VISOKA POSLOVNA ŠKOLA
STRUKOVNIH STUDIJA
ČAČAK

DIPLOMSKI RAD

Izrada racunalnog programa za prikupljanje, obradu, pohranu i graficki prikaz biloških podataka

Mentor:_____ Student: _____
Profesor: _____ Br.Indeksa: _____

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

IZRADA RAČUNALNOG PROGRAMA ZA PRIKUPLJANJE, OBRADU, POHRANU I GRAFIČKI PRIKAZ BIOLOŠKIH PODATAKA

Janko Diminić, 3925/PI
Prehrambeno-biotehnološki fakultet
Pierottijeva 6, 10000 Zagreb

Sažetak: U tijeku je razvoj integralnog računalnog programskog paketa, ***CompGen***, za modeliranje procesa homologne rekombinacije između poliketidnih/peptidnih genskih nakupina u uvjetima *in silico*. Okosnica će programskog paketa ***CompGen*** biti specifično strukturirana baza podataka koja povezuje biosintetski put sinteze sa sekvencijama DNA genskih nakupina. U ovom je radu strukturirana specifična baza podataka poliketidnih/ peptidnih genskih nakupina po uzoru na BioSQL. Razvijeni su i računalni programi za: prikupljanje i strukturiranje bioloških podataka; detaljan opis ("anotaciju") modula, domena i njihovih poveznica; anotaciju biosintetskih putova i grafičko prikazivanje dvodimenzionalnih i trodimenzionalnih kemijskih struktura linearnih lanaca poliketida/peptida objektno orijentiranim računalnim jezikom Java i skriptnim računalnim jezikom JavaScript.

Ključne riječi: baza podataka, modularne genske nakupine PKS i NRPS, računalni programski paket, računalni program, poliketidi i neribosomalno sintetizirani peptidi

Rad sadrži: 69 stranica, 25 slika, 1 tablicu, 84 literaturnih navoda, 3 priloga

Jezik izvornika: hrvatski

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

Mentor: dr. sc. Daslav Hranueli, red. prof.;

Pomoć pri izradi: Antonio Starčević, dipl. inž. biotehnol.

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

Datum obrane: 5, prosinac 2007.

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

DEVELOPMENT OF COMPUTER PROGRAM FOR THE ACQUISITION, PROCESSING, STORAGE AND GRAPHICAL PRESENTATION OF BIOLOGICAL DATA

Janko Diminić, 3925/PI
Faculty of Food Technology and Biotechnology
Pierottijeva 6, 10000 Zagreb

Abstract: An integral generic program package, ***CompGen***, is under development to model the process of homologous recombination between polyketide/peptide gene-clusters *in silico*. The heart of ***CompGen*** will be a specially structured database that will connect the biosynthetic order to the DNA sequences. In this work the specific database was structured based on BioSQL. The computer programs for the acquisition and structuring of biological data; the annotation of modules, domains and their dockers and linkers; the annotation of biosynthetic order and for the graphical presentation of two-dimensional and three-dimensional chemical structures of linear polyketide/peptide chains were developed using computer languages Java and JavaScript.

Keywords: polyketides and non-ribosomal synthesized peptides, modular PKS and NRPS gene-clusters, database, computer program package, computer program

Paper contains: 69 pages, 25 figures, 1 table, 84 references, 3 supplements

Original in: Croatian

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

Mentor: Dr. Daslav Hranueli, Full Professor;

Technical support and assistance: Antonio, Starčević, B.Sc.

Reviewers:
1. Dr. Želimir Kurtanjek, Full Professor
2. Dr. Daslav Hranueli, Full Professor
3. Dr. Ana Vukelić, Assistant Professor

Thesis defended: 5th of December 2007

Sadržaj

Stranica

1. Uvod 1

2. Teorijski dio

2.1. Biološke i kemijske osnove 2

2.2. Baze podataka i bioinformatički alati 9

2.2.1. Baza podataka GenBank 10

2.2.2. Baza podataka sintaza modularnih poliketida 11

2.2.3. Baza podataka Pfam 13

2.2.4. Baza podataka PubChem Compound 14

2.2.5. Bioinformatički program BLAST 15

2.2.6. Bioinformatički programski paket HMMER 18

2.2.7. Kemoinformatički program ChemAxon 21

2.3. Programski jezici i informatički alati 22

2.3.1. Projekt BioSQL 23

2.3.2. Sustav za upravljanje bazama podataka PostgreSQL 23

2.3.3. Računalni jezik SQL 24

2.3.4. Računalni jezik Java 26

2.3.5. Računalni jezik JavaScript 29

2.3.6. Računalni jezici HTML i CSS 30

2.3.7. Pomoćni program Eclipse 31

2.3.8. Programski paket Apache Tomcat 31

3. Eksperimentalni dio

3.1. Materijal 33

3.1.1. Sklopovlje i operativni sustav 33

3.1.2. Informatička programska podrška 33

3.1.2.1. Shema baze podataka Bio SQL 34

3.1.2.2. Sustav za upravljanje bazama podataka PostgreSQL 34

3.1.2.3. Računalni jezik SQL 34

3.1.2.4. Računalni jezik Java 34

3.1.2.5. Računalni jezik JavaScript 35

3.1.2.6. Računalni jezik HTML 35
3.1.2.7. Pomoćni program Eclipse 36
3.1.2.8. Programski paket Apache Tomcat 36
3.2. Metode 36
3.2.1. Strukturiranje sheme baze podataka po uzoru na BioSQL 36
3.2.2. Prikupljanje i strukturiranje bioloških i kemijskih podataka 38
3.2.3. Anotacija modula, domena i njihovih poveznica 42
3.2.4. Anotacija biosintetskih puteva poliketida/peptida 43
3.2.5. Grafičko prikazivanje kemijskih struktura poliketida/peptida 44

4. Rezultati
4.1. Baza podataka poliketida/peptida 45
4.2. Prikupljanje podataka 46
4.3. Obrada podataka 49

5. Rasprava 54

6. Zaključci 59

7. Literatura 60

8. Prilozi
8.1 Popis korištenih kratica u radu 68
8.2 Sadržaj kompakt diska 69

1. Uvod

Poliketidi i neribosomalno sintetizirani peptidi su vrlo važne kemijske supstancije za farmaceutsku industriju i agroindustriju. Njihov biosintetski put obuhvaća spajanje jednostavnih građevnih jedinica u složene kemijske strukture katalitičkim djelovanjem enzimskih kompleksa poliketid sintaza (PKS) ili sintetaza neribosomalnih peptida (NRPS). U posljednjih nekoliko godina u znanstvenoj javnosti postoji osobito zanimanje za oblikovanje novih supstancija u proizvodnji novih lijekova manipulacijom genskih nakupina tih enzimskih kompleksa u uvjetima *in vitro*, postupcima kombinatorne biosinteze. Međutim, značajna je prepreka napretku na tom području što većina promjena u uvjetima *in vitro* ne dovodi do sinteze produkta ili su mu prinosi vrlo mali. Jedno od mogućih rješenja toga problema bilo bi oblikovanje novih genskih nakupina homolognom rekombinacijom u uvjetima *in vivo* jer bi se tako omogućilo spajanje identičnih sekvencija i smanjile poteškoće zbog pojave nefunkcionalnih čvorišta (1).

U tijeku je razvoj integralnih računalnih programskih paketa, ***CompGen*** i ***ClustScan***, za modeliranje tih procesa u uvjetima *in silico*. Okosnica programskog paketa ***CompGen*** biti će specifično strukturirana baza podataka koja povezuje biosintetski put sinteze sa sekvencijama DNA genskih nakupina. Povezanost sekvencija DNA s biosintetskim putem omogućava njezinu povezanost sa strukturom produkta. Jedna je od funkcija računalnoga programskog paketa, temeljena na toj povezanosti, sposobnost oblikovanja virtualnih rekombinanata između genskih nakupina. To će se obavljati pomoću modela rekombinacije da bi se predvidjele sekvencije DNA u kojima dolazi do homologne rekombinacije. ***CompGen*** će iz tih podataka predviđati kemijske strukture novih supstancije i spremati ih u virtualnu bazu podataka kemijskih struktura radi daljnjega molekularnog modeliranja. U cjelini, ***CompGen*** će omogućiti oblikovanje baze podataka novih prirodnih kemijskih entiteta u uvjetima *in silico*, koja će se zatim moći upotrijebiti za istraživanja na području tehnologije računalnoga dizajna novih lijekova. Drugi će integralni generički računalni programski paket, ***ClustScan***, moći prepoznati i detaljno opisati, "anotirati", nove genske nakupine iz sekvencija cjelovitih mikrobnih genoma ili genskih nakupina u metagenomima mikroorganizama što žive u tlu ili u simbiozi s morskim organizmima (2).

