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, listopad 2012. **Marija Duvnjak, 4497/Bi**

PREDVIĐANJE SPECIFIČNOSTI IZBORA SUPSTRATA DOMENA ZA ADENILACIJU

Ovaj je rad izrađen u Kabinetu za bioinformatiku, Zavoda za biokemijsko inženjerstvo, Prehrambeno-biotehnološkog fakulteta, Sveučilišta u Zagrebu pod vodstvom dr.sc. Antonia Starčevića, docenta te uz svesrdnu pomoć dr.sc. Jurice Žučka, višeg asistenta

Zahvaljujem se mentoru, docentu dr. sc. Antoniu Starčeviću na pomoći i vodstvu tijekom izrade ovog diplomskog rada. Isto tako se zahvaljujem dr. sc. Jurici Žučku na konstruktivnim savjetima i pomoći.

Veliko hvala profesore dr. sc. Daslav Hranueli na svemu!

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

PREDVIĐANJE SPECIFIČNOSTI IZBORA SUPSTRATA DOMENA ZA ADENILACIJU

Marija Duvnjak, 4497/Bi

Sažetak: U posljednje se vrijeme, u znanstvenoj javnosti, pojavilo veliko zanimanje za oblikovanje novih lijekova. Genske nakupine čiji proteinski produkti sudjeluju u sintezi biološki aktivnih supstancija za oblikovanje novih lijekova nalaze se u genomima sekvenciranih mikroorganizama. Razvijen je čitav niz javno dostupnih baza podataka i bioinformatičkih alata za njihovu analizu. To se posebno odnosi na bioinformatičke alate koji analiziraju genske nakupine, čiji produkti sudjeluju u sintezi sekundarnih metabolita u uvjetima *in silico*. Jedna od skupina zanimljivih prirodnih spojeva jesu neribosomalno sintetizirani peptidi. Adenilacijske domene (A-domene), obavezne podjedinice modula neribosomalno sintetiziranih peptid sintetaza (NRPS), prepoznaju i aktiviraju aminokiseline koje moraju biti ugrađene u konačni produkt, neribosomalno sintetizirani peptid. Poznavanje specifičnosti A-domena za supstrat omogućit će predviđanje konačne kemijske strukture produkata genskih nakupina NRPS i dizajn novih biološki aktivnih prirodnih spojeva. U ovom su radu predviđanja aminokiselinskih supstrata, koje aktiviraju A-domene, provedena pomoću višestrukih poravnanja upotrebom bioinformatičkog programa Clustal W i filogenetskih analiza upotrebom bioinformatičkog programa za molekularne evolucijske genetske analize, MEGA, te bioinformatičkog programa HMMER 3 za konstrukciju HMM profila za prepoznavanje supstrata.

Ključne riječi: A-domene, višestruka poravnanja, filogenetske analize, predviđanje specifičnosti

Rad sadržava: 66 stranica, 30 slika, 2 tablice, 53 literaturna navoda i 5 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. Antonio Starčević, docent
Pomoć pri izradi: dr.sc. Jurica žučko, viši asistent

Stručno povjerenstvo za ocjenu i obranu:
1. Dr.sc. Daslav Hranueli, red. prof.
2. Dr.sc. Antonio Starčević, doc.
3. Dr.sc. Ana Vukelić, izv. prof.

Datum obrane: 09. listopad, 2012.

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

PREDICTION OF ADENILATION DOMAINS SUBSTRATE SPECIFICITY

Marija Duvnjak, 4497/Bi

Abstract: During the last several years there is an increased scientific interest for the design of novel drugs. Gene clusters whose protein products are involved in the synthesis of these biologically active substances used for novel drug design are present in sequenced microbial genomes. For that reason a number of publically available databases and bioinformatic tools were developed to analyse gene clusters whose products are involved in the biosynthesis of natural products *in silico*. One class of interesting natural products is nonribosomally synthesized peptides. Adenylation domains (A-domains), obligatory subunits of modules of non-ribosomally synthesised peptide synthetases (NRPS), are recognising and activating amino acid that are incorporated into the final product, non-ribosomally sythesised peptide. Knowing the A-domain substrate specificity for every sequenced A-domain would enable the prediction of final chemical structure of NRPS gene cluster product and the design of novel biologically active natural products. In this work, the multiple alignments by the bioinformatics program Clustal W, followed by phylogenetic analysis by the bioinformatics program for the molecular evolutionary genetics analysis, MEGA and bioinformatics program HMMER 3 for the construction of HMM profiles were used for the prediction of the amino acid substrates which are activated by A-domains.

Keywords: A-domains, multiple alignments, phylogenetic analysis, specificity prediction

Thesis contains: 66 pages, 30 figures, 2 tables, 53 references, 5 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. Antonio Starčević, Assistent Professor
Technical support and assistance: Ph. D. Jurica Žučko, Senior Assistant

Reviewers:
1. Ph.D. Daslav Hranueli, Full Professor
2. Ph.D. Antonio Starčević, Assistant Professor
3. Ph.D. Ana Vukelić, Associate Professor

Thesis defended: 9th October, 2012

SADRŽAJ

1. U V O D

Brzi razvoj tehnologija za sekvenciranje molekula DNA doveo je do eksponencijalnog rasta baza podataka, ali je porast eksperimentalnih podataka o funkcijama proteina mnogo sporiji. U brojnim slučajevima standardne bioinformatičke metode mogu pridružiti gene porodicama proteina, ali točna funkcija proteina najčešće ostaje nepoznata. Na primjer, geni se za hidrolizu šećera mogu lako prepoznati, ali se njihov supstrat ne može utvrditi. U slučajevima kada su eksperimentalni podaci dostupni, postojeći algoritmi mogu, iz višestrukih poravnanja sekvencija proteina, prepoznati aminokiseline koje su važne za pridruživanje proteina određenoj pod-porodici i na taj način predvidjeti funkciju proteina iz sekvencije DNA.

Dobar je primjer takvog pristupa predviđanje specifičnosti supstrata domena aciltransferaza (AT) modularnih poliketid sintaza (PKS) (Starcevic i sur., 2008). Ti složeni enzimski kompleksi sintetiziraju poliketidne prirodne spojeve uzastopnim kondenzacijama, pri čemu se svaka kondenzacija provodi pomoću jednog od modula enzima, od kojih se svaki sastoji od više katalitičkih domena. Domene AT svakog modula određuju tip ili vrstu supstrata koji se ugrađuje. U poliketidni se lanac može ugraditi svaki od 5 različitih tipova supstrata pomoću modula za rast poliketidnog lanca (engl. "extender modules"), a tip se supstrata može predvidjeti upotrebom "otiska" (engl. "fingerprint") aminokiselina koje su specifične za određivanje aminokiselinskog ostatka.

Modularne sintetaze neribosomalno sintetiziranih peptida (NRPS) imaju sličnu arhitekturu genskih nakupina i enzima, a izbor je aminokiselinskih supstarata koji se ugrađuju određen aktivnošću i specifičnošću domena za adenilaciju (A). Problem je predviđanja izbora supstrata, međutim, mnogo složeniji nego u poliketida zbog toga što postoji izuzetno veliki broj, od gotovo 500 različitih tipova supstrata (Strieker i sur., 2010). Do sada je postignut samo ograničeni uspjeh predviđanja aminokiselinskih supstrata upotrebom pristupa pomoću metode SVD (akronim engl. izraza "support vector machine") (Rausch i sur., 2005). Taj pristup, međutim, zahtijeva detaljno poznavanje domena A pa je analiza ograničena na mali broj aminokiselinskih ostataka.

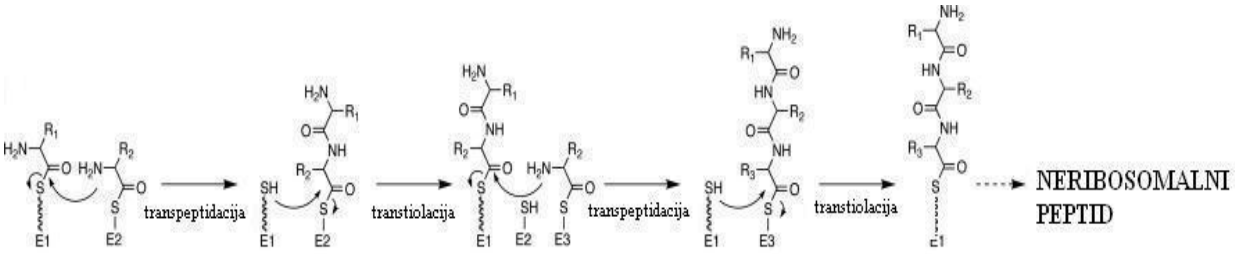
Izradom ovog diplomskog rada obavljena su predviđanja aminokiselinskih supstrata, koje aktiviraju A-domene, uz pomoć višestrukih poravnanja upotrebom bioinformatičkog programa Clustal W (Larkin i sur., 2007), filogenetskih analiza upotrebom bioinformatičkog programa za molekularne evolucijske genetske analize, MEGA (Tamura i sur., 2007), te bioinformatičkog programa HMMER 3 (Finn i sur., 2011) za konstrukciju HMM profila za prepoznavanje supstrata.

2. TEORIJSKI DIO

2.1. NERIBOSOMALNO SINTETIZIRANI PEPTIDI

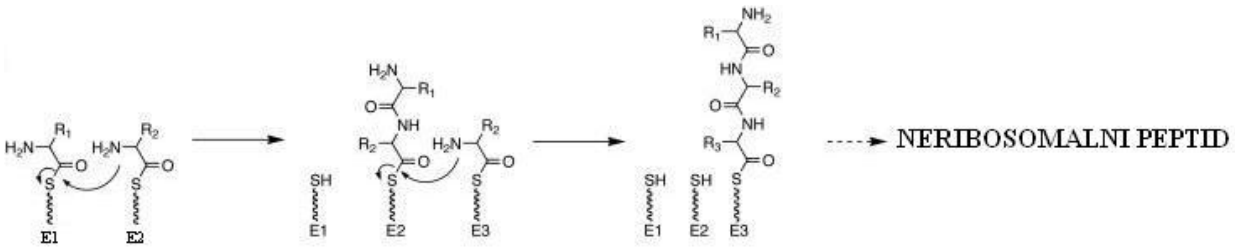
Prirodni spojevi čine važan element biološki aktivnih supstancija. Oni su, u širem smislu, molekule i smjese molekula koje su sintetizirane od strane živih organizama. Mogu se podijeliti s obzirom na funkciju (antibiotici, vitamini, toksini, imunosupresori itd.) ili kemijsku strukturu (peptidi, poliketidi, polisaharidi itd.) (Schwarzer i sur., 2003). Veliki broj biološki aktivnih polipeptida bakterijskog i fungalnog podrijetla sintetizirano je enzimima nazvanim neribosomalne peptid sintetaze (engl. "Nonribosomal Peptide Synthetases", NRPS). Neribosomalne peptid sintetaze djeluju istovremeno kao biološki mehanizam i kao kalup za sintezu polipeptida. Sustavi NRPS su jednostavno "šifra" za sintezu rastućeg polipeptidnog lanca. Nakon što je sredinom dvadesetog stoljeća razjašnjen mehanizam sinteze proteina na ribosomima, Tatum i suradnici prvi su dokazali da se sinteza biološki aktivnih polipeptida odvija mehanizmom koji je različit od sinteze polipeptida na ribosomima (Slika 1) (Felnagle i sur., 2008). Danas se zna da enzimi NRPS pripadaju familiji megasintetaza. Megasintetaze su jedni od najvećih enzima, s molekulskim masama od ~2,3 MDa (~21.000 aminokiselinskih ostataka) (Stachelhaus i sur., 1999). Kloniranjem i sekvencioniranjem gena koji sadržavaju genetisku uputu za sustave NRPS, otkriveni su konzervirani dijelovi u lancu i modularna organizacija tih enzima.

Modul je autonoman dio unutar sustava NRPS koja sadržava potrebnu informaciju za prepoznavanje, aktivaciju, prijenos (thiolaciju) i kondenzaciju, te u nekim slučajevima modifikaciju monomera (aminokiselina). Svaki modul dugačak je približno 1.000 aminokiselina i odgovoran je za prepoznavanje, te vezanje pojedinačnog monomera u polipeptidni lanac. Tako bi sustav NRPS koji se sastoji od 10 modula, sintetizirao polipeptid dugačak 10 aminokiselina. Primarna struktura, veličina i složenost polipeptidnog produkta određena je brojem i organizacijom modula sustava NRPS, tzv. "pravilo kolinearnosti". Specifični redoslijed modula u genomskoj DNA određuje broj i redoslijed aminokiselina koje čine sintetizirani polipeptid pomoću sustava NRPS. Takav tip organizacije sličan je multienzimskim kompleksima koji su odgovorni za biosintezu masnih kiselina i poliketida (Conti i sur., 1997). Na tim enzimskim sustavima kemijski redoslijed reakcija sinteze odvija se po rasporedu modula. Takav tip sinteze poznat je kao modularni sustav tio-kalupa (engl. "Thiotemplate Modular Systems", TMS). Sustav je prvi put pretpostavio Lipmann 1971. godine tijekom proučavanja mehanizma biosinteze antibiotika gramicidina S i tirocidina.



Slika 1. Mehanizam sinteze modularnog sustava tio-kalupa za sustav NRPS; E1 predstavlja dio enzima koji sadrži 4'-fosfopanteteinsku skupinu; E2 i E3 predstavljaju module multienzima NRPS koji sadržavaju pripadajući aminoacil tioester; E1 se upotrebljava prilikom sinteze peptida NRPS više puta, dok se E2 i E3 upotrebljava samo jednom (Felnagle i sur., 2008).

Modularni sustav NRPS enzima postao je važan faktor za interpretaciju genetskih informacija nakon što je sekvencioniranje DNA postala rutina. Nove sekvence dovele su do revizije tadašnjeg modularnog sustava tio-kalupa NRPS enzima, jer se otkrilo da svaki modul ima zasebnu 4-fosfopanteteinsku skupinu koja je odgovorna za vezanje aminokiseline unutar modula za razliku od prijašnjeg modela u kojem se govorilo o zajedničkoj 4-fosfopanteteinskoj skupini za cijeli sustav NRPS. Novi model je poznat kao sustav modela višekratnih nosača (engl. "Multiple carrier model") (Slika 2) (Felnagle i sur., 2008).



Slika 2. Mehanizam sinteze modela višekratnih nosača za sustav NRPS; E1, E2 i E3 predstavljaju individualne module koji sadržavaju pripadajući aminoacil tioester (Felnagle i sur., 2008).

Sekvencijske i strukturalne analize ukazale su na modularnu strukturu ovih enzima, no osim modularne organizacije analize su ukazale na još jedan tip podorganizacije. Naime,

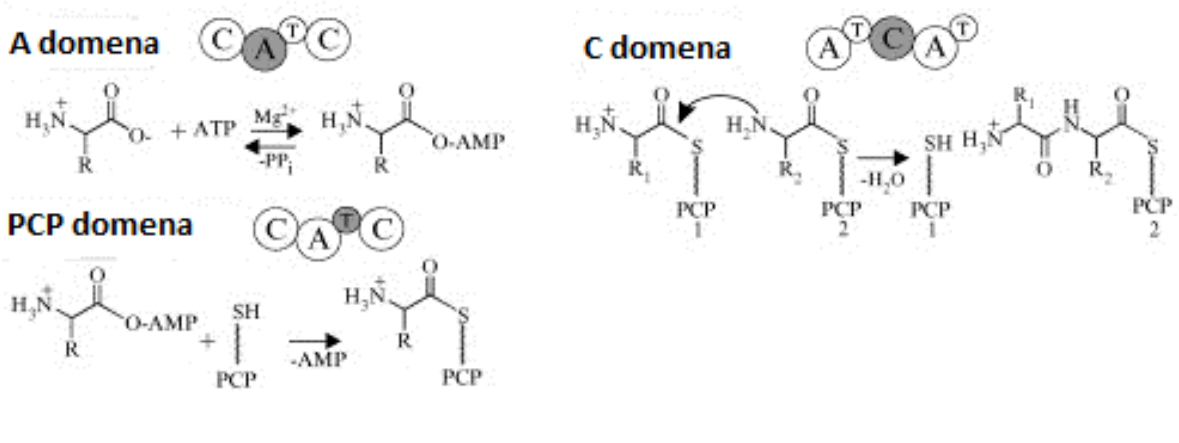
svaki modul sastoji se od nekoliko domena. Svaka pojedina domena katalizira specifičnu reakciju unutar modula. Definiranje domena unutar modula olakšano je i time što svaki tip domene sadržava specifične konzervirane dijelove tzv. "otiske" bez obzira na podrijetlo enzima (Konz i Marahiel, 1999; Schwarzer i sur., 2003).

2.1.1. Domene sustava NRPS

Razumijevanje aktivnosti sustava NRPS kao polazišnu točku ima razumijevanje aktivnost pojedinih domena koje sačinjavaju pojedini modul sustava NRPS. Domene su strukturalno i funkcionalno zasebne cjeline modula, cjeline koje su međusobno povezane malim i fleksibilnim regijama, tzv. poveznicama (engl. "linker") (Weber i Marahiel, 2001).

Osnovni modul definiraju domene koje su odgovorne za aktivaciju i tioesterifikaciju pojedinog supstrata te sintezu jedne peptidne veze između tioesterificiranog supstrata prethodnog i pripadajućeg modula. Osnovni modul sustava NRPS (Slika 3) tako sadržava 3 domene:

- domenu za adenilacij aminokiselina (engl." Adenylation domain", domenu A)
- domenu za kondenzaciju (engl. "Condensation domain", domenu C)
- domenu za prijenos (engl. "Peptidyl Carrier Protein", domenu PCP ili domenu T).



Slika 3. Domene osnovnog modula s pozicijama unutar modula i reakcije koje kataliziraju.

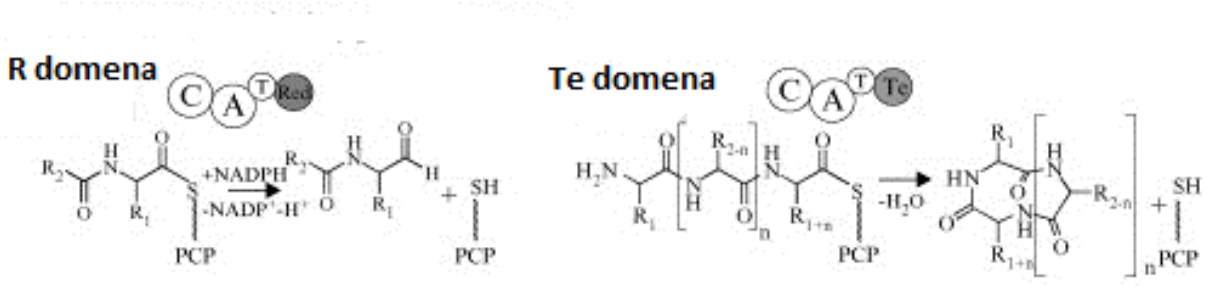
Domena A regulira ulazak monomera u proces sinteze na sustavu NRPS. Ta 556 aminokiselina dugačka domena prvi put je identificirana kroz biokemijska istraživanja 1995 godine (Stachelhaus i Marahiel, 1995). Domena A prepoznaje određenu aminokiselinu, većinom specifično, te ju aktivira adenilacijom. To je reakcija između α-karboksilne skupine supstrata i α-fosfata molekule ATP (Weber i Marahiel, 2001). Pri adenilaciji se troši jedna molekula ATP. Aktivirana aminokiselina se nakon toga u obliku tioestera kovalentno veže na 4-fosfopanteteinsku ruku domene PCP (Slika 3).

Domena PCP predstavlja transportnu jedinicu koja omogućava prijenos aktiviranih aminokiselina i elongacijskih međuprodukata. **Domena C** koja se većinom nalazi na početku modula katalizira stvaranje peptidne veze između aminoacil tioestera pripadajuće domene PCP (akceptor pozicija, nukleofil) i aminoacil tioestera ili peptidilacil tioestera domene PCP prethodnog modula (donor pozicija, elektrofil) (Challis i sur., 2000). U novijim istraživanjima pokazalo se da domena C također pokazuje određenu razinu specifičnosti za supstrat i to na strani akceptora (nukleofila). Sa strane donora (elektrofila), domena C ne pokazuje specifičnost (Schwarzer i sur., 2003).

Iznimku od osnovnog rasporeda predstavlja prvi modul i zadnji modul:

- prvi modul se najčešće sastoji samo od domene A i domene PCP. Tako kondenzacijska domena slijedećeg modula (drugog po redu modula) katalizira nastajanje peptidne veze između monomera vezanog na domenu PCP prvog modula i monomera vezanog uz domenu PCP drugog modula.
- zadnji modul sadržava, osim osnovnih domena, i dodatnu tioesterazu tipa I (engl. "Thioesterase domain type I", domenu Te tipa I) ili reduktaza domenu (engl. "Reductase domain", domena R) (Slika 4). Te domena se nalazi na C-terminalnom kraju zadnjeg modula. Taj tip domene odgovoran je za odvajanje polipeptidnog produkta s multienzima NRPS. Kada **domena Te** katalizira nukleofilni napad vode onda je rezultat linearni proizvod, a kada katalizira napad unutarnjeg nuleofila rezultat je ciklički polipeptid. Tioesteraza tipa II (engl. "Thioesterase domain type II", domena Te tipa II) je domena koja ne igra veliku ulogu u sintezi polipeptida. Njena uloga je uklanjanje krivo vezane 4-fosfopanteteinske ruke domene PCP (Strieker i sur., 2010).

Osim osnovnih domena, moduli mogu sadržavat i neke netipične domene. Moduli u kojima dolazi do vezanja N-metiliranih aminokiselina imaju prisutnu i metilacijsku domenu (engl. "Methyltransferase domain", **domena M** ili **domena MT**) (Slika 5). Domena je dugačka otprilike 420 aminokiselina. Biokemijska istraživanja su pokazala da domena MT katalizira reakciju N-metilacije monomera prije sinteze peptidne veze od strane domene C (Konz i Marahiel, 1999; Challis i Naismith, 2004).



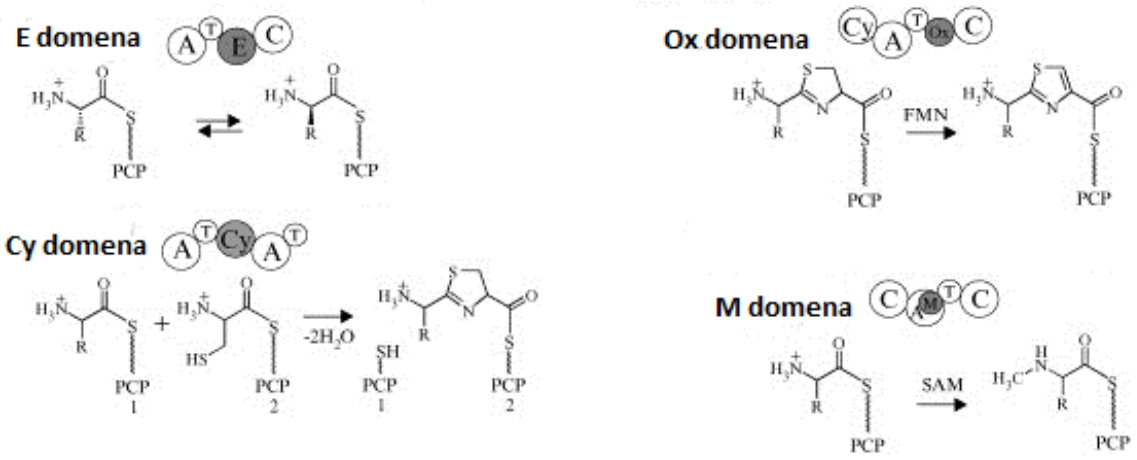
Slika 4. Domene koje kataliziraju odvajanje polipeptidnog produkta sa multienzima NRPS s pozicijama unutar modula i pripadajućim reakcijama.

Slična situacija prisutna je i u nekim modulima koji vežu D-aminokiseline. Takvi moduli na C-terminalnom dijelu domene PCP imaju prisutnu dodatnu domenu epimeraza koja katalizira racemizaciju iz L-aminokiseline u D-aminokiselinu (engl. "Epimerization domain", **domena E**) (Slika 5). Reakcija se događa dok je aminokiselina vezana na domenu PCP. Domena E je dugačka otprilike 400 aminokiselina. Postoji i druga dva tipa modula koji vežu D-aminokiseline ali u kojima nije prisutna domena E. U jednom je prisutna domena E koja nije dio sustava NRPS, ali katalizira racemizaciju supstrata vezanog na modul sustava NRPS. U drugom tipu modula prisutna je domena A koja specifično prepoznaje i aktivira D-aminokiseline (Konz i Marahiel, 1999).

