu metodologiju i povezuje je s grafičkim prikazom upućujući na prisutnost domena unutar gena. Program *MAPSI* pruža mogućnost da se predvidi linearni poliketidni lanac, pri čemu korisnik mora odabrati građevnu jedinicu s ponuđene liste. Program, međutim, postojeću stereokemiju ne uzima u obzir. Tvrtka ECOPIA je razvila alat, *DecipherITTM* (Zazopoulos i sur., 2003), koji može pomoći pri anotaciji novih genskih nakupina temeljem usporedbe s poznatim genskim nakupinama. Razvijen je i sličan alat pod nazivom *Biogenerator* (Zotchev i sur., 2006), koji dozvoljava predviđanje novih poliketida na temelju poznatih modula u uvjetima *in silico*.

Svi su ovi programi ovisni o usporedbi novih sekvencija s poznatim sekvencijama i prikladani su za slične genske nakupine. Svi se oni, međutim, temelje na usporedbi sličnosti sekvencija i ne predviđaju aktivnost i specifičnost domena, te stereokemiju. Nedavno su opisana još dva računalna programa *CLUSEAN* (Weber i sur., 2009) i *NP.searcher* (Li i sur., 2009) koje za potrebe ovoga diplomskoga rada nisu analizirani.

2.3.2. 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, sekvencija DNA se automatski prevodi u šest otvorenih okvira čitanja sekvencije proteina pomoću programa *Transeq* (Rice i sur., 2000). Programski paket omogućuje traženje gena na temelju sekvencije 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 proteinskih profila preuzetih iz baze podataka Pfam (Finn i sur., 2008), ili pomoću vlastitih proteinskih profila, strogo definiranim ili relaksiranim parametrima. Nakon obavljene anotacije genske nakupine i biosintetskoga 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, 2009) odnosno EMBL (EBI, 2009) (Starcevic i sur., 2008).

Starčević i suradnici (2008) su usporedili uspješnost generičkog programskog paketa *ClustScan* (Tablica 1) 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 sekvencija DNA. Točnost predviđanja je demonstrirana na primjeru jedne dobro opisane genske nakupine što sadržava genetičku uputu za biosintezu poliketidnog antibiotika eritromicina (Starcevic i sur., 2008).

Tablica 1. Usporedba generičkog računalnog programskog paketa *ClustScan* sa sustavima *SEARCHPKS* i *MAPSI*.

SVOJSTVA	***CLUSTSCAN***	***SEARCHPKS***	***MAPSI***
Učitavanje sekvencije DNA	Da	Samo protein	Da[1]
Prepoznaje specifičnost domena AT	Da	Da	Da
Prepoznaje stereokemiju domena KR	Da	Ne	Ne
Prepoznaje neaktivne domene KR	Da	Ne	Ne
Prepoznaje neaktivne domene DH	Da	Ne	Ne
Prepoznaje 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 sekvencije 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 slijedećeg sklopovlja: prijenosno računalo HP: procesor Intel®Celeron® M CPU 530 @ 1,7 GHz, radne memorije 0,99 GB. Na čvrstom disku upotrijebljenog računala instaliran je operativni sustav "Microsoft Windows XP Professional".

3.1.2. Baze podataka

3.1.2.1. Baza podataka GenBank

Baza podataka GenBank osnovana je u listopadu 1992. godine i održava je Nacionalni centar za biotehnološke informacije (engl. "National Center for Biotechnology Information", NCBI). Ona sadržava dobro opisane, tj. anotirane, sekvencije DNA. U njoj se može pronaći 85.759.586.764 baza iz 82.853.685 analiziranih sekvencija DNA. Cjeloviti se pregled trenutne verzije baze podataka GenBank nalazi na poslužitelju centra NCBI (NCBI, 2009). Baza podataka GenBank dio je međunarodne suradnje (engl. "International Nucleotide Sequence Database Collaboration") koja pruža mogućnost neprestane razmjene i dostupnosti sekvencija 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 Nacionalnom centru za imunologiju (engl. "National Institute of Immunology", NII), New Delhi, Indija, sadržava računalni program za analiziranje velikih više-enzimskih, više-domenskih megasintetaza uključenih u biosintezu farmaceutski važnih prirodnih produkata (Ansari i sur., 2004). Razvoj baze podataka NRPS-PKS temelji se na sveobuhvatnim analizama sekvencija i strukturnih svojstava nekoliko eksperimentalno karakteriziranih biosintetskih genskih nakupina. Na internetskoj stranici baze podataka NRPS-PKS nudi se mogućnost jednostavne izolacije različitih domena iz vlastite polipeptidne sekvencije i određivanja katalitičkih aktivnosti domena, aktivnih mjesta, specifičnosti za supstrat te omogućuje usporedbu domena neopisanih NRPS/PKS genskih nakupina na temelju njihove sličnosti, specifičnosti za supstrat te motiva aktivnog mjesta.

Rezultati tih analiza organizirani su u četiri inegrirane 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. Međusobno su povezane kako bi se mogle provoditi analize genskih nakupina odgovornih za biosintezu poliketidno/peptidnih hibridinih 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 specifičnosti adenilacijskih i acetiltransferaznih domena sa relativno velikom točnošću. Ova svojstva bazu podataka NRPS-PKS čine vrijednim izvorom za identifikaciju prirodnih produkata biosintetiziranih pomoću NRPS/PKS genskih nakupina pronađenih u novosintetiziranim genomima. Baza NRPS-PKS preko sučelja omogućuje korisnicima povezivanje kemijskih struktura prirodnih spojeva sa 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 biosintezu novih prirodnih produkata. Bazi podataka NRPS-PKS moguće je pristupiti preko internetske stranice http://www.nii.res.in/nrps-pks.html (Anonymous 1, 2004).

3.1.2.3. Baza podataka Pfam

Baza podataka Pfam je velika zbirka višestruko poravnatih domena proteina koje su organizirane u obitelji proteina, u obliku skrivenih Markovljevih modela. Za svaku obitelj proteina 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 obitelji baze podataka Pfam. 74% sekvencija proteina ima barem jednu sličnu sekvenciju proteina unutar baze podataka Pfam. Taj broj se zove pokrivenost baze podataka. Baza podataka Pfam-A je ručno izrađeni dio baze podataka koji sadržava više od 9.000 unosa. Za svaki unos pohranjena su poravnanja proteinskih sekvencija i skiveni Markovljev model. Postoji još i dio baze podataka Pfam-B, koji sadržava veliki broj malih obitelji proteina slabije kvalitete podataka. Taj se dio baze podataka upotrebljava u slučajevima kada nisu pronađene sekvencije sa sličnošću unutar baze podataka Pfam-A (Finn i sur., 2008).

3.1.2.4. Institucija "The Sanger Institut"

Institucija "The Sanger Institut" je jedan od vodećih genetičkih centara, posvećen analiziranju i razumijevanju genoma (The Wellcome Trust Sanger Institute 1, 2009). Ova institucija podupire biološka i medicinska istraživanja putem projekata i suradnji, te sadržava bazu podataka o sekvencijama velikog broja genoma različitih organizama. Osnovana je 1993. godine od strane zaklade "The Wellcome Trust" i britanskog medicinskog vijeća (engl. "Medical Research Council", MRC). "The Wellcome Trust Sanger Institute" je neprofitna organizacija koja sve novčane donacije usmjerava u biomedicinska istraživanja.

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, archaea i virusa. Razvijen je i prvotno upotrijebljen u instituciji "The Institute for Genomic Reasearch", (TIGR), za anotaciju više od 100 različitih bakterijskih vrsta. Programom *Glimmer* može se konstruirati model za dijelove DNA s genetičkom uputom koristeći druge otvorene okvire čitanja u učitanoj (analiziranoj) sekvenciji DNA kao podatak za testiranje (engl. "training data"). Manje je učinkovit prilikom analize kratkih sekvencija DNA, kao i u sekvencija DNA s visokim G+C sastavom baza. Takvu DNA imaju upravo genomi streptomiceta. To su dugački dijelovi DNA koji ne sadržavaju genetičku uputu, pa mogu smanjiti preciznost predviđanja u dijelova DNA s genetičkom uputom (Delcher i sur., 2007).

3.1.3.2. Računalni program GeneMark-PS

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

3.1.3.3. Programski paket HMMER

Programski paket *HMMER* upotrebljava, za analizu pozicija pojedinih aminokiselina u sekvenciji proteina, profile proteina skrivenih Markovljevih modela (engl. "Hidden Markov

Models", HMM). Zbog toga 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 delecije, te im pridodaje stupanj vjerojatnosti (Eddy, 1998). Funkcija *hmmpfam* (sastavni dio paketa *HMMER*) čita sekvenciju po sekvenciju upita i traži sličnost sa nekim od proteinskih profila u bazi podataka. Profili proteina, koji se nalaze u bazi podataka koju funkcija *hmmpfam* čita, organizirani su u obitelji proteina (npr. obitelj proteina globina). Sve sekvencije unutar jedne obitelji proteina pokazuju značajan stupanj sličnosti. Uspješnost određivanja sličnosti točnija je prilikom usporedbe dviju sekvencija proteina nego u usporedbi dviju sekvencija DNA. Za promatranu sekvenciju moguće je, pomoću parametriziranih modela, odrediti koji model najvjerojatnije objašnjava podatke, tj. promatranu sekvenciju, te koji je najvjerojatniji put kroz tzv. "stanja" za sekvenciju i za model.

3.1.3.4. Programski paket ClustScan

Računalni generički programski paket *ClustScan* je razvijen 2008. godine u Kabinetu za bioinformatiku, Prehrambeno-biotehnloškog fakulteta, Sveučilišta u Zagrebu. Generički programski paket *ClustScan* upotrebljava se za anotaciju modularnih genskih nakupina sa na znanju utemeljenim predviđanjima aktivnosti i specifičnosti njihovih modula i katalitički aktivnih domena (Starcevic i sur., 2008). Tim se programskim paketom mogu anotirati genske nakupine u novosekvenciranim genomima i metagenomima i predvidjeti kemijsku strukturu produkta biosinteze iz sekvencije DNA genske nakupine. Računalni generički programski paket *ClustScan* može se preuzeti sa Web stranice http://bioserv.pbf.hr/cms/index.php?page=clustscan (Bioinformatics group, 2009).

3.2. METODE

3.2.1. Prikupljanje literaturnih podataka

Prilikom prikupljanja literaturnih podataka upotrebljavani su "online" pretraživači literature kao što su: "Google Scholar" (Google, 2009) i "Science direct" (Anonymous 3, 2009).

3.2.2. Prikupljanje sekvencija DNA i proteina

Kod prikupljanja primarnih sekvencija DNA i proteina, pristupano je već postojećim bazama podataka GenBank i NRPS-PKS putem Interneta. Za pristup bazama podataka

upotrebljavan je Web preglednik "Google Chrome". Baze podataka koje su uključene u "International Nucleotide Sequence Database Collaboration", a u koje spada i baza podataka GenBank, povezuje sustav za pretraživanje i prikupljanje podataka, Entrez. Odabirom "Entrez home" (NCBI; 2009) i upisivanjem ključne riječi u polje "Search across databases" pokreće se željeno pretraživanje (Slika 11).

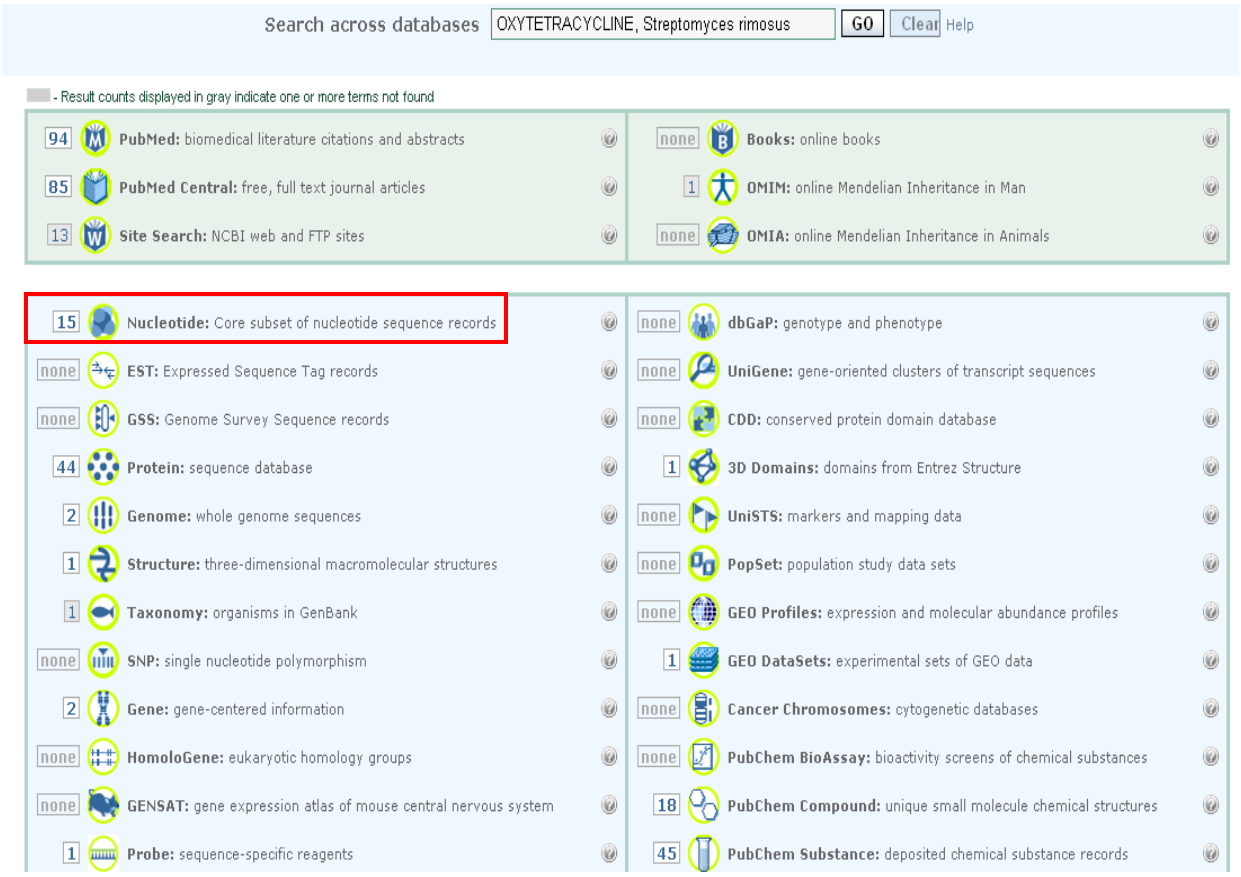
Odabirom sučelja "Nucleotide" (Slika 11) dobio se ispis svih pronađenih sekvencija DNA. Odabirom svakog pojedinog ispisa otvara se zapis baze podataka GenBank. Upisivanjem pristupnog broja (engl. "accession number") u polje "Search across databases" drugi je način na koji je pretraživana baza podataka GenBank pomoću sustava Entrez. Na primjer, upisivanjem pristupnog broja DQ143963 (pristupni broj baze podataka GenBank za poliketid sintazu oksitetraciklina, pronađen u prikupljenim literaturnim podacima) i pokretanjem pretraživanja, odmah se učitava traženi zapis baze podataka GeneBank.