Za potrebe prikupljanja, obrade, pohrane i grafičkog prikaza bioloških i kemijskih podataka, u ovom će diplomskom radu biti razvijen računalni program objektno orijentiranim računalnim jezikom Java i skriptnim računalnim jezikom JavaScript.

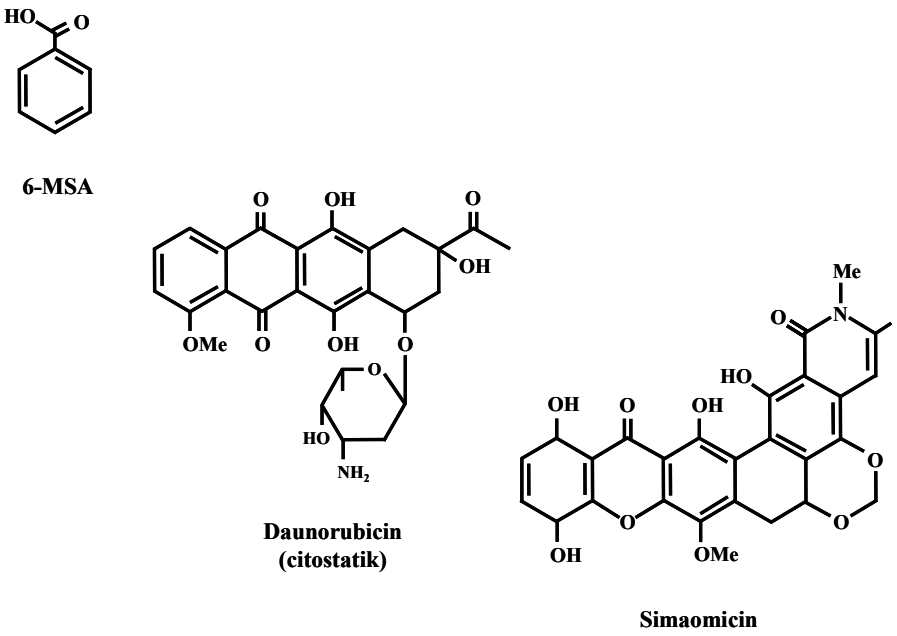
2. Teorijski dio

2.1. Biološke i kemijske osnove

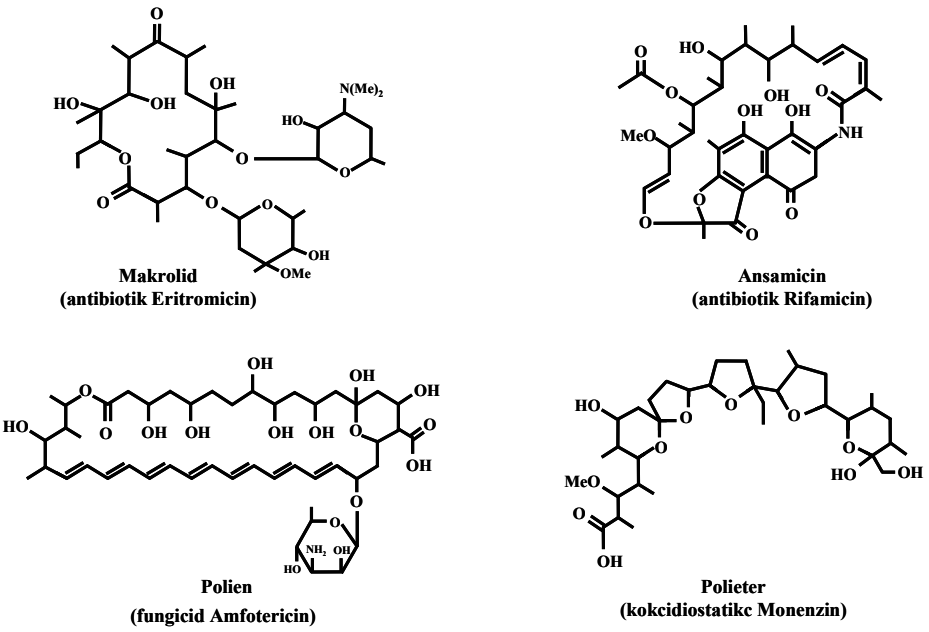
Mikroorganizmi i biljke sintetiziraju veliki broj različitih metabolita koji se proizvoljno mogu podijeliti na takozvane primarne i sekundarne metabolite. Primarni su metaboliti prijeko potrebni za njihov rast i razmnožavanje, dok su sekundarni metaboliti bitni za preživljavanje u prirodnim staništima (3, 4). S obzirom na strukturalni potencijal biološke raznolikosti, mikroorganizmi koji sintetiziraju najveći broj sekundarnih metabolita, od kojih su mnogi biološki aktivni, pripadaju bakterijama roda *Streptomyces* i srodnih rodova. Ilustracije radi, od otprilike dvadesetak tisuća antibiotički aktivnih supstancija, izoliranih od bakterija do sisavaca, streptomicete sintetiziraju gotovo 8.000. Štoviše, vrste roda *Streptomyces* proizvode 75 % svih klinički važnih antibiotika (5). Nedavno su Watve i suradnici pokušali matematički procijeniti broj antibiotika koje ove bakterije mogu sintetizirati. Njihov je model pokazao da je taj broj oko 100.000, (6), od kojih je manje od 10 % do sada otkriveno. Svaka pojedina vrsta roda *Streptomyces* može sintetizirati više od jednog sekundarnog metabolita. Tako, na primjer, bakterija *S. coelicolor* sintetizira najmanje 4 antibiotika: aktinorodin, undecilprodigiosin, metilenomicin i lipopeptidni antibiotik CDA. Štoviše, sekvenciranje je genomskih DNA bakterija *S. coelicolor*, *S. avemitilis* i *Saccharopolyspora erythraea* pokazalo da svaka od tih bakterija sadržava 18, 35 i 25 genskih nakupina za biosintezu sekundarnih metabolita (7, 8, 9).

Ti sekundarni metaboliti nisu samo značajni u njihovim prirodnim staništima već i za zdravlje ljudi (što je u suprotnosti s njihovim nazivom koji upućuje na "sekundarno značenje"), jer gotovo 50 % najvažnijih lijekova, koji se danas klinički primjenjuju, kao ljekovite supstancije sadržavaju prirodne spojeve (3). Osim antibiotika, fungicida, antivirusnih sredstava i citostatika, u kliničkoj se primjeni nalaze i imunosupresori, antihipertenzivna sredstva, antidiabetici, antimalarici, te antiholesterolemici (10, 11). Komercijalno su važni i proizvodi za agroindustriju, kao što su to antiparazitici, kokcidiostatici, životinjski promotori rasta i prirodni insekticidi. Među sekundarnim metabolitima poliketidi su najveća skupina. Svoje su ime dobili po keto skupinama koje nakon sinteze zaostaju u molekuli. Sa gledišta kemije, poliketidi su strukturno vrlo heterogena skupina prirodnih spojeva. Može ih se proizvoljno podijeliti u dvije skupine. Prva su skupina aromatski ili jednostavni poliketidi, koji mogu sadržavati od jednog do četiri aromatska prstena (Slika 1A). Njihova se struktura, međutim, ne može predvidjeti iz sekvencije DNA gena što sadržavaju genetičku uputu za njihovu biosintezu. Druga su skupina složeni poliketidi, koji se dalje dijele na makrolide i ansamicine, s laktonskim i laktamskim prstenovima, te na poliene i polietere (Slika 1B).