Još jedan tip netipične domene je, domena za ciklizaciju dugačka otprilike 450 aminokiselina (engl. "Cyclization domain", **domena C_y**). Domena C_y se može naći umjesto domene C, i to u modulima koji ugrađuju serin, treonin ili cistein. Domena C_y katalizira

formiranje hetero-cikličkih prstenova, kao što su oksazolinski ili tiazolinski prsten, i to prilikom stvaranja peptidne veze (Konz i Marahiel, 1999; Schwarzer i sur., 2003).

Kao pratilac domena C_y u modulima sustava NRPS javlja se i oksidacijska domena (engl. "Oxidation domain", **domena Ox**) (Slika 5). Ona može biti smještena na dva različita mjesta u modulima NRPS, unutar domene A nekog modula ili na C-terminalnom kraju domene PCP. Domena Ox katalizira oksidaciju tiazolinskog prstena (Schwarzer i sur., 2003).



Slika 5. Netipične domene sustava NRPS s pozicijama unutar modula i reakcijama.

Pokazalo se da je za aktivnost domena unutar sustava NRPS važno i okruženje u kojem se domene nalaze. Tako imamo dva tipa utjecaja. Utjecaj *Cis*, domene se nalaze jedna do druge na istom polipeptidnom lancu. Utjecaj *Trans*, domene se nalaze na dva različita proteina. Neki sustavi NRPS su izgrađeni i od više različitih polipeptidnih lanaca. Jedan primjer je tirocidin sintetaza, sustav NRPS koji se sastoji od tri polipeptidna lanca (TycA, TycB i TycC) (Lautru i Challis, 2004). Strukturalna raznolikost produkata multienzima NRPS proizlazi iz broja i rasporeda modula unutar sustava NRPS, supstrata koje aktiviraju njihove domene A i prisutnosti dodatnih modificirajućih domena, kao što je npr. domena C_y (Felnagle i sur., 2008). Nakon sinteze i odvajanja sa sustava NRPS polipeptid može biti dodatno modificiran u postsintetskim reakcijama halogenacije, hidroksilacije ili glikolizacije. Enzimi koji kataliziraju te reakcije djeluju u utjecaju *trans*. To su samostalni enzimi koji su udruženi s kompleksom NRPS, tzv. **prateći enzimi**. Geni koji sadržavaju genetsku uputu za

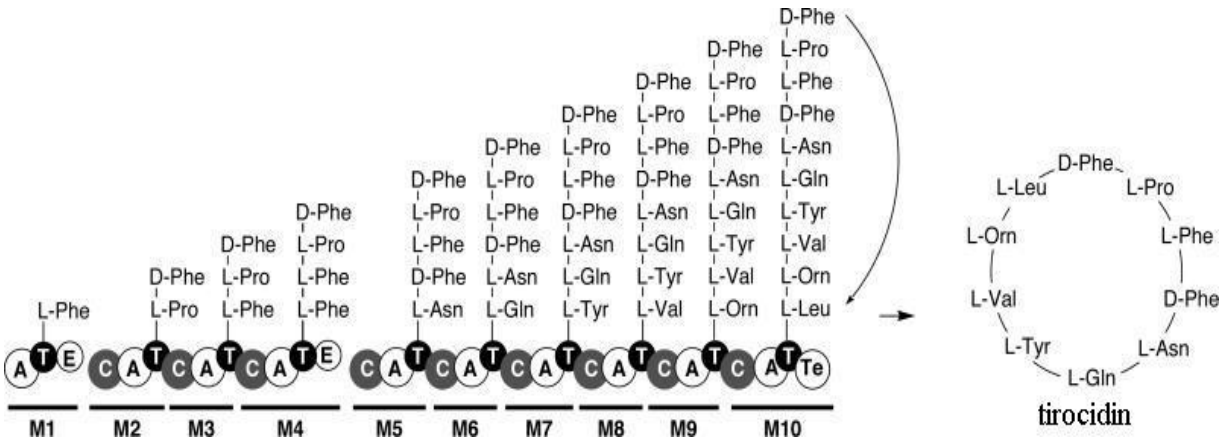
takav tip enzima najčešće su dio genske nakupine NRPS, što omogućuje zajedničku kontrolu transkripcije tih enzima i pripadajućeg kompleksa NRPS. Primjer takvih enzima su glikozil transferaza, halogenaza i oksidaza (Samel i sur, 2008; Schwarzer i sur., 2003).

2.1.2. Arhitektura genskih nakupina i enzima

Tri su tipa organizacije modula unutar sustava NRPS: LINEARNA, PONAVLJAJUĆA i NELINEARNA.

2.1.2.1. Linearne sintetaze neribosomalnih peptida

U linearnom sustavu NRPS, primarna struktura neribosomalno sintetiziranog polipeptida odgovara broju i rasporedu modula koji sačinjavaju taj sustav NRPS, tzv. "pravilo kolinearnosti". U takvom linearnom sustavu svaki modul se upotrebljava samo jednom tijekom sinteze. Broj modula označava i broj aminokiselina koje će sačinjavati sintetizirani polipeptid. Organizacija modula takvog sustava može se prikazati kao A-PCP-$[C\text{-}A\text{-}PCP]_n$-C-A-PCP-Te, pri čemu modul A-PCP predstavlja inicijacijski modul. Elongacijski modul je C-A-PCP, a terminacijski modul je oblika C-A-PCP-Te. Osim osnovnih domena, ovi moduli mogu sadržavati i netipične domene (Schwarzer i sur., 2003). Primjer proizvoda ovakvog tipa sustava NRPS su: β-laktami, daptomicin, ciklosporin A, tirocidin (Slika 6) itd.

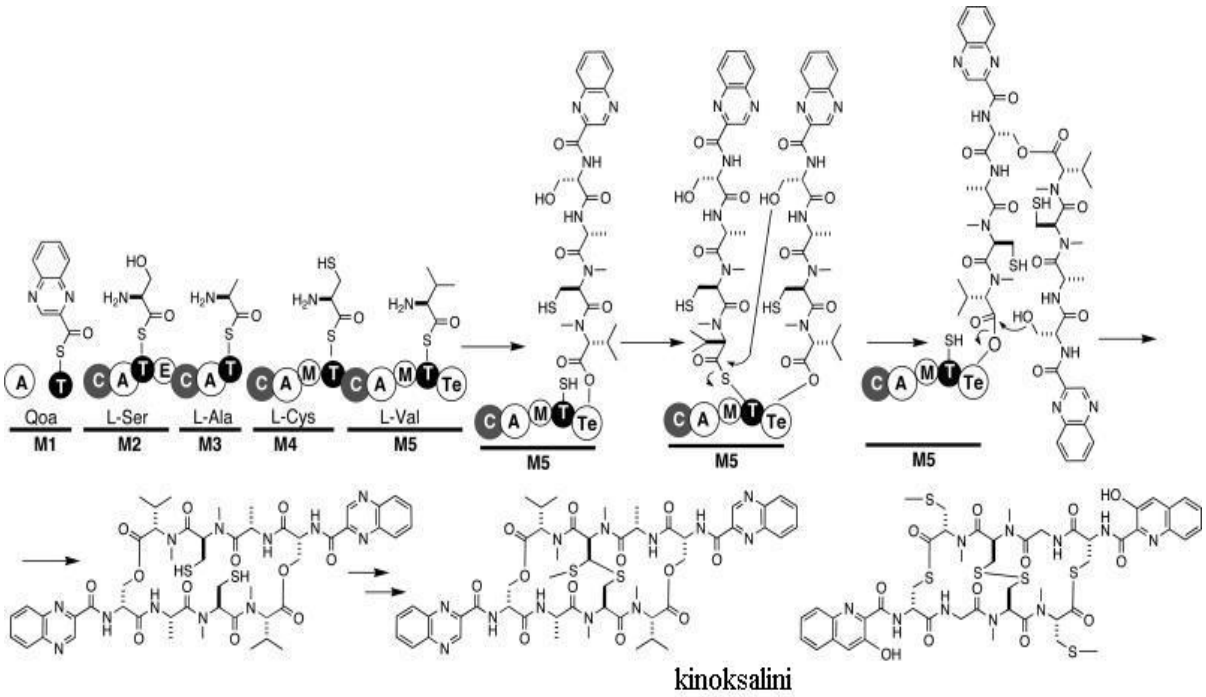


Slika 6. Mehanizam sinteze antibiotika tirocidina pomoću linearne sintetaze neribosomalnih peptida (Felnagle i sur., 2008).

Pokazalo se da je otprilike 80% svih genskih nakupina sustava NRPS, prisutnih u bazi NORINE, ko-linearno i modularno. To ukazuje na to da se transkripcija i translacija većine sekvencioniranih genskih nakupina sustava NRPS događa u jednom smjeru (Li i sur., 2009).

2.1.2.2.Ponavljajuće sintetaze neribosomalnih peptida

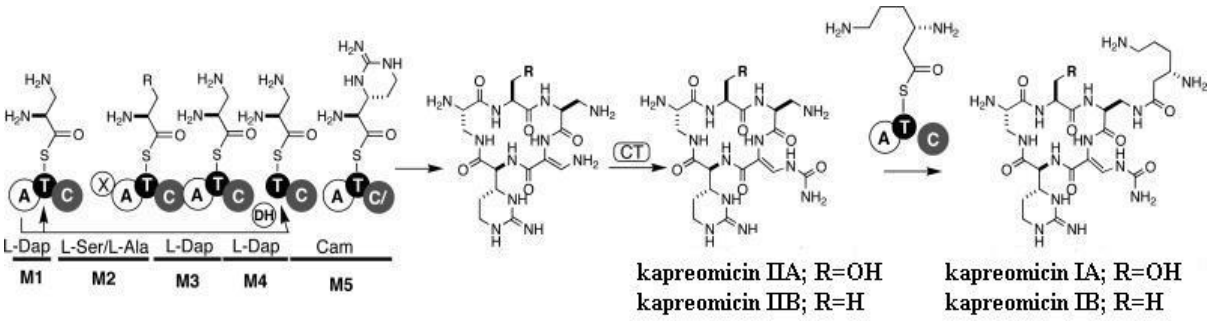
Ovaj tip sustava NRPS upotrebljava module i domene više puta tijekom sinteze jednog polipeptidnog lanca. Primjer takve sinteze je sinteza peptida enterobaktina. Enterobaktin sintetaza se sastoji od dva modula koja se upotrebljavaju 3 puta (moduli Dhb-Ser). Svaki od ta tri nastala međuprodukta se veže na krajnju Te domenu, na kojoj se događa oligomerizacija u konačni polipeptidni produkt koji se oslobađa nakon ciklizacije s tog sustava NRPS (Schwarzer i sur., 2003). Zaključak o linearnosti sustava NRPS moguće je izvesti iz primarne strukture samo u nekim slučajevima. Teško je razlikovati linearne i ponavljajuće sustave NRPS, jer je teško predvidjeti aktivnost domene Te samo iz njene primarne strukture. Tako ponavljajuće sustave NRPS možemo smatrati linearnim sustavima NRPS s ponavljajućom aktivnošću domene Te (Slika 7) (Schwarzer i sur., 2003).



Slika 7. Mehanizam sinteze kinoksalina pomoću ponavljajuće sintetaze neribosomalnih peptida (Felnagle i sur., 2008).

2.1.2.3.Nelinearne sintetaze neribosomalnih peptida

Odlika nelinearnih sistema je da sadržavaju netipična svojstva arhitekture sustava NRPS. Tako je prisutno barem jedno odstupanje od organizacije osnovnog modula C-A-PCP, i to kroz ponavljajuću aktivnost pojedinih domena a ne svih domena kao što je to slučaj s ponavljajućim sustavima NRPS. U većini slučajeva nelinearnost se može predvidjeti iz primarne strukture takvih sustava NRPS, jer je vidljivo odstupanje od osnovne organizacije modula A-PCP-[C-A-PCP]$_n$-C-A-PCP-Te (Schwarzer i sur., 2003). Isto tako ovi sustavi često sadržavaju netipične domene čija je funkcija često nepoznata. Hibridni sustavi NRPS-PKS također pripadaju ovoj skupini sustava NRPS. Primjer takvog nelinearnog sustava NRPS je sustav odgovoran za sintezu kapreomicina (Slika 8). Sustav NRPS za sintezu kapreomicina ima četiri domene A, iako se polipeptidna jezgra kapreomicina sastoji od 5 aminokiselina i prisutno je 5 domena PCP. Pretpostavka je da se jedna od domena A upotrebljava 2 puta tijekom sinteze. Također 2. modul sadržava netipičnu domenu nepoznate funkcije, ta domena vjerojatno ima ulogu u diferencijaciji između aminokiselina L-Ser i L-Ala jer se na drugoj poziciji jezgre polipeptida kapreomicina mogu naći obje aminokiseline. Još su dvije karakteristike koje ga svrstavaju u skupinu nelinearnih sustava NRPS, kao što je prisustvo modificirane domene C umjesto terminalne domene Te na kraju sustava NRPS kapreomicina (Felnagle i sur., 2008). Nelinearni sustavi NRPS čine najčešći oblik sustava NRPS u prirodi. Produkti takvih sustava su nelinearni, često s neobičnim cikličkim i razgranatim strukturama.



Slika 8. Mehanizam sinteze četiri komponente antibiotika kapreomicina pomoću nelinearne sintetaze neribosomalnih peptida (Felnagle i sur., 2008).

2.1.3. Supstrati sustava NRPS

Peptidi sintetizirani sustavima NRPS posjeduju izvanrednu strukturnu raznolikost. Ti sustavi variraju od jednostavnih linearnih, pa do cikličkih i razgranatih polipeptida te sadržavaju netipične strukture u lancu polipeptida kao što su razni heterociklički prsteni. Ta izvanredna raznolikost može se objasniti kroz aktivnost multienzimskih sustava NRPS koji ugrađuju ne samo 20 proteinogenih (uobičajenih) aminokiselina, nego i neproteinogene (neuobičajene) aminokiseline. Više od 500 takvih neuobičajenih monomera je identificirano, kao što su D-aminokiseline, N-metilirane aminokiseline, različite varijante hidroksi kiselina i aminokiselina koje mogu prolaziti i složenije modifikacije kao što su acilacija i glikolizacija (Stachelhaus i Marahiel, 1995; Strieker i sur., 2010). Osim toga, često su aminokiseline u takvim polipeptidima vezane i dodatnim vezama osim peptidne veze i disulfidnog mosta (Challis i Naismith, 2004). Osim aminokiselina, lanac sintetiziran sustavom NRPS može sadržavati i neke druge jedinice kao što je acetat ili propionat. Acetat i propionat su integrirani u lanac pomoću multienzima poliketid sintetaza (PKS). U tom slučaju riječ je o velikim multienzimskim kompleksima NRPS-PKS. U tim kompleksima odnos NRPS i PKS varira od kompleksa do kompleksa. Na primjer, u slučaju antibiotika bleomicina A2 prisutan je samo jedan modul PKS, a za kancerostatik epotilon minimalno 9 modula PKS i samo jedan modul NRPS (Schwarzer i sur., 2003).

2.1.4. Domene za adenilaciju

Domene za adenilaciju (domene A) su većinom visoko selektivne i toleriraju vrlo malo strukturalne varijabilnosti od svojih supstrata. Zbog te specifičnosti domene A su podvrgnute intenzivnom proučavanju kako bi se razjasnio temelj njihove selektivnosti za određenu aminokiselinu (Challis i Naismith, 2004). Razumijevanje mehanizama selektivnosti domena A omogućit će pretpostavke o mogućim proizvodima novo sekvencioniranih multienzima NRPS i inženjering novih farmaceutski važnih spojeva. Prvu identifikaciju domene A uz pomoć biokemijskih postupaka objavili su Stachelhaus i Marahiel 1995. godine. U tom radu prvi put je identificiran 556 aminokiselina dug fragment sustava NRPS koji je odgovoran za prepoznavanje i aktivaciju aminokiselina uz pomoć ATP-a. Strukturalna analiza domene A pokazala je sličnost domene A s luciferazom iz mušice *Photinus pyralis*, acetil-CoA sintetazom iz bakterije *Salmonella enterica* i kvasaca, 4-klorobenzoat-CoA ligazom iz bakterije *Alcaligenes sp.*, te acil-CoA sintetazom iz bakterije *Thermus thermophilus* (Yonus i

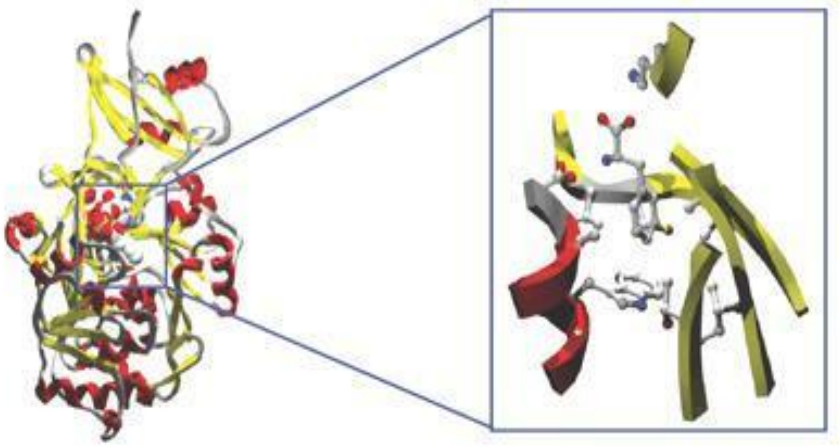
sur., 2008). Svi enzimi pripadaju adenilat formirajućoj superobitelji koja ima više od 17.000 sekvenci iz više od 1.550 vrsta koje pripadaju carstvima bakterija, praživotinja ili eukariota. Unatoč niskoj sličnosti sekvencija, ta superobitelj pokazuje visoku razinu strukturalne i funkcionalne sličnosti (Li i sur., 2009). Svi enzimi kataliziraju ATP ovisnu aktivaciju supstrata u obliku acil-adenilata. Taj tip reakcije prisutan je i u sintezi polipeptida na ribosomima, ali domena A ne pokazuje nikakvu evolucijsku ili strukturalnu sličnost s enzimima koji učestvuju u sintezi polipeptida na ribosomima, bilo da je riječ o aminoacil-tRNA sintetazi I ili aminoacil-tRNA sintetazi II. Općenito, sinteza polipeptida kontrolirana sustavima NRPS pokazuje "manje strogu" selekciju supstrata nego li sinteza polipeptida na ribosomima.

Proučavanje strukture i aktivnosti domene A olakšano je nakon otkrića kristalne strukture domene A prvog modula gramicidin S sintetaze, GrsA, koja aktivira aminokiselinu L-Phe, tzv. domena PheA. Kristalnu strukturu domene PheA otkrili su Conti i suradnici, 1997. godine (Slika 10). Antibiotik gramicidin S je sekudarni metabolit gram(+) bakterije *Bacillus brevis*, a sintetizira se uz pomoć multienzimskog kompleksa NRPS gramicidin S sintetaze. Geni koji sadržavaju genetičku uputu za taj multienzim organizirani su u obliku velikog *grs* operona dužine 19 kb koji uključuje gene *grsA*, *grsB* i *grsT*. Gen *grsA* odgovoran je za transkripciju i translaciju 1.098 aminokiselina dugog proteina GrsA. Taj je protein odgovoran za prepoznavanje, aktivaciju i vezanje, te racemizaciju aminokiseline L-Phe. Dio modula proteina GrsA predstavlja i domena PheA. Domena PheA odgovorna je za prepoznavanje aminokiseline L-Phe, te aktivaciju te iste aminokiseline u reakciji adenilacije α-karboksilne skupine L-Phe uz pomoć α-fosfata ATP-a. Kristalizacija domene PheA zajedno sa supstratima, aminokiselinom L-Phe, AMP-om i Mg^{2+} omogućila je identifikaciju aktivnog mjesta za vezanje supstrata, kao i aminokiselinske ostatke te domene koji sudjeluju u reakcijama prepoznavanja i aktiviranja supstrata (Conti i sur., 1997; Lautru i Challis 2004).

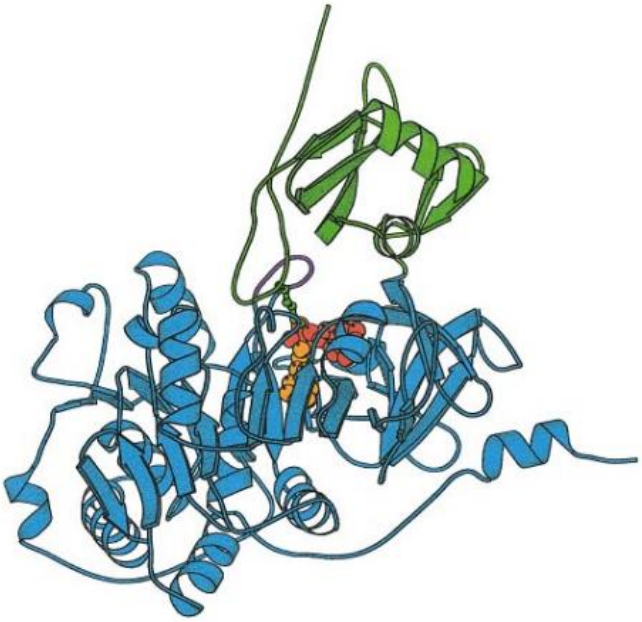
2.1.4.1. Kristalna struktura domena za adenilaciju

Domena A je otprilike veličine 550 aminokiselina, podijeljenih u dvije strukturalne pod-domene. Velika N-terminalna pod-domena sastoji se od oko 420 aminokiselina, a manja C-terminalna pod-domena od oko 110 aminokiselina. Pod-domene su povezane pomoću malog mosta koji se sastoji od 5-10 aminokiselina. Između tih dviju strukturalnih pod-domena prisutno je vrlo malo izravnih interakcija protein-protein. Većina komunikacije

između strukturalnih pod-domena dešava se uz pomoć vodikovih veza aminokiselinskih ostataka, te uz pomoć tankog sloja molekula vode na površini peptida (Slike 9 i 10) (Conti i sur., 1997).

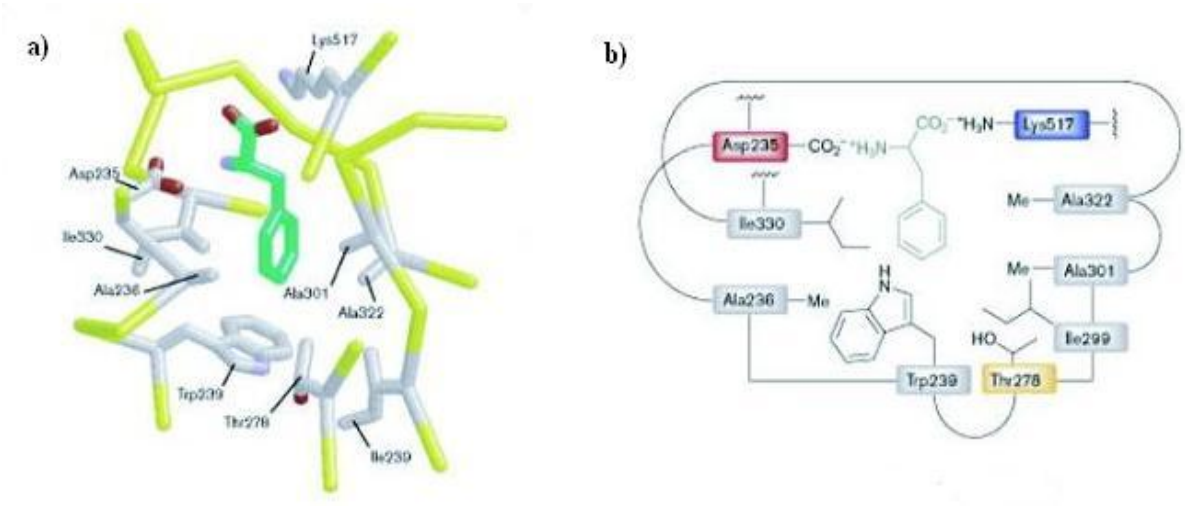


Slika 9. Struktura pod-domena domene PheA s vezanim supstratom L-Phe i uvećanim mjestom za vezenje supstrata (Lautru i Challis, 2004).



Slika 10. Vrpčasti 3D prikaz domene A s velikom N-terminalnom pod-domena (plave boje) i manjom C-terminalnom pod-domena (zelene boje). Prikazani supstrati: L-Phe (narančastom bojom) i AMP (crvenom bojom) (Conti i sur., 1997).