Zapis baze podataka GeneBank (Slika 12) sadržava mnoštvo informacija, od raznih opisa sekvencije, pa do same sekvencije nukleotida. Ovaj oblik zapisa uz podatke sadržava i komentare, tj. pojašnjenja, koja pružaju mogućnost jednostavnijeg čitanja, identifikacije i upotrebe različitih tipova podataka iz baze.

Odabirom zapisa FASTA u izborniku "Format" (Slika 12) promijenjen je oblik zapisa (Slika 13). Zapis FASTA je jednostavan, minimalistički, zapis koji se sastoji od prvog retka zvanog zaglavlje (engl. "header"). Zaglavlje započinje sa znakom ">" koji slijedi kratki opis sekvencije. Nakon toga, u novom retku slijedi sama sekvencija DNA ili proteina. Zapis FASTA je računalni standard, stoga su sekvencije sačuvane u takvom obliku. Pretraživač Entrez također pruža mogućnost preuzimanja sekvencija proteina vezanih za svaku sekvenciju DNA (Slika 12) tako da su i one pohranjene.

Baza podataka NRPS-PKS poslužila je kao znanstveno utemeljeni izvor podataka za analizu poliketid sintaza i neribosomalno sintetiziranih peptida kao i poliketidno/peptidnih hibridnih produkata. Primjerice, pretraživanje baze PKSDB, koja se nalazi u sklopu baze NRPS-PKS, započinje odabirom jedne od ponuđenih biološki aktivnih supstancija s lijeve strane početne stranice (Slika 14). Moguće je pregledati kemijsku strukturu svake supstancije, te literaturne podatke (iz baze podataka PubMed), grafički prikaz modula i domena poliketid sintaza i neribosomalnih peptid sintetaza odgovornih za nastanak spoja. Za svaki se odabrani element (gen, modul, domenu i poveznicu) može dobiti sekvencija proteina u obliku zapisa

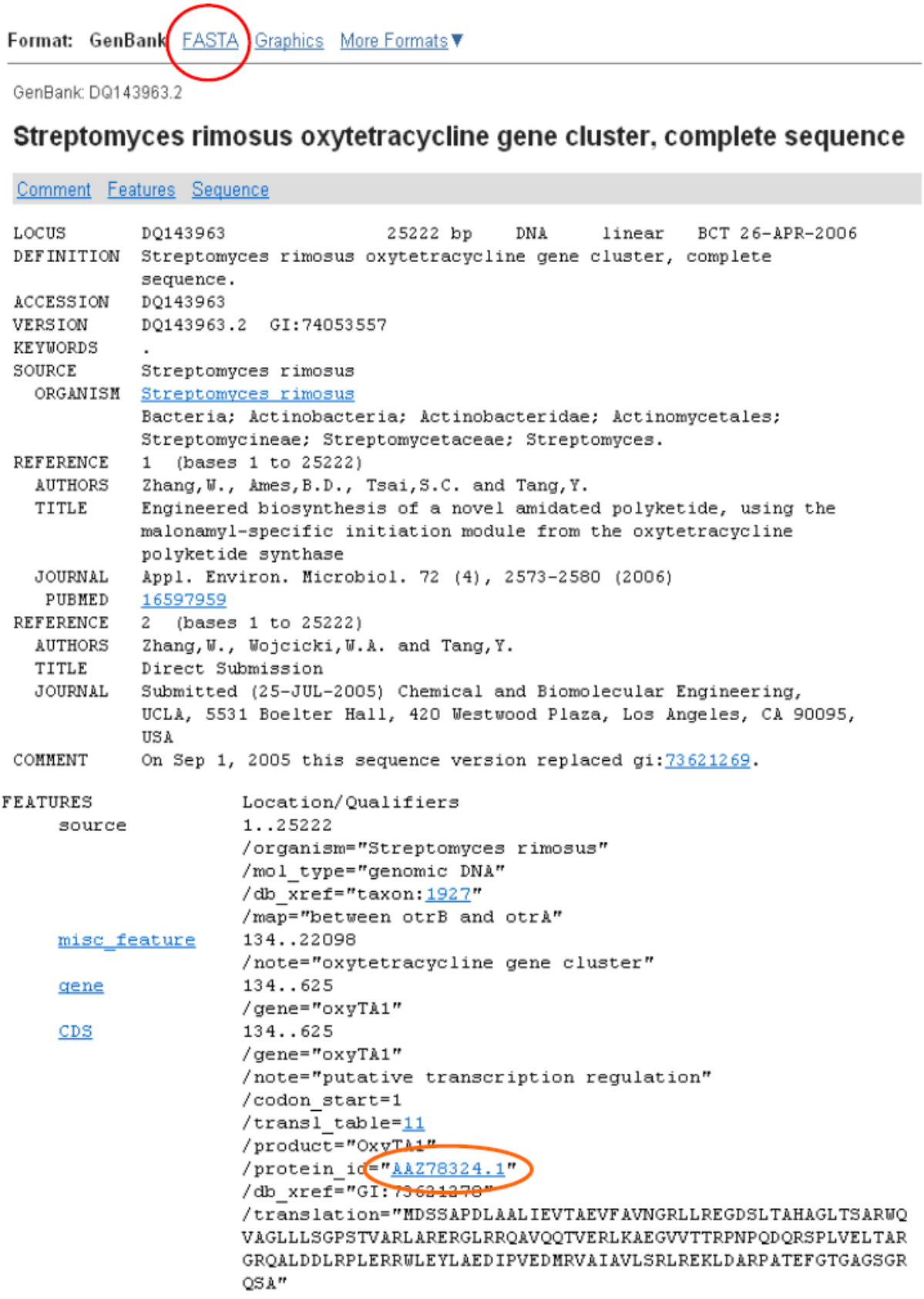
FASTA. Sve prikupljene sekvencije DNA i proteina u obliku zapisa FASTA nalaze se u prilogu (vidi: podpoglavlje 8.2.2.).



Slika 11. Izgled početne stranice sustava za pretraživanje i prikupljanje podataka, Entrez. Rezultati pretraživanja prikazuju se kao brojevi pronađenih pogodaka ispisani uz svaku bazu podataka.

3.2.3. Analiza sekvencija proteina programskim paketom *HMMER*

Sve prikupljene sekvencije proteina koje pripadaju istoj klasifikacijskoj skupini (unutar klasifikacijske tablice izrađene prema prethodno obrađenim literaturnim podacima, vidi: podpoglavlje 4.1., Tablica 2; podpoglavlje 8.2.2., Tablica 2P) pohranjene su u jedan zapis FASTA sa zajedničkim zaglavljem, spajanjem pojedinačnih sekvencija pomoću alata "WordPad". Nakon spajanja, svaki je zapis FASTA obrađen računalnim programom *Readseq* (EBI, 2009). Program *Readseq* upotrebljen je za validiranje zapisa FASTA (Slika 15). Sve se datoteke nalaze u prilogu (vidi: podpoglavlje 8.2.2.).



Format: GenBank FASTA Graphics More Formats▼

GenBank: DQ143963.2

Streptomyces rimosus oxytetracycline gene cluster, complete sequence

Comment Features Sequence

```
LOCUS       DQ143963               25222 bp    DNA     linear   BCT 26-APR-2006
DEFINITION  Streptomyces rimosus oxytetracycline gene cluster, complete
            sequence.
ACCESSION   DQ143963
VERSION     DQ143963.2  GI:74053557
KEYWORDS    .
SOURCE      Streptomyces rimosus
  ORGANISM  Streptomyces rimosus
            Bacteria; Actinobacteria; Actinobacteridae; Actinomycetales;
            Streptomycineae; Streptomycetaceae; Streptomyces.
REFERENCE   1  (bases 1 to 25222)
  AUTHORS   Zhang,W., Ames,B.D., Tsai,S.C. and Tang,Y.
  TITLE     Engineered biosynthesis of a novel amidated polyketide, using the
            malonamyl-specific initiation module from the oxytetracycline
            polyketide synthase
  JOURNAL   Appl. Environ. Microbiol. 72 (4), 2573-2580 (2006)
   PUBMED   16597959
REFERENCE   2  (bases 1 to 25222)
  AUTHORS   Zhang,W., Wojcicki,W.A. and Tang,Y.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-JUL-2005) Chemical and Biomolecular Engineering,
            UCLA, 5531 Boelter Hall, 420 Westwood Plaza, Los Angeles, CA 90095,
            USA
COMMENT     On Sep 1, 2005 this sequence version replaced gi:73621269.

FEATURES             Location/Qualifiers
     source          1..25222
                     /organism="Streptomyces rimosus"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:1927"
                     /map="between otrB and otrA"
     misc_feature    134..22098
                     /note="oxytetracycline gene cluster"
     gene            134..625
                     /gene="oxyTA1"
     CDS             134..625
                     /gene="oxyTA1"
                     /note="putative transcription regulation"
                     /codon_start=1
                     /transl_table=11
                     /product="OxyTA1"
                     /protein_id="AAZ78324.1"
                     /db_xref="GI:73621270"
                     /translation="MDSSAPDLAALIEVTAEVFAVNGRLLREGDSLTAHAGLTSARWQ
                     VAGLLLSGPSTVARLARERGLRRQAVQQTVERLKAEGVVTTRPNPQDQRSPLVELTAR
                     GRQALDDLRPLERRWLEYLAEDIPVEDMRVAIAVLSRLREKLDARPATEFGTGAGSGR
                     QSA"
```

Slika 12. Dio zapisa baze podataka GeneBank za sekvenciju DNA oksitetraciklina.

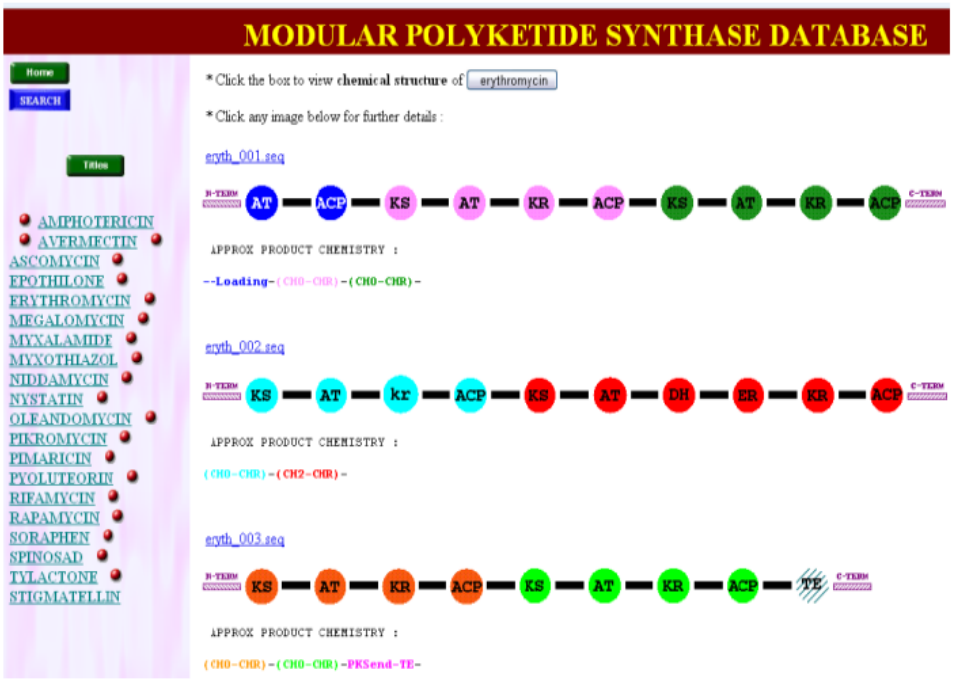
Format: GenBank FASTA Graphics More Formats▼