A

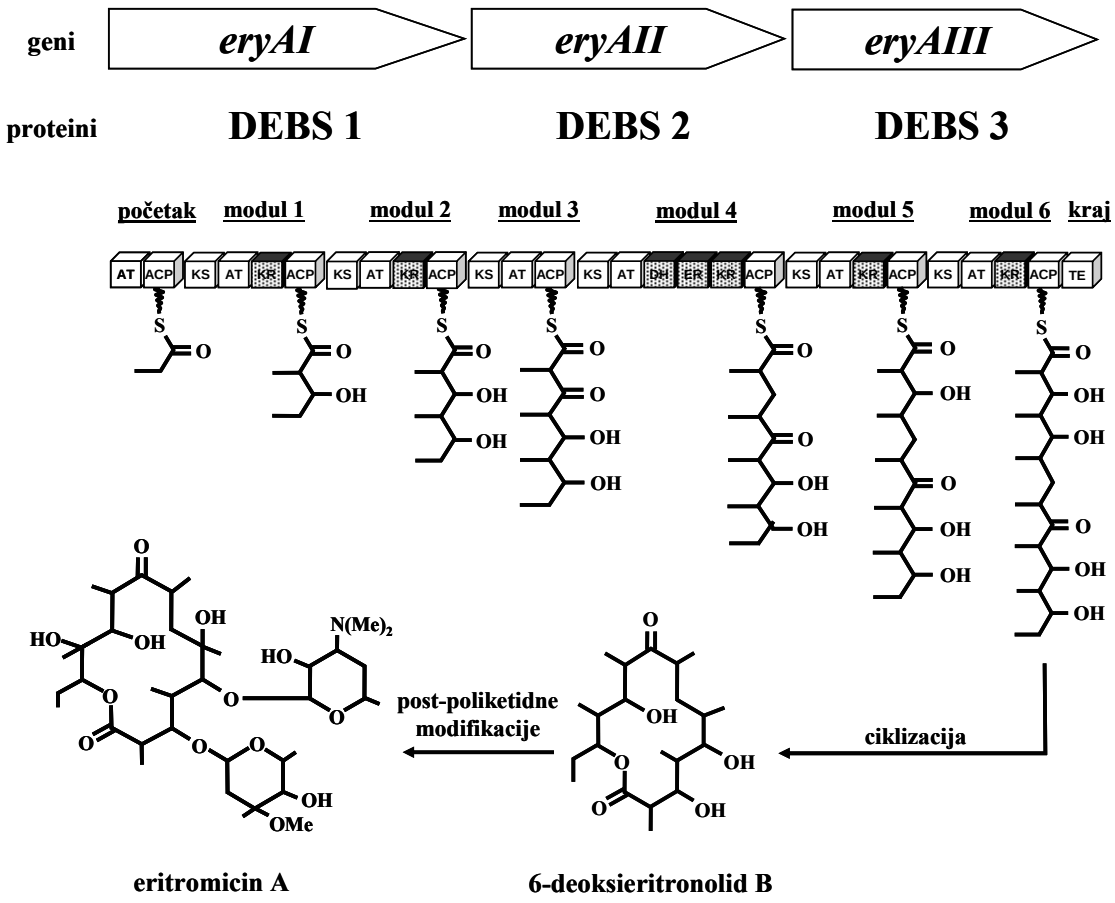


B



Slika 1. Kemijske strukture odabranih poliketidnih supstancija. **(A)** *jednostavni* aromatski poliketidi: 6-metilsalicilna kiselina, daunorubicin i simaomicin α. **(B)** *složeni* reducirani poliketidi podijeljeni na makrolide (eritromicin A), poliene (amfotericin B), ansamicine (rifamicin B) i polietere (monenzin A).

Unatoč strukturnoj raznolikosti, poliketidi se sintetiziraju pomoću enzimskih kompleksa nazvanih poliketid sintaze (PKS) (Slika 2), biosintetskim putem sličnim biosintezi masnih kiselina. Rast poliketidnog ugljikova lanca započinje kondenzacijom početne građevne jedinice s prvom produžnom građevnom jedinicom, i tako dalje sve do kraja sinteze poliketidnoga lanca. Građevne se jedinice u stanicama bakterija nalaze u obliku tioestera koenzima-A. Najčešće početne i produžne građevne jedinice su acetil i malonil, ali se u tu svrhu mogu upotrijebiti i ostaci drugih organskih kiselina, kao što su to metilmalonil, etilmalonil, propionil, butiril, 2-metilbutiril i neke druge (11).

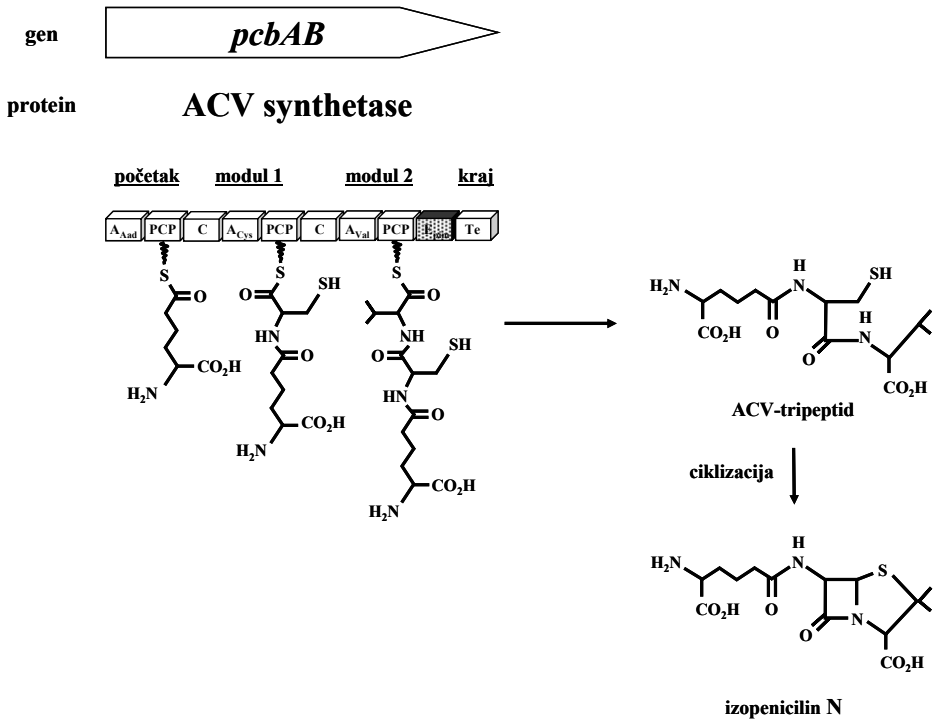


Slika 2. Shematski prikaz organizacije gena i proteina modularne poliketid sintaze tipa I odgovorne za biosintezu eritromicina A. Linearni raspored gena (*ery*AI, *ery*AII i *ery*AIII) i proteina [DEBS1, DEBS2 i DEBS3; koji sadržavaju aktivne domene za acil-transferaze (AT), male polipeptide, nosače acila (ACP), ß-ketoacil-sintaze (KS), ketoreduktaze (KR), dehidratazu (DH) alkenreduktazu (ER) i tioesterazu (TE)], te njihovih supstrata/produkata, poliketidnog aglikona 6-deoksieritronolida B (6-dEB) i makrolidnog antibiotika eritromicina A

Mikroorganizmi sintetiziraju poliketidne supstancije pomoću različitih tipova enzima PKS. U bakterija su to multimodularne PKS tipa I, ponavljajuće PKS tipa II i monomodularne ponavljajuće PKS tipa III (1, 12, 13). Plijesni poliketide sintetiziraju pomoću monomodularnih ponavljajućih PKS tipa I i monomodularnih ponavljajućih PKS tipa III (14, 15). Sekvenciranje genskih nakupina ponavljajućih PKS tipa II, kao što je to PKS aromatskoga poliketidnog antibiotika aktinorodina, upućuje na prisutnost od tri do šest monofunkcionalnih ili bifunkcionalnih proteina čija se aktivna mjesta ponovljeno upotrebljavaju tijekom prije-poliketidne sinteze poliketidnoga kostura. Sve su PKS tipa II vrlo slične, pa se kemijska struktura poliketida ne može predvidjeti iz strukturne organizacije gena PKS (11). Isto vrijedi i za monomodularne ponavljajuće PKS tipa I i tipa III (14, 15). Zbog toga te genske nakupine, to jest ti geni, neće biti detaljno razmatrani. Nasuprot tome, sekvenciranje je genskih nakupina PKS tipa I, kao što je to PKS makrolidnog antibiotika eritromicina u bakterije *Sac. erythraea* (Slika 2; vidi: 11), pokazalo da su to multifunkcionalni enzimi multimodularne organizacije. Svaki je modul odgovoran za rast poliketidnog lanca za jednu produžnu građevnu jedinicu i sadržava domene potrebne za redukciju β-ugljika. Moduli su najčešće raspoređeni na nekoliko polipeptida, od koji svaki sadržava više od jednog modula. Unutar modula PKS genskih nakupina nalazi se niz domena koje provode različite stupnjeve procesa biosinteze. Tako, naprimjer, aciltransferaze (AT) dovode građevne jedinice na male polipeptide nosače acila (ACP). Ketosintaze (KS) prenose rastuće poliketidne lanace do nove građevne jedinice i provode kondenzacije.