Aktivno mjesto domene A nalazi se u utoru između dvije strukturalne pod-domene. Reakcija prepoznavanja supstrata određena je uz pomoć mreže vodikovih veza između L-Phe i ostataka aminokiselina koji čine aktivno mjesto. Ti aminokiselinski ostaci nalaze se u području dugom otprilike 100 aminokiselina većinom na N-terminalnoj pod-domeni, uz neke iznimke. Jedna od tih iznimka je i aminokiselina Lys-517 koja se nalazi na C-terminalnoj pod-domeni. Ta aminokiselina odgovorna je za stabilizaciju α-karboksilne skupine supstrata i AMP-a. Ona fiksira supstrate u aktivnom mjestu i zatvara N-terminalnu pod-domene u produktivan oblik. Bočni lanac ostatka aminokiseline Asp-235 zajedno sa kisikom iz karboksilnih skupina glavnih lanaca aminokiselina Gly-324 i Ile-330, odgovorni su za stvaranje vodikovih veza sa α-amino grupom supstrata. Bočni lanac supstrata (L-Phe) okružen je s jedne strane aminokiselinama Ala-236, Ile-330 i Cys-331, te s druge strane aminokiselinama Ala-322, Ala-301, Ile-229 i Thr-278. Te dvije strane odvojene su s aminokiselinskim ostatkom Trp-239, koji se nalazi na dnu aktivnog mjesta (Slika 11) (Stachelhaus i sur., 1999). Na drugom kraju aktivnog mjesta, nalazi se kanal ispunjen vodom koji spaja aktivno mjesto s okolinom. Dvije strane aktivnog mjesta odvojene su dovoljno da bi se omogućio i olakšao ulazak aromatskog ostatka supstrata.



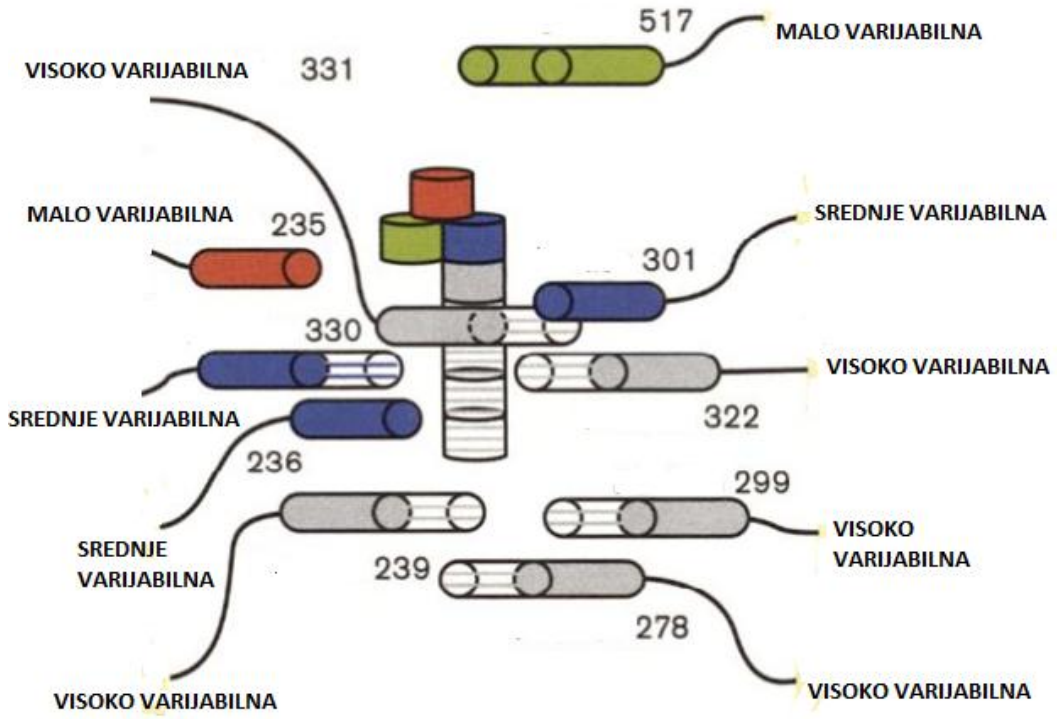
Slika 11. Struktura 3D aktivnog mjesta domene PheA sa supstratom L-Phe i aminokiselinskim ostacima odgovornim za vezanje L-Phe (a). Struktura 2D aktivnog mjesta domene PheA sa supstratom L-Phe i aminokiselinskim ostacima odgovornim za vezanje L-Phe (b). Ostaci su označeni: kiseli, crveno; bazični, plavo; neutralni polarni, žuto; hidrofobni, sivo (Challis i sur., 2000).

2.1.4.2. Kod sustava NRPS

Zbog međusobne sličnosti sekvencija različitih domena A, koje variraju od 26 % do 56 %, izvjesno je da je i konformacija glavnog lanca tih enzima vrlo slična. U skladu s time, razlike u aktivaciji različitih supstrata proizlaze iz prirode aminokiselinskih ostataka koji se nalaze u aktivnom mjestu. Stachelhaus i suradnici (1999), te Challis i suradnici (2000), pretpostavili su da bi pozicije aminokiselinskih ostataka unutar aktivnog mjesta domene PheA, koji su odgovorni za reakciju s aminokiselinom L-Phe, odgovarale isto pozicioniranim ostacima i u drugim domenama A. Upotrebom analize sekvencija aktivnoga mjesta domene PheA s ostalim domenama A, identificirano je 10 pozicija aminokiselinskih ostataka (Challis i suradnici su identificirali 8 pozicija), koje su najodgovornije za prepoznavanje supstrata od strane domena A, tzv. "kod sustava NRPS" ili "10AA kod" (Challis i sur. 2000; Stachelhaus i sur., 1999). Pozicije tih ostataka su: 235, 236, 239, 278, 299, 301, 322, 330, 331 i 517. Naziv "kod sustava NRPS" proizlazi iz činjenice da su ti aminokiselinski ostaci odgovorni za selekciju supstrata sustava NRPS. Isti princip nalazimo i u translaciji proteina na ribosomima, pri čemu se ključni ostaci koji su odgovorni za selekciju supstrata nalaze na t-RNA, to je tzv. "antikodon". Upotrebljavajući filogenetske analize dobili su odnose između dijelova domena A, koji sadrže, te ostatke i supstrata koje one aktiviraju, tj. domene koje prepoznaju iste supstrate su se zajedno grupirale. Iz tako dobivenih različitih grupa domena A, ovisno o tome koji supstrat aktiviraju, odredili su se specifični ostaci koje se nalaze na tim pozicijama a koji su odgovorni za vezanje određenog supstrata.

Prema radu Stachelhaus i suradnika iz 1999. godine, te ostatke moguće je klasificirati na tri tipa ostataka unutar "koda sustava NRPS". Tipovi ovise o stupnju varijabilnosti aminokiselina unutar pojedinih grupa odgovornih za aktivaciju iste aminokiseline a koji se nalaze na tim pozicijama. Pokazalo se da su pozicije 517 i 235 **malo varijabilne pozicije** unutar domena A. Pozicija Asp-235 prisutna je u svim domenama koje prepoznaju i aktiviraju supstrate koji imaju α-amino skupinu. Pozicija Lys-517 je strogo konzervirana unutar svih domena A. Taj aminokiselinski ostatak ulazi u interakcije sa α-karboksilnom skupinom supstrata te 4-OH i 5-OH skupinom riboze molekula ATP ili AMP. Pozicija Lys-517 djeluje kao "vratar" aktivnog mjesta domena A. Pozicije Asp-235 i Lys-517 iako važni ostaci za vezanje supstrata, ne mogu se smatrati ostacima koji su odgovorni za prepoznavanje točno određene aminokiseline. Tako se u mnogim modelima za prepoznavanje supstrata od strane domena A pozicija Lys-517 zanemaruje (Bushley i sur., 2008). **Srednje varijabilne**

pozicije 236, 301 i 330 većinom imaju hidrofobne alifatske ostatke. **Visoko varijabilne pozicije** 239, 278, 299, 322 i 331 ovise izravno o supstratu koji se aktivira od strane određene domene A. Srednje i visoko varijabilne pozicije daju ostatke koji su odgovorni za oblikovanje aktivnog mjesta i omogućuju prepoznavanje bočnog lanca točno određenog supstrata (Slika 12) (Stachelhaus i sur., 1999). Tako domene A koje aktiviraju supstrate s nabijenim bočnim lancima, kao što su to aminokiseline Gln, Orn, Asn na pozicijama 239 i 278, imaju suprotno nabijene ostatke nego što je supstrat kojeg te domene A prepoznaju (Conti i sur., 1997).



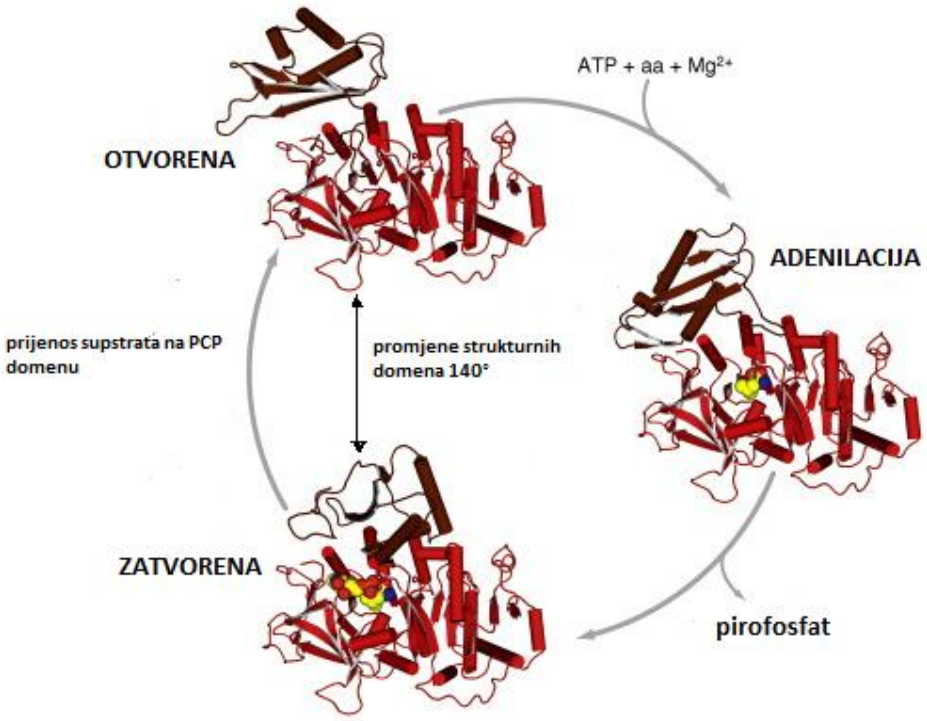
Slika 12. Podjela tipova pozicija u "kodu sustava NRPS". Grupe ovise o varijabilnosti aminokiselina unutar pojedinih grupa odgovornih za aktivaciju iste aminokiseline a koji se nalaze na tim pozicijama (Stachelhaus i sur., 1999).

2.1.4.3. Konformacijske promjene domene A

Domena A prilikom svoje aktivnosti prolazi nekoliko konformacijskih stanja. Yonus i suradnici su 2008. godine objavili istraživanje domene A DltA, domene koja aktivira aminokiselinu D-alanin. U tom istraživanju proveli su iscrpno modeliranje odnosa

strukturalnih dijelova domene A. Na taj su način postavili temelje za model ciklusa promjenjivih konformacijskih stanja prilikom aktivnosti domena A (Yonus i sur., 2008).

Ciklus započinje otvorenom konformacijom domena A. U toj konformaciji C-terminalna strukturna domena je odmaknuta od aktivnog mjesta, te je prisutno malo izravnih interakcija između dvije strukturalne domene. Takav raspored strukturnih domena omogućuje lakše vezanje supstrata domene A, aminokiseline i ATP-a te Mg^{2+}. Nakon vezanja supstrata dolazi do reakcije adenilacije aminokiseline koja za posljedicu ima i oslobađanje pirofosfata. Posljedica tih reakcija je promjena konformacije domene A iz otvorene u zatvorenu konformaciju. U toj konformaciji nastali aminoacil-AMP je zaštićen od okoline. U sljedećem koraku aktivirana aminokiselina se prenosi na domenu PCP sustava NRPS što dovodi do ponovne promjene konformacije domene A iz zatvorene u otvorenu konformaciju. Istraživanja upućuju na činjenicu da su male energetske promjene odgovorne za pojedina konformacijska stanja (Slika 13). Blage promjene unutar aktivnog mjesta su dovoljne za reorganizaciju velike i male strukturalne domene i samog aktivnog mjesta, a te reorganizacije su važne za potpuni ciklus aktivnosti domene A.



Slika 13. Konformacijske promjene domene A.

Novija istraživanja pokazuju da N-terminalni strukturni dio domena A ulazi i u jaku interakciju s domenom C. To ukazuje na činjenicu da aktivnost i konformacijske promjene domene A ovise izravno i o promjenama prisutnim na okolnim domenama unutar pojedinog modula (Strieker i sur., 2010).

2.1.4.4. Funkcionalna analiza aktivnog mjesta

Unutar aktivnog mjesta domena A, koje aktiviraju supstrate s polarnim bočnim lancima, mogu se definirati jedan ili dva konzervirana polarna aminokiselinska ostatka. Ti konzervirani polarni ostaci ulaze u vodikove i ionske veze s bočnim lancem supstrata. Ostali ostaci aktivnog mjesta su hidrofobni i variraju između domena koje aktiviraju čak i iste supstrate. S druge strane domene A koje aktiviraju supstrate s hidrofobnim bočnim lancima u aktivnim mjestima imaju samo hidrofobne bočne lance, osim u domenama koje aktiviraju prolin i pipekolat. Domene A koje aktiviraju hidrofobne supstrate pokazuju i nižu supstratnu selektivnost nego što je to slučaj s domenama A koje aktiviraju polarne supstrate. Većina domena A pokazuju i "relaksiranu specifičnost", tj. svojstvo da aktiviraju više od jednog supstrata ovisno o koncentracijama supstrata u njihovom okruženju ili o selekcijskim mehanizmima prisutnim niže na lancu sustava NRPS. To je česti slučaj sa domenama A koje aktiviraju hidrofobne supstrate. Jedan primjer takve "relaksirane specifičnosti" su domene SyrE-M8, BacA-M4 i SrfAA-M1 koje mogu aktivirati aminokiseline Glu i Asp (Challis i sur., 2000).

Također, u nekim slučajevima domene A koje imaju različita aktivna mjesta aktiviraju isti supstrat. To upućuje na određenu degeneraciju "koda sustava NRPS". Primjer takve degeneracije su četiri domene koje aktiviraju aminokiselinu L-Pro; domene A RedM, ProB i PltF imaju različite ostatke u "kodu sustava NRPS" od domene TycB. Različiti ostaci nalaze se na dnu aktivnog mjesta tih domene A i vidljivo je da nisu u direktnom kontaktu sa supstratom. Ta očita degeneracija najvjerojatnije proizlazi iz činjenice da je domena PheA, domena koja prepoznaje aminokiselinu s razmjerno velikim bočnim lancem (L-Phe) loš model za domene A koje prepoznaju aminokiseline s malim bočnim lancem kao što su aminokiseline Pro, Gly, Ala, Thr, Ser i Val. Veličina aktivnog mjesta je najvjerojatnije jedan od najvažnijih kriterija specifičnosti pojedinih domena A (Lautru i sur., 2004). Prilikom funkcionalne analiza problem predstavljaju i domene A koje pokazuju nisku razinu homologije unutar grupe domena A odgovornih za aktiviranje određenog supstrata, primjer

su domene za aminokiseline Ala i Gly. Drugi je slučaj kada je prisutna jedna, ili mali broj, domena A koje aktiviraju određeni supstrat (Challis i sur., 2000).

2.2. BIOINFORMATIČKI ALATI ZA ANOTACIJU DOMENA ZA ADENILACIJU

Testiranje bioloških aktivnosti farmaceutski važnih supstancija nije se puno promijenilo od vremena Aleksandra Fleminga. Da li neka supstancija ima antibiotsko djelovanje još uvijek se temelji na principu da li inhibira ili ne inhibira rast stanica. To je dugotrajan i financijski iscrpljujući postupak. Razvoj računalnih programa za pretraživanje baza podataka modularnih biosintetskih genskih nakupina, te predviđanje strukture takvih multienzima i njihovih potencijalnih proizvoda, ubrzalo bi i olakšalo utvrđivanje sekundarnih metabolita s potencijalnom biološkom aktivnošću. Bolje rečeno, računalno simuliranje modela novootkrivenih multienzimskih kompleksa odgovornih za sintezu sekundarnih metabolita, te predviđanje strukture i aktivnosti takvih metabolita, pokazat će se kao važan korak u razvoju biološki važnih spojeva i odmak od dosadašnje metode ispitivanja odnosno pokušaj-pogreška metode.

Sinergija dviju znanstvenih disciplina, računarstva i molekularne biologije, svoje početke vuče iz sredine devedesetih godina 20. stoljeća. Godine 1995. Fleischmann i suradnici sekvencionirali su prvi put u povijesti cijeli mikrobni genom. To je bio genom bakterije *Haemophilus influenza.* To otkriće se poklopilo s razvojem važnog računalnog grafičkog alata, GUI (engl. "Graphic User Interface"). Ta dva događaja označila su početak bioinformatičke ere. Bioinformatička era je era u kojoj je sekvencioniranje novih genoma i naknadno pretraživanje uz pomoć specijaliziranih računalnih programa predstavilo važan alat za znanost. Računalni programi, kao što su to programi za višestruko poravnanje sekvenci (eng. "Multiple sequence alignment"), programi za kristalografsku analizu, te programi za kreiranje proteinskih profila omogućavaju lakše određivanje svojstava i mogućih funkcija sekundarnih metabolita. U tu raznovrsnu i važnu skupinu spojeva pripadaju i neribosomalno sintetizirani polipeptidi (Li i sur., 2009).

2.2.1. Programski paketi za anotiranje

Za analizu genskih nakupina poliketida i neribosomalno sintetiziranih peptida razvijeno je nekoliko bioinformatičkih alata:

- **ClustScan**
- **NP.searcher**
- **SEARCHPKS** (engl. "Structure Based Sequence Analysis of Polyketide Synthases")
- **MAPSI** (engl. "Management and Analysis for Polyketide Synthase Type I")
- **Biogenerator**
- **NRPS.predictor**
- **CLUSEAN** (engl. "Cluster Sequence Analyzer")

ClustScan

ClustScan je programski paket koji omogućava brzu, poluautomatsku anotaciju modularnih biosintetskih genskih nakupina i predviđanje proizvoda tih nakupina, primarno metabolita sintetiziranih sustavima NRPS i PKS. Ta mogućnost širom otvara vrata za identifikaciju novih biološki važnih molekula. Program generira tzv. formate SMILES sekundarnih metabolita iz sekvencija genskih nakupina NRPS ili PKS. Pri tome sekvencije generirane pomoću programskog paketa ClustScan daju osnovne linearne ili cikličke konformacije metabolita NRPS/PKS (Starcevic i sur., 2008).

NP.searcher

Program NP.searcher također identificira i opisuje gene za sintezu skundarnih metabolita kao što su to metaboliti genskih nakupina sustava NRPS, PKS ili hibridnih PKS/NRPS. I taj program generira formate SMILES tih metabolita. Programi ClustScan i NP.searcher su dva komplementarna programa. Razlika između programa proizlazi iz fokusa aktivnosti i same provedbe. Tako je program ClustScan primarno orijentiran na metabolite sustava PKS, a NP.searcher na metabolite sustava NRPS (Li i sur., 2009).

SEARCHPKS

Programski paket SEARCHPKS omogućava analizu genskih nakupina sustava PKS i predviđanje domena istih. Aktivnost programa je temeljena na anotiranim sekvencama baze podataka PKSDB (Yadav i sur., 2003).

MAPSI

Programski paket MAPSI PRUŽA mogućnost analize genskih nakupina koje sadržavaju gene PKS tip I. Unutar programskog paketa korisniku je dostupna i baza podataka MapsiDB koja sadržava dostupne literaturne informacije o poliketidima (Hongseok i sur., 2009).

Biogenerator

Programski paket Biogenerator simulira *in silico* aktivnost sustava PKS i generira moguće metabolite tih sustava kao i njihovu moguću biološku aktivnost. U sklopu programskog paketa dostupna je baza tako generiranih sekundarnih metabolita (Zotchev i sur., 2006).

NRPS.predictor

Programski paket NRPS.predictor, i njegova novija verzija programski paket NRPS.predictor2, su programski paketi koji pružaju mogućnost određivanje specifičnosti adenilacijski domena sustava NRPS (Rausch i sur., 2005; Röttig i sur., 2011).

CLUSEAN

Programski paket Clusean pruža mogućnost anotiranja i analize genskih nakupina NRPS i PKS, kao i predviđanje specifičnosti domena A sustava NRPS (Weber i sur., 2009).

2.2.2. Skriveni Markovljevi modeli

Počevši od druge polovice dvadesetog stoljeća pa sve do danas, ali i u budućnosti, vidljiv je trend enormnog porasta broja podataka unutar biologije. Ta velika masa podataka označila je ulazak računarstva u biologiju. Pojavilo se novo znanstveno područje nazvano bioinformatika. Dva su aspekta od interesa za bioinformatiku: razvoj programa za manipulaciju velikom masom podataka (engl. "large scale data management") i razvoj programa za strojno učenje (engl. "machine learning").

Skriveni markovljevi modeli (engl. "Hidden Markov Models", HMM) predstavljaju jedan od principa strojnog učenja i formalna su osnova za modele vjerojatnosti linearnih sekvencija. Temelj su za različite tipove računalnih programa kao što su programi za pronalaženje gena (engl. "genefinding"), pretraživači profila (engl. "profile searches"), programi za poravnavanje višestrukih sekvencija (engl. "multiple sequence alignment") i programi za pronalaženje aktivnih mjesta (engl. "active site identification"). Jednostavnije rečeno, oni danas predstavljaju osnovu za računalnu analizu linearnih sekvencija (Eddy, 2004).

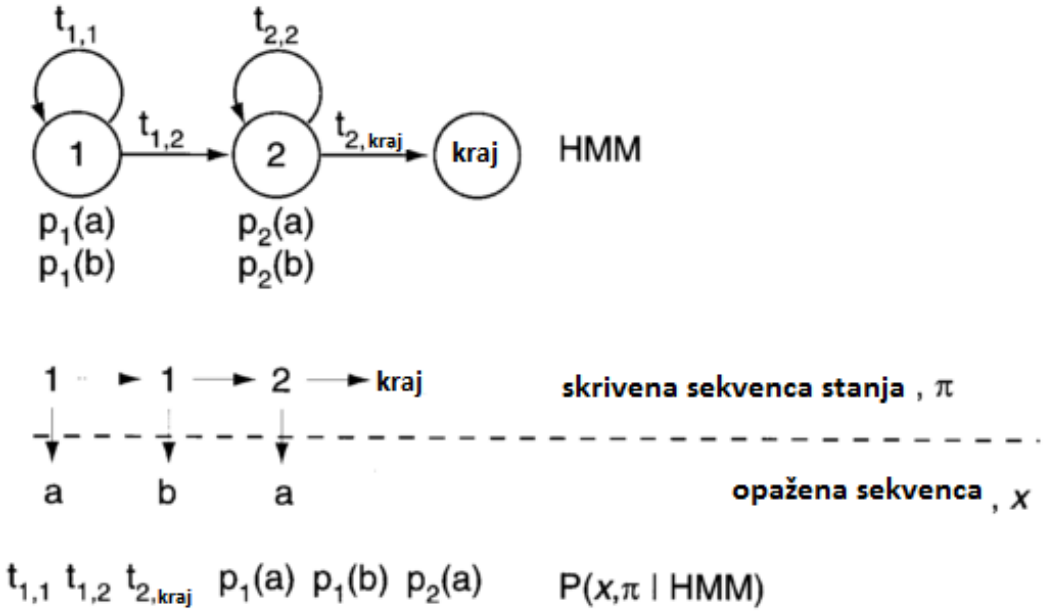
Pet je osnovnih područja unutar biologije koji danas upotrebljavaju modele vjerojatnosti ili modele HMM:

- prevođenje genetičke upute iz DNA u strukturu proteina,
- usporedba sekvencije s homolognim sekvencijama korištenjem profila HMM,
- poravnanje dviju različitih sekvencija, npr. proteina (rjeđe se koriste HMM),
- generiranje filogenetskih stabala, i
- za analizu RNA, upotrebom tzv. SCFG (engl. "stochastic.context-free grammars") (Birney, 2001)

Model HMM predstavlja graf povezanih stanja, a svako stanje ima potencijal da emitira nekoliko opažanja (pozicija) koje su vidljive. Model predstavlja vjerojatnost za stanje u vremenu t+1, s time da model HMM ne zahtijeva saznanja o svim stanjima koja su bila prije stanja t+1, nego samo stanje koje je neposredno prethodilo stanju t+1. Kako proces napreduje u vremenu kroz različita stanja, svako stanje može potencijalno emitirati nekoliko opažanja. Stanje ostaje skriveno a vidimo opažanja, otud naziv "Skriveni Markovljevi modeli". Takve modele se najčešće prikazuje grafički, u kojima su stanja prikazana kao krugovi a tranzicije (prijelazi) između stanja kao strelice između krugova. S time da je dimenzija vremena u biološkim podacima zamijenjena s pozicijom unutar sekvencije. Vjerojatnost $P(S,\pi|HMM, \theta)$ da model HMM s parametrom θ generira stanja π i opaženu sekvencu S je rezultat svih emitiranih i tranzicijskih vjerojatnosti unutar modela (Slika 14.) (Birney, 2001; Eddy, 2004).