GenBank: DQ143963.2

Streptomyces rimosus oxytetracycline gene cluster, complete sequence

```
>gi|74053557|gb|DQ143963.2| Streptomyces rimosus oxytetracycline gene cluster,
complete sequence
GCGCGCGGCTGGGATCCCCCCCCGTACACACACGCTTGTTCACCATCGGCACCGAGGCCGGGTGCCGAGCG
TAATCACCGCCGCGGCTTTTGACAAGGTCTTGTCGTTCTCCGGGGAGTACATACGCTGTGGCCATGGATT
CCTCAGCCCCTGACCTGGCCGCTCTGATCGAGGTGACCGCCGAGGTCTTCGCGGTCAACGGCCGCCTGCT
CCGCGAAGGCGACAGCCTCACCGCCCACGCGGGGCTGACCTCGGCGCGCTGGCAGGTGGCCGGACTGCTG
CTGAGCGGCCCCTCGACGGTCGCCCGCCTGGCCCGCGAGCGGGGGCTGCGGCGGCAGGCGGTCCAGCAGA
CCGTCGAGCGGCTGAAGGCCGAGGGCGTCGTCACGACCCGGCCCAACCCGCAGGACCAGCGCAGCCCCCT
GGTCGAGCTCACCGCACGCGGCCGGCAGGCGCTGGACGACCTGCGTCCCCTGGAACGGCGGTGGCTGGAG
TATCTGGCCGAGGACATTCCGGTCGAGGACATGCGCGTGGCGATCGCGGTGCTGAGCCGCCTGCGGGAGA
AGCTGGACGCCCGTCCGGCGACGGAGTTCGGGACCGGGGCCGGGTCCGGGCGGCAGTCCGCCTGAGCGGC
CGCCGCCCCGCCCGGCCGTCTCTCCGGAGTGGCCCGGGCGTCGAATCCGCCTCCAAACCGCCTTGAGGTC
CGAAGGTCACTCTGTGCCTCGACGCGCTCGTGCCGCCCCCGGTGGCGCGACGACGGCAGAGGAGACTCAA
TGTCCAAGATCCATGACGCGCGACGCGTCGTAATCACAGGGATCGGCGTGGTGGCTCCGGGAGACGTGGG
CACCAAACCGTTCTGGGAGATGCTCACCGCGGGGCCGTACCGCCACCCGCCCGATCTCGTCCTTCGACGCC
TCGCCCTTCCGCTCCCAGGTGGCGGCGGAGTGCGACTTCGACCCGGCGGCGGAGGGGCTGAGCCAGCGGC
AGGTCCGCGCCTGGGACCGCACCATGCAGTTCGCTTACGTGGCCGCGCGGGAGGCGCTGGCGGACAGCGG
TGTGACGGGCGAGGCGGACCCGCTGCGCACCGGCGTCATGGCCGGCACCGCCTGCGGCATGACGATGAGC
CTGGACCGCGAGTACGCGGTGGTCAGCGACGAGGGCCGGCTGTGGCAGGTGGACGACGCCCATGGGGTGC
CGTACCTCTACGACTACTTCGTGCCCTCGTCGATGGCGGCCGAGATCGCCTGGCTGGCGGAGGCGGAGGG
CCCGGCCGGGGTGGTCTCGGCCGGCTGCACCTCCGGCATCGACGTGCTCACCCACGCGGCGGACCTGGTA
```

Slika 13. Dio zapisa FASTA baze podataka GeneBank za sekvenciju DNA oksitetraciklina.



Slika 14. Prikaz modula, domena i njihovih poveznica za eritromicin u bazi podataka PKSDB. Geni su prikazani odvojeno s modularnom organizacijom i specifično obojenim nazivima domena. Domene su šrafirane različitim uzorcima, ovisno o njihovom stanju (aktivna, neaktivna). Poveznice između domena i modula prikazane su punim crnim crtama.

Spremljene proteinske sekvencije u obliku točnog zapisa FASTA analizirane su korištenjem programskog paketa *HMMER* (Anonymous 4, 2009), točnije njegove funkcije *hmmpfam*. Pretraživanje je izvršeno čitavom bazom podataka profila proteina Pfam. Bazi podataka Pfam pristupljeno je putem Interneta (The Wellcome Trust Sanger Institute 2, 2009). Baza podataka je preuzeta u obliku jedne datoteke sa nekoliko stotina profila proteina HMM poznatih proteinskih domena. Upute za preuzimanje baze nalaze se na Internetskoj stranici.

Rezultati analize funkcijom *hmmpfam* sastoje se od nekoliko dijelova:

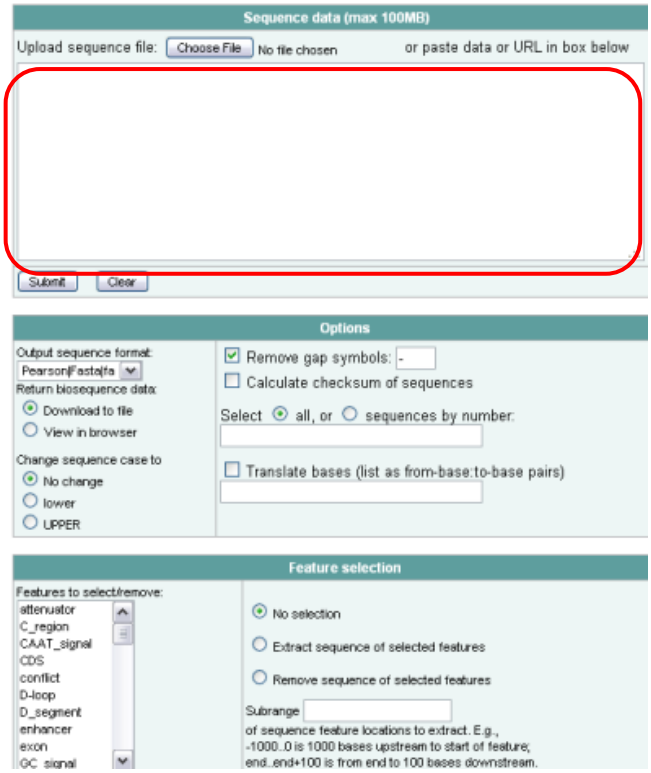
- zaglavlje,
- klasifikacijska lista analiziranih sekvencija,
- detaljan prikaz analiziranih domena redom kako se pojavljuju unutar sekvencije proteina, i
- izlazno poravnanje.

Zaglavlje donosi nekoliko osnovnih podataka o upotrijebljenom programu i sekvencijama proteina (engl. "sequence file"). Sekvencije su poredane na temelju vrijednosti E (engl. "E-value“, Slika 16, označeno crveno) i uspjehu pogotka (engl. "Score", Slika 16, označeno zeleno). Što je uspjeh pogotka viši, a vrijednost E manja, znači da je pronađen visok stupanj sličnosti između analizirane sekvencije i profila proteina, odnosno jedna ili više domena iz analizirane sekvencije pripadaju domenama iz obitelji proteina. Prikazuje se i broj pojavljivanja "N" (Slika 16, označeno plavo) pojedine domene unutar sekvencije.

Za svaku se domenu može saznati:

- broj domene (Slika 17, označeno crveno), npr. 1/54 za domenu ketoacyl-synt znači da je to prva domena po redu od 54 ukupno pronađene,
- početak (seq-f) i kraj (seq-t) domene unutar analizirane sekvencije, izraženo u parovima baza (Slika 17, označeno plavo),
- da li sekvencija pokazuje potpunu ili djelomičnu sličnost, na što ukazuju točkice (poklapanje sekvencija ne ide od početka do kraja) ili uglate zagrade (poklapanje sekvencija je potpuno) (Slika 17, označeno zeleno). Ukoliko se pojave dvije točkice znači da je poklapanje lokalno, negdje unutar sekvencije. Ukoliko se pojave dvije uglate zagrade poklapanje sekvencija je globalno, odnosno obuhvaćena je čitava sekvencija domene.

- početak, odnosno kraj poklapanja analizirane sekvencije sa sekvencijom u bazi podataka, odnosno modelom, što nam govore vrijednosti hmm-f i hmm-t, izražene u parovima baza (Slika 17, označeno ljubičasto).



Slika 15. Prozor programa *Readseq*. Spojene sekvencije proteina kopiraju se u prozor. Kao izlazni oblik sekvencije proteina (engl. "Output sequence format") odabran je zapis "Pearson|Fasta|fa", te opcija uklanjanja praznina (engl. "Remove gap symbols"). Sekvencija se nakon obrade neposredno sprema na čvrsti disk u obliku točnog zapisa FASTA.

Prvu liniju čini konsenzus model HMM (Slika 18, označeno crveno). Aminokiselina prikazana u sekvenciji konsenzus modela ima najvišu vjerojatnost da će se pojaviti na toj poziciji sudeći prema modelu HMM (nije nužno da je aminokiselina s najvišim uspjehom pogotka). Velika štampana slova označavaju visoko konzerviran dio čija je vjerojatnost > 0,5 za modele proteina odnosno > 0,9 za modele DNA. Središnja linija (Slika 18, označeno plavo) pokazuje samo aminokiseline u kojima se sekvencije poklapaju ili + kada pogodak ima pozitivan uspjeh pogotka i stoga se smatra konzerviran (sačuvan). Treća linija (Slika 18, označeno zeleno) pokazuje čitavu analiziranu sekvenciju. Izvješće o poravnanju može sadržavati dodatnu liniju "CS" (engl. "consensus structure", Slika 18, označeno ljubičasto)

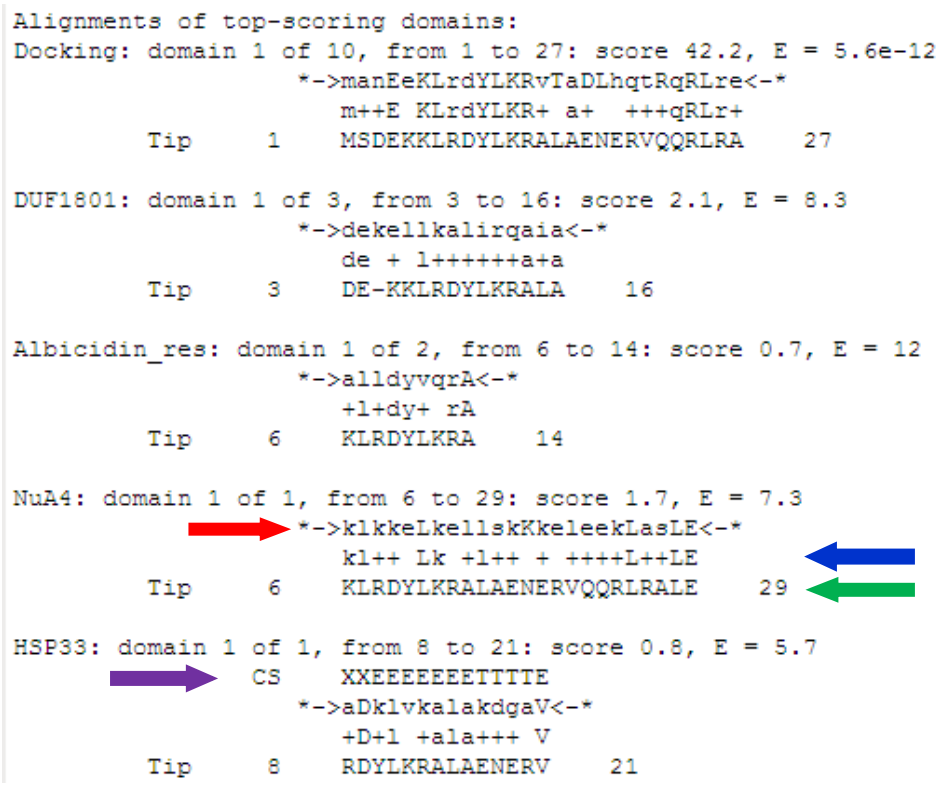
koja pokazuje da li se radi o strukturi zavojnice (engl. "helix", H), uzvojnice (engl. "coil", C) ili ploče (engl. "sheet", E).

```
hmmpfam - search one or more sequences against HMM database
HMMER 2.3.2 (Oct 2003)
Copyright (C) 1992-2003 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HMM file:                 databases\Pfam_fs
Sequence file:            sequences\all_PKS.fasta
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query sequence: PKS_Hibridi
Accession:      [none]
Description:    204784 bp

Scores for sequence family classification (score includes all domains):
Model           Description                               Score    E-value  N
--------        -----------                               -----    -------  ---
ketoacyl-synt   Beta-ketoacyl synthase, N-terminal do 26183.7          0   78
AMP-binding     AMP-binding enzyme                     25372.9          0   67
Ketoacyl-synt_C Beta-ketoacyl synthase, C-terminal do 14785.5          0   80
KR              KR domain                              14327.3          0   60
Condensation    Condensation domain                    13457.5          0   59
adh_short       short chain dehydrogenase              11339.1          0   69
Acyl_transf_1   Acyl transferase domain                10308.5          0   50
PP-binding      Phosphopantetheine attachment site      8527.1          0  161
```

Slika 16. Prikaz rezultata programskog paketa *HMMER*, zaglavlje i klasifikacijska lista analiziranih sekvencija.

```
Parsed for domains:
Model           Domain  seq-f seq-t    hmm-f hmm-t       score  E-value
--------        -------  -----  -----    -----  -----      -----  -------
Docking           1/10       1    27 [.       1    27 []    42.2  5.6e-12
DUF1801           1/3        3    16 ..      96   110 .]     2.1      8.3
Albicidin_res     1/2        6    14 ..      69    77 .]     0.7       12
NuA4              1/1        6    29 ..       1    24 [.     1.7      7.3
HSP33             1/1        8    21 ..       1    14 [.     0.8      5.7
BAR               1/2        9    27 ..     246   264 .]     0.5      8.4
VirE3             1/1       14    31 ..     292   309 ..     2.0      3.4
bZIP_2            1/3       16    31 ..      40    57 .]     2.0       14
Terminase_GpA     1/1       19    46 ..     572   601 ..    -0.7      5.3
Sigma70_ner       1/1       20    29 ..     218   227 .]    -0.2      9.3
PRP4              1/5       21    29 ..       1     9 [.     2.3       18
ketoacyl-synt     1/54      34   284 ..       1   300 []   437.9 1.5e-128
XRCC4             1/1       50    57 ..     327   334 .]    -0.3      9.5
GoLoco            1/2       52    61 ..       1    10 [.     0.8       38
P53_TAD           1/11      53    62 ..      16    25 .]     0.8       28
DUF681            1/1       55    62 ..      94   101 .]     0.4        9
Isochorismatase   1/2      120   128 ..       1     9 [.     1.4      4.3
TetR_N            1/18     127   146 ..       1    20 [.     2.4       10
Thiolase_N        1/54     198   233 ..      83   118 ..    18.6  0.00019
ACP_syn_III       1/34     204   229 ..       5    30 ..     8.4     0.16
TB                1/24     223   234 ..       1    13 [.     2.3        6
DUF1197           1/11     262   285 ..      57    80 .]     1.3       14
Trehalase         1/5      269   282 ..     566   579 .]     0.1      5.6
MdcE              1/8      284   300 ..     114   130 ..     2.0      3.2
Ketoacyl-synt_C   1/54     292   408 ..       1   126 []   240.5  4.3e-69
```

Slika 17. Prikaz rezultata programskog paketa *HMMER* i analize domena (engl. "domain parse") redom kako se domene pojavljuju unutar sekvencije proteina.