Prema tome, osnovni moduli sadržavaju samo domene KS-AT-ACP i ugrađuju keto skupine u poliketidne lance. Međutim, većina modula sadržava i druge domene potrebne za redukciju keto skupina. Ako su prisutne ketoreduktaze (KR), keto skupine se reduciraju u hidroksilne skupine. Ako su prisutne i dehidrogenaze (DH), hidroksilne će skupine biti reducirane u alkene, dok će prisutnost enoilreduktaza (ER) reducirati alkene u alkane. Postoji, dakle, neposredna veza između sekvencija DNA tih genskih nakupina, njihovih proteinskih produkata, modula i katalitički aktivnih domena sa strukturom njihovih poliketidnih produkata. Zbog toga se kemijska struktura poliketidnoga produkta može predvidjeti iz sekvencije DNA gena PKS. Posljednji, šesti modul, završava domenom tioesteraza (TE) koja je odgovorna za odvajanje poliketidnog ugljikova lanca od PKS i ciklizaciju, pri čemu nastaje poliketidni aglikon 6-deoksieritronolid B. Nakon toga, post-poliketidne hidroksilaze i glikozilaze sudjeluju u sintezi makrolidnog antibiotika eritromicina A (11).

Neribosomalno sintetizirani peptidi su, osim poliketida, druga velika skupina sekundarnih metabolita. Među neribosomalnim peptidima također se nalaze poznati antibiotici (npr. preteča penicilina ACV-tripeptid), fungicidi (npr. fengycin) i imunosupresori (npr. ciklosporin). Osim toga, opisane su i miješane poliketidno/peptidne supstancije poput citostatika bleomicina i epotilona. Oni se sintetiziraju pomoću sintetaza neribosomalno sintetiziranih peptida (NRPS), kao što je NRPS peptidnog antibiotika penicilina u plijesni *Penicillium chrysogenum* (Slika 3; vidi: 16). Genske nakupine NRPS imaju sličnu organizaciju onoj genskih nakupina PKS, samo što se kao građevne jedinice upotrebljavaju aminokiseline. Kao i u enzima PKS, osnovni moduli enzima NRPS sadržavaju domene za adenilaciju aminokiselina (A), kondenzaciju (C) i male polipeptide nosače peptidila (PCP). Neki moduli enzima NRPS također sadržavaju i domene za izmjenu struktura građevnih jedinica, to jest domene potrebne za formiranje heterocikličkih prstenova (Cy), epimerizaciju (E_{pim}) (Slika 3), redukciju (R), *N*-, *C*- ili *O*-metilaciju (MT) i druge.

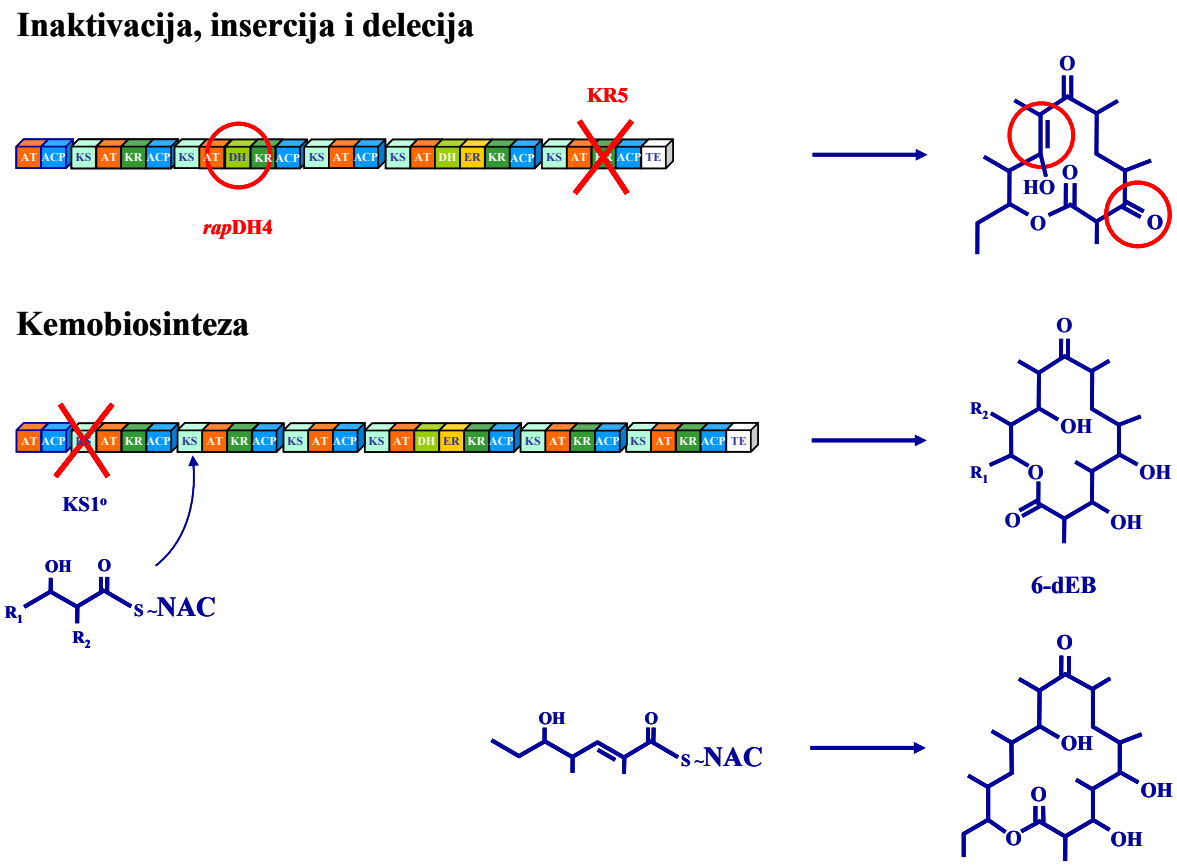


Slika 3. Shematski prikaz organizacije gena i proteina multimodularne peptid sintetaze odgovorne za biosintezu preteče penicilina (ACV). Prikazan je linearni gen (*pcb*AB) i protein (AVC sintetaza), koji sadržava katalitički aktivne domene za adenilaciju (A), male polipeptide nosače peptidila (PCP), kondenzaciju (C) i tioesterazu (Te), te peptidna preteča ACV i peptidni antibiotik izopenicilin N. Osjenčana je domena koja sudjeluje u epimerizaciji građevne jedinice (E_{pim}).

Na kraju se posljednjeg modula, kao i u enzima PKS, nalazi domena tioesteraza (Te) odgovorna za odvajanje linearnoga peptidnog lanca od enzima i ciklizaciju. Osim 20 aminokiselina (koje se pojavljuju u staničnim proteinima), u biosintezi neribosomalno sintetiziranih peptida, kao građevne jedinice, mogu poslužiti i brojne druge aminokiseline (oko 200 različitih aminokiselina). Prema tome, biosintetski im je potencijal još raznovrsniji od potencijala poliketid sintaza (17).