Generiranje modela HMM (Slika 14) znači odrediti:

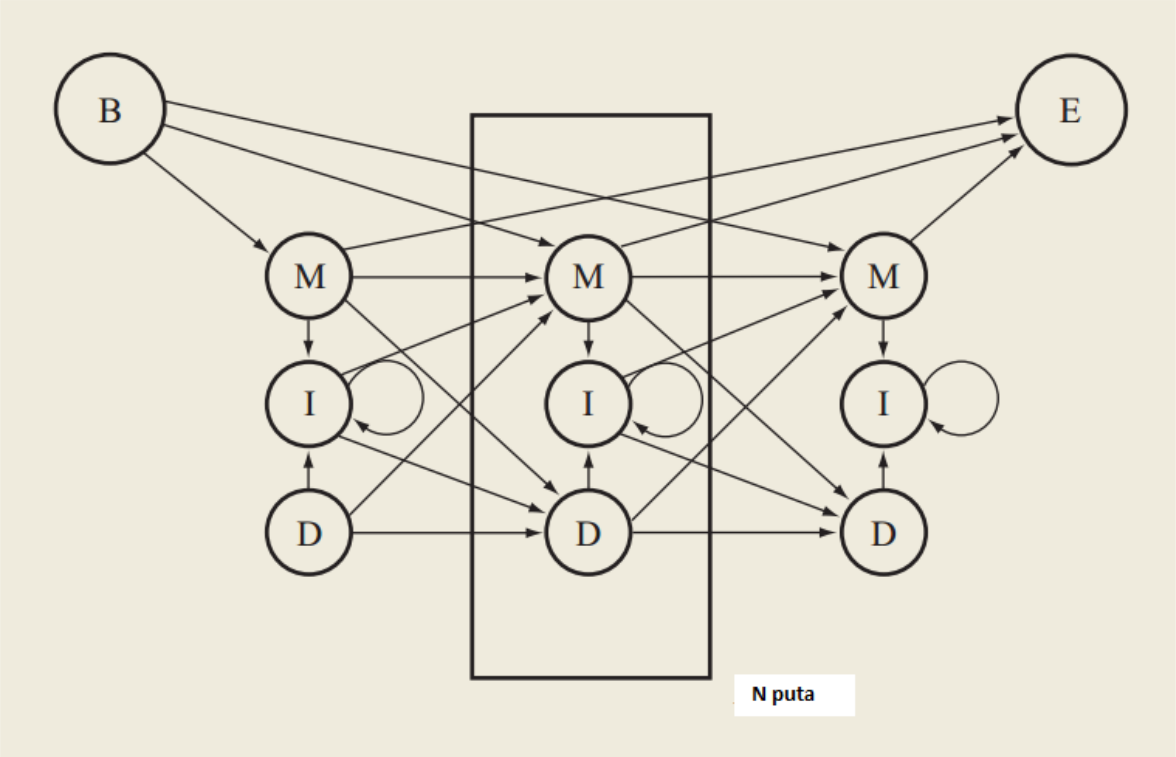
- opažanja i njihove simbole te broj,
- broj stanja u modelu,
- vjerojatnosti svakog opažanja za svako stanje modela, i
- tranzicijske vjerojatnosti za prijelaz s jednog stanja modela na bilo koje drugo (uključujući i prijelaz sa stanja na isto stanje). Tranzicijske vjerojatnosti moraju uključivati sva stanja modela.



Slika 14. Prikaz modela HMM.

Anders Krogh i suradnici su razvili model HMM koji je pogodan za analizu profila porodica proteina, nazvanog profil HMM. Profili HMM upotrebljavaju informaciju iz poravnanja višestrukih sekvencija i prenose je u pozicijski specifični bodovni sustav koji predstavlja profil tih sekvencija. Generiranje takvih profila HMM omogućava pretraživanje baza sekvencija i identifikaciju sekvencija koje su homologne generiranim profilima HMM (Eddy, 1998). Jednostavan profil HMM ima strukturu u kojoj postoji lijevo-desno tranzicija između tri osnovna stanja. Stanja su pogodak (stanje M, engl. "match"), delecija (stanje D, engl. "delete") i insercija (stanje I, engl. "insert"). Stanje M označava konsenzus aminokiselinu unutar porodice proteina, stanje D označava ne emitirajuću poziciju koja se

preskače unutar porodice. I na kraju stanje I koje označava ubacivanje nepoznatog broja ostataka nakon konsenzus pozicije (Slika 15).



Slika 15. Prikaz jednostavnog profila HMM.

Postoji nekoliko dostupnih računalnih programa koji se temelje na profilima HMM ili modelima sličnim modelu HMM. Glavna razlika tih programa proizlazi iz arhitekture modela kojeg primjenjuju. Dvije su osnovne arhitekture modela HMM, profil i motiv. Modele profil označavaju modeli koji imaju stanje insercije i delecije blisko povezano sa svakim stanjem pogotka. Takva arhitektura omogućava inserciju i deleciju bilo gdje unutar ciljane sekvencije. Modele profil koriste programi HMMER i PFTOOLS.

Modele motiv odlikuju međusobno povezana stanja pogotka koja mogu biti odvojena s malim brojem inserata. Modele motiv koriste programi PROBE, META-MEME i BLOOCKS. Modele motiv mogu kreirati i profil model programi kao što su HMMER i PFTOOLS (Eddy, 1998).

Već generirani profili HMM dostupni su u unutar dviju velikih baza podataka:

- Baze podataka Pfam (http://www.sanger.ac.uk/Pfam) (Punta i sur., 2012),
- Baze podataka PROSITE (http://ulrec3.unil.ch/software/PFSCAN_form.html) (Sigrist i sur., 2010).

Upotreba profila HMM se značajno unaprijedila razvojem projekta HMMER od strane znanstvenika Seana Eddya. Cilj projekta HMMER je generiranje široko upotrebljivog i naprednog programskog paketa za analizu linearnih sekvencija i njihove homologije. Danas programski paket HMMER ima široku primjenu. Ta primjena uključena je i u važne proteinske baze podataka, kao što su baze podataka Pfam i InterPro (Finn i sur., 2011). Profili HMM toleriraju varijabilnost u duljinama kodirajućih sekvencija. Većina bioloških sekvencija ima varijabilne dužine. Računalni programi koji zahtijevaju točno određene dužine sekvencija za analiziranje, kao što su NN (engl. "neural networks") i SVM (engl. "support vector machines") su se pokazali manje uspješnim od projekta HMMER u analizi bioloških sekvencija (Birney, 2001). "Kod sustava NRPS" predstavlja aminokiselinske ostatke koji su odgovorni za selekciju supstrata sustava NRPS. Ta ugrađena specifičnost u obliku "koda" unutar sekvencija domena A sustava NRPS može se iskoristiti za generiranje skrivenih Markovljevih profila multienzimskoga sustava NRPS (Li i sur., 2009).

2.3 FILOGENETSKA ANALIZA

Filogenija je znanost o evolucijskim odnosima, a filogenetska analiza predstavlja sve postupke koji procjenjuju evolucijske odnose. Zaključci izvedeni iz filogenetske analize prikazuju se kao dijagrami u obliku drveta s granama, pri čemu grane služe za prikaz procijenjenih evolucijskih odnosa između podataka koji se filogenetski analiziraju (Baxevanis i Oullette, 2001). Filogenetska analiza danas ima široko područje upotrebe: od rekonstrukcije sekvencije gena zajedničkog pretka analiziranih skupina gena, proučavanja podrijetla i epidemiologije ljudskih bolesti, definiranja funkcija proteina, praćenja razvoja genotipskih i fenotipskih obilježja itd. (Hillis, 1997). S teoretskom bazom čvrsto utemeljenom u molekularnoj evoluciji i populacijskoj genetici, filogenetske analize sekvencija nukleinskih kiselina i sekvencija proteina igraju ključnu ulogu u praćenju povijesti evolucije vrsta (Sjölander, 2004).

Kako je sve više genoma sekvencionirano, tako se razvijao interes za pručavanje evolucije gena i proteina. Filogenetska analiza gena i proteina uključuje usporedbu homologa tj. sekvencija koje imaju isto podrijetlo ali mogu i ne moraju predstavljati istu funkciju. Homolozi se mogu svrstati u tri skupine: ortolozi, paralozi i ksenolozi. Ortolozi su homolozi koji su nastali specijacijom (razdvajanjem vrsta) i često takav tip homologa ima sličnu funkciju. Paralozi su homolozi koji su nastali duplikacijom gena i u pravilu ovaj tip homologa ima različitu funkciju. Ksenolozi su homolozi koji su nastali kao rezultat horizontalnog prijenosa gena između organizama i većinom imaju slične funkcije. Prilikom same provedbe filogenetske analize moraju se zadovoljiti osnovni koraci: izgradnja višestrukog poravnanja homolognih sekvencija, određivanje sekvencija supstitucijskim modelom, izgradnja filogenetskog stabla i evaluacija stabla.

Metode koje se koriste za izgradnju filogenetskog stabla mogu se svrstati u dvije skupine: metode temeljene na udaljenosti (engl. "distance based methods") i metode bazirane na znakovima (engl. "character base methods"). Metode temeljene na udaljenosti proračunavaju matricu uparenih udaljenosti (engl. "pairwise distances") između sekvencija u poravnanju, te nakon toga zanemaruju same sekvencije, konstruirajući stablo u potpunosti temeljeno na proračunima izvorne udaljenosti. Metode temeljene na znakovima konstruiraju stablo koje optimizira distribuciju podataka za svaku poziciju unutar analizirane sekvence. Ovaj tip metoda ne daje fiksne udaljenosti, nego su one rezultat same topologije stabla. Filogenetske metode koje ulaze u metode temeljene na udaljenostima su: UPGMA (engl. "Unweighted Pair Group Method with Aritmetic Mean"), NJ (engl. "Neighbor Joining"), FM (engl. "Fitch-Margoliash"), ME (engl. "Minimum Evolution"). Po nekim autorima ME i FM metode daju najbolje rezultate od svih metoda temeljenih na udaljenosti ali to treba uzeti s rezervom, jer uspješnost same metode prilikom izrade filogenetskog stabla ovisi o tipu i broju podataka koji se analiziraju (Baxevanis i Oullette, 2001; Desper i Gascuel, 2002). U skupinu filogenetskih metoda koje se temelje na znakovima ulaze: MP (engl. "Maximum Parsimony") i ML (engl. "Maximum Likelihood"). Te dvije metode nemaju puno zajedničkih karakteristika, osim činjenice da se obje temelje na znakovima.

3. EKSPERIMENTALNI DIO

3.1. MATERIJAL

3.1.1. Računalna podrška i operativni sustav

Rad je izrađen na računalu slijedećeg sklopovlja: prijenosno računalo HP, procesor Intel® CoreTM 2Duo CPU T5470 1.60GHz 782MHz, radne memorije 0,99GB. 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

GenBank je baza nukleotidnih sekvencija, koja obuhvaća uz sekvencije prateće biološke i bibliografske informacije. GenBank je generiran i distribuiran od strane institucije NCBI (engl. "National Center for Biotechnology Information") i dio je suradnje INSDC (engl. "International Nucleotide Sequence Database Collaboration") zajedno s institucijama ENA (engl. "European Nucleotide Archive") i DDBJ (engl. "DNA Dana Bank of Japan"). Svaki zapis unutar baze podataka GenBank ima svoj identifikacijski broj, nazvan pristupni kod koji je jedinstven unutar cijelog INSDC. Suradnja INSDC omogućava svakodnevnu razmjenu sekvencija i informacija između baza podataka, što omogućava uniformnost i točnost sekvencija nukleotida na svjetskoj razini. Baza podataka GenBank je dostupna posredstvom sustava NCBI Entrez koji omogućuje pretraživanje različitih baza podataka DNA i proteinskih sekvencija kao i brzu informaciju o taksonomiji, genomu, mapiranju te informacije o domenama i proteinskoj strukturi kao i informacije o vezanoj biomedicinskoj literaturi posredstvom baze podataka PubMed.

Unutar baze podataka GenBank dostupno je više od 380.000 sekvenci različitih organizama definiranih na razini roda ili nižoj, sa zabilježenim rastom od 3800 sekvencija mjesečno i rastom od 12,6 % unutar godine dana. Većina sekvencija dodana je od strane različitih laboratorija ili velikih sekvencijskih projekata (Benson i sur., 2011).

3.1.2.2. Baza podataka NRPS-PKS

Baza podataka NRPS-PKS dio je računalnog programa za analiziranje velikih multienzimskih kompleksa koji su odgovorni za sintezu bioloških proizvoda. Biološki

proizvodi danas čine važan dio farmaceutski važnih tvari. Baza podataka NRPS-PKS organizirana je u četiri dijela prikladna za pretraživanje i klasifikaciju domena multienzima neribosomalnih peptid sintetaza i tri tipa poliketid sintaza. Unutar baze podataka omogućeno je i određivanje specifičnosti za supstrate domena. Baza podataka NRPS-PKS podijeljena je obzirom na mehanizam biosinteze i tip prirodnih produkata:

- NRPSDB (engl. "Database of Nonribosomal Peptide Synthetases"),
- PKSDB (engl. "Database of Modular Polyketide Synthases"),
- ITERDB (engl. "Database of Type-I Iterative Polyketide Synthase"), i
- CHSDB [engl. "Database of Type-III Polyketide Synthase (Chalcone Synthases)"].

NRPS-PKS baza podataka pruža mogućnost usporedbe neopisanih NRPS/PKS genskih nakupina s podacima dostupnim u bazi podataka NRPS-PKS na temelju njihove sličnosti, specifičnosti za supstrat te motiva aktivnog mjesta. Baza podataka NRPS-PKS je međusobno povezana kako bi se omogućile i analize genskih nakupina odgovornih za biosintezu hibridnih poliketidno/peptidnih produkata. Baza podataka NRPS-PKS pruža mogućnost i predviđanje specifičnosti domena za adenilaciju, domena A, multienzima NRPS i domena za prijenos acetila, domena AT, multienzima PKS što čini ovaj program vrijednim izvorom za identifikaciju prirodnih produkata sintetiziranih od strane novo definiranih multienzima NRPS/PKS pronađenih u novo sekvencioniranim mikrobnim genomima (Ansari i sur., 2004).

3.1.2.3. Baza podataka Pfam

Baza podataka Pfam sadržava porodice proteinskih sekvencija, s trenutačno više od 13.000 ručno generiranih proteinskih porodica u svojoj zadnjoj verziji 26.0. Svaka porodica označava nakupinu proteinskih sekvencija koje između sebe imaju visoki stupanj sličnosti ili homologije. Definirana je sa statističkim modelom poznatim kao profil skrivenog Markovljevog modela (profil HMM). Svaki profil HMM određen je poravnanjem homolognih sekvencija iz kojih proizlazi profil. Baza podataka Pfam se sastoji od dva osnovna tipa proteinskih porodica, Pfam-A i Pfam-B. Pfam-A porodice su generirane ručno. Takav profil HMM proizlazi iz ručnog poravnanja homolognih sekvencija. S druge strane Pfam-B označava porodice generirane automatiziranim programom za generiranje profila, koji proizlazi iz algoritma ADDA. Profili HMM porodica baze podataka Pfam korisni su instrument prilikom anotiranja nepoznatih proteinskih sekvencija, te kao svojevrsni vodič

prilikom eksperimentalnog rada koji uključuje proteinske sekvence. Baza podataka Pfam je dostupna kroz tri osnovna poslužitelja:

- UK (http://(pfam.sanger.ac.uk)
- USA (http://pfam.janelia.org)
- Švedska (http://pfam.sbc.su.se)

U svojoj zadnjoj verziji, 26.0 baza podataka Pfam je proširena za funkcionalnu analizu porodica proteina uvrštavanjem informacija i s Wikipedije. Trenutačno je 4.909 porodica Pfam povezano sa 1.016 člankom Wikipedije.

Kao referentnu bazu proteinskih sekvencija baza podataka Pfam koristi bazu podataka UniProtKB. U lipnju 2011. godine unutar baze podataka UniProtKB bilo je definirano 15.9 milijuna proteinskih sekvencija. Trenutačna pokrivenost baze podataka UniProtKB sekvencija od strane baze podataka porodica Pfam je otprilike 80 % (Punta i sur., 2012).

3.1.2.4. Baza podataka UniProt

Baza podataka UniProt predstavlja opsežan katalog proteinskih sekvencija i njihovih funkcionalnih anotacija, te literaturne informacije vezane uz sekvencije. Cilj je baze podataka osigurati centralno mjesto potpore za rad istraživača. Baza podataka je podijeljena na četiri zasebne komponente:

- UniProtKB (engl. "UniProt Knowledgebase"),
- UniParc (engl. "UniProt Archive"),
- UniRef (engl. "UniProt Reference Clusters"), i
- UniMES (engl. "UniProt Metagenomic and Environmental Sequence Database").

Komponenta UniProtKB predstavlja centralno mjesto dostupnih informacija o proteinima sa referencama iz više različitih izvora. Komponenta UniParc je sveobuhvatan izvor do sada sekvencioniranih proteinskih sekvenci. Komponenta UniRef sadržava grupe sličnih i srodnih sekvencija što omogućava bržu pretragu baze podataka. Komponenta UniMES je generirana radi rastuće količine podataka unutar ekološke genetike (Magrane i UniProt Consortium, 2011). Baza je dostupna na Web stranici: www.uniprot.org.

3.1.3. Bioinformatički računalni paketi i programi

3.1.3.1. Programski paket CLUSTAL

Programski paket Clustal predstavlja alat za izradu višestrukih poravnanja nukleinskih i proteinskih sekvencija. Višestruka poravnanja sekvencija osnova su za većinu bioinformatičkih analiza. Analize koje uključuju višestruka poravnanja su: pronalazak karakterističnih otisaka unutar sekvencija, pronalaženje regija homologije između sekvencija, predviđanje tercijarne i kvatarne strukture sekvencija, predviđenje oligonukleotidnih početnica za metodu PCR itd. (Thompson *i sur*., 1994). Clustal serija programa je najstariji i trenutno najviše upotrebljavani alat za višestruka poravnanja sekvencija. Višestruka poravnanja se izgrađuju postepeno kroz seriju parnih poravnanja koja slijede granjanje filogenetskog stabla svih sekvencija koje se analiziraju višestrukim poravnanjem (Larkin i sur., 2007).

Prva verzija programskog paketa Clustal pojavila se u kasnim osamdesetim godinama 20 stoljeća. Prva verzija je obuhvaćala četiri zasebna programa Clustal 1 - Clustal 4, koja su se izvodila samo na računalima kompatibilnim onima tvrtke IBM. Kasnije su ta četiri zasebna programa prepisana u programskom jeziku C++ i spojena u novu verziju nazvanu Clustal V. Treću generaciju Clustal serije predstavlja verzija Clustal W, u koju su inkomponirana poboljšanja algoritama za poravnanje, koja uključuju novi pozicijsko-specifični bodovni i važnosni sustav. Nakon nje verzija programa Clustal X dodatno uključuje i grafički susretljivo sučelje (engl. "Graphical user interface") koje omogućuje lakše korištenje programa od strane korisnika. Novije verzije Clustal W2 i Clustal X2 su prepisane u programskom jeziku C++, te je tako omogućeno lakše održavanje paketa, uvođenje novih algoritama ali i veća brzina prilikom izvođenja programa te manipulacija sa većim brojem sekvencija. Najnovija verzija Clustal Ω, omogućuje znatno ubrzanje i poravnanje gotovo bilokojeg broja sekvenci te se uvodi takozvano poravnanje vanjskog profila (engl. "External profile alingnment"). Ta karakteristika omogućuje da se prilikom dodavanja novih sekvencija u višestruko poravnanje ne provodi novo poravnanje od početka nego se prvotno višestruko poravnanje prevodi u vanjski profil koji se zatim poravna sa novim sekvencijama. Time se postigla znatna ušteda vremena. Također se vanjski profil može definirati i iz postojećih profila HMM prisutnih u proteinskim bazama podataka čime je znatno poboljšana upotreba postojećih višekratnih poravnanja prisutnih u znanstvenoj literaturi (Thompson i sur., 1994; Thompson i sur., 1997; Larkin i sur., 2007; Sievers i sur., 2011).

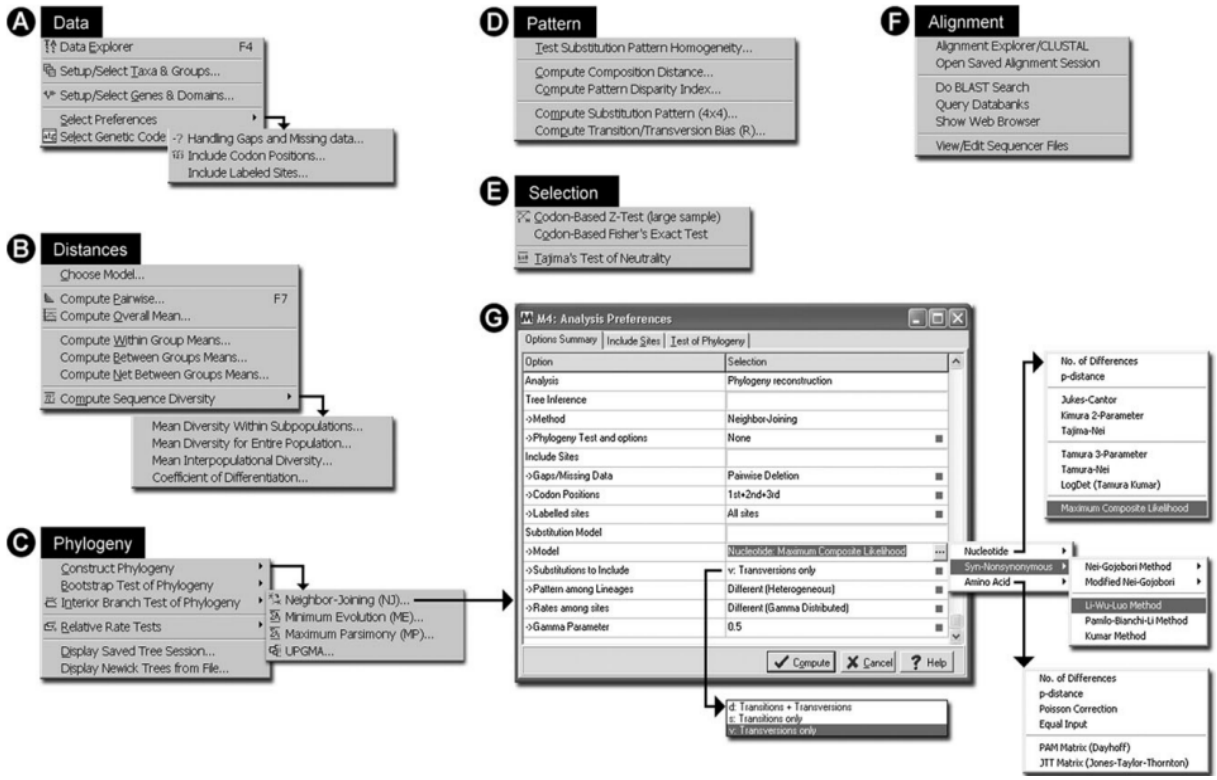
3.1.3.2. Programski paket JALVIEW

Višestruka poravnanja nukleotidnih i proteinskih sekvencija predstavljaju osnovu prilikom istraživanja nukleotidnih i proteinskih sekvencija. Danas postoje razni automatizirani alati za višestruka poravnanja sekvencija ali se pokazalo da niti jedan alat nije apsolutno točan i pouzdan. Tako i najtočniji automatizirani alati postižu manje od 50 % točnosti prilikom poravnanja sekvencija koje imaju sličnost sekvencija manju od 20 % (Clamp i sur., 2004). Istraživači zato često koriste dostupne znanstvene informacije o proučavanim sekvencijama kako bi ručno poboljšali generiranje višestrukih poravnanja.

Programski paket Jalview ("Java alignment editor") je programski paket za analizu, uređivanje i anotaciju višestruko poravnatih nukleotidnih i proteinskih sekvencija. Programski paket, Jalview verzija 1.0, prvi put se pojavio 1996. godine. Prva verzija programa sadržavala je metode za vizualizaciju i analizu poravnanja, koje su uključivale izrezivanje, dodavanje, bojanje sekvencija. Taj je programski paket isto tako sadržavao metodu za filogenetske analize i pregled 3D struktura analiziranih sekvencija. Program je podržavao razne formate poravnanja, kao što su Clustal, MSF, Fasta itd. Zbog svojih svojstava programski paket Jalview je postao alat za vizualizaciju i analizu sekvencija unutar važnih baza podataka kao što je to npr. baza podataka Pfam. No s vremenom je verzija 1.0 programskog paketa Jalview postala nedostatna za detaljnu analizu većih i duljih sekvencija, koja je postala neophodna zbog rastućeg broja sekvencija u znanstvenoj litelaturi. Zato je generirana verzija 2.0 programskog paketa Jalview koja omogućuje detaljnu analizu velikog broja sekvencija. Programski paket Jalview 2.0 sastoji se od dva zasebna programa: JalviewLite (JVL) i JalviewDesktop (JVD). Program JVL je mrežna verzija koja omogućuje jednostavnu analizu sekvencija, dok je program JVD tzv. "desktop" verzija koja omogućuje opsežniju analizu višestrukih poravnanja. Program JVL je zadržao karakteristike programskog paketa Jalview 1.0. Nasuprot tome, program JVD omogućuje dodatno: veći broj nezavisnih analiza istog poravnanja, skrivanje ili dodavanje odabranih sekvencija, zasebnu analizu odabranih sekvencija bez potrebe da se otvara dodatni prozor, anotirani redovi i konzervirani dijelovi mogu biti prikazani ispod stupova višestrukog poravnanja itd. Programski paket Jalview je dostupan na Web stranici www.jalview.org (Clamp i sur., 2004; Waterhouse i sur., 2009).