Slika 18. Prikaz rezultata programskog paketa *HMMER*, izlazno poravnanje (engl. "alignment output").

Analizom dobivenih rezultata izdvojene su domene koje su zadovoljavale postavljene uvjete: vrijednost $E < 10^{-5}$ i uspjeh pogotka > 0, bez obzira da li je obuhvaćena cijela domena ili je ona obuhvaćena samo djelomično. Na taj su način klasifikacijske tablice nadopunjene specifičnim domenama za svaku klasifikacijsku grupu i izrađeni profili proteina koji će kasnije poslužiti pri anotaciji genoma (vidi: podpoglavlje 4.2., Tablica 3; podpoglavlje 8.2.3., Tablica 3P; podpoglavlje 8.2.4.).

3.2.4. Pronalaženje strukturnih gena

Sekvencija genoma vrste *S. scabies* preuzeta je sa "Blast Servera" Sangerova Instituta, tj. sa njihove "FTP" stranice (The Wellcome Trust Sanger Institute 4,2009). Programski paket *ClustScan* dostupan je putem Interneta (Bioinformatics group, 2009) nakon registracije korisnika.

Nakon učitavanja DNA genoma vrste *S. scabies* u programski paket *ClustScan* odabirom opcije "*File/Import DNA*", DNA je automatski prevedena u šest otvorenih okvira čitanja pomoću programa *Transeq* (Rice i sur., 2000). Sama analiza sekvencije DNA odvija se

na Linux serveru [svaki korisnik ima svoju lozinku (engl. "password")]. Lozinka korisniku omogućuje: pristup vlastitoj radnoj površini (engl. "workspace"), učitavanje sekvencije DNA i provođenje analize, dok se rezultati pohranjuju na serveru ili na čvrstom disku. Stoga je svaku analizu potrebno provesti samo jednom. To je važno jer pretraživanje jednog genoma, poput genoma vrste *S. scabies*, može trajati nekoliko sati.

Genom je zatim obrađen bioinformatičkim programom *Glimmer*, koji je dio programskog paketa *ClustScan*, odabirom opcije "*Tools/Search for genes with custom model*". Nakon analize rezultata koje je ponudio program *Glimmer*, kao model je odabrana bakterija *S. avermitilis* i provedena je analiza programom *GeneMark-PS*. Analiza genoma pomoću programa *GeneMark* pokrenuta je odabirom opcije "*Tools/Search for a genes with a model/Species: Streptomyces avermitilis*".

3.2.5. Preuzimanje gotovih profila proteina iz baze podataka Pfam

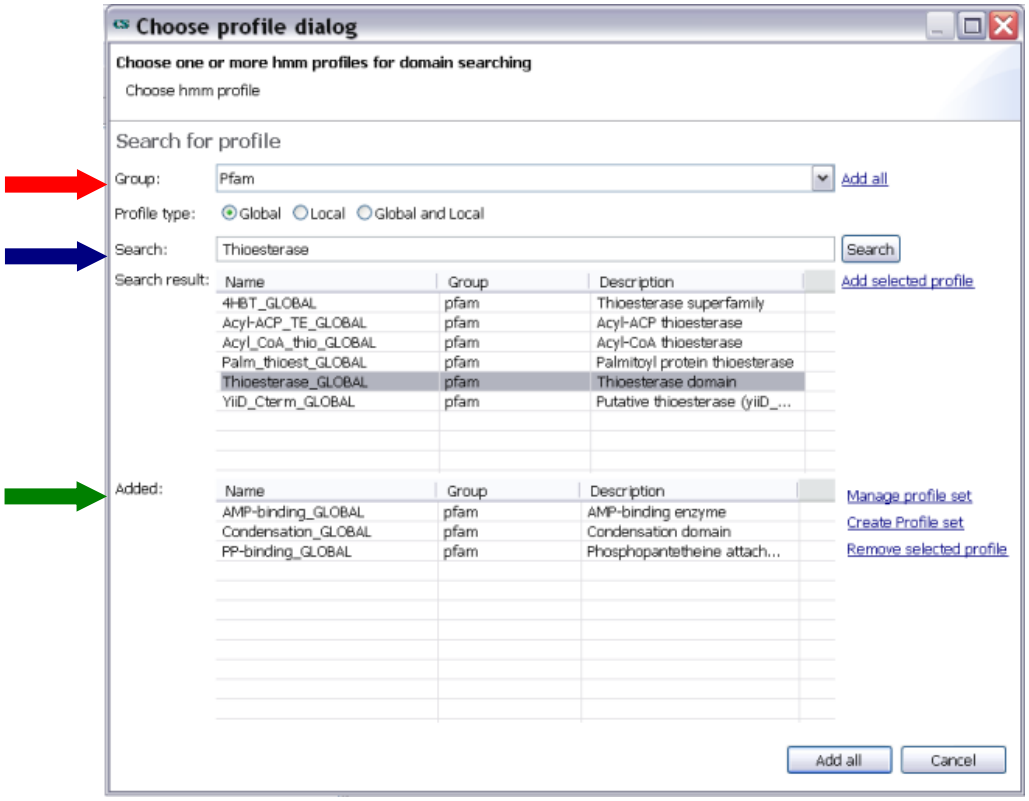
Baza profila proteina za pretraživanje genoma vrste *S. scabies* izrađena je upotrebom baze podataka Pfam pomoću programskog paketa *ClustScan* odabirom opcije "*Tools/Search for domains*". Na učitanoj kartici izabrana je DNA (bakterije *S. scabies*) i opcija "*Add profiles/Search*" (Slika 19).

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 prihvaća, označava se klikom miša i odabire se opcija "Add selected profile", te automatski prikazuje pod izbornikom "Added" (Slika 19, označeno zeleno). Nakon učitavanja svih domena, odabrana je opcija kreiranja profila proteina (engl. "create profile set"). Profil proteina zatim je učitan odabirom opcije "Manage profile set". Podešeni su parametri programskog paketa *HMMER* (opcija "stringent") i pokrenuto je pretraživanje.

3.2.6. 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 sekvencija 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 je sadržavala nekoliko primjeraka homolognih sekvencija. Sekvencije koje čine jednu datoteku zalijepljene su u za to predviđeni prozor (Slika 20, označeno crveno) i odabrana je opcija provjere zapisa (sekvencije 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 sekvencija DNA ili proteina, *ClustalW* (Anonymous 2, 2009). Nakon poravnanja program sam izrađuje profil proteina i daje mogućnost njegovog imenovanja.



Slika 19. Kartica za odabir domena i izradu profila proteina programom *ClustScan*.

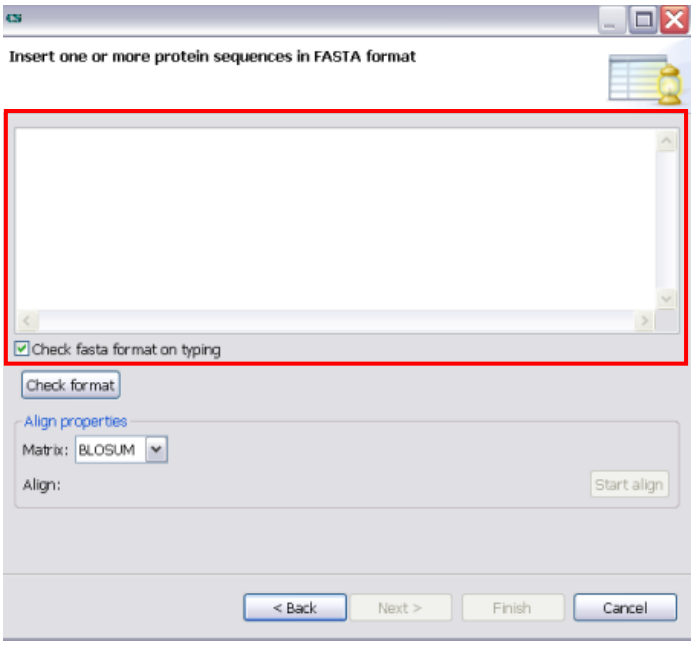
3.2.7. Anotacija genoma bakterije *Streptomyces scabies* pomoću programa *ClustScan*

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

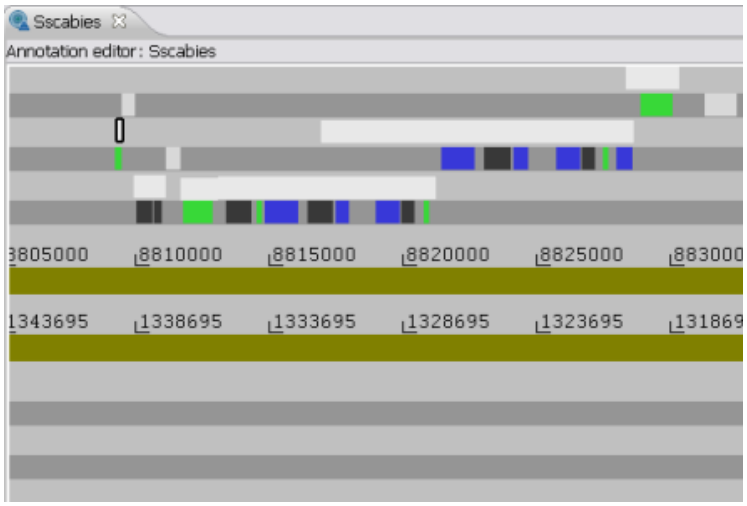
- liste ili stabla, tj. prozor radne površine (engl. "Workspace window") prikazuje sve pronađene proteine, strukturirane u tri okvira čitanja s lijeva na desno (engl. "forward") i

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, bijeli pravokutnici) i proteinske domene (Slika 21, svaka proteinska domena prikazana je drugom bojom pravokutnika).



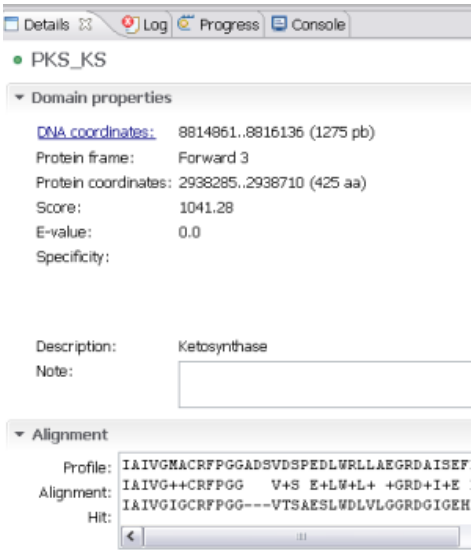
Slika 20. Kartica programskog paketa *ClustScan* za izradu vlastitih profila proteina.



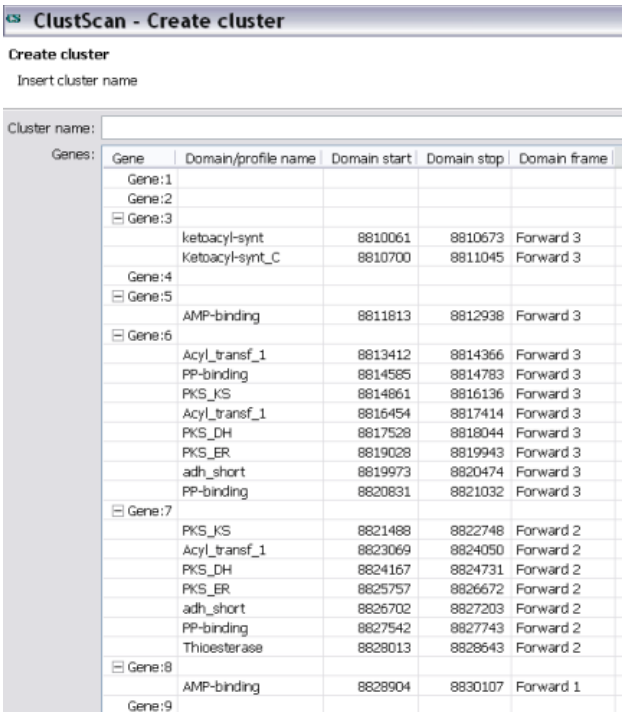
Slika 21. Prozor programskog paketa *ClustScan* za uređivanje rezultata anotacije.