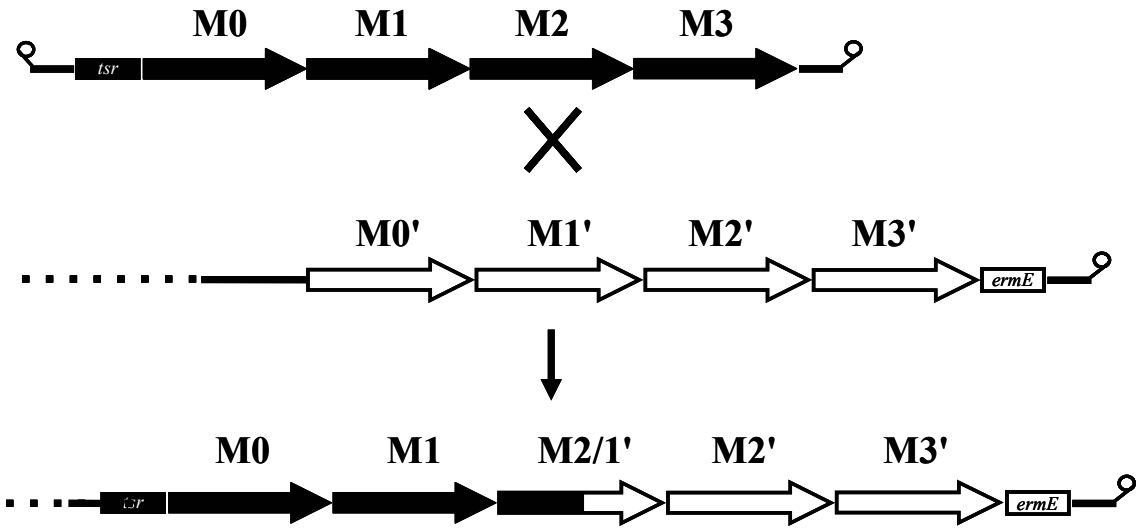
Priroda je već upotrijebila svojstva kombinatorne biosinteze i sintetizirala preko 170.000 poznatih prirodnih spojeva (vidi: "Chapman & Hall dictionary of natural products"; Dictionary of Natural Products, 18). Do danas izolirani prirodni poliketidi/peptidi predstavljaju, međutim, samo mali dio kombinatornog potencijala koji se može postići permutacijama gena PKS/NRPS.

Na primjer, ako se u 6 modula PKS eritromicina permutira svih 7 aciltransferaza i svih 5 domena koje reduciraju β-ugljike, teorijski bi se moglo dobiti 10^7 poliketidnih struktura. Štoviše, cjelokupna bi permutacija istih domena u PKS rapamicina, koja sadržava 14 modula, dovela do sinteze nevjerojatnog broja od 10^{14} poliketida (11, 19). Zbog toga su znanstvenici u složenih poliketida, poput makrolida, započeli s modifikacijom gena u uvjetima *in vitro*. Započeli su s inaktivacijom aktivnih mjesta u PKS ili dodavanjem aktivnih mjesta iz drugih PKS, radi sinteze novih makrolida s različito reduciranim β-ugljicima i različitom stereokemijom. Ne smije se zaboraviti da i tako male kemijske promjene mogu značajno promijeniti biološku aktivnost. Osim toga, provode se i delecije cijelih modula multimodularnih gena PKS ili se upotrebljava takozvana kemijska biosinteza, radi sinteze novih makrolida s različitom veličinom ugljikovih kostura (Slika 4) (11, 20, 21). Taj pristup, međutim, ima i svojih ograničenja. On zahtijeva vrlo specifična znanja i razmjerno mnogo vremena. Do danas je genetičkim manipulacijama u uvjetima *in vitro* konstruirano svega dvjestotinjak genskih nakupina koje sintetiziraju potpuno nove u prirodi nepoznate poliketide (22). Osim toga, mnoge od tako konstruiranih genskih nakupina ne dovode do sinteze poliketida ili je prinos poliketida vrlo mali. Zbog toga je predložena konstrukcija novih poliketidnih genskih nakupina procesom rekombinacije u uvjetima *in vivo*, to jest homolognom rekombinacijom (23) (Slika 5) koja u bakterija nastaje nakon prirodnog prijenosa genetičkoga materijala procesima konjugacije ili transformacije. Vjeruje se da će taj pristup omogućiti rekombinaciju između sekvencija DNA koje se i inače događaju u prirodi. Tako bi se smanjila mogućnost da nastanu nefunkcionalna neprirodna čvorišta, kao tijekom rekombinacije u uvjetima *in vitro* (1).



Slika 4. Shematski prikaz ciljane genetičke manipulacije genskim nakupinama složenih poliketida radi sinteze novih poliketidnih spojeva. Shema prikazuje linearni raspored proteina [DEBS1, DEBS2 i DEBS3; koji sadržava aktivne domene za acil-transferaze (AT), male polipeptide, nosače acila (ACP), ß-ketoacil-sintaze (KS), ketoreduktaze (KR), dehidratazu (DH) alkenreduktazu (ER) i tioesterazu (TE)] koji sudjeluju u biosintezi poliketidnog aglikona 6-deoksieritronolida B (6-dEB). Prikazana je inaktivacija aktivne domene (**X**), dodavanje aktivne domene (O) i delecija cjelovitih modula uz premještanje aktivne domene TE, te postupak kemobiosinteze.

Da bi se pronašle genske nakupine s potencijalno novom genetikom (tj. s genima za koje u bazama podataka ne postoji homologija) ili kemijom (što sadržavaju domene AT/A koje dovode nepoznate građevne jedinice), potrebno je anotirati genske nakupine PKS/NRPS, prisutne u genomima nedovoljno dobro opisanih organizama (npr. u genomu amebe *Dictyostelium discoideum*; 24, 25), kao i genske nakupine PKS/NRPS u metagenomima mikroorganizama koji žive u tlu ili u simbiozi s morskim organizmima (26, 27).



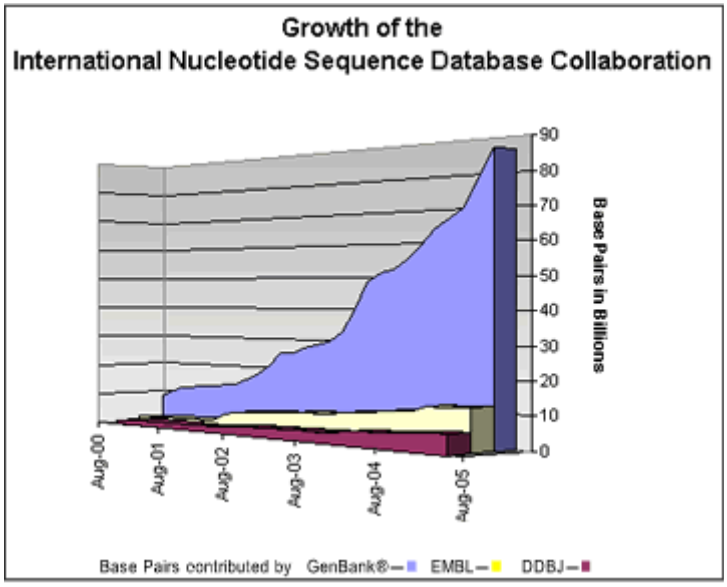
Slika 5. Predložena strategija za homolognu rekombinacije između dvaju poliketidnih genskih nakupina u vrsta roda *Streptomyces*. Jedna bi genska nakupina (➡) trebala biti klonirana na linearni plazmidni vektor što sadržava genetički biljeg za otpornost prema antibiotiku tiostreptonu (*tsr*). Druga bi genska nakupina (⇨) trebala biti klonirana u blizini kraja linearnog kromosoma što sadržava genetički biljeg za otpornost prema antibiotiku eritromicinu (*erm*E). Nova bi poliketidna genska nakupina (➡) nastala jednostrukim premošćenjem između dviju genskih nakupina. Ona bi se selekcionirala prijenosom rekombinantnoga plazmida, što sadržava oba genetička biljega, u drugu vrstu roda *Streptomyces*.