3.1.3.3. Programski paket MEGA

Filogenetska analiza predstavlja metodu i način funkcionalne klasifikacije sekvencija nukleotida i proteina s čvrstom osnovom u molekularnoj evoluciji i populacijskoj genetici. Pretpostavka razvoja filogenetskih istraživanja su razvoj automatskih metoda sekvencioniranja DNA te novih statističkih i računalnih metoda, koje obuhvaćaju programe kao što su to npr. program za molekularnu genetičku analizu evolucije poravnanjem nukleotidnih i proteinskih sekvencija (engl. "Molecular Evolutionary Genetic Analysis", MEGA). Programski paket MEGA sadržava metode za filogenetsku analizu utemeljene na udaljenosti i metode temeljene na znakovima. Četvrta verzija programa, MEGA 4.0, pruža mogućnost višestruke postupke sa ispitivanim sekvencijama, postupci su prikazani na Slici 16.

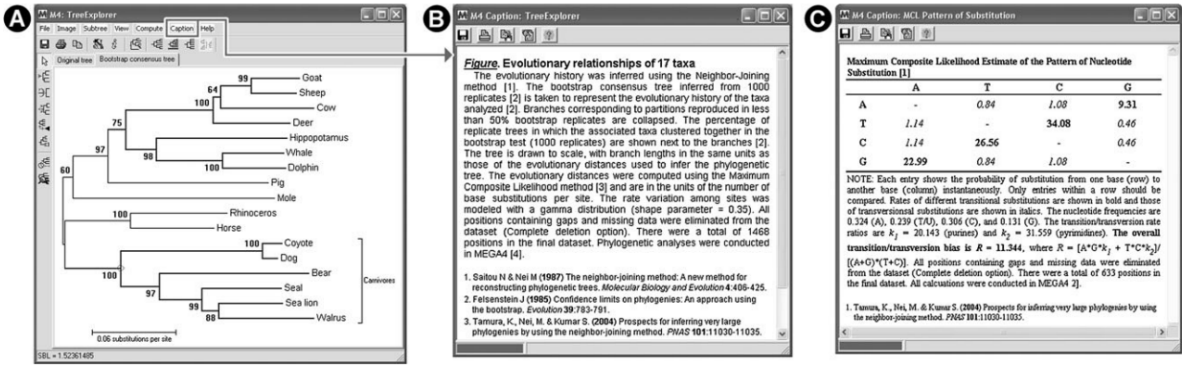


Slika 16. Mogućnosti programskog paketa MEGA verzija 4.0. Skup postupaka (od A do F) za analizu sekvencija: A različiti tipovi unosa podataka; B procjena evolucijske udaljenosti; C izrada filogenetskih stabala; D testovi homogenosti supstitucijskih postupaka i njihove procjene; E testovi selekcije; F poravnanja nukleotidnih i proteinskih sekvencija; G meni koji omogućuje korisniku selekciju modela supstitucije i raznih podsetova (Tamura i sur., 2007).

Od novina koje su uvedene u ovoj verziji programskog paketa potrebno je naglasiti tzv. "Caption Expert", to je programski modul koji generira opis provedenih postupaka od strane korisnika i navodi ga na daljnje mogućnosti za analizu (Slika 17) (Tamura i sur., 2007).

3.1.3.4. Programski paket HMMER 3

Posljednja verzija programskog paketa HMMER, nazvana HMMER 3, rezultirala su 100x većem ubrzanju rada u odnosu na prijašnje verzije. Također su omogućila i brza "online" pretraživanja pojedinačnih sekvencija proteina, višestruko poravnanje sekvencija proteina, te profila HMM unutar ciljane baze podataka proteinskih sekvencija, npr. baze podataka Pfam.



Slika 17. Prikaz sučelja "Caption Expert" (Tamura i sur., 2007).

Programski paket HMMER 3 sastoji se od četiri zasebna programa za analizu proteinskih sekvencija: *phmmer*, *hmmscan*, *hmmsearch* i *jackhmmer*. Programski paket HMMER 3 dostupan je na Web stranici (http://hmmer.janelia.org). Trenutačno su zasebni programi *phmmer*, *hmmscan* i *hmmsearch* dostupni korisnicima.

Prilikom upotrebe zasebnog programa *phmmer*, potrebno je unijeti zasebnu proteinsku sekvenciju u obliku formata FASTA. Zasebni program *phmmer* prevodi tu sekvenciju u profil HMM i uz korištenje tog profila pretražuje bazu podataka proteina. Trenutačno se tako može pretražiti šest baza podataka: NR, UniProtKB, SwissProt, PDB, UniMes i okolišni dio NR. Te su baze podataka izabrane jer sadržavaju ili sveobuhvatne zbirke proteinskih sekvencija (NCBI, NR i UniProtKB), anotirane i strukturalno definirane sekvencije (SwissProt i PDB) ili velike metagenomske sekvencije (UniMess i env NR). Nasuprot tome, zasebni program *hmmscan* koristi isto zasebnu sekvenciju koju uspoređuje s bazom podataka Pfam profila

HMM. U prethodnoj verziji programskog paketa HMMER 2, zasebni se program zvao *hmmpfam*. Zasebni program *hmmsearch* koristi profil HMM koji je izgrađen iz višestrukih poravnanja sekvencija i uz pomoć njega pretražuje baze podataka. Korisnik može unijeti ili prethodno generiran profil HMM ili poravnanje višestrukih sekvencija. Program dopušta različite formate poravnanja: MSF, SELEX, STOCKHOLM ili poravnati FASTA format. Nakon unosa poravnanja višestrukih sekvencija, one se automatski uz pomoć zasebnog programa *hmmbuilt* prevode u profil HMM. Zasebni program *jackhammer* je četvrti algoritam za pretraživanje sekvencija, koji omogućava ponavljajuća pretraživanja baza sekvencija, temeljena na profilu HMM počevši od jednog upita. Zasebni program *jackhammer* predstavlja jedan od osnovnih postupaka za kreiranje porodica proteina Pfam. Početna sekvencija predstavlja upit za minimalno 3 puta ponavljajuće pretraživanje sekvencija od strane zasebnog programa *jackhammer*. Krajnje poravnanje sekvencija za generiranje profila HMM proizlazi iz poravnanja sekvencija dobivenih zasebnim programom *jackhammer*. Trenutačno još nije dostupan za korisnike (Finn i sur., 2011). Usporedba programskog paketa HMMER 3 (Tamura i sur., 2007) i programa BLAST (Altschul i sur., 1990) prikazani su u Tablici 1.

Tablica 1. Usporedba programskog paketa HMMER 3 i programa BLAST (Finn i sur., 2011).

	HMMER		**BLAST**	**KOMENTAR**
PROGRAM **UPIT** **BAZA**	*phmmer*	Jedna sekvencija Baza sekvenca	blastp	daju slične rezultate homolognosti sekvenca
PROGRAM **UPIT** **BAZA**	*hmmscan* Baza profila HMM (Pfam)	Jedna sekvenca	rpsblast PSSM baza (CDD)	većinom se upotrebljavaju za detekciju pojedinih domena na sekvencama
PROGRAM **UPIT** **BAZA**	*hmmsearch*	Profil HMM Baza sekvenca		u programskom paketu BLAST ne postoji homolog hmmsearch
PROGRAM **UPIT** **BAZA**	*jackhmmer*	Jedna sekvencija Baza sekvencija	psi-blast	oba dva programa se koriste za iterativno pretraživanje baza sekvenci

3.1.3.5. Programski paket Microsoft Office

Unutar programskog paketa Microsoft Office (Anonymous 3, 2011) nalazi se računalni program Microsoft Excel koji je upotrebljen u izradi ovog diplomskoga rada. To je program za tablične proračune (engl. "spreadsheet"). Upotrebljena je verzija Microsoft Office Excel 2007.

3.2. METODE RADA

3.2.1. Prikupljanje literaturnih podataka

Prilikom prikupljanja literaturnih podataka upotrebljavani su "online" pretraživači literature kao što su: "Google Scholar" (Anonymous 2, 2012) i "Science direct" (Anonymous 4, 2012).

3.2.2. Prikupljanje sekvencija proteina

Skup sekvencija za generiranje profila proteina sastoji se od 397 domena za adenilaciju s poznatim supstratima. Domene su preuzete iz dodatnih materijala Rausch i suradnika (2005). U radu Rausch i suradnika definirani su pristupni kodovi svih 397 domena za adenilaciju. Pristupni kodovi su zapisani u zasebni dokument pomoću kojeg su se automatskim postupkom uz pomoć uslužnog programa "Entrez Batch" preuzele proteinske sekvencije sustava NRPS koje su sadržavale svih 397 domena A. Sekvencije su se uz pomoć programa "MATLAB" bioinformatičke funkcije *fastawrite* zapisale u obliku formata FASTA. Format FASTA jest računalni standard za zapis sekvencija nukleotida ili proteina. On je jednostavan, minimalistički zapis koji započinje sa zaglavljem. Zaglavlje je definirano na početku sa simbolom ">" nakon čega slijedi kratki opis sekvencije. Nakon zaglavlja, u novom redu slijedi sama sekvencija DNA ili proteina. Cijele sekvencije domena A su bile definirane i izrezane iz sekvencija sustava NRPS uz pomoć programskog paketa HMMER 3 i pripadajućeg profila HMM domena A.

3.2.3. Definiranje radnih sekvencija

Temeljeno na kristalnoj strukturi aktivirajuće domene za adenilaciju L-Phe prvog modula gramicidin S sintetaze GrsA, tzv. domene PheA definiralo se aktivno mjesto odgovorno za vezanje supstrata, kao i 10 aminokiselinskih ostataka unutar aktivnoga mjesta koji sudjeluju u reakcijama prepoznavanja i aktiviranja supstrata (vidi 2.1.4.2 Slika 12). Zbog međusobne sličnosti sekvencija različitih domena za adenilaciju, variraju od 26 % do 56 %. Stachelhaus i suradnici (1999) te Challis i suradnici (2000) su zaključili da će 10 ostataka aminokiseline PheA odgovarati i isto pozicioniranim ostacima u drugim domenama A prilikom višestrukog poravnanja domena A. Ti su ostaci definirani kao tzv. "kod sustava NRPS".

Isto tako, ako su ostaci aminokiselina prisutni u "kodu sustava NRPS" odgovorni za selekciju supstrata, za pretpostaviti je da će "kodovi sustava NRPS", koji su odgovorni za selekciju istog supstrata prilikom filogenetske analize svih domena A, biti zajedno grupirani. Na temelju gore navedenih saznanja definirale su se u ovom radu dvije grupe proteinskih sekvencija koje su se upotrijebile za filogenetske analize:

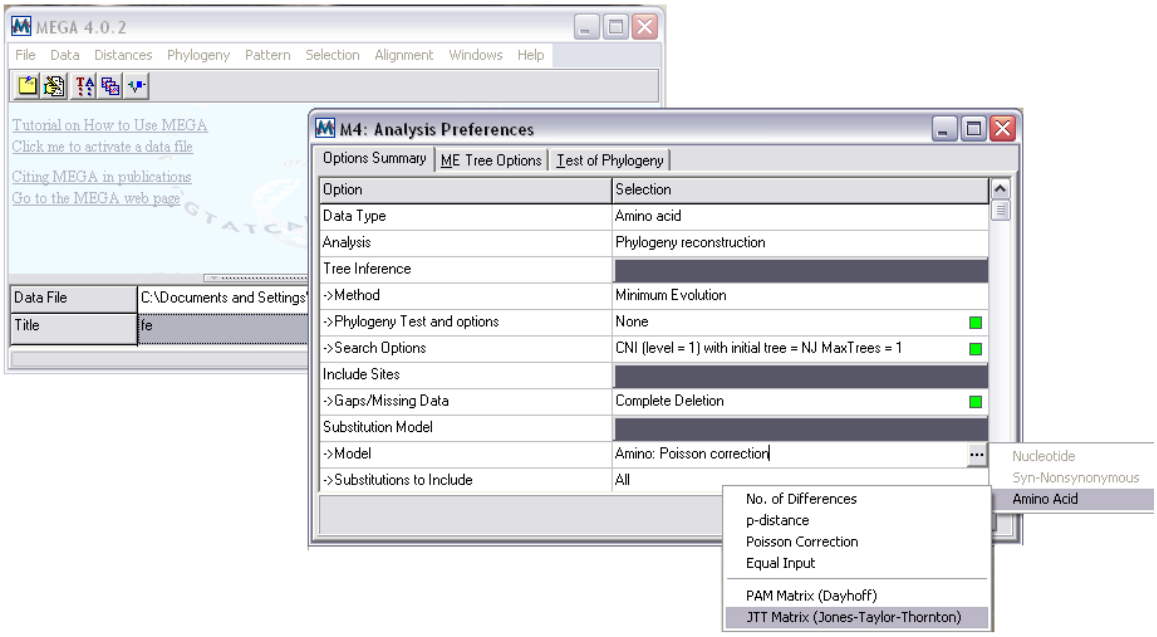
- cijele sekvencije domena A
- sekvencije aktivnih mjesta

Cijele sekvencije domena A su u prosjeku duge 400 aminokiselina. Sekvencije aktivnog mjesta predstavljaju sekvencije dužine 136 do 150 aminokiselina, koje su dobivene nakon višestrukog poravnanja svih 397 domena A s domenom PheA i izrezivanjem u odnosu na poziciju aktivnog mjesta u domeni PheA. Izrezano aktivno mjesto sadržavalo je 9 od 10 ostataka aminokiselina iz "koda sustava NRPS". Iz koda se izbacila aminokiselina Lys na poziciji 517 jer je ta aminokiselina visoko sačuvana unutar porodice domena A, te je za pretpostaviti da nema ulogu u selekciji supstrata.

3.2.4. Izrada filogenetskog stabla

Generiranje filogenetskog stabla provedeno je uz pomoć programskog paketa MEGA 4.0.2. Prije izrade samog stabla morala su se generirati višestruka poravnanja sekvencija proteina. Višestruka poravnanja domena A su provedena uz pomoć programskog paketa Clustal W uz korištenje matrice BLOSUM za poravnanje. Izradilo se potpuno poravnanje, dok su za kaznene bodove za otvaranje, zatvaranje i produljivanje praznina upotrijebljene standardne vrijednosti definirane u samom programskom paketu (Larkin i sur., 2007).

Poravnate cijele sekvencije domena A korištene su zatim za filogenetsku analizu u kojoj su se koristile dvije filogenetske metode: "Minimum Evolution" i "Maximum Parsimony" (Tamura i sur., 2007). Prvo su se unijele sekvencije u program MEGA, a generiranje filogenetskog stabla provedeno je odabirom opcije "Phylogeny→ Minimum Evolution" (ME) i kasnije "Phylogeny→Maximum Parsimony" (MP). Pri tome se kod ME koristio "model" JTT sa "rates" GAMMA, te je bilo provedeno 1000 "bootstrap" poduzorkovanja. Isto tako pri MP bilo je provedeno 1000 "bootstrap" poduzorkovanja. Ostale su opcije bile zadržane kao dio originalnih postavki. Dobiveno filogenetsko stablo iskoristilo se radi daljnje obrade i analize (Slika 18).



Slika 18. Prikaz izbornika programskog paketa MEGA 4.0.2 prilikom odabira svojstava za konstrukciju filogenetskog stabla iz poravnatih sekvencija.

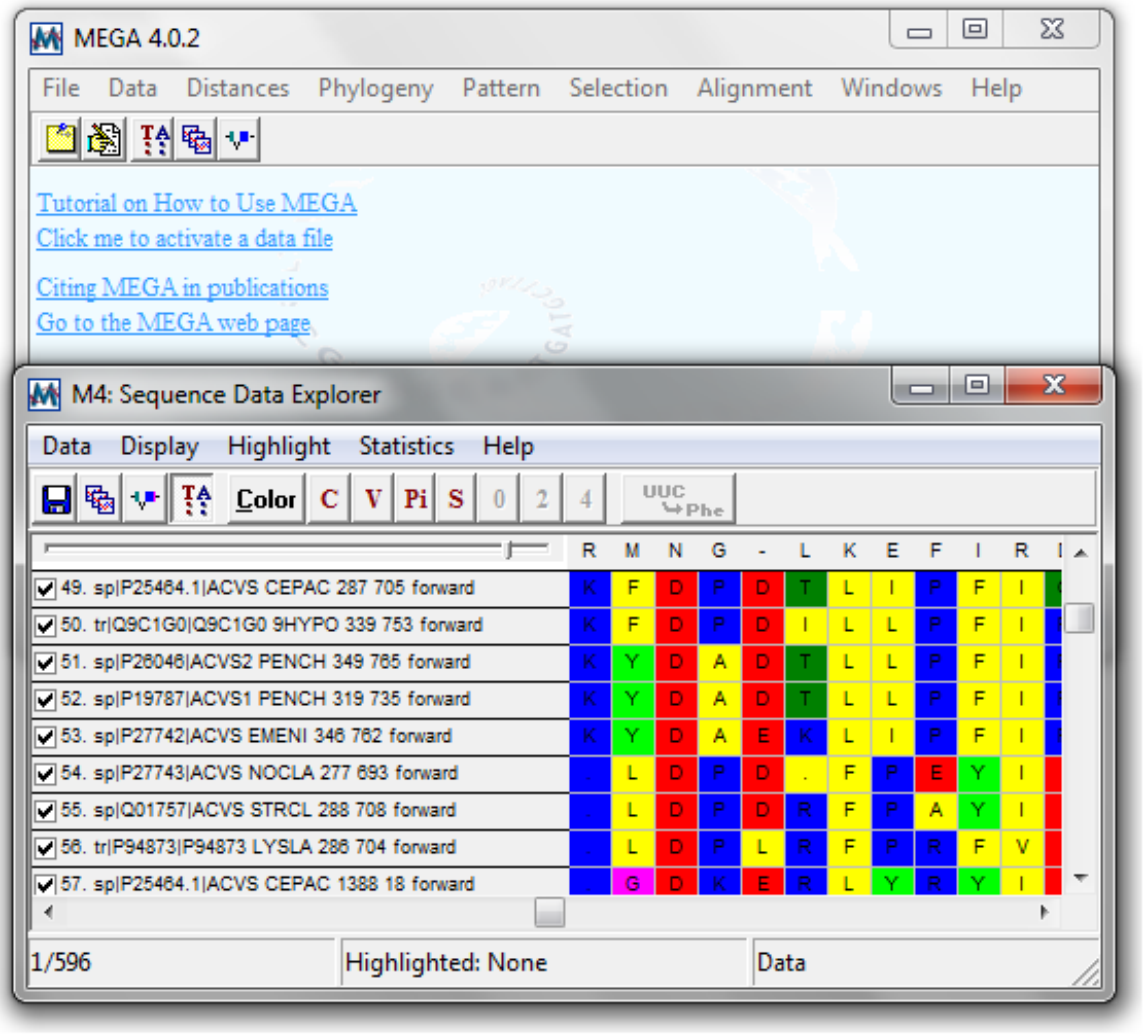
Grupiranje domena A, do kojeg je došlo u filogenetskom stablu cijelih sekvencija, iskoristilo se za izrezivanje sekvencija na principu homologije s domenom PheA. Sekvencije višestrukog poravnanja svake pojedine grupe domena s domenom PheA izrezane su u odnosu na pozicije aminokiselina u "kodu sustava NRPS" unutar aktivnog mjesta domene PheA. Rezultati poravnanja vizualizirani su uz pomoć programskog paketa Jalview 2.7 kako bi se olakšalo izrezivanje željenih sekvencija. Programski paket Jalview 2.7. (Clamp i sur., 2004; Waterhouse i sur., 2009) je bioinformatički paket koji omogućava vizualizaciju i lakšu manipulaciju sa poravnatim sekvencama (vidi: 3.1.3.2).

Nakon izrezivanja, svih je 397 domena A objedinjeno kako bi se generirala velika skupina sekvencija aktivnih mjesta svih domena koja se iskoristila za izradu drugog filogenetskog stabla. Prilikom izrade drugog stabla višestruko poravnanje svih sekvencija napravljeno je pomoću programa Clustal W ali koji je bio integriran u programski paket MEGA 4.0.2 (Tamura i sur., 2007). Prilikom generiranja samog stabla koristile su se iste postavke kao i kod generiranja filogenetskog stabla cijelih sekvencija.

3.2.5. Izrada profila specifičnosti domena A

Grupiranje domena A do kojeg je došlo u filogenetskom stablu sekvencija aktivnih mjesta iskoristilo se za izradu profila HMM specifičnosti domena A pomoću programskog paketa HMMER 3 (Tamura i sur., 2007). Generirana su 42 specifična profila HMM iz 42 grupe. Pri izradi svakog pojedinog profila napravilo se višestruko poravnanje sekvencija koje su se definirale u svakoj pojedinoj grupi kako bi se omogućilo generiranje profila (Slika 19). Upotrjebljena je slijedeća naredba:

*for i in *.aln;do hmmbuild/home/marija/PROFILI/$i.hmm$i; done*



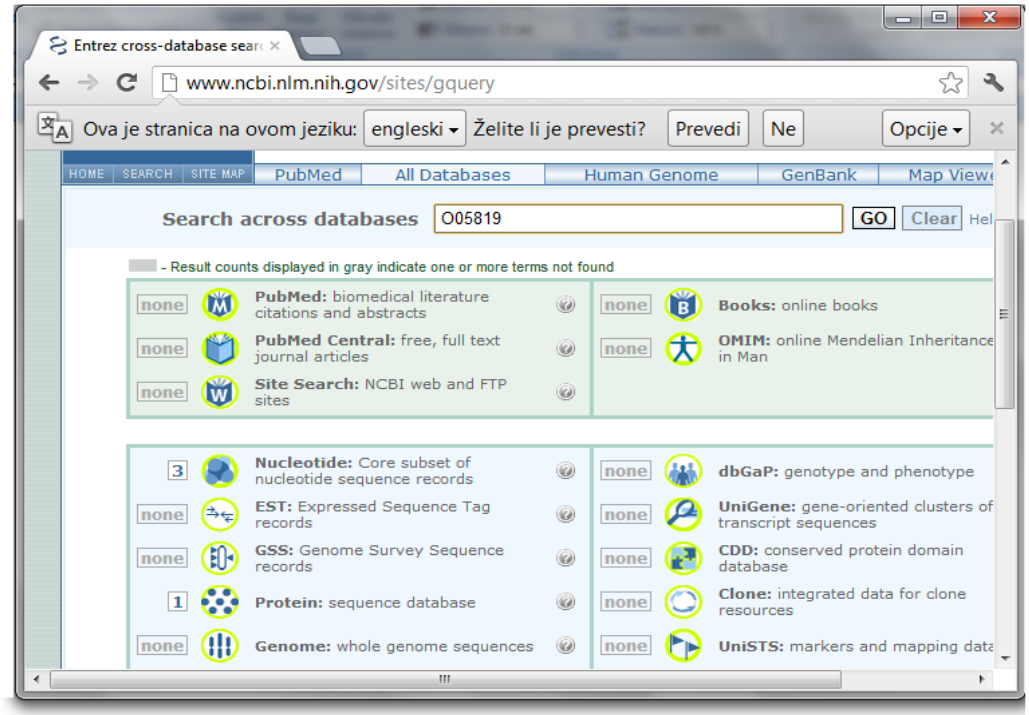
Slika 19. Prikaz izbornika programskog paketa MEGA 4.0.2 prilikom izrade višestrukog poravnanja sekvencija aktivnog mjesta svih 397 domena A upotrebom integriranog programa Clustal W.