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

Sve domene čiji parametri programskog paketa *HMMER* ne zadovoljavaju postavljane uvjete (vrijednost E, uspjeh pogotka) su izbrisane, a zatim su definirane genske nakupine. Postupak definiranja genske nakupine je vrlo jednostavan. Najprije se označi lijevom tipkom miša prvi gen, zatim se pritisne tipka "ctrl" i označi zadnji gen u genskoj nakupini. Klikom desne tipke miša na zadnji gen pojavljuje se opcija "Create cluster" (Slika 23). Na kartici (Slika 23) su ispisani svi geni obuhvaćeni u gensku nakupinu, proteinske domene unutar njih te položaj unutar sekvencije DNA. Kreiranje genske nakupine završava se njezinim imenovanjem (upisom imena u polje "Cluster name"). Nakon kreiranja genske nakupine učitava se prozor za uređivanje genskih nakupina, takozvani "Cluster editor" (Slika 24).

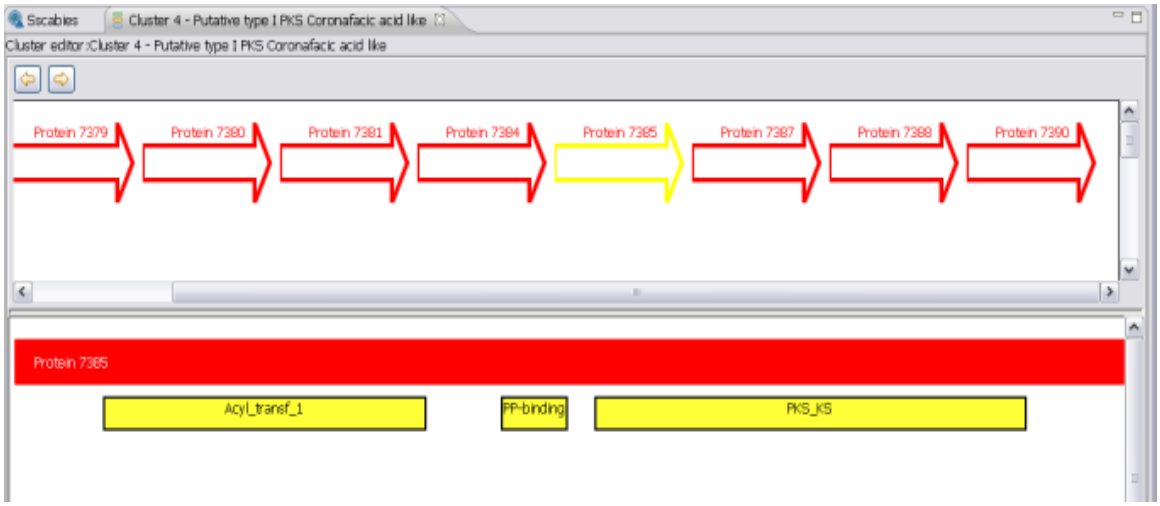


Slika 22. Prozor programa *ClustScan* s detaljnim informacijama o svojstvima pojedine domene.



Gene	Domain/profile name	Domain start	Domain stop	Domain frame
Gene:1				
Gene:2				
Gene:3				
	ketoacyl-synt	8810061	8810673	Forward 3
	Ketoacyl-synt_C	8810700	8811045	Forward 3
Gene:4				
Gene:5				
	AMP-binding	8811813	8812938	Forward 3
Gene:6				
	Acyl_transf_1	8813412	8814366	Forward 3
	PP-binding	8814585	8814783	Forward 3
	PKS_KS	8814861	8816136	Forward 3
	Acyl_transf_1	8816454	8817414	Forward 3
	PKS_DH	8817528	8818044	Forward 3
	PKS_ER	8819028	8819943	Forward 3
	adh_short	8819973	8820474	Forward 3
	PP-binding	8820831	8821032	Forward 3
Gene:7				
	PKS_KS	8821488	8822748	Forward 2
	Acyl_transf_1	8823069	8824050	Forward 2
	PKS_DH	8824167	8824731	Forward 2
	PKS_ER	8825757	8826672	Forward 2
	adh_short	8826702	8827203	Forward 2
	PP-binding	8827542	8827743	Forward 2
	Thioesterase	8828013	8828643	Forward 2
Gene:8				
	AMP-binding	8828904	8830107	Forward 1
Gene:9				

Slika 23. Kartica programa *ClustScan* sa detaljnim prikazom gena koji čine jednu gensku nakupinu.



Slika 24. Prozor programskog paketa *ClustScan* za uređivanje genskih nakupina.

Svaki gen (Slika 24) prikazan je crvenom strelicom koja ujedno i pokazuje njegov smjer čitanja, odabirom pojedinog gena učitava se dodatni prozor koji prikazuje domene unutar označenog gena. Nakon anotacije genoma bakterije *S. scabies*, "Workspace" sa pronađenim genskim nakupinama sačuvan je na čvrstom disku (vidi: podpoglavlje 8.2.5.).

4. REZULTATI

4.1. REZULTATI ANALIZE LITERATURNIH PODATAKA

Pretraživanjem literature, te baza podataka GenBank i NRPS-PKS, prikupljeno je 85 sekvencija DNA. U Tablici 2 prikazani su neki tipični primjeri. Cjelovit se popis prikupljenih sekvencija nalazi u Tablici 2P zajedno sa sekvencijama DNA u obliku zapisa FASTA (vidi: podpoglavlje 8.2.2.). Na sustave PKS otpada 73 primjera. Tip I čini 18 sekvencija DNA, od toga je 6 sekvencija iz modularnog, a 12 iz ponavljajućeg podsustava (7 bakterijskih i 5 fungalnih sekvencija). Na tip II otpada 13 primjera, dok su 42 primjera tipa III. Od toga je 21 sekvencija DNA biljnog, a 21 bakterijskog porijekla (7 RppA i 14 Naringenin sekvencija). Preostali dio (12 sekvencija) čine hibridni sustavi PKS/NRPS.

Na temelju prikupljenih sekvencija DNA izrađena je klasifikacijska podjela sustava PKS (Slika 25). Podjela se temelji na zajedničkim enzimskim sustavima koji sudjeluju u biosintezi konačnih produkata. Tako su sustavi PKS podijeljeni na tri glavna tipa – tip I, II i III, ovisno o enzimskom sastavu genskih nakupina u prikupljenim sekvencijama poliketid sintaza (Slika 25). Sustavi PKS tipa I dijele se na ponavljajuće i modularne podsustave. Prilikom prikupljanja sekvencija DNA i proteina enzima ponavljajućeg tipa I prikupljeni su enzimi bakterijskog i fungalnog porijekla, ali u Tablici 2 nisu odvojeni. Enzimi PKS tipa III podijeljeni su na temelju porijekla sekvencija – na bakterijske i biljne, s naglaskom na to da se bakterijski enzimi PKS tipa III dodatno razlikuju po funkciji konačnih produkata (ovisno o tome da li kataliziraju biosintezu pigmenata ili antioksidansa) pa je i ta činjenica istaknuta u tablici. U posebnu su skupinu u tablici smješteni hibridni sustavi PKS/NRPS.

4.2. REZULTATI ANALIZE SEKVENCIJA PROTEINA PROGRAMSKIM PAKETOM *HMMER*

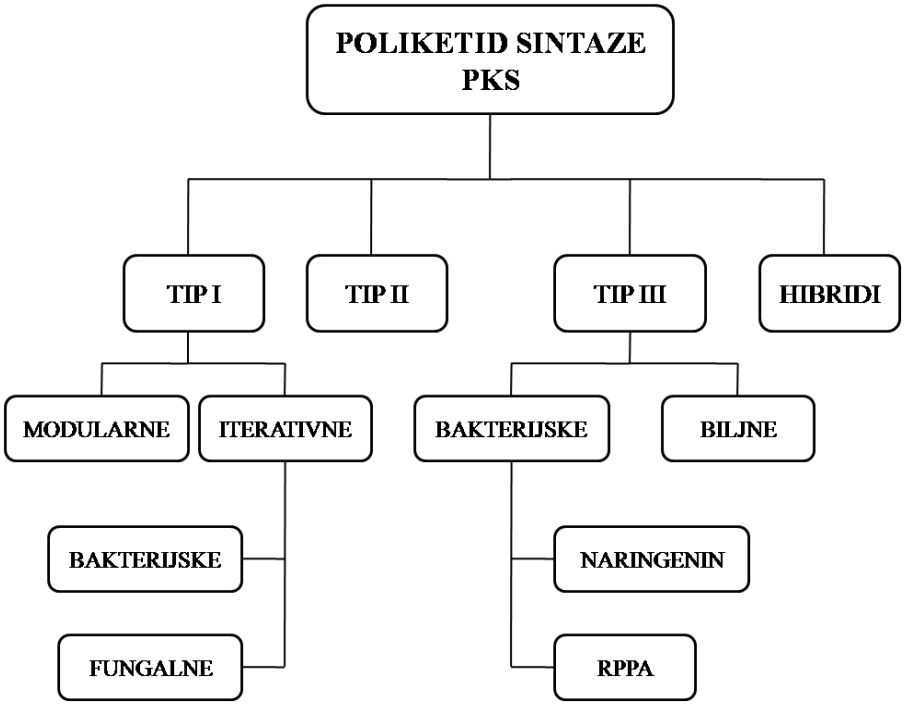
Analizom sekvencija proteina pomoću programskog paketa *HMMER* dobiveni su rezultati na temelju kojih je izrađena tablica profila proteina. U Tablici 3 prikazani su samo neki tipični primjeri. Cjelovit se popis profila proteina nalazi u Tablici 3P zajedno sa pripadajućim matricama (vidi: podpoglavlje 8.2.3. i 8.2.4.). Tablica 3P obuhvaća domene svojstvene za sve skupine iz klasifikacijske tablice. Prilikom analize rezultata u obzir su uzete samo domene čiji su parametri zadovoljavali zadane uvjete: vrijednost $E < 10^{-5}$ i uspjeh pogotka > 0. Na taj su način izdvojene sve zadovoljavajuće domene koje su zatim raspoređene u tablicu domena podijeljenu prema prethodno napravljenoj klasifikaciji sustava PKS.

Kao primjer rezultata dobivenih programskim paketom *HMMER*, naveden je rezultat dobiven za enzime PKS tipa II (Tablica 3; Slika 26). Ostatak rezultata nalazi se u prilozima (vidi: podpoglavlje 8.2.3. i 8.2.4.).

Tablica 2. Primjer klasifikacije poliketida i hibridnih sustava PKS/NRPS.

TIP	SEKUNDARNI METABOLIT	PROIZVODNI MIKROORGANIZAM	GENBANK PRISTUPNI BR.	BIOLOŠKA AKTIVNOST
Tip I				
Modularne				
	Ansamitocin	*Actinosynnema pretiosum*	AF453501	Citostatik
	Aurafuron	*Stigmatella aurantiaca*	AM850130	Fungicid
	Lipomicin	*Streptomyces aureofaciens*	DQ176871	Antibiotik
Ponavljajuće				
	Maduropeptin	*Actinomadura madurae*	AY271660	Citostatik
	Kompactin	*Penicillium citrinum*	AB072893	Antibiotik
	Lovastatin	*Aspergillus terreus*	AF151722	Antikolesterolemik
Tip II				
	Aklacinomicin	*Streptomyces galilaeus*	AF257324	Citostatik
	Daunorubicin	*Streptomyces peucetius*	L35560	Citostatik
	Oksitetraciklin	*Streptomyces rimosus*	DQ143963	Antibiotik
Tip III				
Bakterijske				
	-	*Streptomyces griseus*	AB018074	Smeđi pigment
	Tetrahidroksinaftalen	*Streptomyces lividans*	AB084491	Melanin
	Balhimicin	*Amycolatopsis balhimycina*	Y16952	Antibiotik
Biljne				
	Valerofenon	*Humulus lupulus*	AB015430	Komponenta arome
	Benzofenon	*Hypericum androsaemum*	AF352395	Antivirotik
	Akridon	*Ruta graveolens*	AJ297788	Antivirotik
PKS/NRPS hibridi				
	Barbamid	*Lyngbya majuscula*	AF516145	Antibiotik
	Bleomicin	*Streptomyces verticillus*	AF210249	Citostatik
	Epotilon	*Sorangium cellulosum*	AF217189	Citostatik
	Jersiniababaktin	*Yersinia pestis*	AF091251	Siderofor
	Mikosubtilin	*Bacillus subtilis*	AF184956	Antibiotik
	Zwittermicin A	*Bacillus cereus*	AF155831	Antibiotik
	Iturin	*Bacillus subtilis*	AB050629	Antibiotik

Iz Tablica 3 i 3P moguće je iščitati specifične domene za svaku klasifikacijsku skupinu, ali i koliko se puta određena sekvencija proteina ponavlja. Također je moguće iščitati da li je promatrana domena u dobivenim rezultatima cjelovita ili djelomična. Primjer za više puta ponovljenu domenu je domena AT (Acyl_transf_1) ponovljena 9 puta u rezultatima za enzime PKS tipa II.