2.2. Baze podataka i bioinformatički alati

Opće baze podataka, kao što su to EMBL (engl. "EMBL Nucleotide Sequence Database"), GenBank i DDBJ (engl. "DNA Database of Japan") (28), sadržavaju sve podatke povezane s različitim sekvencijama DNA i proteina genskih nakupina, pa tako i za sintaze/sintetaze modularnih poliketida/peptida. Ti podaci, međutim, nisu dovoljno detaljno opisani, te su organizirane specijalizirane baze podataka sintaza modularnih poliketida kao što je to baza podataka sintaza modularnih poliketida PKSDB (eng. "Database of Modular Polyketide Synthases").

2.2.1. Baza podataka GenBank

Baza podataka GenBank (29) je baza podataka dobro opisanih, to jest anotiranih, sekvencija DNA koju održava Nacionalni centar za biotehnološke informacije (NCBI). U njoj se može pronaći 65,369,091,950 baza u 61,132,599 zapisa. Cjeloviti se pregled trenutne verzije baze podataka GenBank nalazi na poslužitelju centra NCBI za prijenos podataka (engl. "NCBI ftp site"). Novo se izdanje baze podataka GenBank izdaje svaka dva mjeseca. Baza podataka GenBank dio je međunarodne suradnje (engl. "International Nucleotide Sequence Database Collaboration") između baza podataka sekvencija DNA koja, pored baze podataka GenBank, uključuje i baze podataka EMBL te DDBJ. Te tri organizacije izmjenjuju podatke na dnevnoj razini (30).



Slika 6. Prikaz porasta broja sekvencija DNA (u parovima baza) koji potječe iz međunarodne suradnje između baza podataka GenBank, EMBL i DDBJ.

Podacima baze podataka GenBank može se pristupiti na nekoliko načina. Dva najčešće upotrebljavana pristupa su sustav za pristup podacima Entrez, koji povezuje GenBank sa svim ostalim bazama podataka koje održava centar NCBI, te pretraživanje baze podataka pomoću bioinformatičkog programa BLAST (vidi: poglavlje 2.2.5).

Mnogi znanstveni časopisi zahtijevaju da se prije objavljivanja znanstvenog rada, koji sadržava sekvenciju DNA, sekvencija DNA pohrani u bazi podataka i dobije pristupni broj (engl. "accession number") koji u radu mora biti naveden. Za tu svrhu baza podataka GenBank posjeduje računalne alate za pohranu sekvencija DNA pod nazivom "BankIt" i "Sequin". Računalni alati BankIt i Sequin se upotrebljavaju i za reviziju, to jest ažuriranje, pohranjenih zapisa što se može obaviti u bilo koje vrijeme.

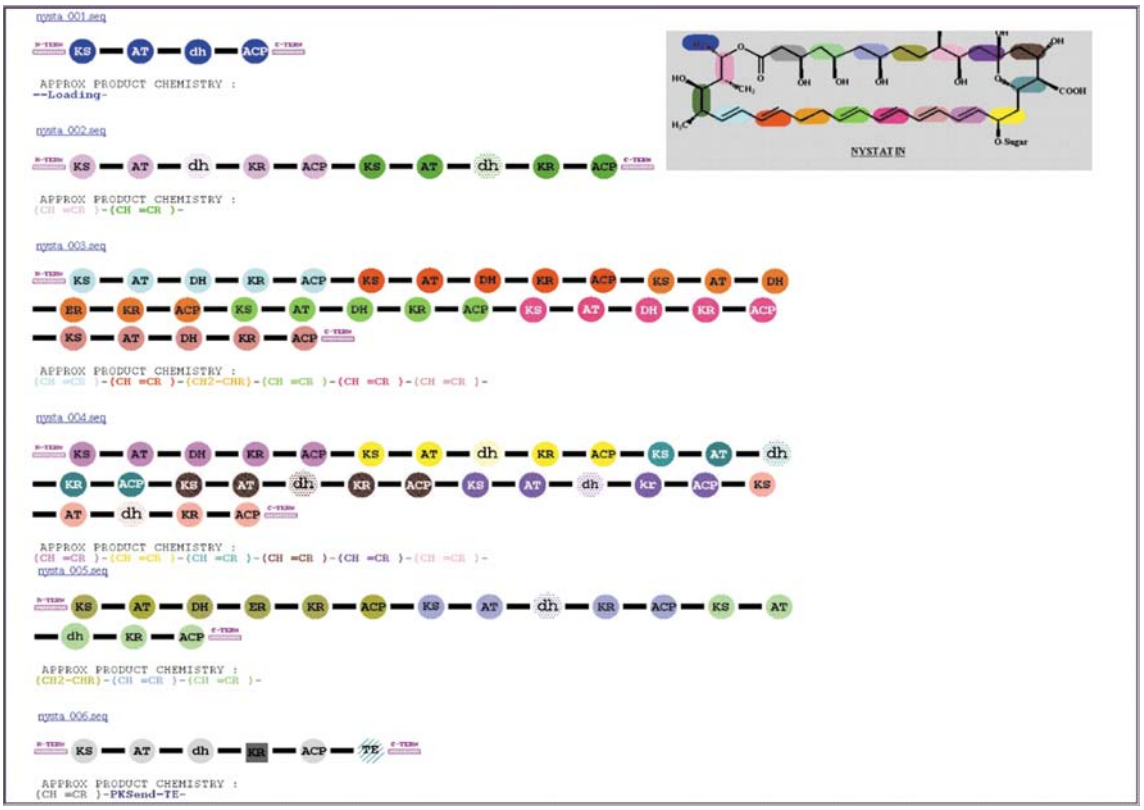
Nacionalni centar za biotehnološke informacije (NCBI) trajno razvija nove alate i zamjenjuje postojeće u svrhu efikasnijeg pristupa i pohranjivanja podataka o sekvencijama DNA u bazu podataka GenBank. Najjednostavniji način pristupa informacijama o razvoju baze podataka GenBank jest uključivanje u listu elektroničke pošte "NCBI Announce", redovito praćenje "*NCBI News*", te provjerom sekcije "What's New" na Web stranicama centra NCBI (30).

2.2.2. Baza podataka sintaza modularnih poliketida

Baza podataka sintaza modularnih poliketida PKSDB (eng. "A Database of Modular Polyketide Synthases") je specijalizirana baza podataka sintaza modularnih poliketida tipa I, koja je utemeljena pri Nacionalnom centru za imunologiju (eng. "National Institute of Immunology", NII), New Delhi, Indija (31). Ona služi kao znanstveno utemeljeni izvor podataka za analizu poliketid sintaza. Cilj baze je da osigura znanstvenu osnovu za predviđanje organizacije domena i specifičnosti supstrata u neopisanim genskim nakupinama PKS-a (32). Korisnik ne pristupa bazi neposredno, već upotrebljava grafičko sučelje putem Internet stranica na udaljenom poslužitelju. Sama baza sadržava 20 različitih, dobro opisanih (anotiranih) sintaza poliketida.

Upotreba baze započinje na taj način da se, nakon početne stranice, odabere jedna od ponuđenih biološki aktivnih supstancija. Nakon odabira, korisniku je prikazana kemijska struktura odabranog spoja. Nakon toga se može odabrati grafički prikaz gena, modula, domena i poveznica poliketid sintaze odgovorne za nastanak spoja. Geni su prikazani odvojeno jedan od drugog, sa svakim modulom unutar gena obojenim njemu specifičnom bojom. Domene su unutar modula prikazane geometrijskim simbolima s imenima domena upisanim u njih različitim veličinama slova. Domene su šrafirane različitim uzorcima, ovisno o njihovom stanju (aktivna, neaktivna, uspješno ili neuspješno prepoznata, nefunkcionalna). Poveznice između domena i modula su prikazane punim crnim crtama, dok su C-terminalna i N-terminalna poveznica

prikazane ljubičastom osjenčanom crtom (Slika 7). Za svaki se odabrani element (gen, modul, domenu i poveznicu) može dobiti proteinska sekvencija u zapisu oblika FastA (33). Na stranici za domene se, između dvije PKS domene nađene u bazi PKSDB, može izvršiti usporedba homologije prema sličnosti, identičnosti ili očekivanoj vrijednosti. Uz to se za svaku sekvenciju pojedinačne domene mogu usporediti motivi aktivnih mjesta i specifičnosti za supstrat, te se odabrana domena može sravniti s bilo kojom drugom domenom, okarakteriziranom u bazi PKSDB. Na stranici se može dobiti i grafički prikaz kemijske strukture spoja koji je povezan sa modularnom genskom strukturom sintaze, te literaturni podatak o odabranom spoju i sintazi putem hiperveze prema bazi podataka PubMed. Baza podataka PKSDB sadržava i stranicu za pretraživanje, koja korisniku pruža mogućnost da prepozna moguće domene poliketida u modularnim genskim nakupinama PKS s nepoznatim poliketidnim produktom ili u nedovoljno dobro opisanoj genskoj nakupini sličnoj PKS. Važno je primijetiti da baza podataka PKSDB ne sadržava primarne sekvencije DNA, već isključivo one primarnih sekvencija proteina (31).