Kako bi se olakšalo kasnije pretraživanje baze podataka s generiranim profilima domena A, svi su profili spremljeni u isti dokument i indeksirani. Indeksiranje je omogućilo brže pretraživanje baze podataka UniProt. Upotrjebljene su slijedeće naredbe:

cat.hmm > /home/marija/BAZA_PROFILA/grupe*

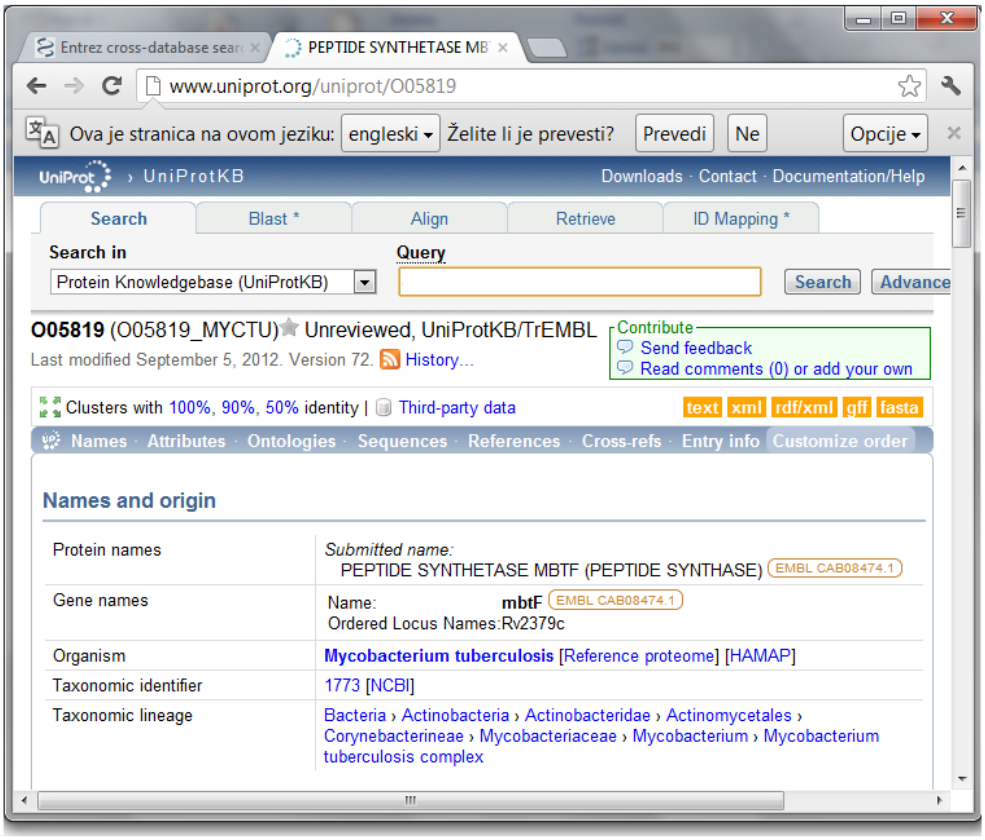
hmmpress/home/marija/BAZA_PROFILA/grupe

Uspješnost pretraživanja pomoću generiranih profila HMM provjerena je statistički. U izračun su ulazile domene u kojima je pretraga pokazala: vrijednost $E < 10^{-5}$ (engl. "E value") i uspjeh pogotka > 50 (engl. "Score"). Što je uspijeh pogotka viši, a vrijednost E manja, znači da je pronađen visok stupanj sličnosti između analizirane sekvence i profila HMM. Isto tako se točnost prepoznavanja supstrata provjerila u literaturnim izvorima. Naime, domene A u većini slučaja pokazuju specifičnost za više od jednog supstrata. Provjera je napravljena uz pomoć baza podataka GenBank i UniProt putem interneta. U bazu podatake GenBank pristupljeno je kroz sustav Entrez, sustav za pretraživanje i prikupljenje podataka. Odabirom izbornika baze podataka "Entrez home" (Anonymus 1, 2012) i upisivanja pristupnog koda (engl."acession number") u polje "Search across database" pokrenulo se pretraživanje (Slika 20).



Slika 20. Prikaz izbornika "Entrez home" za pretraživanje, s upisanim kodom domene A proteina u polju "Search across database" i s rezultatima pretraživanja.

U bazu podataka UniProt pristupilo se izravno (Anonymus 5, 2012), te se upisivanjem pristupnog koda u polje "Query" pokrenulo pretraživanje (Slika 21).



Slika 21. Prikaz izbornika UniProt baze za pretraživanje s rezultatima pretraživanja.

Statistička provjera napravljena je preko izračuna slijedećih grupa:

- TP - stvarno pozitivni (engl. "True Positive")
- TN - stvarno negativni (engl. "True Negative")
- FP - lažno pozitivni (engl. "False Positive")
- FN - lažno negativni (engl. "False Negative")

Grupa TP označava točno prepoznavanje od strane modela dok grupa TN označava krivo prepoznavanje. Grupe FP i FN označavaju grešku modela, tako su domene grupe FP, koje je model krivo prepoznao kao točne, a domene grupe FN su domene koje su u rezultatima pretraživanja označene kao krive a rezultat je zapravo točan. Pogreška modela se izračunala po sljedećoj formuli: (FP + FN)/(FP + FN + TP + TN) (Rausch i sur., 2005).

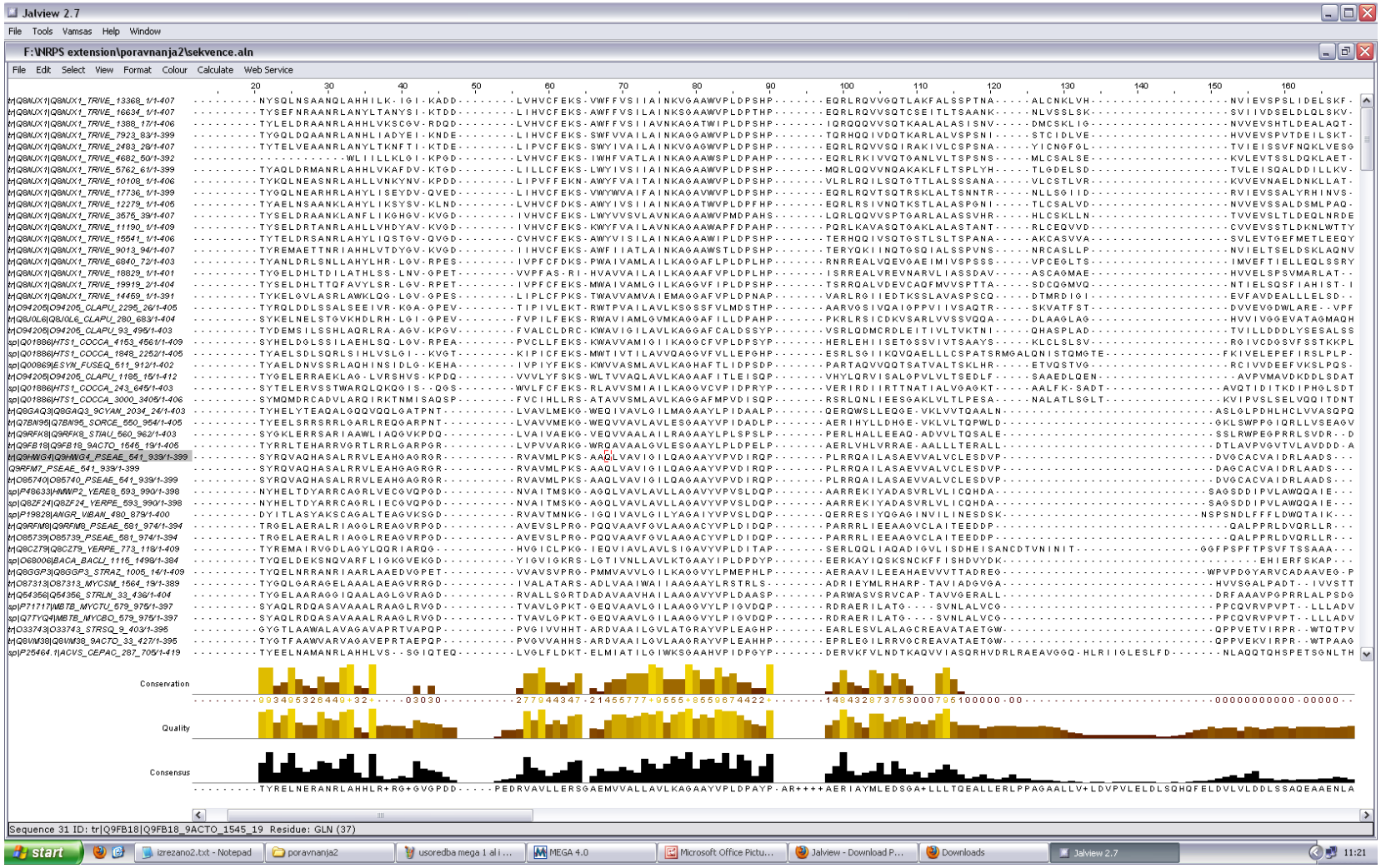
4. REZULTATI

4.1 REZULTATI FILOGENETSKE ANALIZE PRIKUPLJENIH SEKVENCIJA PROTEINA DOMENA ZA ADENILACIJU

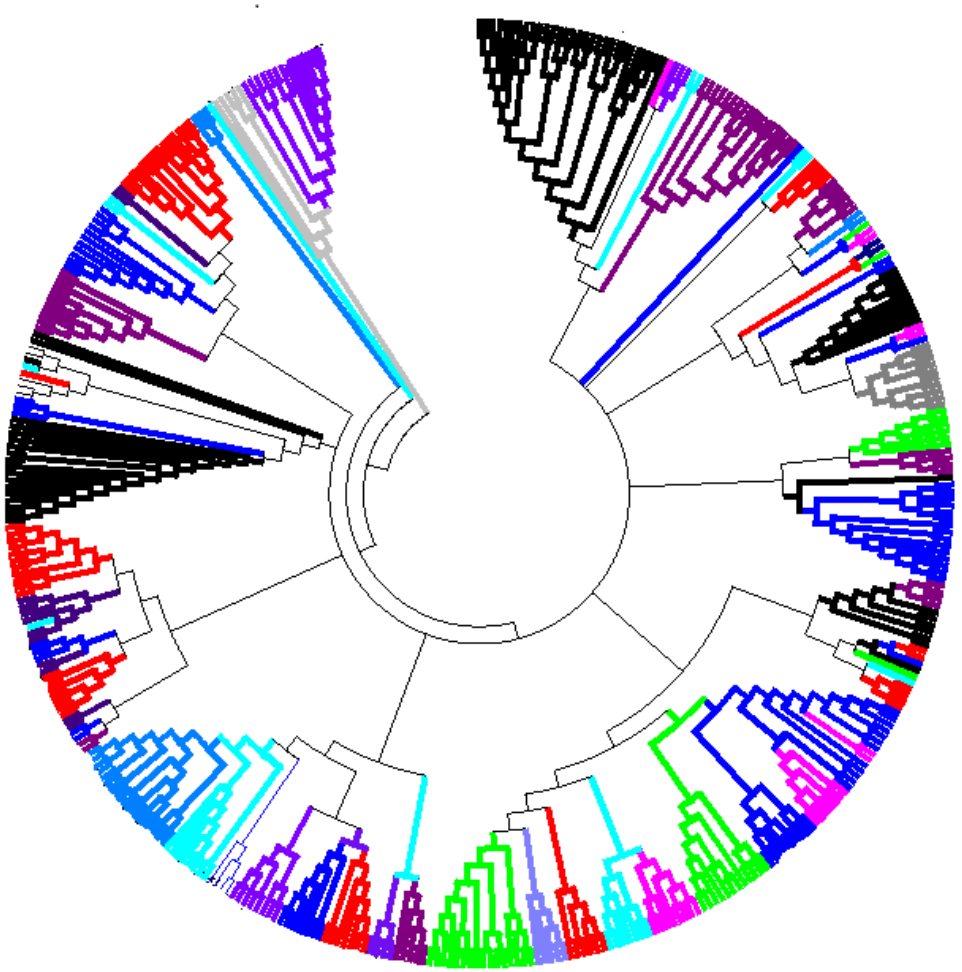
Svih 397 proteinskih sekvencija domena A preuzetih iz rada Rausch i suradnici (2005.), poravnano pomoću programa ClustalW (Larkin i sur., 2007) i prikazano programom Jalview 2.7 (Clamp i sur., 2004; Waterhouse i sur., 2009), prikazano je na Slici 22. Na slici su vidljive konzervirane regije unutar cijele skupine upotrebljene za testiranje.

Radi izrade filogenetskog stabla programom MEGA 4.0.2 (Tamura i sur., 2007) provedena je analiza višestruko poravnatih cijelih sekvencija uzorka upotrebljenog za testiranje. Na Slici 23 prikazano je umanjeno filogenetsko stablo cijelih sekvencija domena za adenilaciju koje je generirano metodom ME programom MEGA 4.0.2. Na stablu su grupe, zbog boljeg prikaza rezultata, grupirane ručno i istaknute upotrebom boja. Zapisi su opisa domena izbrisani radi boljeg prikaza grupa. Cijelo stablo nalazi se u prilogu (vidi: podpoglavlje 8.2.2). Ukupno je generirano preko četrdest grupa sekvencija koje su, u analizi filogenetskog stabla, pokazale loše grupiranje. Detaljno je objašnjenje raspravljeno u poglavlju 5. Primjer grupa sekvencija unutar filogenetskog stabla cijelih sekvencija prikazan je na Slici 24. Posljedično je, zbog loše generiranih grupa, obavljena filogenetska analiza proteinskih sekvencija aktivnih mjesta koje sadržavaju "kod sustava NRPS". Izostavio se jedino ostatak aminokiseline Lys na položaju 517 koji je sačuvan unutar porodice domena A. Time su se sekvencije smanjile na optimalnu veličinu za filogenetsku analizu (vidi: Slika 25.).

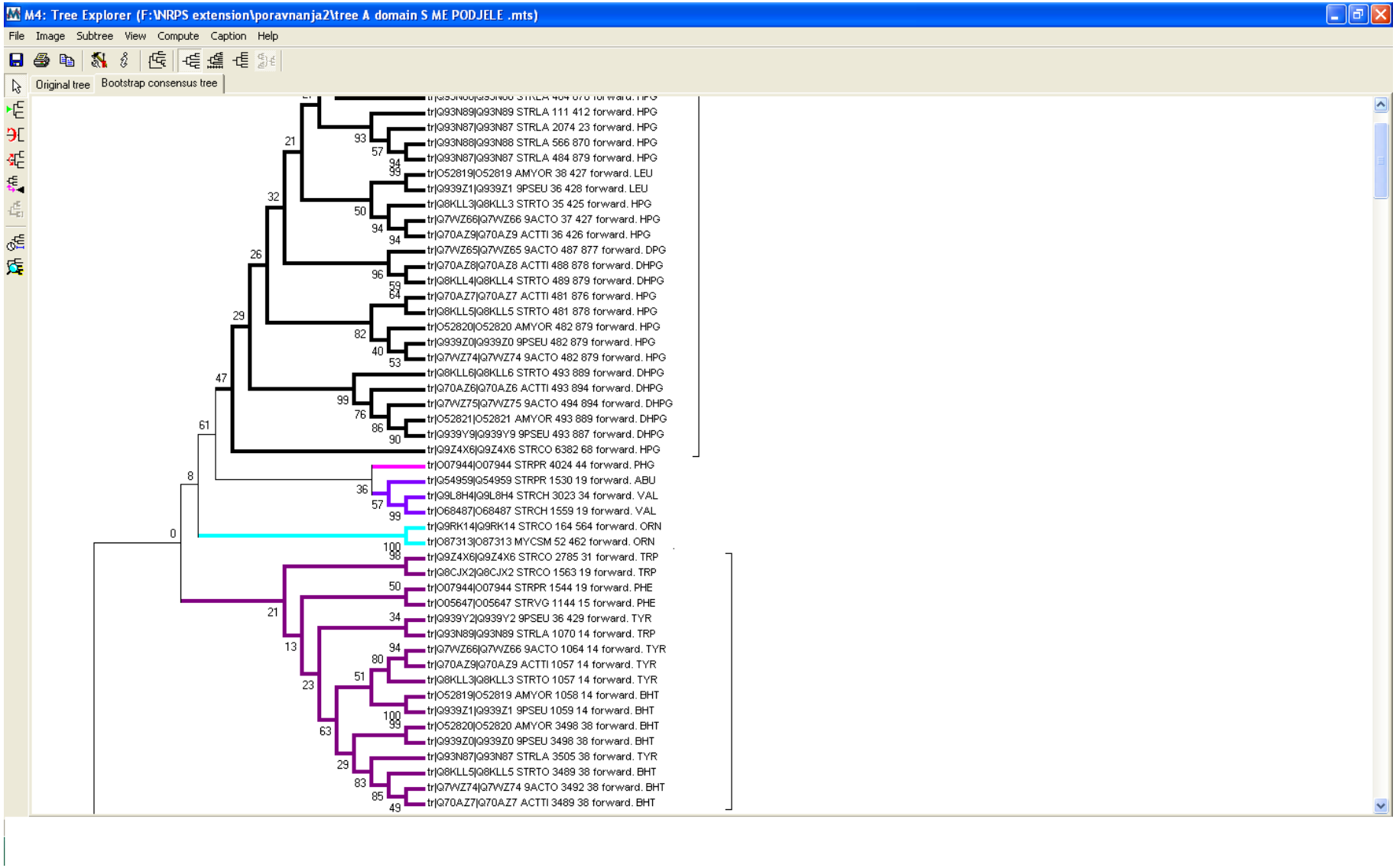
Iz višestrukog poravnanja sekvencija aktivnih mjesta (sekvencija dužine 136 do 150 aminokiselina) generirano je, istom metodom, drugo filogenetsko stablo. U stablu su ručno anotirane velike i manje grupe. Prikaz svih grupa vidljiv je na Slikama 26 i 28. Na Slikama je prikazana umanjena radijalna verzija stabla. Cijelo stablo nalazi se u prilogu (vidi: podpoglavlje 8.2.4.). Ukupno su definirane 42 grupe: 19 velikih i 23 manje. Velike grupe su označene na filogenetskom stablu na Slici 26. Od ukupno 19 velikih grupa, 10 grupa je bilo samostalno dok je 9 grupa sadržavalo manje grupe. Prikaz velikih grupa sa sadržanim manjim grupama vidljiv je na Slici 27. Prosječno su velike grupe sa sadržanim manjim grupama imale 21 domenu, dok su samostalne velike grupe imale 12 domena. Grupa 7 je imala najveći broj domena, njih 30.



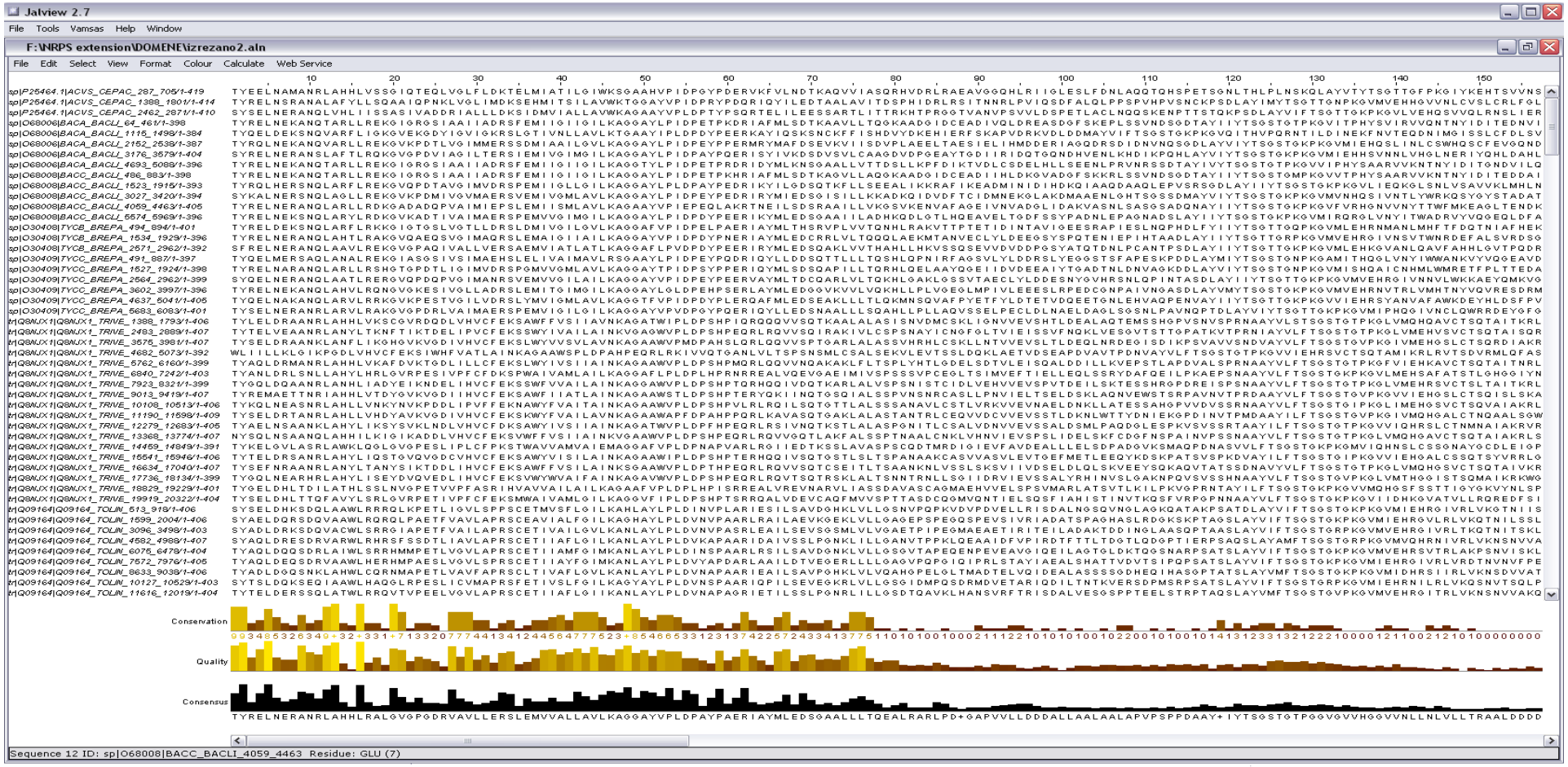
Slika 22. Prikaz poravnanja svih 397 sekvenci domena A u programu Jalview. Prikaz konzerviranosti sekvencija vidljiv je na dnu slike.



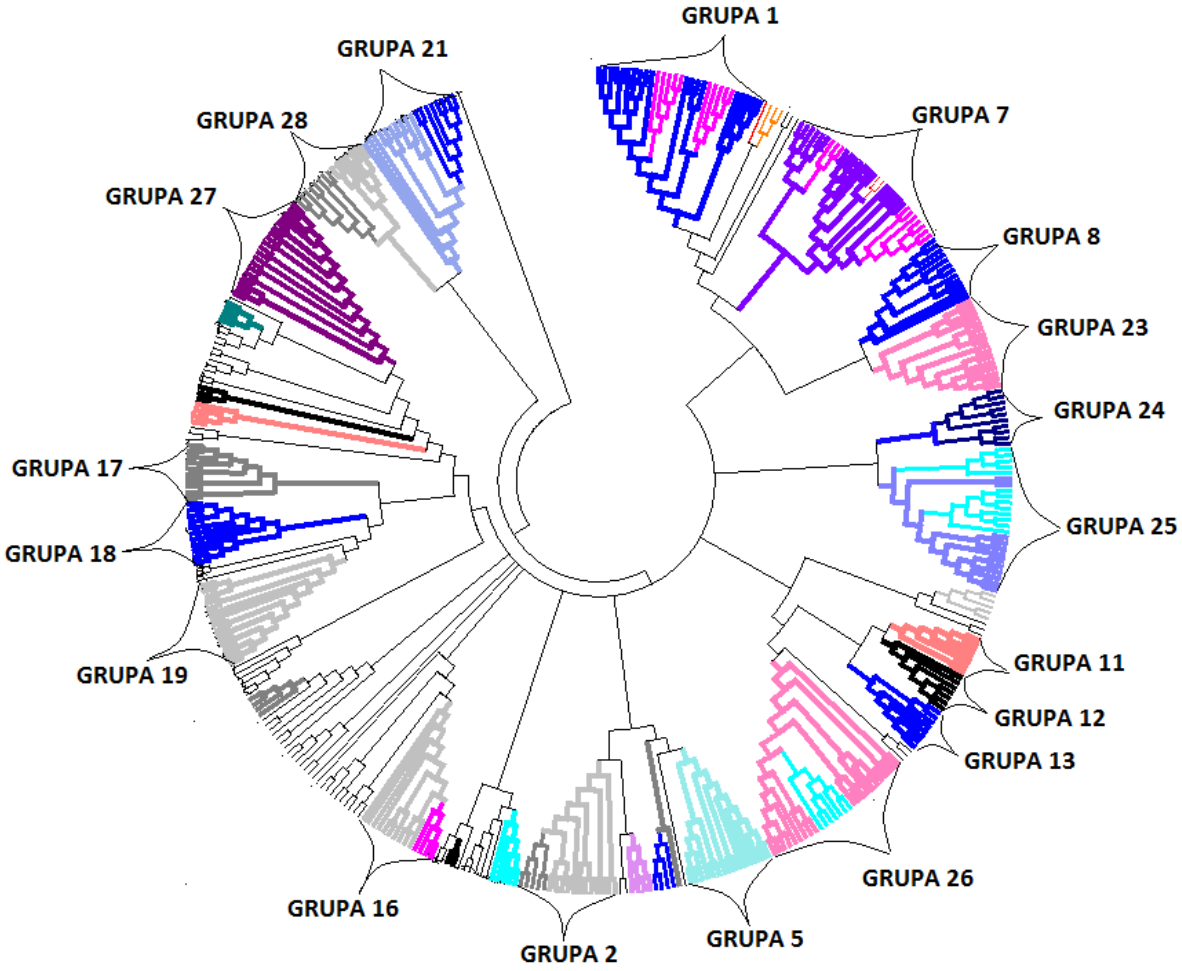
Slika 23. Prikaz filogenetskog stabla generiranog iz cijelih sekvencija, filogenetska grupiranja sekvencija naglašena su upotrebom boja (s time da su zapisi opisa domena izbrisani radi boljeg prikaza grupa).