Slika 25. Klasifikacija sustava PKS.

```
Query sequence: PKS_Tip
Accession:      [none]
Description:    II 36728 bp

Scores for sequence family classification (score includes all domains):
Model            Description                              Score    E-value  N
--------         -----------                              -----    ------- ---
ketoacyl-synt    Beta-ketoacyl synthase, N-terminal do  6516.1          0  25
Ketoacyl-synt_C  Beta-ketoacyl synthase, C-terminal do  3931.8          0  34
FAD_binding_3    FAD binding domain                     2361.8          0  10
adh_short        short chain dehydrogenase              1493.7          0  14
Methyltransf_2   O-methyltransferase                    1030.6          0   4
Acyl_transf_1    Acyl transferase domain                 632.6   3.9e-187   9
Cyclase          Putative cyclase                        624.2   1.3e-184   2
PP-binding       Phosphopantetheine attachment site      593.8   1.8e-175  12
ACP_syn_III_C    3-Oxoacyl-[acyl-carrier-protein (ACP)   496.8   2.9e-146  24
Polyketide_cyc2  Polyketide cyclase / dehydrase and li   472.0   8.3e-139  17
Cyclase_polyket  Polyketide synthesis cyclase            461.2   1.5e-135   3
Polyketide_cyc   Polyketide cyclase / dehydrase and li   404.9   1.3e-118  18
AMP-binding      AMP-binding enzyme                      364.6   1.9e-106   4
Asn_synthase     Asparagine synthase                     363.4   4.1e-106   1
ABM              Antibiotic biosynthesis monooxygenase   356.8   3.9e-104   9
Amidase          Amidase                                 345.3   1.2e-100   2
Methyltransf_12  Methyltransferase domain                322.9    6.7e-94  11
CPSase_L_D2      Carbamoyl-phosphate synthase L chain,   321.6    1.6e-93   1
Epimerase        NAD dependent epimerase/dehydratase f   321.0    2.5e-93  20
BTAD             Bacterial transcriptional activator d   307.0    3.9e-89   4
Methyltransf_11  Methyltransferase domain                300.2    4.3e-87  10
dTDP_sugar_isom  dTDP-4-dehydrorhamnose 3,5-epimerase    274.3    1.3e-81   1
KR               KR domain                               281.2    1.3e-81  13
ACP_syn_III      3-Oxoacyl-[acyl-carrier-protein (ACP)   226.6    6.2e-65   8
Biotin_carb_C    Biotin carboxylase C-terminal domain    196.6      2e-60   2
DAO              FAD dependent oxidoreductase            207.4    3.7e-59  13
DUF1205          Protein of unknown function (DUF1205)   180.2    3.6e-58   1
SnoaL            SnoaL-like polyketide cyclase           169.1    1.4e-51   6
CPSase_L_chain   Carbamoyl-phosphate synthase L chain,   162.6    1.6e-46   1
```

Slika 26. Enzimski sastav analiziranih sekvencija proteina programskim paketom *HMMER* za sustave PKS tipa II ukazuje na postojanje jasno definiranih katalitički aktivnih podjedinica – domena.

Tablica 3. Prikaz primjera domena PKS tipa II za izradu vlastitih profila (**Napomena:** nazivi domena preuzeti su neposredno iz programskog paketa *HMMER*).

	VIŠE PUTA PONOVLJENE DOMENE, CIJELE SEKVENCIJE	**JEDINSTVENE DOMENE, CIJELA SEKVENCIJA**	**JEDINSTVENE ILI VIŠE PUTA PONOVLJENE DOMENE, DJELOMIČNA SEKVENCIJA**
PKS			
TIP II	ketoacyl-synt Ketoacyl-synt_C FAD_binding_3 adh_short Methyltransf_2 Cyclase PP-binding ACP_syn_III_C Polyketide_cyc2 Cyclase_polyket Polyketide_cyc AMP-binding ABM Methyltransf_12 BTAD Methyltransf_11 ACP_syn_III Biotin_carb_C SnoaL Trans_reg_C ECH Pentapeptide Lactamase_B Thiolase_C Cupin_2	Asn_synthase CPSase_L_D2 dTDP_sugar_isom DUF1205 CPSase_L_chain Aminotran_1_2 Omt_N BBE Biotin_lipoyl Pyridox_oxidase	Acyl_transf_1 Amidase Epimerase FAD_binding_4 Carboxyl_trans Glyco_transf_28 ACCA

To znači da je programski paket *HMMER* u spojenoj sekvenciji proteina, sastavljenoj od svih prikupljenih sekvencija proteina za enzime PKS tipa II, pronašao točno toliko domena AT. Isto tako, sekvencije proteina ponovljene samo jednom čine jedinstvene domene (primjerice domena Asn_synthase (AS) u enzima PKS tipa II). Cjelovite sekvencije su sekvencije koje se podudaraju sa profilima proteina u bazi podataka Pfam i obuhvaćaju čitavu duljinu sekvencije. Djelomične sekvencije su sekvencije koje se samo djelomično preklapaju sa profilima proteina u bazi podataka Pfam, te ne obuhvaćaju cijelu dužinu sekvencije.

A

Family: *ketoacyl-synt* (PF00109)

Beta-ketoacyl synthase, N-terminal domain

Seed source:	Dotter
Previous IDs:	none
Type:	Domain
Author:	Sonnhammer ELL, Griffiths-Jones SR
Number in seed:	167
Number in full:	6443
Average length of the domain:	214.7 aa
Average identity of full alignment:	29%
Average coverage of the sequence by the domain:	11.25%

B

```
HMMER2.0  [2.3.2]
NAME  ketoacyl-synt
ACC   PF00109.18
DESC  Beta-ketoacyl synthase, N-terminal domain
LENG  300
ALPH  Amino
RF    no
CS    yes
MAP   yes
COM   hmmbuild -F HMM_ls.ann SEED.ann
COM   hmmcalibrate --seed 0 HMM_ls.ann
NSEQ  167
DATE  Fri Apr 25 14:33:57 2008
CKSUM 4328
GA    -73.6000 -73.6000;
TC    -73.3000 -73.3000;
NC    -73.8000 -73.8000;
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT      -4  -8455
NULE     595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313     45    531    201    384  -1998   -644
EVD   -189.870453   0.132467
HMM        A      C      D      E      F      G      H      I      K      L      M      N      P      Q      R      S      T      V      W      Y
         m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
          -18      *  -6340
     1   -925   -496    587   2214  -1217  -1044  -1716  -5070    269  -5015    200   -106   -396  -2700   1746   -308     -1  -1081  -5182  -1884      2
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11488 -12530   -894  -1115   -701  -1378    -18      *
     2  -1266  -4999    603    355  -5320   -792  -3159  -5071   -725  -3571  -4088  -3136   2813   -972   2215   -268  -1461  -2581  -1148  -2040      3
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -    -45 -11488  -5037   -894  -1115   -701  -1378      *      *
     3   -380  -5155  -8368  -8056  -5889  -8228  -8301   2794  -8035  -1284  -1909  -7880  -7959  -8016  -8223  -7611  -2151   2889  -7668  -7124      4
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     E     -1 -11443 -12485   -894  -1115   -671  -1427      *      *
```

Slika 27. Prikaz detalja o porodici proteina *ketoacyl_synt* (PF00109): broj sekvencija proteina koje čine temeljno poravnanje, broj sekvencija proteina koje čine potpuno poravnanje, prosječna duljina domene izražena u broju aminokiselina, prosječna identičnost potpunog poravnanja; prosječna pokrivenost domene sekvencijom proteina (A). Dio profila programskog paketa *HMMER* za protein *ketoacyl_synt* (PF00109) u obliku matrice (B).

Iz baze podataka Pfam ekstrahirani su profili proteina za sve katalitički aktivne domene svih opisanih tipova enzima PKS i sintaza poliketidno/peptidnih hibrida navedenih u Tablici 3P (vidi: podpoglavlje 8.2.3. i 8.2.4.). Na Slici 27 su, kao primjer, prikazani detalji o porodici domena "beta-ketoacyl synthase", to jest N-terminalnog kraja te domene [*ketoacyl_synt* (PF00109)]. Broj sekvencija domena prisutnih u potpunom poravnanju iznosi 167. Broj sekvencija domena prisutnih u temeljnom poravnanju iznosi 6.443. Prosječna dužina domene jest 214.7 amino kiselina. Prosječna identičnost amino kiselina u potpunom poravnanju iznosi 29%, dok je prosječna pokrivenost sekvencije proteina tom domenom 11,25% (Slika 27A). Također je prikazan i dio profila programskog paketa *HMMER* za domenu *ketoacyl_synt* (PF00109) u obliku matrice (Slika 27B).

4.3. REZULTATI ANOTACIJE GENOMA BAKTERIJE *Streptomyces scabies* POMOĆU PROGRAMSKOG PAKETA *ClustScan*

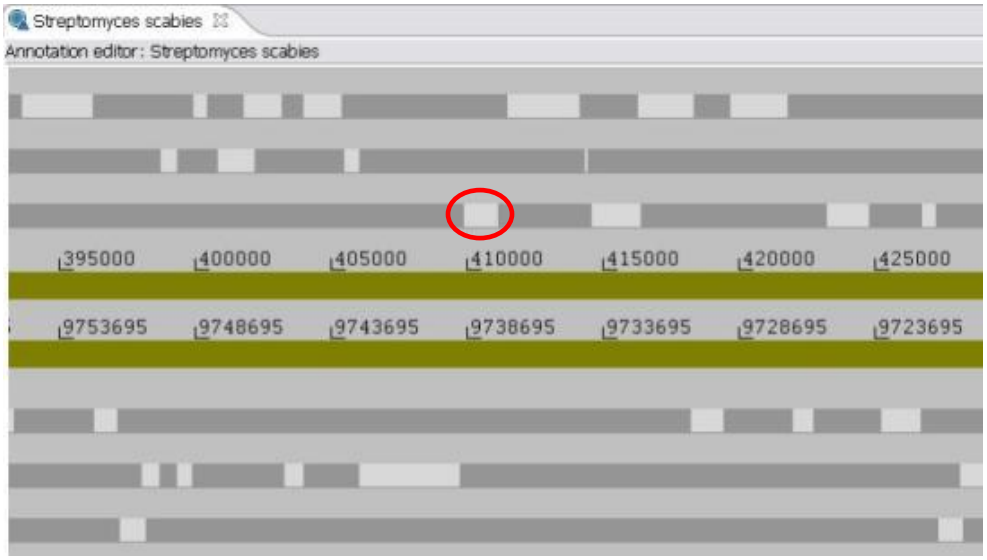
Primjenom programskog paketa *ClustScan* dobiveni su slijedeći rezultati:

- obradom genoma bakterije *S. scabies* bioinformatičkim alatom za predviđanje gena *Glimmer*, pronađeno je 9462 gena (Slika 28), a upotrebom programa *GenMark-PS* 8446 gena,
- uz upotrebu prethodno izdvojenih domena, u svih 6 otvorenih okvira čitanja, izdvojeno je 1322 proteina (Tablica 4) kao kandidata za hibridne sustave PKS/NRPS.

Anotirano je ukupno 7 sustava, od toga je:

- 3 poliketid sintaza,
- 4 hibridnih sustava.

Detaljan prikaz nalazi se u Tablici 5.



Slika 28. Prikaz rezultata programskog paketa *ClustScan*. Grafički rezultat prikazuje položaj gena, koji su prikazani kao bijeli pravokutnici unutar šest otvorenih okvira čitanja. Okviri čitanja prikazani su kao sive linije dok zelene linije predstavljaju čitavu sekvenciju DNA. Položaj gena na sekvencijama izražen je u parovima baza.

Tablica 4. Broj proteina pronađen u svih 6 otvorenih okvira čitanja.

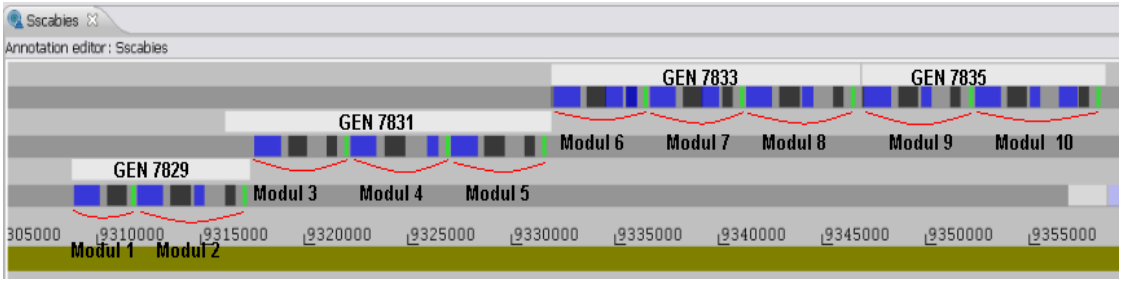
TIP OTVORENOG OKVIRA ČITANJA	BROJ ODREĐENIH PROTEINA
S DESNA NA LIJEVO 1	241
S DESNA NA LIJEVO 2	235
S DESNA NA LIJEVO 3	269
S LIJEVA NA DESNO 1	210
S LIJEVA NA DESNO 2	182
S LIJEVA NA DESNO 3	185

Tablica 5. Raspodjela genskih nakupina sekundarnih metabolita u genomu bakterije *S. scabies*.