Slika 7. Shematski prikaz genske nakupine nistatina. Prikaz Web stranice baze podataka PKSDB.

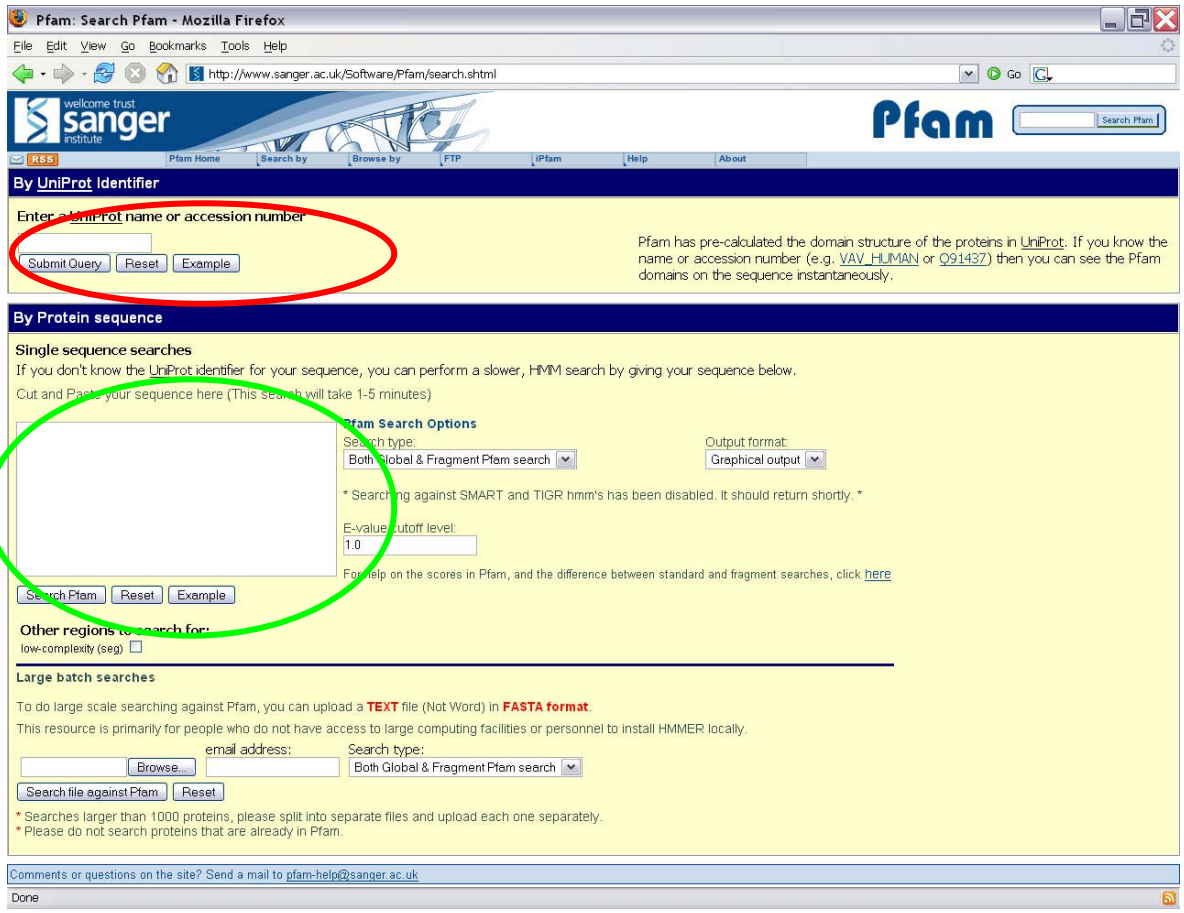
2.2.3. Baza podataka Pfam

Baza podataka Pfam (34) je baza podataka koja sadržava podatke o obiteljima proteinskih domena. Ova baza sadržava pregledana višestruka poravnanja sekvencija smještenih u pojedinu proteinsku obitelj (engl. "protein family"), kao i profile skrivenih Markovljevih modela (profili – HMM) koji se upotrebljavaju pri traženju domena u novim proteinskim sekvencama. Baza podataka Pfam sadržava funkcionalnu anotaciju, literaturne citate i hiperveze (engl. "hyperlinks") za svaku od obitelji.

Postoje dva višestruka poravnanja za svaku Pfam obitelj: osnovno poravnanje (engl. "seed") koje sadržava razmjerno malen broj reprezentativnih članova obitelji i ručno je pregledan, te cjelovito poravnanje (engl. "full") koje sadržava sve članove obitelji koji se u bazi mogu pronaći. Sva poravnanja upotrebljavaju sekvencije preuzete iz baze "pfamseq", koja je skup proteinskih sekvenci baza Swiss–Prot i SP-TrEMBL bez ponavljanja istih sekvencija (engl. "non–redundant"). Profil – HMM je izgrađen od osnovnog poravnanja upotrebom programskog paketa HMMER (vidi: poglavlje 2.2.6.), te se upotrebom tako kreiranog profila pretražuje "pfamseq" baza sekvencija. Dobivene podudarnosti iznad određenog praga poravnate su i tvore cjelovito poravnanje.

U verziji 10.0 baza podataka Pfam sadržava 6190 obitelji Pfam (35), te ima podudarnost sa 75% sekvencija iz baza podataka Swiss–Prot i TrEMBL. Baza podataka Pfam sadržava vrlo točne opise proteinskih domena. Postoje dva osnovna načina na koje se ove informacije mogu pretraživati: traženjem informacija o opisanim sekvencijama baze podataka Swiss-Prot, kada je struktura od prije poznata, ili upotrebom potpuno novih proteinskih sekvencija tražeći od baze podataka Pfam da proračuna mjesta na kojima se pretpostavlja da se domene nalaze.

U slučaju da se upotrebljavaju ustanovljene i opisane sekvencije baza podataka Swiss–Prot/TrEMBL, potrebno je znati pristupni broj (engl. "accession number"), kojega se unese u gornje tekstualno polje (Slika 8, zaokruženo crvenom elipsom). Kao rezultat se dobije struktura domena proteina koji je bio upotrijebljen kao upit. Kada je upit do sada neistražena sekvencija proteina, potrebno je takvu sekvenciju u obliku zapisa FastA upisati kao upit u polje, koje je na Slici 8, zaokruženo zelenom elipsom.