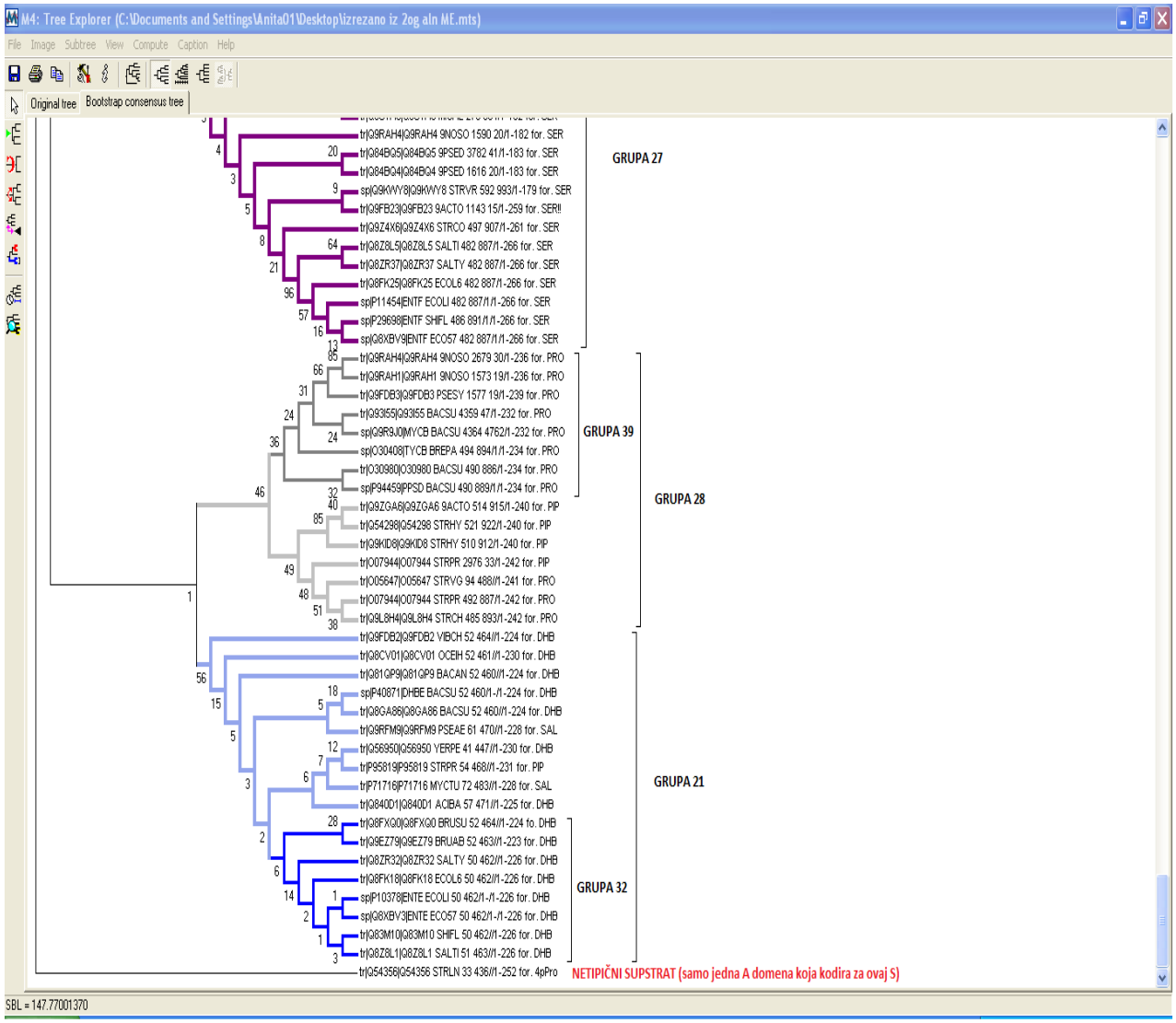
Slika 24. Bliži prikaz primjera grupa sekvencija unutar filogenetskog stabla cijelih sekvencija.



Slika 25. Prikaz izrezanih sekvencija aktivnih mjesta u programu Jalview.

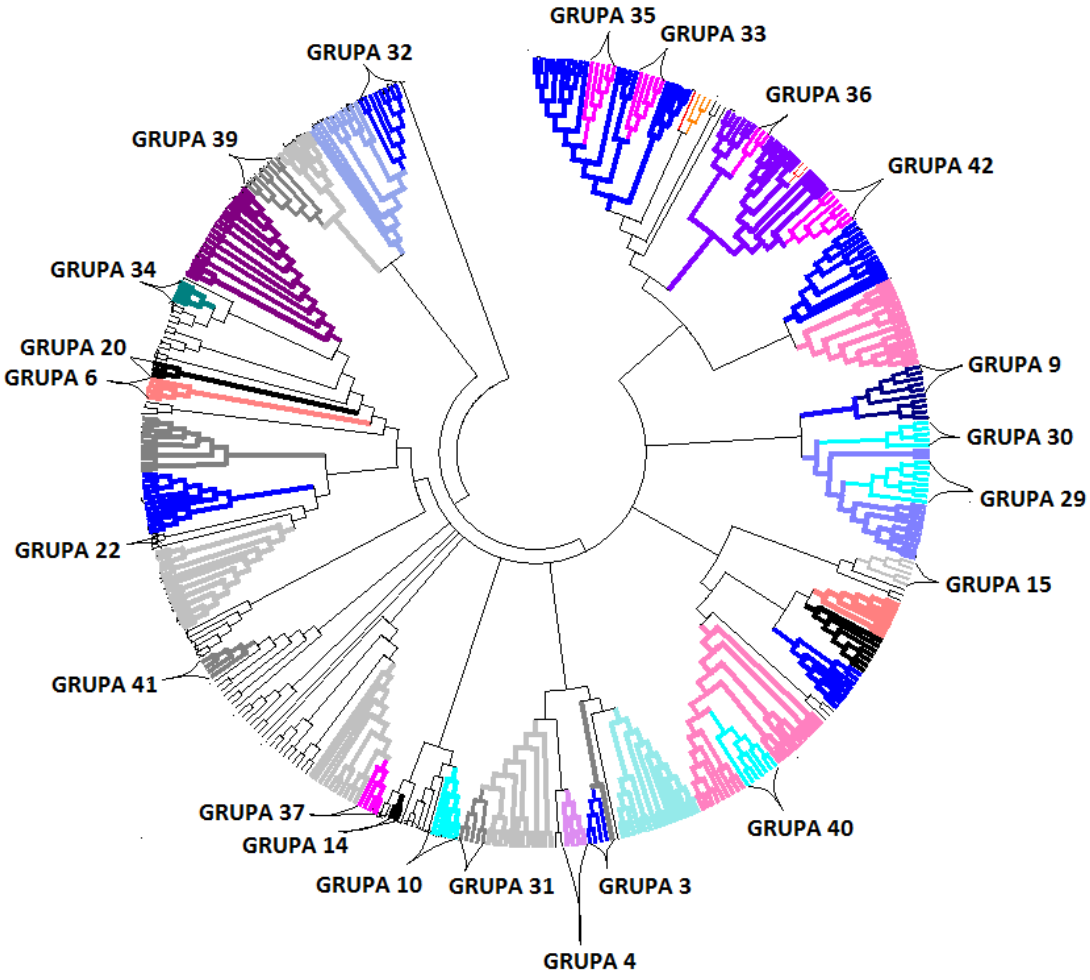


Slika 26. Prikaz filogenetskog stabla generiranog iz sekvencija aktivnih mjesta. Filogenetska grupiranja sekvencija naglašena su upotrebom boja. Prikazane su **velike grupe** (s time da su zapisi opisa domena izbrisani radi boljeg prikaza grupa).



Slika 27. Prikaz dijela filogenetskog stabla generiranog iz sekvencija aktivnih mjesta. Prikaz velikih 27, 28 i 21 te malih 39 i 32 grupa.

Manje grupe su označene na filogenetskom stablu na Slikama 27 i 28. Filogenetsko stablo je sadržavalo ukupno 23 manje grupe. Prosječno su manje grupe imale 5 domena, s time da su grupe 14 i 22 sadržavale minimalni broj domena. Ukupno se 71 domena nije grupiralo prilikom filogenetske analize.



Slika 28. Prikaz filogenetskog stabla generiranog iz sekvencija aktivnih mjesta. Filogenetska grupiranja sekvencija naglašena su upotrebom boja. Prikazane su **manje grupe** (s time da su zapisi opisa domena izbrisani radi boljeg prikaza grupa).

U Tablici 2 prikazane su sve grupe i supstrati za koje su specifične domene A u svakoj pojedinoj grupi. Objašnjenje grupiranja i tipova grupa nalazi se u poglavlju 5.

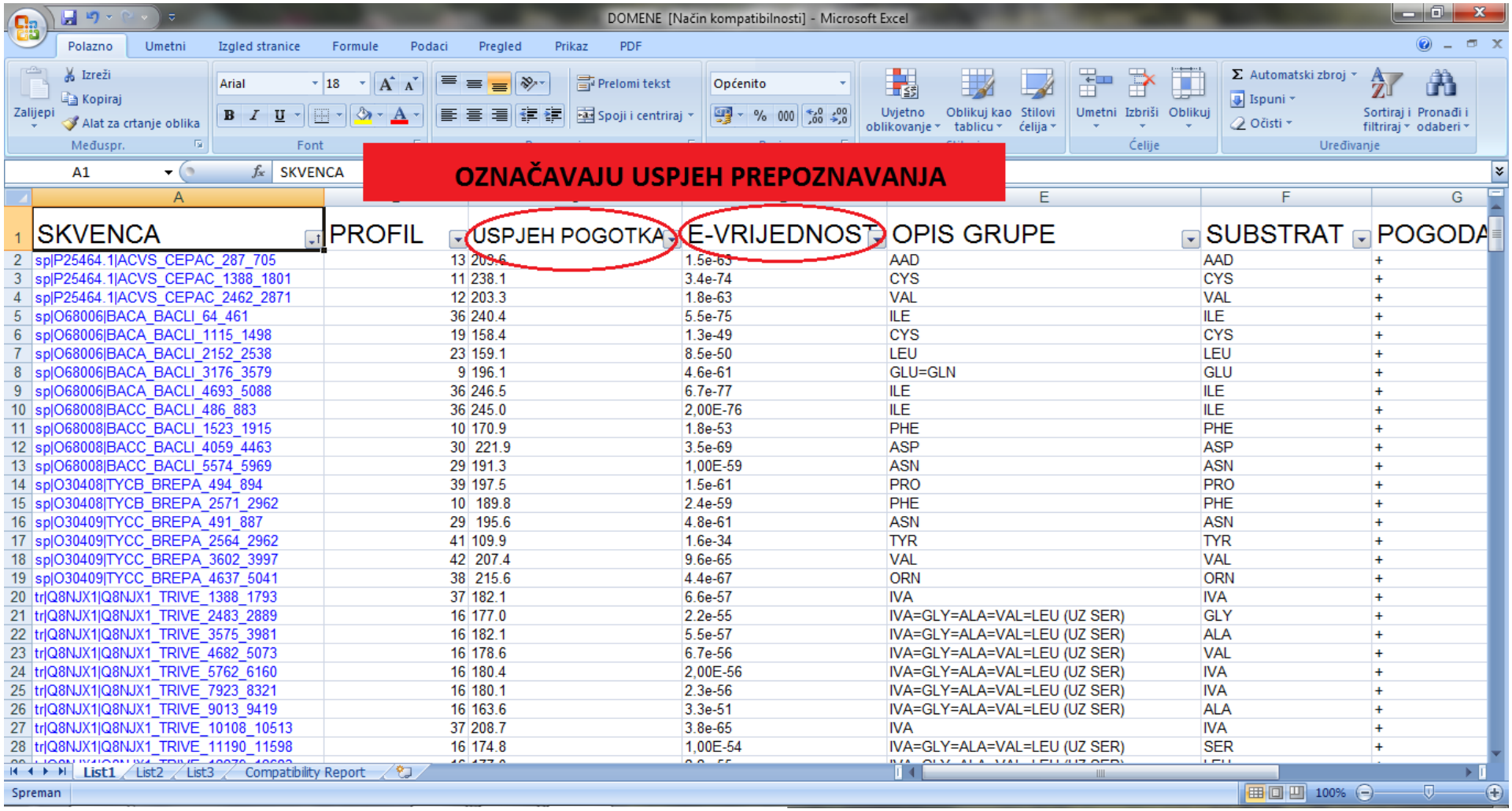
Tablica 2. Grupiranje domena A i klasifikacija grupa.

GRUPA	TIP GRUPE	SUPSTRATI
1	VELIKA	Hpg, Dhpg, Dpg
2	VELIKA	Tyr, Trp, Bht, Phe
3	MALA	Orn, Arg
4	MALA	Dab
5	VELIKA	Ala
6	MALA	Arg
7	VELIKA	Leu, Ile, Val, Phe, Tyr
8	VELIKA	Ala-D, Abu, Val, Leu, Ala, Gly, Valhyphaa, Bmt
9	MALA	Glu, Gln
10	MALA	Phe
11	VELIKA	Cys
12	VELIKA	Val
13	VELIKA	Aad
14	MALA	Lys-B
15	MALA	Orn, Lys
16	VELIKA	Iva, Gly, Ala, Ser, Val, Leu
17	VELIKA	Gly
18	VELIKA	Ala
19	VELIKA	Cys
20	MALA	Ala-B
21	VELIKA	Dhb, Sal, Pip
22	MALA	Ser-Thr
23	VELIKA	Leu
24	VELIKA	Glu, Gln, Aad
25	VELIKA	Asp, Asn, 3-Me-Glu
26	VELIKA	Thr, Dht
27	VELIKA	Ser
28	VELIKA	Pro, Pip
29	MALA	Asn
30	MALA	Asp
31	MALA	Bht
32	MALA	Dhb
33	MALA	Dhpg
34	MALA	Glu
35	MALA	Hpg
36	MALA	Ile
37	MALA	Iva
38	MALA	Orn
39	MALA	Pro
40	MALA	Thr
41	MALA	Tyr
42	MALA	Val

4.2. REZULTATI PRETRAGE BAZA PROTEINSKIH SEKVENCIJA S GENERIRANIM PROFILIMA HMM DOMENA ZA ADENILACIJU

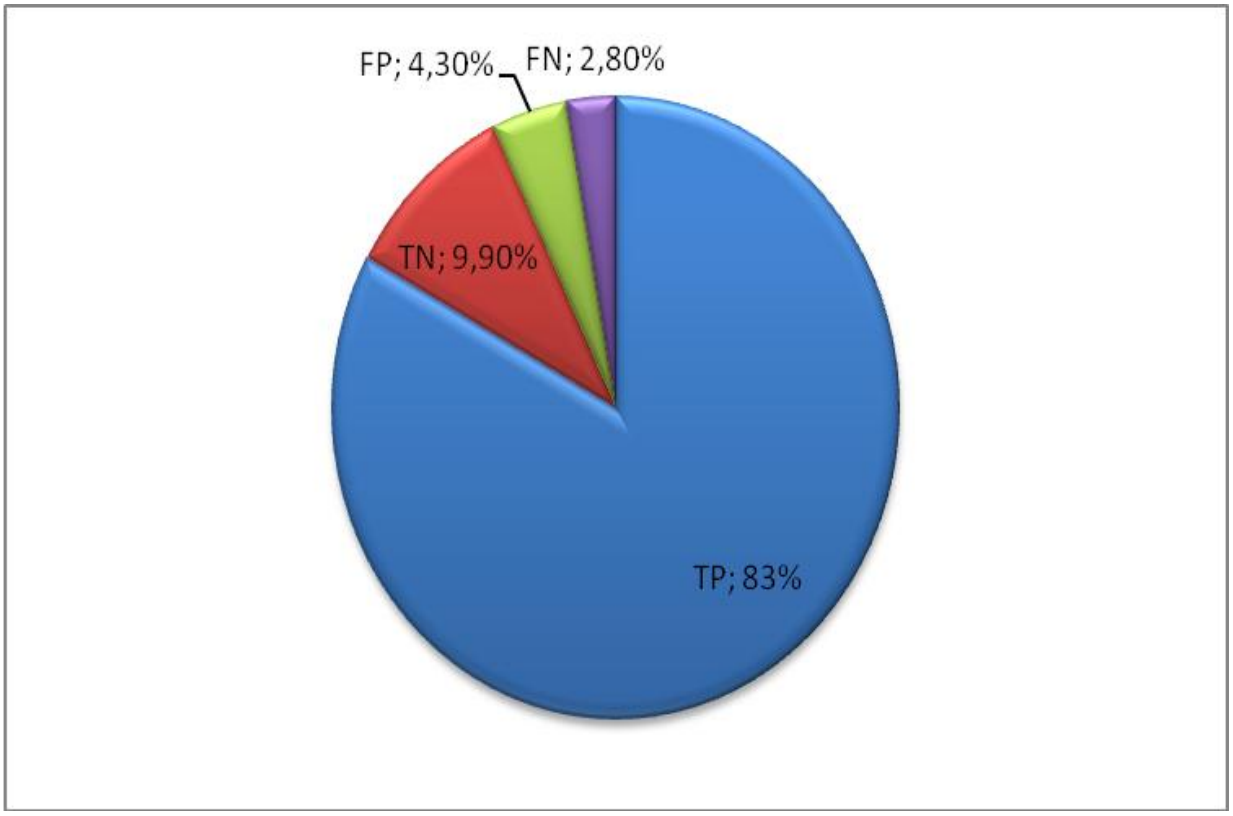
Potpuni rezultati pretrage baze proteinskih sekvencija sa 42 generirana profila HMM domena za adenilaciju nalaze se u prilozima (vidi: podpoglavlje 8.2.5). Djelomični prikaz rezultata prikazan je na Slici 29. Prilikom analize rezultata u obzir su uzete domene u kojima je pretraga pokazala: vrijednost $E < 10^{-5}$ i uspjeh pogotka > 50. Što je uspijeh pogotka viši, a vrijednost E manja, znači da je pronađen visok stupanj sličnosti između analizirane sekvencije i profila HMM.

U rezultatima je pretraživanja, radi boljeg prikaza, točno prepoznavanje od strane generiranih profila HMM prikazano plavom bojom, tzv. TP. Krivo prepoznavanje modela označeno je crvenom bojom, tzv. TN, dok je FP označeno zelenom bojom i FN ljubičastom bojom (za objašnjenje kratica vidi podpoglavlje 8.1). Od ukupno 414 sekvencija domena A: za 351 je točno prepoznata specifičnost, za 59 sekvencija domena A krivo, dok za 4 domene model nije dao nikakvo predviđanje. Pogreška generiranog modela iznosila je 7,1%. Na Slici 30 prikazani su postotni odnosi TP, TN, FN i FP.



	A	B	C	D	E	F	G
1	SKVENCA	PROFIL	USPJEH POGOTKA	E-VRIJEDNOST	OPIS GRUPE	SUBSTRAT	POGODA
2	sp|P25464.1|ACVS_CEPAC_287_705	13	209.6	1.5e-63	AAD	AAD	+
3	sp|P25464.1|ACVS_CEPAC_1388_1801	11	238.1	3.4e-74	CYS	CYS	+
4	sp|P25464.1|ACVS_CEPAC_2462_2871	12	203.3	1.8e-63	VAL	VAL	+
5	sp|O68006|BACA_BACLI_64_461	36	240.4	5.5e-75	ILE	ILE	+
6	sp|O68006|BACA_BACLI_1115_1498	19	158.4	1.3e-49	CYS	CYS	+
7	sp|O68006|BACA_BACLI_2152_2538	23	159.1	8.5e-50	LEU	LEU	+
8	sp|O68006|BACA_BACLI_3176_3579	9	196.1	4.6e-61	GLU=GLN	GLU	+
9	sp|O68006|BACA_BACLI_4693_5088	36	246.5	6.7e-77	ILE	ILE	+
10	sp|O68008|BACC_BACLI_486_883	36	245.0	2,00E-76	ILE	ILE	+
11	sp|O68008|BACC_BACLI_1523_1915	10	170.9	1.8e-53	PHE	PHE	+
12	sp|O68008|BACC_BACLI_4059_4463	30	221.9	3.5e-69	ASP	ASP	+
13	sp|O68008|BACC_BACLI_5574_5969	29	191.3	1,00E-59	ASN	ASN	+
14	sp|O30408|TYCB_BREPA_494_894	39	197.5	1.5e-61	PRO	PRO	+
15	sp|O30408|TYCB_BREPA_2571_2962	10	189.8	2.4e-59	PHE	PHE	+
16	sp|O30409|TYCC_BREPA_491_887	29	195.6	4.8e-61	ASN	ASN	+
17	sp|O30409|TYCC_BREPA_2564_2962	41	109.9	1.6e-34	TYR	TYR	+
18	sp|O30409|TYCC_BREPA_3602_3997	42	207.4	9.6e-65	VAL	VAL	+
19	sp|O30409|TYCC_BREPA_4637_5041	38	215.6	4.4e-67	ORN	ORN	+
20	tr|Q8NJX1|Q8NJX1_TRIVE_1388_1793	37	182.1	6.6e-57	IVA	IVA	+
21	tr|Q8NJX1|Q8NJX1_TRIVE_2483_2889	16	177.0	2.2e-55	IVA=GLY=ALA=VAL=LEU (UZ SER)	GLY	+
22	tr|Q8NJX1|Q8NJX1_TRIVE_3575_3981	16	182.1	5.5e-57	IVA=GLY=ALA=VAL=LEU (UZ SER)	ALA	+
23	tr|Q8NJX1|Q8NJX1_TRIVE_4682_5073	16	178.6	6.7e-56	IVA=GLY=ALA=VAL=LEU (UZ SER)	VAL	+
24	tr|Q8NJX1|Q8NJX1_TRIVE_5762_6160	16	180.4	2,00E-56	IVA=GLY=ALA=VAL=LEU (UZ SER)	IVA	+
25	tr|Q8NJX1|Q8NJX1_TRIVE_7923_8321	16	180.1	2.3e-56	IVA=GLY=ALA=VAL=LEU (UZ SER)	IVA	+
26	tr|Q8NJX1|Q8NJX1_TRIVE_9013_9419	16	163.6	3.3e-51	IVA=GLY=ALA=VAL=LEU (UZ SER)	ALA	+
27	tr|Q8NJX1|Q8NJX1_TRIVE_10108_10513	37	208.7	3.8e-65	IVA	IVA	+
28	tr|Q8NJX1|Q8NJX1_TRIVE_11190_11598	16	174.8	1,00E-54	IVA=GLY=ALA=VAL=LEU (UZ SER)	SER	+

Slika 29. Djelomični prikaz rezultata pretraživanja baze proteinskih sekvencija generiranih profilima HMM domena za adenilaciju. Prikazan je dio tablice Excel s rezultatima.



Slika 30. Uspješnost modela za predviđanje specifičnosi izbora supstrata domena za adenilaciju.

5. R A S P R A V A

Prisutan je rastući interes za genetičko inženjerstvo sintetaza neribosomalno sintetiziranih peptida (NRPS) kako bi se sintetizirale nove biološki aktivne supstance prikladne za generiranje novih lijekova koji su potrebni i medicini i farmaceutskoj industriji. Mogućnost za tako nešto uvelike ovisi o našem razumijevanju kako sustavi NRPS vrše selekciju između raznoraznih monomera koji su prisutni u neribosomalno sintetiziranim proteinima. Metode za analiziranje i predviđanje supstrata domena A, ali i ostalih domena sustava NRPS su od vitalnog interesa za moguće genetičko inženjerstvo novih neribosomalno sintetiziranih proteina (Lautru i Challis, 2004). Mogućnost *in silico* predviđanja supstrata domena A iz primarne strukture proteina još biokemijski neopisanih domena A ubrzala bi tehnološki razvoj novih farmaceutski važnih spojeva. Također, otvorila bi se mogućnost za racionalnu promjenu specifičnosti i aktivnosti poznatih domena A, a saznanja o novim supstratima koje bi takve domene aktivirale dovele bi do modifikacije već postojećih neribosomalno sintetiziranih proteina (Stachelhaus i sur., 1999).