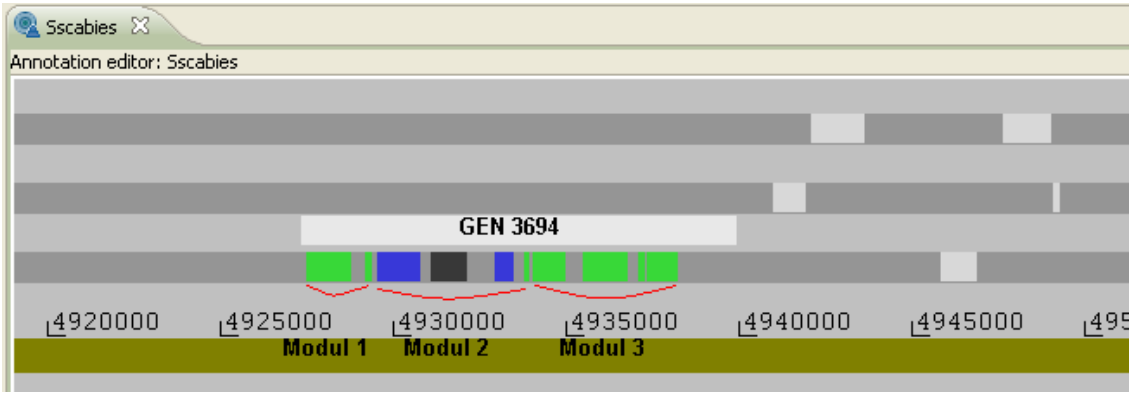
GENSKA NAKUPINA		POČETAK	KRAJ	BROJ GENA	DUŽINE GENSKIH NAKUPINA (pb)
POLIKETID SINTAZE					
1.	Pretpostavljeni modularni tip I	9308805	9379965	6	71160
2.	Pretpostavljeni ponavljajući tip I	8720511	8724597	1	4086
3.	Pretpostavljeni tip II	4866093	4871179	6	5086

HIBRIDI					
4.	Pretpostavljeni PKS tip II/NRPS	7000340	7023427	18	23087
5.	Pretpostavljeni PKS tip I/PKS tip II	8809334	8830368	8	21034
6.	Pretpostavljeni PKS tip I/NRPS	4927380	4939979	1	12599
7.	Pretpostavljeni PKS tip I/NRPS/PKS tip II	6962553	6972788	6	10235

Prikazan je po jedan primjer od svake skupine anotiranih sustava (Slike 29 i 30) dok se cjelokupni rezultati anotacije genoma bakterije *S. scabies* pomoću generičkog računalnog programskog paketa *ClustScan* nalaze u prilozima (vidi: podpoglavlje 8.2.5.).



Slika 29. Prikaz dijela anotirane genske nakupine br. 1. Pretpostavljeni modularni PKS tip I anotiran pomoću programskog paketa *ClustScan*.



Slika 30. Prikaz anotirane genske nakupine br. 6. Pretpostavljeni PKS tip I/NRPS hibrid anotiran pomoću programskog paketa *ClustScan*.

Na Slikama 29 i 30 bijeli pravokutnici predstavljaju gene unutar kojih su smještene domene proteina označene različitim bojama. Crvenim su linijama označeni moduli. U nastavku se nalaze shematski prikazi pomoću koji pokazuju od kojih se gena, modula i domena pronađene genske nakupine sastoje. Domene unutar istog modula označene su istom

bojom, a moduli koji se nalaze na jednome genu, nalaze se u istom retku pod zajedničkim brojem gena. Potrebno je istaknuti da su u shematskom prikazu navedeni samo geni koji sadržavaju module i domene, dok "prazne" gene bez modula i domena, koji su također u nekim slučajevima sastavni dio genskih nakupina, nisu posebno navedeni. Takvi "prazni geni" vjerojatno su uključeni u određene poslije poliketidne/peptidne modifikacije ili sadržavaju neke nove i do sada nepoznate funkcionalnosti. Problematična domena obojana je žutom bojom, a potencijalni uzrok uočenog problema navodi se u nastavku (vidi: podpoglavlje 5.3.). Ispod shematskog prikaza pronađenih genskih nakupina nalazi se legenda sa kraticama naziva domena (Tablica 6).

SHEMATSKI PRIKAZ PRONAĐENIH GENSKIH NAKUPINA

Genska nakupina br. 1: Pretpostavljeni modularni PKS tip I

GEN 7829: -KS-AT-ACP- -KS-AT-DH-ADH-ACP-
GEN 7831: -KS-AT-ADH-ACP- -KS-AT-KR-ACP- -KS-AT-ADH-ACP-
GEN 7833: -D-KS-AT-KR-ACP- -KS-AT-ADH-ACP- -KS-AT-DH-ADH-ACP-
GEN 7835: -D-KS-AT-DH-ADH-ACP- -KS-AT-DH-ER-ADH-ACP-
GEN 7836: -D-KS-AT-KR-ACP- -KS-AT-KR-ACP- -KS-AT-DH-ADH-ACP-
GEN 7839: -KS-AT-DH-ADH-ACP-TE-

Genska nakupina br. 2: Pretpostavljeni ponavljajući PKS tip I

GEN 7311: -ACP-KS-AT-ACP-TE-

Genska nakupina br. 3: Pretpostavljeni PKS tip II

GEN 3624: -KS-
GEN 3625: -KS-
GEN 3626: -ACP-
GEN 3627: -PCy-
GEN 3628: -CyP-
GEN 3629: -M2-

Genska nakupina br. 4: Pretpostavljeni hibrid PKS tip II/NRPS

GEN 5658: -TrC-BTAD-
GEN 5659: -TE-
GEN 5663: -ACPS-ACPSC-
GEN 5665: -A-
GEN 5666: -PCP-
GEN 5667: -ACPS-ACPSC-
GEN 5671: -PCy-
GEN 5670: -FAD3-
GEN 5677: -Cy-
GEN 5678: -FAD2-

Genska nakupina br. 5: Pretpostavljeni hibrid PKS tip I/PKS tip II

GEN 7378: -PP-
GEN 7380: -KS-
GEN 7384: -A-
GEN 7385: -AT-PP- -KS-AT-DH-ER-ADH-ACP-
GEN 7387: -KS-AT-DH-ER-ADH-ACP-TE-
GEN 7388: -A-

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

GEN 3694: -A-PCP- -KS-AT-KR-ACP- C-A-PCP-C-

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

GEN 5629: -A-
GEN 5630: -ACM-AC1-
GEN 5631: -ACM-AC1-
GEN 5632: -PCP-
GEN 5633: -KS-AT-ACP-
GEN 5634: -ACPS-ACPSC-

Tablica 6. Legenda sa kraticama naziva domena.

ACPS = ACP_syn_III	C = Condensation	KS = ketoacyl-synt
ACPSC = ACP_syn_III_C	Cy = Cyclase	KR = KR
AT = Acyl_transf_1	CyP = Cyclase_polyket	M2 = Methyltransf_2
AC1 = Acyl-CoA_dh_1	D = Docking	PCy = Polyketide_cyc
ACM = Acyl-CoA_dh_M	DH = dehidrogenase	PP = PP-binding ACP/PCP
ADH = adh_short	ER = Epimerase	TE = Thioesterase
A = AMP-binding	FAD2 = FAD_binding_2	TrC = Trans_reg_C
BTAD = BTAD	FAD3 = FAD_binding_3	

5. R A S P R A V A

5.1. PRETRAŽIVANJE LITERATURNIH PODATAKA

Prilikom pretraživanja literaturnih podataka pronađeni su različiti tipovi klasifikacija sustava PKS i samo jedan rad u kojem se detaljnije obrađuje tematika klasifikacije tih sustava. Dostupna literatura, o zanimljivim sekundarnim metabolitima, isključivo se odnosi na metode izolacije sekundarnog metabolita (iz hranjive podloge na kojoj raste proizvodni mikroorganizam ili iz stanice producenta), te identifikaciju genskih nakupina koje sadržavaju genetičku uputu za te enzime. Činjenice upućuju da su ovi biosintetski mehanizmi nedovoljno istraženi (Shen, 2003).

Tijekom izrade klasifikacijskih tablica najviše problema uzrokovali su PKS/NRPS hibridni sustavi. U literaturnim podacima hibridni sustavi ubrajani su u sustave NRPS tipa C ili su stavljeni u zasebnu skupinu (Challis, 2005). Zbog velike raznolikosti u organizaciji genskih nakupina hibridnih sustava PKS/NRPS, dolazi se do zaključka da bi takvi sustavi trebali predstavljati zasebnu skupinu. Većina je detaljno opisanih modularnih sustava PKS tipa I. U ovom radu pažnja je usmjerena na prikupljanje manje poznatih sustava i analizu njihovih domena, umjesto bavljenja takvim dobro istraženim sustavima. Zbog toga se u tablici nalazi relativno malen broj primjera modularnih sustava PKS tipa I (vidi: podpoglavlje 4.1., Tablica 3).

Prilikom prikupljanja sekvencija iz baza podataka pomoću pristupnog broja uočeno je da za neke upite postoji samo jedna sekvencija DNA i jedna sekvencija proteina (na primjer ansamitocin; GenBank pristupni broj sekvencije DNA: AF453501.1). Međutim, pronađeni su i slučajevi kada se unatoč jednoj sekvenciji DNA, koja pokriva čitavu gensku nakupinu, pojavljuje više proteinskih sekvencija kojima se pristupa preko različitih pristupnih brojeva (na primjer oksitetraciklin; GenBank pristupni broj sekvencije DNA: AF453501.1) (NCBI, 2009). Takvi se slučajevi mogu protumačiti kao posljedice različitog pristupa sekvencioniranju, tj. u prvom slučaju je sekvencionirana cijela genska nakupina odjednom, a u drugom pojedinačni geni. Literaturni podaci često nude po nekoliko mikroorganizama producenata za isti sekundarni metabolit, no ovdje su prikupljani samo pojedinačni primjeri, po mogućnosti iz skupine aktinobakterija (oslanjajući se na svojstvo sličnosti sekvencija) (Austin i Noel, 2002).

5. 2. OBRADA PRIKUPLJENIH SEKVENCIJA DNA I PROTEINA

Iz rezultata dobivenih programskim paketom *HMMER* (Eddy, 1998) izdvajane su domene čije su vrijednosti parametara zadovoljavale uvjete: vrijednost $E < 10^{-5}$ i uspjeh pogotka > 0. Te vrijednosti odabrane su kao granične zbog pretpostavke da domene čija je vrijednost E viša od 10^{-5} i čiji je pogodak blizu nule ili negativan ne zadovoljavaju traženo svojstvo sličnosti. Izdvajane su sve domene koje su zadovoljavale postavljene uvjete parametara bez obzira da li je domena cjelovita ili djelomična, zbog opravdane bojazni da bi isključivanje jedne vrste domena moglo dovesti do gubitka dragocjenih informacija. Ovisno o tome da li su domene cjelovite ili djelomične, te da li se ponavljaju više puta ili su jedinstvene, u tablici profila proteina prikazanoj u rezultatima (vidi: podpoglavlje 4.2., Tablica 3) smještene su u odvojene kolone. Višestruko su se ponavljale domene koje čine osnovne module enzima PKS, odosno hibridnih sustava PKS/NRPS. Jedinstvene domene su specifične za pojedini tip sustava PKS/hibrida. Pojavljivanje djelomičnih domena u rezultatima dobivenim programskim paketom *HMMER* moguće je objasniti na dva načina:

- prilikom prikupljanja sekvencija nije obuhvaćena čitava genska nakupina, ili
- analizirana sekvencija i profili proteina iz baze podataka Pfam pokazuju djelomičnu sličnost.

Analizom tablice profila proteina (vidi: podpoglavlje 4.2., Tablica 3; podpoglavlje 8.2.3., Tablica 3P; podpoglavlje 8.2.4.), uočene su neke domene svojstvene za pojedine tipove sustava PKS/hibrida (nazivi su domena preuzeti neposredno iz programskog paketa *HMMER* pa su zbog toga napisane engleskim jezikom):

- domene ADH_N, ADH_zinc_N, Thioesterase, Methyltransf_12 i Condensation omogućuju razlikovanje ponavljajućih fungalnih enzima PKS tipa I od bakterijskih,
- domene ADH_zinc_N, AMP-binding, Docking, ADH_N, Methyltransf_12, Thioesterase, NPD i FAD_binding_2 čine razliku između modularnih enzima PKS tipa I od ponavljajućih bakterijskih,
- domene AMP-binding, Docking, NPD i FAD_binding_2 razlikuju enzime modularnih PKS tipa I od ponavljajućih fungalnih, i
- domene Methyltransf_12, AMP-binding, adh_short, PP-binding, ketoacyl-synt, Acyl_transf_1 i Ketoacyl-synt_C su zajedničke enzimima PKS tipa I i tipa II.

U slučajevima kada su se domene preklapale, odnosno kada su se nalazile na djelomično istim sekvencijama DNA, izdvojena je samo ona domena koja ima bolje parametre. Na primjer, domena KR 37745 pb - 37923 pb (vrijednost E: $4.6e^{-92}$, uspjeh pogotka: 316.8) preklapa se sa domenom adh_short 37745 pb – 37910 pb (vrijednost E: $4.4e^{-63}$, uspjeh pogotka: 220.5). Pošto ima bolje parametre, domena KR je izdvojena, a domena adh_short je obrisana.

5.3. ANOTACIJA GENOMA BAKTERIJE *Streptomyces scabies* POMOĆU PROGRAMSKOG PAKETA *ClustScan*

Prilikom odabira modela za određivanje gena u genomu bakterije *S. scabies* izabran je model vrste *S. avermitilis*. Broj gena koje sadržavaju genomi vrste *S. avermitilis* (8446 gena; Anonymous 5, 2009) i vrste *S. coelicolor* (8350 gena; The Wellcome Trust Sanger Institute 3, 2007) preuzet je iz literaturnih podataka (Challis i Hopwood, 2003; Ōmura i sur., 2001). Broj gena u genomu bakterije *S. scabies* određen je pomoću bioinformatičkog alata *Glimmer* (Delcher i sur., 2007). Iznosio je 9462 gena u genomu. Budući da je model vrste *S. avermitilis* sličniji po broju gena, on je odabran kao model za analizu genoma programom *GeneMark* (Besemer i Borodovsky, 2005). Analizom rezultata uočeni su slučajevi u kojima su domene bile izvan gena. Pretpostavlja se da je to posljedica pomaka okvira čitanja što je najvjerojatnije pogreška prilikom sekvencioniranja DNA. Pošto programski paket *ClustScan* (Starcevic i sur., 2008) ima mogućnost prilagođavanja koordinata, u nekoliko slučajeva geni su pomaknuti iz jednog okvira čitanja u drugi, kako bi se domene smjestile unutar gena. Ako domena nije unutar gena nemoguće ju je obuhvatiti u gensku nakupinu.

Prilikom anotiranja uočeno je nekoliko problema koji su onemogućili da se precizno klasificira svaki tip genske nakupine (vidi: podpoglavlje 4.3., Tablica 5). Uočeno je da domena ACP unutar gena 7378 iz genske nakupine br. 5, iako se nalazi unutar gena, nakon kreiranja genske nakupine nije prikazana u prozoru za uređivanje genskih nakupina. Pretpostavka je da gen ne pokriva cijelu domenu pa je ni ne prikazuje kao njegov sastavni dio. U mnogim je genskim nakupinama uočena neuobičajena organizacija domena unutar modula. Primjerice, u genskoj nakupini br. 6, u zadnjem modulu gena 3694, domena C nalazi na samom kraju. Ovaj neobični smještaj domene C može biti posljedica delecije narednih dijelova modula ili više uzastopnih djelomičnih modula (Challis, 2005).

Prilikom analiziranja genskih nakupina hibridnih sustava pronađenih u genomu bakterije *S. scabies*, uočeno je mnogo odstupanja od do sada anotiranih hibridnih genskih nakupina. Genska nakupina br. 7 prema sastavu gena predstavlja hibridni sustav sastavljen od domena svojstvenih za sustave NRPS, PKS i hibride. Pronađen je i hibridni sustav sastavljen od domena uobičajenih za enzime PKS tipa I i tipa II (genska nakupina br. 5) (Shen, 2003). Genska nakupina br. 1 po broju gena, organizaciji modula i sastavu domena nalikuje genskoj nakupini odgovornoj za sintezu spoja Konkanamicin A. U genima 7833 i 7836 iste genske nakupine zamijećeno je odstupanje u kojima je na dva mjesta došlo do preklapanja domena AT i DH. Budući da su domene DH u takvim slučajevima bile neaktivne i sa lošijim parametrima, uklonjene su. Prema organizaciji domena u genu 7311, genska nakupina br. 2 svrstana je u ponavljajuće sustave PKS tipa I. Budući da položaj domene ACP na početku modula nije uobičajen kod takvog tipa sustava PKS, moguće je da je ovo zapravo signal nekog modularnog sustava PKS tipa I. U genu 3624 genske nakupine br. 3 nalazi se domena KS s malim uspjehom pogotka (vrijednost E: $4.18585e^{-33}$, uspjeh pogotka: 8,543) pa postoji mogućnost da se radi o lažnom pozitivnom pogotku. Također, da bi se ova genska nakupina mogla sa sigurnošću klasificirati kao sustav PKS tipa II, nedostaje joj domena AT svojstvena za ovaj tip.

Iz rezultata pretraživanja genoma bakterije *S. scabies* (The Wellcome Trust Sanger Institute 3, 2009) pomoću vlastitih profila proteina koji su učitani u programski paket *ClustScan* (Starcevic i sur., 2008) mogu se sa sigurnošću klasificirati genske nakupine do određene razine, odnosno sa visokom sigurnošću se može ustanoviti da li genska nakupina spada u sustav PKS ili je hibridni sustav. Međutim, vrlo je teško precizno odrediti kojoj klasifikacijskoj grupi unutar tih sustava pronađene genske nakupine pripadaju jer su zapažena velika odstupanja u organizaciji domena unutar modula, a dodatna otežavajuća okolnost bila je nemogućnost provjere rezultata budući da je genom tek nedavno sekvencioniran, a još nije opisan.

6. Z A K L J U Č C I

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

1. Na temelju prikupljene i analizirane literature izrađena je klasifikacija sustava PKS upotpunjena profilima proteina pripadajućih domena.

2. Hibridne bi sustave zbog raznolikosti organizacije unutar genskih nakupina trebalo izdvojiti u zasebnu skupinu.

3. Programskim paketom *ClustScan* u genomu bakterije *S. scabies* opisano je 7 genskih nakupina:

   - produkti triju genskih nakupina sintetiziraju poliketide, a
   - produkti četiriju genskih nakupina sintetiziraju poliketidno/peptidne hibride.

4. Daljnja istraživanja genskih nakupina koje sadržavaju genetičku uputu za sustave PKS osigurala bi precizniju anotaciju i na razini podgrupa, a ne samo glavnih tipova.

7. POPIS LITERATURE

1. Altschul, S. F., Miller, G. W., Myers, E. W., Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol*. **215**, 403-410.
2. Anonymous 1 (2004) A knowledge based resource for analysis of Non-ribosomal Peptide Synthetases and Polyketide Synthases, http://www.nii.res.in/nrps-pks.html. Pristupljeno 21. 09. 2009.
3. Anonymous 2 (2009) Clustal: Multiple Sequence Alignment, http://www.clustal.org/. Pristupljeno 23. 09. 2009.
4. Anonymous 3 (2009) Science direct, http://www.sciencedirect.com/ Pristupljeno 14. 09. 2009.
5. Anonymous 4 (2009) HMMER - biosequence analysis using profile hidden Markov models, http://hmmer.janelia.org/. Pristupljeno 12. 09. 2009.
6. Anonymous 5 (2009) Genome project of *Streptomyces avermitilis*, http://avermitilis.ls.kitasato-u.ac.jp/. Pristupljeno 19. 09. 2009.
7. 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**, 405-13.
8. Austin, M. B., Noel., J. P. (2002) The chalcone synthase superfamily of type III polyketide synthases. *Nat. Prod. Rep*. **20**, 79–110.
9. Besemer, J., Borodovsky, M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*. **33**, 451-454.
10. Bioinformatics group, (2009) Thiotemplate Modular Systems Studies, http://bioserv.pbf.hr/cms/index. php?page=clustscan. Pristupljeno 23. 09. 2009.
11. Caboche, S., Pupin, M., Leclére, V., Fontaine, A., Jacques, P., Kucherov, G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res*. **36**, 326-331.
12. Challis, L. G. (2005) A widely distributed bacterial pathway for siderophore biosynthesis independent of nonribosomal peptide synthetases. *ChemBioChem* **6**, 601-611.
13. Challis, L. G., Hopwood, D. A., (2003) Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proc. Natl. Acad. Sci. USA* **100**, 14555-14561.
14. Challis, L. G., Naismith, J. H., (2004) Structural aspects of non-ribosomal peptide biosynthesis. *Curr. Opin. Chem. Biol*. **14**, 748-756.
15. 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.

16. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry III, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., Barrell, B. G. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544.
17. Delcher, A. L., Bratke, K. A., Powers, E. C., Salzberg, S. L. (2007) Indentifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673-679.
18. Demain, A. (1999) Pharmaceutically active secondary metabolites of microorganisms. *Appl. Microbiol. Biotechnol.* **52**, 455-463.
19. Du, L., Sánchez, C., Shen, B. (2001) Hybrid peptide-polyketide natural products: biosynthesis and prospects toward engineering novel molecules. *Metab. Eng*. **3**, 78-95.
20. EBI (2009) European Bioinformatics Institute - European Molecular Biology Laboratory EMBL, http://www.ebi.ac.uk/. Pristupljeno 15. 09. 2009.
21. EBI (2009) European Bioinformatics Institute – Readseq, http://www.ebi.ac.uk/cgi-bin/readseq.cgi. Pristupljeno 13.09.2009.
22. Eddy, S. R. (1998) Profile Hidden Markov models. *Bioinformatics* **14**, 755-763.
23. 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.
24. 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.
25. Google (2009) Google Scholar, http://scholar.google.hr/. Pristupljeno 27. 09. 2009.
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. Hranueli, D., Cullum, J. (2001) Novi hibridni poliketidi dobiveni kombinatornom biosintezom. *Kem. Ind*. **50**, 381-411.
28. 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.

29. Hutchinson, R. C. (2003) Polyketide and non-ribosomal peptide synthases: Falling together by coming apart. *Proc. Natl. Acad. Sci. USA* **100**, 3010-3012.
30. Komaki, H., Harayama., S. (2006) Sequence diversity of type-II polyketide synthase genes in *Streptomyces*. *Actinomycetologic.* **20**, 42–48.
31. Lambert, D. H., Loria, R. (1989) *Streptomyces scabies* sp. nov., nom. rev. *Int. J. Syst. Bacteriol.* **39**, 387-392.
32. 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.
33. Marahiel, M. A., Mootz, H. D., Schwarzer, D. (2002) Ways of Assembling Complex Natural Products on Modular Nonribosomal Peptide Synthetases. *ChemBioChem* **3**, 490-504.
34. NCBI (2009) National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/. Pristupljeno 21. 09. 2009.
35. Ōmura, S., Ikeda, H., Ishikawa, J., Hanamoto, A., Takahashi, C., Shinose, M. (2001) Genome Sequence an Industrial Microorganism *Streptomyces avermitilis*: *Streptomyces avermitilis*: Deducing the Ability of Producing Secondary Metabolites. *Proc. Natl. Acad. Sci. USA* **98**, 12215-12220.
36. Reeves, G. A., Talavera, D., Thornton, J. M. (2009) Genome and proteome annotation: organization, interpretation and integration. *J. R. Soc. Interface* **6**, 129-147.
37. Reva, O., Tümmler, B. (2008) Think big – giant genes in bacteria. *Environ. Microbiol.* **10**, 768-777.
38. Rice, P., Longden, I., Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. **16**, 276-277.
39. Shen, B. (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Curr. Opin. Chem. Biol*. **7**, 285–295.
40. Starčević, A., Žučko, J., Šimunković, J., Long, P. F., Cullum, J., Hranueli, D. (2008) ClustScan: an integrated program package for semi-authomatic annotation of modular biosynthetic gene cluster and *in silico* prediction of novel chemical structures. *Nucleic Acids Res.* **10**, 1-11.
41. Tae, H., Kong, E. B., Park, K. (2007) ASMPKS: an analysis system for modular polyketide synthases. *BMC Bioinformatics* **8**, 327.
42. 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.

43. The Wellcome Trust Sanger Institute 1 (2009), http://www.sanger.ac.uk/. Pristupljeno 16.09.2009.
44. The Wellcome Trust Sanger Institute 2 (2009) Pfam, http://www.sanger.ac.uk/Pfam. Pristupljeno 20. 09. 2009.
45. The Wellcome Trust Sanger Institute 3 (2007) Genome project of *Streptomyces coelicolor*, http://www.sanger.ac.uk/Projects/S_coelicolor/. Pristupljeno 19. 09. 2009.
46. The Wellcome Trust Sanger Institute 4 (2009) Genome project of *Streptomyces scabies*, http://www.sanger.ac.uk/Projects/S_scabies/. Pristupljeno 19. 09. 2009.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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 librariesgenerated 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 polipeptid nosač acila (PKS)
ARO	- domena aromataze (PKS)
AT	- domena aciltransferaze (PKS)
bp	- parovi baza
C	- domena za kondenzaciju (NRPS)
CHS like	- šalkon sintazi slični sustavi (PKS)
CLF	- faktor koji određuje duljinu ugljikova lanca (PKS)
CYC	- domena ciklaze (PKS)
DH	- domena dehidrataze (PKS)
DNA	- deoksiribonukleinska kiselina
E	- očekivana vrijednost
E	- domena za epimerizaciju (NRPS)
ER	- domena enoilreduktaze (PKS)
KR	- domena ketoreduktaze (PKS)
KS	- domena ketosintaze (PKS)
MT	- domena za *N*-, *C*- ili *O*-metilaciju (NRPS)
NRPS	- sintetaza neribosomalno sintetiziranih peptida
ORF	- otvoreni okvir čitanja
PCP	- mali polipeptid nosač peptidila (NRPS)
PKS	- poliketid sintaza
PKS/NRPS	- mješovita poliketid sintaza i sintetaza neribosomalno sintetiziranih peptida
Te	- domena tioesteraze (NRPS)
TE	- domena tioesteraze (PKS)

8.2. SADRŽAJ KOMPAKTNOG DISKA

Svi navedeni prilozi, uključujući i cjelovit tekst Diplomskog rada (Diplomski rad IT.pdf), nalaze se na kompaktnom disku (CD-R) nazvanom "Diplomski rad IT: Prilozi".

8.2.1. Sekvencije DNA cjelovitih genskih nakupina, te sekvencije DNA i proteina pojedinačnih gena, poliketida i poliketidno/peptidnih hibrida u obliku zapisa FASTA

8.2.2. Sekvencije proteina iz priloga 8.2.1., u obliku zapisa FASTA, pripremljene za višestruko poravnavanje sekvencija pomoću programskog paketa *HMMER*

8.2.3. Rezultati višestrukih poravnanja sekvencija, iz priloga 8.2.2., pomoću programskog paketa *HMMER*

8.2.4. Profili proteina za sve katalitički aktivne domene svih opisanih tipova enzima PKS i sintaza poliketidno/peptidnih hibrida ekstrahirani iz baze podataka Pfam

8.2.5. Rezultat anotacije genoma bakterije *Streptomyces scabies* pomoću programskog paketa *ClustScan*

Priložena je radna površina programskog paketa *ClustScan* (*S. scabies* rezultat anotacije_Ida). Napomena: za njeno je pregledavanje potreban klijent programskog paketa na osobnom računalu. Kabinet za bioinformatiku će na zahtjev instalirati programski paket *ClustScan* na računalu korisnika.