Filogenetska analiza provedena je kako bi se ustanovili evolucijski odnosi i povezanost između sekvencija ispitivanih adenilacijskih domena i njihove specifičnosti. Temelj za takav pristup bili su radovi Stachelhaus i suradnici (1999) te Challis i suradnici (2000). Ako su ostaci aminokiselina, na točno određenim pozicijama unutar aktivnog mjesta domena za adenilaciju, odgovorni za selekciju supstrata te ako su sve domene iste specifičnosti nastale od jedne domene zajedničkoga pretka, za pretpostaviti je da će prilikom filogenetske analize domene odgovorne za selekciju istog supstrata biti zajedno grupirane u jednu filogenetsku grupu. Stachelhaus i suradnici (1999) su pozicije ostataka aminokiselina unutar aktivnog mjesta odgovornih za selekciju supstrata definirali kao "kod sustava NRPS". Za filogenetske analize upotrebljene su metode: "Minimum Evolution" i "Maximum Parsimony". Ove metode su upotrebljene zbog visoke razine homologije sekvencija unutar skupine domena za adenilaciju. Pored toga, u rezultatima su prikazana samo filogenetska stabla generirana metodom "Minimum Evolution", jer su metode ME i MP iz analiziranih sekvencija cijelih domena za adenilaciju (vidi: Sliku 23; Poglavlje 8.2., Stablo 1) i sekvencija aktivnih mjesta domena za adenilaciju (vidi: Slike 26. i 28.; Poglavlje 8.2., Stablo 2) generirale slične rezultate filogenetske analize. Detaljnije objašnjenje analiza filogenetskih stabala nalazi se u poglavlju 3.2.3. Filogenetsko stablo cijelih sekvencija (vidi: Sliku 23 i Poglavlje 8.2., Stablo 1) pokazalo je često grupiranje domena temeljeno na podrijetlu domena za adenilaciju a ne na temelju specifičnosti za određeni supstrat. Jednostavnije, domene A iz istog mikroorganizma su se zajedno grupirale. Primjer takvog grupiranja su domena za adenilaciju unutar

ciklosporin sintetaze iz plijesni *Tolypocladium niveum* (UniProt pristupni kod: Q09164). Domene za adenilaciju unutar sustava NRPS ciklosporin sintetaze, odgovorne su za selekciju i aktivaciju: četiri aminokiseline Leu, dvije aminokiseline Val, te po jedne aminokiseline Ala, Ala-D, Abu, Gly i Bmt. Iako je riječ o domenama za adenilaciju koje prepoznaju različite supstrate s različitim fizikalno-kemijskim svojstvima, te imaju različita aktivna mjesta, unutar filogenetskog stabla one su se prikazale kao jedna grupa. To je zasigurno zato jer sekvencije cijelih domena zadržavaju evolucijski signal koji je više posljedica specifičnih mutacija genoma nego li funkcionalne evolucije domena koje se ispituju. Ovi rezultati potvrđuju važnost analize samih sekvencija aktivnih mjesta prilikom definiranja specifičnosti domena za adenilaciju, što su ustanovili i Stachelhaus i suradnici (1999) te Challis i suradnici (2000). Zato se krenulo u filogenetsku analizu proteinskih sekvencija aktivnih mjesta koje sadržavaju "kod sustava NRPS", izostavio se jedino ostatak aminokiseline Lys na poziciji 517 koji je visoko sačuvan unutar porodice domena A i za koji se smatra da nema ulogu pri selekciji supstrata od strane domena za adenilaciju.

Funkcionalno grupiranje u filogenetskom stablu sekvencija aktivnih mjesta domena A pokazalo je puno bolje rezultate od filogenetskog stabla cijelih sekvencija domena za adenilaciju. Ukupno je generirano 19 velikih i 23 manjih grupa (vidi: Slike 26. i 28. te Poglavlje 8.2., Stablo 2). Supstrati koje aktiviraju domene za adenilaciju svake pojedine grupe prikazani su u Tablici 2. Od ukupno 19 velikih grupa, 10 grupa je bilo samostalno dok je 9 grupa sadržavalo manje grupe. Prikaz velikih grupa sa sadržanim manjim grupama vidljiv je na Slici 27. Prosječno su velike grupe sa sadržanim manjim grupama imale 21 domenu, dok su samostalne velike grupe imale 12 domena. Manje grupe su u prosjeku imale 5 domena, s time da su grupe 14 i 22 sadržavale minimalni broj domena, njih dvije. Ukupno se od 397 domena A njih 71 nije grupiralo. Većinom su to bile domene koje aktiviraju netipične supstrate kao što je npr. aminokiselina 4pPro (vidi: Slika 27.). Za takve netipične supstrate još uvijek u literaturi ne postoji dovoljan broj sekvencija da bi se mogla postići adekvatna analiza. Filogenetsko grupiranje velikih grupa pokazalo je slično grupiranje s grupama ustanovljenima od strane Rausch i suradnika (2005). Oni su grupirali domene za adenilaciju na principu fizikalno-kemijskih svojstava supstrata i aktivnog mjesta. Uzeli su u obzir: broj mogućih vodikovih veza, polarnost, hidrofobnost, izoelektričnu točku, volumen aktivnog mjesta, te pojavljivanje dodatnih sekundarnih struktura. Grupiranje je slično kod u domena za adenilaciju koje prepoznaju aminokiseline: Tyr=Trp=Bht=Phe, Cys, Iva=Gly=Ala=Ser=Val=Leu, Dhb=Sal, Pro=Pip, Ser i Asp=Asn. Generiranje sličnih grupa

navodi na već prije pretpostavljenu pretpostavku, da zbog visoke sličnosti između domena za adenilaciju, 26 % do 56 %, na pozicijama ostataka unutar aktivnih mjesta koji su odgovorni za selekciju supstrata sličnih fizikalno-kemijskih svojstava, nalaze iste ili fizikalno-kemijski slične aminokiseline. Challis i suradnici (2000) su utvrdili da se "kodovi sustava NRPS" unutar domena za adenilaciju koje aktiviraju supstrate fizikalno-kemijski sličnih svojstava često razlikuju samo u jednom ostatku aminokiseline. Eksperimentalna potvrda ovoj pretpostavci je rad Stachelhausa i suradnika (1999) u kojem su neposrednom mutagenezom unutar domene za adenilaciju prvog modula gramicidin S sintetaze GrsA, tzv. domena PheA, mutacija aminokiseline Ala na poziciji 322 unutar "koda sustava NRPS" u aminokiselinu Gly, povećala specifičnost domene PheA za aminokiselinu L-Trp. Inače domena PheA pokazuje veću specifičnost za aminokiselinu L-Phe. Grupe domena za adenilaciju generirane filogenetskom analizom potvrda su tih pretpostavki. Dodatno, grupe domena za adenilaciju koje su specifične za supstrate sa sličnim fizikalno-kemijskim svojstvima unutar filogenetskog stabla grupirane su blizu. Primjer takvih grupa su grupe 24 i 25. Grupa 24 specifična je za aminokiseline: Glu, Gln, Aad, a grupa 25 za aminokiseline Asp, Asn i 3-Me-Glu. Jednostavnije rečeno, obadvije grupe domena za adenilaciju predstavljaju domene koje su specifične za alifatske aminokiseline, čija R skupina završava donorom vezanim s vodikom.

Iz rezultata pretraživanja baza podataka proteinskih sekvencija dobivenih programskim paketom HMMER3 (Finn i sur., 2011) izdvojene su domene čija je podudarnost s generiranim profilima HMM zadovoljavala slijedeće uvjete: vrijednost $E < 10^{-5}$ i uspjeh pogotka > 50. Ti uvjeti su odabrani kao granični zbog pretpostavke da profili HMM, čija je vrijednost E viša od 10^{-5} i čiji je uspjeh pogotka ispod 50, ne zadovoljavaju traženo svojstvo sličnosti. Od ukupno 414 sekvencija domena A za 351 je točno predviđena specifičnost, za 59 je specifičnost krivo predviđena, dok za 4 domene model nije dao nikakvo predviđanje. Pogreška generiranog modela iznosila je 7,1 %. Pretpostavka za dobro generiranje modela predviđanja specifičnosti domena za adenilaciju uz pomoć specifičnih profila je dobar početni skup sekvencija za svaki određeni profil specifičnosti (Rausch i sur., 2005; Röttig i sur., 2011). Profili specifičnosti, generirani programskim paketom HMMER 3 na temelju grupa generiranih filogenetskom analizom aktivnih mjesta domena za adenilaciju, pokazali su se manje uspješnima u slučajevima kada je profil generiran iz manjeg broja sekvencija, kao što je to slučaj s profilom HMM za domene koje su specifične za aminokiselinu Glu. Profil HMM je bio generiran iz manje grupe koja je sadržavala četiri sekvencije, te je

prepoznavanje ovog profila iznosilo 70 %. Od ukupno 10 domena koje u literaturi pokazuju specifičnost za aminokiselinu Glu, profilom HMM je točno prepoznato njih 7 (vidi: Poglavlje 8.2., Tablicu 3).

Jedan od problema prilikom izrade modela specifičnosti profila HMM za domene za adenilaciju bile su domene s tzv. "relaksiranom" specifičnosti. Većinom domene A pokazuju kospecifičnost, kao što je slučaj s domenom za adenilaciju prvog modula mikobaktin sintetaze koja aktivira aminokiseline L-Ser i L-Thr (Challis i sur., 2000). Još jedan primjer takve kospecifičnosti je domena za adenilaciju trećeg modula tirocidin sintetaze iz bakterije *Bacillus brevis* koja pokazuje aktivaciju aminokiseline L-Thr od 100 % i aktivaciju aminokiseline L-Phe od 48 % prilikom biokemijskih testova specifičnosti, ali je anotirana kao domena specifična za aminokiselinu L-Phe jer je aminokiselina D-Phe otkrivena u polipeptidnom lancu (Mootz i Marahiel, 1997). Isto tako, moguće je da je prilikom biokemijskih testova specifičnosti otkrivena visoka kospecifičnost testirane domene A na određeni supstrat, ali se taj supstrat ne otkriva u neribosomalno sintetiziranom polipeptidu koji je sintetiziran pomoću neribosomalne peptid sintetaze koja sadržava tu domenu za adenilaciju. Jedan od razloga je da prilikom sinteze neribosomalnog peptida pomoću neribosomalne peptid sintetaze, zbog aktivnosti domena prisutnih "nizvodno" na lancu, ne dolazi do vezanja tog supstrata ili ne dolazi do vezanja zbog steričkih razloga, tj. zbog okoline u kojoj se nalazi domena za adenilaciju u neribosomalnoj peptid sintetazi. Primjer takve domene za adenilaciju je domena BarD iz barbamid sintetaze. Prilikom biokemijskih testova specifičnosti, ta domena pokazuje specifičnost od 100 % za aminokiseline Leu i Val, te specifičnost od 80 % za aminokiselinu Tcl. Istovremeno u uvjetima *in vivo* nije otkrivena ugradnja aminokiseline Val od strane domene BarD (Chang i sur., 2002). Rješenje za ovaj tip problema prilikom određivanja specifičnosti domena A neribosomalnih peptid sintetaza predložili su Rausch i suradnici (2005). Oni su zaključili da će generiranje profila iz grupa domena za adenilaciju koje su specifične za supstrate fizikalno-kemijski sličnih svojstava na odgovarajući način riješiti problem. Time se povećao spektar sekvencija koje određuju određeni tip profila, to jest prostor varijabilnosti unutar određenog profila, i otvorila se mogućnost da takvi profili prepoznaju supstrate koji su fizikalno-kemijski slični supstratima profila. Filogenetska analiza aktivnih mjesta domena za adenilaciju generirala je slične rezultate. Filogenetsko grupiranje velikih skupina pokazalo je slično grupiranje s velikim skupinama koje su opisali Rausch i suradnici (2005). Drugi problema prilikom izrade modela bile su domene koje su specifične za netipične supstrate kao što su to aminokiseline: L-Dab, Aeo, Me-Asp, 4pPro, D-Lyserg itd. Za takav tip domena za adenilaciju bilo je nemoguće

generirati profil HMM jer u literaturi trenutno postoji premali broj sekvencija (Röttig i sur., 2011). Domene koje prepoznaju netipične supstrate su zato prilikom pretraživanja ulazile u netočno prepoznate domene, ulazile su u grupu TN (stvarno negativni engl. "True Negative"). Da bi se postiglo što bolje prepoznavanje specifičnosti domena za adenilaciju od strane modela profila HMM, potrebno je model stalno obnavljati sa sekvencijama novo anotiranih domena za adenilaciju.

6. ZAKLJUČCI

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

- Filogenetska analiza aktivnih mjesta domena za adenilaciju potvrdila je pretpostavku da se na pozicijama ostataka aminokiselina, unutar aktivnih mjesta koji su odgovorni za odabir supstrata, sličnih fizikalno-kemijskih svojstava nalaze iste ili fizikalno-kemijski slične aminokiseline što ukazuje na djelovanje principa konvergentne evolucije.

- Minimalni broj sekvencija domena za adenilaciju, specifičnih za određeni supstrat, prilikom izrade prihvatljivog profila HMM za određivanje specifičnosti jest 5. Profili HMM koji su generirani iz grupa domena za adenilaciju koje su imale 4 sekvencije ili manje, pokazali su zamjetno lošije rezultate prilikom pretrage baze podataka proteina UniProt od strane modela profila HMM za specifičnost domena za adenilaciju.

- Rezultati analize specifičnosti domena za adenilaciju od strane profila HMM generiranih na temelju filogenetske analize aktivnih mjesta domena za adenilaciju pokazali su da još uvijek postoji neriješeni problem, a to je "relaksirana" specifičnost. Jedan od pokušaja rješenja ovog problema je generiranje grupa specifičnosti domena za adenilaciju koje vežu supstrate fizikalno-kemijski sličnih svojstava. Ovo rješenje ima svoje mane: ne uzima u obzir steričke utjecaje koje okolina domena za adenilaciju ima na konformaciju same domene a samim time i na aktivnost iste, definira specifičnost za grupu fizikalno-kemijski sličnih supstrata a ne za jedan određeni supstrat, čime umanjuje točnost predikcije. Generiranje grupa specifičnosti domena za adenilaciju za izradu profila specifičnosti samo je jedan dio u rješavanju problema "relaksirane" specifičnosti domena za adenilaciju.

- Iako je početni broj sekvencija domena za adenilaciju, iz kojih su generirani profili HMM, bio relativno malen (njih 397), rezultati analize specifičnosti domena za adenilaciju od strane profila HMM pokazuju visoku razinu izvedbe generiranog modela. Daljnjim povećanjem broja anotiranih sekvencija domena za adenilaciju za izradu profila HMM ostavlja se realna mogućnost za poboljšanje izvedbe modela.

7. LITERATURA

Altschul, S.F., Miller, G.W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*. **215**, 403-410.

Anonymous 1 (2012) National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/sites/gquery. Pristupljeno 14. 02. 2012.

Anonymous 2 (2012) Google Scholar, http://scholar.google.hr/. Pristupljeno 14. 04. 2012.

Anonymous 3 (2011) Microsoft Corporation, http://www.microsoft.com/hr-hr/download/default.aspx. Pristupljeno 11. 05. 2011.

Anonymous 4 (2012) Science direct, http://www.sciencedirect.com/. Pristupljeno 14. 04. 2012.

Anonymous 5 (2012) The Universal Protein Resource, http://www.uniprot.org/. Pristupljeno 14. 02. 2012.

Ansari, M.Z., Yadav, G., Gokhale, R.S., Mohanty, D. (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.* **32**, 405-413.

Baxevanis, A.D., Oullette B.F.F. (2001) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2. izd., John Wiley & Sons, New York, str. 323-358.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W. (2011) GenBank. *Nucleic Acids Res.* **39**, 32-37.

Birney, E. (2001) Hidden Markov models in biological sequence analysis. *IBM. J. Res. Dev.* **45**, 449-454.

Bushley, K.E., Ripoll, D.R., Turgeon, B.G. (2008) Module evolution and substrate specificity of fungal nonribosomal peptide synthetases involved in siderophore biosynthesis. *BMC Evol. Biol*. **8**, 328-352.

Challis, G.L., Naismith, J.H. (2004) Structural aspects of non-ribosomal peptide biosynthesis. *Curr. Opin. Struct. Biol.* **14**, 748-756.

Challis, G.L., Ravel, J., Townsend, C.A. (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7**, 211-224.

Chang, Z., Flatt, P., Gerwick, W.H., Nguyen, V.A., Willis, C.L., Sherman, D.H. (2002) The barbamide biosynthetic gene cluster: a novel marine cyanobacterial system of mixed polyketide synthase (pks)-non-ribosomal peptide synthetase (nrps) origin involving an unusual trichloroleucyl starter unit. *Gene* **296**, 235-247.

Clamp, M., Cuff, J., Searle, S.M., Barton, G.J. (2004) The Jalview Java alignment editor. *Bioinformatics* **20**, 426-427.

Conti, E., Stachelhaus, T., Marahiel, M.A., Brick, P. (1997) Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *Embo J.* **16**, 4174-4183.

Desper, R., Gascuel, O. (2002) Fast and accurate phylogeny reconstruction algorithms on the minimum-evolution principle. *J. Comput. Biol.* **9**, 687-705.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* **14**, 755-763.

Eddy, S.R. (2004) What is a hidden Markov model? *Nat. Biotechnol.* **22**, 1315-1316.

Felnagle, E.A., Jackson, E.E., Chan, Y.A., Podevels, A.M., Berti, A.D., McMahon, M.D., Thomas, M.G. (2008) Nonribosomal peptide synthetases involved in production of medically relevant natural products. *Mol. Pharm.* **5**, 191-211.

Finn, R.D., Clements, J., Eddy, S.E. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, 29-37.

Fleischmann R. D., Adams M. D., White O., Clayton R. A., Kirkness E. F., Kerlavage A. R., Bult C. J., Tomb J. F., Dougherty B. A., Merrick J. M. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512.

Hillis, D. (1997) Phylogenetic analysis. *Curr. Biol.* **7**, 129-131.

Hongseok, T., Sohng, J.K., Park, K. (2009) Development of an analysis program of type I polyketide synthase gene clusters using homology search and profile hidden Markov model. *J. Microbiol. Biotechnol.* **19**, 140-146.

Konz, D., Marahiel, M.A. (1999) How do peptide synthetases generate structural diversity? *Chem. Biol.* **6**, 39-48.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

Lautru, S., Challis, G.L. (2004) Substrate recognition by nonribosomal peptide synthetase multi-enzymes. *Microbiology* **150**, 1629-1636.

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**, 185-195.

Magrane, M., UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database* **2011**, 1-13.

Mootz, H.D., Marahiel, M.A (1997) The tyrocidine biosynthesis operon of *Bacillus brevis*: complete nucleotide sequence and biochemical characterization of functional internal adenylation domains. *J. Bacteriol.* **179**, 6843-6850.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer E.L.L., Eddy, S.R., Bateman, A., Finn, R.D. (2012) The Pfam protein families database. *Nucleic Acids Res.* **40**, 290-301.

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., Huson, D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.* **33**, 5799-5808.

Röttig, M., Medema, M.H., Blin, K., Weber, T., Rausch, C., Kohlbacher, O. (2011) NRPSpredictor2 - a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, 1-6.

Samel, S.A., Marahiel, M.A., Essen, L.-O. (2008) How to tailor non-ribosomal peptide products - new clues about the structures and mechanisms of modifying enzymes. *Mol. BioSyst.* **4**, 387-393.

Schwarzer, D., Finking, R., Marahiel, M.A (2003) Nonribosomal peptides: from genes to products. *Nat. Prod. Rep.* **20**, 275-287.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539-544.

Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., Hulo, N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*. **38** (Database issue): D161-D166.

Sjölander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* **20**, 170-179.

Stachelhaus, T., Marahiel, M.A. (1995) Modular structure of peptide synthetases revealed by dissection of the multifunctional enzyme GrsA. *J. Biol. Chem.* **270**, 6163-6169.

Stachhelhaus, T., Mootz, H.D., Marahiel, M.A. (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493-505.

Starcevic, A., Zucko, J., Simunkovic, J., Long, P.F., Cullum, J., Hranueli, D. (2008) *ClustScan*: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. *Nucleic Acids Res.* **36**, 6882-6892.

Strieker, M., Tanović, A., Marahiel, M.A. (2010) Nonribosomal peptide synthetases: structures and dynamics. *Curr. Opin. Struct. Biol*. **20**, 234-240.

Tamura, K., Dudley, J., Nei, M., Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol. Biol. Evol.* **24**, 1596-1599.

The UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40**, 71-75.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876-4882.

Thompson, J.D., Higgins, D.G., Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence algnment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.

Verne Lee, T., Johnson, L.J., Johnson, R.D., Koulman, A., Lane, G.A., Lott, J.S., Arcus, V.L. (2010) Structure of a eukaryotic nonribosomal peptide synthetase adenylation domain that activates a large hydroxamate amino acid in siderophore biosynthesis. *J. Biol. Chem*. **285**, 2415-2427.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J. (2009) Jalview Version 2 - a multiple sequence alingment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191.

Weber, T., Marahiel, M.A. (2001) Exploring the domain structure of modular nonribosomal peptide synthetases. *Structure* **9**, 3-9.
Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, 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.

Yadav, G., Gokhale, R.S., Mohanty, D. (2003) SEARCHPKS: a program for detection and analysis of polyketide synthase domains. *Nucleic Acids Res.* **31**, 3654-3658.

Yonus, H., Neumann, P., Zimmermann, S., May, J.J., Marahiel, M.A., Stubbs, M.T. (2008) Crystal Structure of DltA. *J. Biol. Chem*. **283**, 32484-32491.

Zotchev, S.B., Stepanchikova, A.V., Sergeyko, A.P., Sobolev, B.N., Filimonov, D.A., Poroikov, V.V. (2006) Rational design of macrolides by virtual screening of combinatorial libraries generated through *in silico* manipulation of polyketide synthases. *J. Med. Chem.* **49**, 2077-2087.

8. PRILOZI

8.1. Popis u radu upotrijebljenih kratica

U ovoj su diplomskoj radnji upotrjebljene su sljedeće kratice:

A - domena za adenilaciju aminokiselina (u sustavu NRPS)
AMP – adenozin monofosfat (engl. "Adenosine Monophosphate")
ATP- adenozin trifosfat (engl. "Adenosine Triphosphate")
bp - parovi baza
C - domena za kondenzaciju (u sustavu NRPS)
Cy - domena za formiranje heterocikličkih prstenova (u sustavu NRPS)
DNA - deoksiribonukleinska kiselina
E - domena za epimerizaciju (u sustavu NRPS)
FN – lažno negativni (engl. "False Negative")
FP – lažno pozitivni (engl. "False Positive")
HMM – skriveni Markovljev model (engl. "Hidden Markov Model")
MEGA – računalni program za molekularnu evolucijsku genetičku analizu (engl. "Molecular Evolutionary Genetics Analysis Software")
MT - domena za N-, C- ili O-metilaciju (u sustavu NRPS)
NRPS - sintetaza neribosomalno sintetiziranih peptida (engl. "Nonribosomal Peptide Synthases")
Ox - domena oksidacije (u sustavu NRPS)
PCP - mali polipeptid nosač peptidila (u sustavu NRPS)
PKS - poliketid sintaza (engl. "Polyketide Synthase")
R – domena reduktaze (u sustavu NRPS)
Te - domena tioesteraze (u sustavu NRPS)
TN – stvarno negativni (engl. "True Negative")
TP – stvarno pozitivni (engl. "True Positive")

8.2. Sadržaj kompaktnog diska

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

Ova diplomska radnja sadržava slijedeće priloge:

8.2.1. Cijele sekvencije 397 domena za adenilaciju sustava NRPS u obliku zapisa FASTA pripremljenih za višestruko poravnavanje sekvencija pomoću programskog paketa Clustal W.

8.2.2. Filogenetsko stablo cijelih sekvencija domena A sustava NRPS (Stablo 1) generirano pomoću programskog paketa MEGA 4.0.2.

8.2.3. Sekvencije aktivnih mjesta svih 397 domena A u obliku zapisa FASTA pripremljenih za višestruko poravnavanje sekvencija pomoću programa Clustal W koji je integriran u programski paket MEGA 4.0.2.

8.2.4. Filogenetsko stablo sekvencija aktivnih mjesta domena A sustava NRPS (Stablo 2) generirano pomoću programskog paketa MEGA 4.0.2.

8.2.5. Tablica rezultata pretrage baza sekvencija proteina s generiranim profilima HMM domena za adenilaciju prikazana u programu Excel (Tablica 3)

NAPOMENA:
Za pregledavanje filogenetskih stabla cijelih sekvenci domena A (Stablo 1) i filogenetskog stabla aktivnih mjesta domena A (Stablo 2) potreban je klijent programskog paketa MEGA 4.0.2. na osobnom računalu.