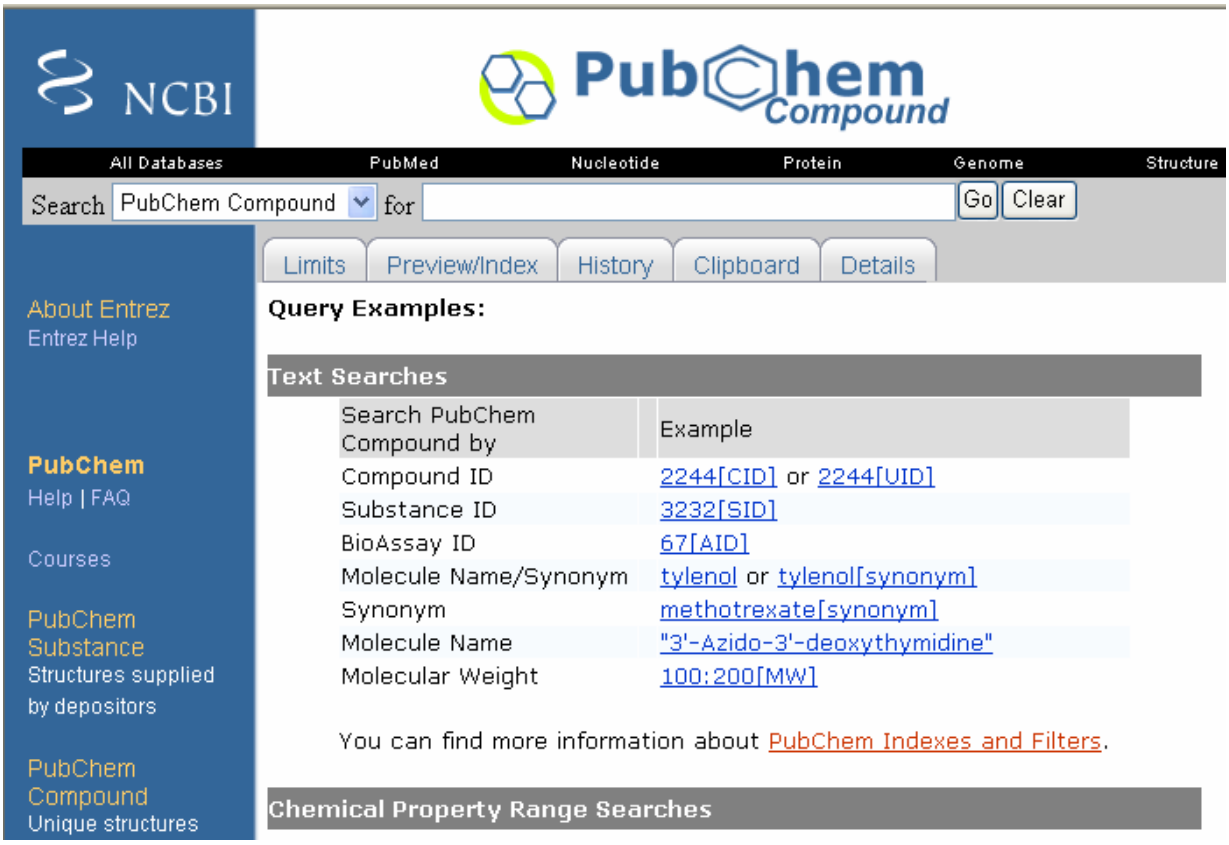
Slika 8. Prikaz Web stranice za pretragu baze podataka Pfam.

2.2.4. Baza podataka PubChem Compound

Servis PubChem (36) pruža informacije u kemijskim strukturama, svojstvima i biološkim aktivnostima malih molekula. Organiziran je u tri povezane baze podataka koje se mogu pretraživati pomoću sustava za pristup podacima Entrez Nacionalnog centara za biotehnološke informacije (NCBI). Te tri baze podataka su: "PubChem_Substance", "PubChem_Compound" i "PubChem_BioAssay". Baza podataka PubChem_Substance sadržava više od 19 miliona zapisa. Baza podataka PubChem Compound sadržava više od 10 miliona jedinstvenih kemijskih struktura, a baza podataka PubChem_BioAssay sadržava više od 500 bioloških testova.

Bazu podataka PubChem Compound može se pretraživati pomoću kemisjke strukture (Slika 9). Kemijska se struktura može nacrtati ili se može izraziti u obliku izomeričkog zapisa SMILES (akronim za "Simplified Molecular Input Line Entry System"; 37). Takav način pretraživanja pruža mogućnost da se pronađu sve početne i produžne građevne jedinice poliketida

i peptida. Kemijske strukture početnih i produžnih građevnih jedinica mogu se pronaći kao zasebne kemijske strukture, ili kao dijelovi složenijih kemijskih struktura iz kojih se zatim mogu izvesti u obliku izomeričkih zapisa SMILES.



Slika 9. Prikaz Web stranice za pretragu baze podataka PubChem Compound.

2.2.5. Bioinformatički program BLAST

Neposredan rezultat molekularne evolucije jesu sekvencije bioloških makromolekula. Sličnost u strukturi sekvencija (sličan slijed nukleinskih baza ili aminokiselina), a samim time i u funkciji molekula, može se vidjeti u sekvencijama koje dijele zajedničkog pretka. Usporedi li se nova, nepoznata, sekvencija (dobivena sekvenciranjem genoma) s bazom podataka dobro opisanih (anotiranih) sekvencija, moguće joj je predvidjeti biološku strukturu. Samim time, moguće joj je predvidjeti i funkciju. Kako je u kasnim '70-ima počeo rasti broj dostupnih sekvencija DNA i proteina, raslo je i zanimanje za razvojem računalnih programa za analizu istih.

Prvotne su metode upotrebljavale grubu usporedbu (poravnanje) i nisu mogle odrediti sličnost između sekvencija koje sadržavaju dijelove koji se ne poklapaju, ili koje sadržavaju insercije i delecije. Također je bilo teško pronaći poravnanja sekvencija koje nisu dovoljno srodne ili su vrlo dugačke (38).

Prilikom traženja sličnosti u sekvencijama DNA ili proteina, standardno programiranje bi zbog količine podataka oduzimalo previše vremena, te ne bi dalo statistički značajne rezultate. Zbog toga se upotrebljava dinamičko programiranje. Dinamičko programiranje je računska metoda koja se upotrebljava za poravnanje dviju sekvencija aminokiselina ili parova nukleotida, te daje optimalno poravnanje dviju ili više sekvencija. Poravnanje se vrši uspoređivanjem svakog para sličnih slova u sekvenciji i dodjeljivanjem bodova svakoj takvoj usporedbi. Pritom se u obzir uzimaju i identičnosti, sličnosti, te praznine i različitosti, tako da se svakoj od tih usporedbi dodjeljuje različit broj bodova. Nakon toga se bodovi zbrajaju, te se u obzir uzima ono poravnanje koje je dobilo statistički najveći broj bodova. Ova je metoda dovoljno pouzdana da biolozima može dati korisne informacije o odnosima između sekvencija. Te su informacije važne za stvaranje funkcionalnih, strukturalnih i evolucijskih predviđanja vezanih uz molekule DNA ili proteina (39).

Najčešće upotrebljavane inačice dinamičkog programiranja su globalno i lokalno poravnanje. Do obje inačice se dolazi jednostavnim promjenama u osnovnim dinamički programiranim algoritmima. Programi za globalno poravnanje vrše poravnanja duž cijele duljine sekvencija i temelje se na algoritmu Needleman-Wunsch (40). Prilikom pronalaženja najboljeg poravnanja, algoritam pokušava obuhvatiti što veće dijelove sekvencija, počevši od krajnje lijevog kraja do krajnje desnog kraja sekvencija. Ustanovljeno je, međutim, da su biološki najvažniji dijelovi u sekvencijama DNA i proteina oni koji se nalaze u specifičnim dijelovima, te koji se gotovo potpuno poklapaju. Također je ustanovljeno da ostali dijelovi (koji ne pokazuju dovoljnu sličnost) nisu od važnosti, te da su insercije i delecije vrlo česta evolucijska promjena. Zbog toga je razvijen novi algoritam, algoritam Smith-Waterman (41) (inačica algoritma Needleman-Wunsch). Programi za lokalno poravnanje upotrebljavaju algoritam Smith-Waterman i uspoređuju samo male, lokalne dijelove obiju sekvencija prilikom pronalaženja lokalnih poravnanja između dviju sekvencija. Rezultat takvog poravnanja su lokalne sličnosti među sekvencijama. U praksi se češće upotrebljava algoritam Smith-Waterman jer stvarne biološke sekvencije često ne pokazuju međusobnu sličnost uzduž cjelovitih dužina već samo na kratkim, lokaliziranim, dijelovima.

Preciznost poravnavanja, koju nam pružaju algoritmi dinamičkog programiranja, u suprotnosti je s brzinom rada. Veličina i opseg baza podataka ne dopušta da se, u današnje vrijeme, upotrebljavaju spore metode. Zbog toga su razvijene, alternativne, brže metode koje upotrebljavaju drugačije algoritme. Takvi novi programi se zasnivaju na heurističkim algoritmima. Pritom je program, dobivši na brzini, izgubio na preciznosti. Najčešće upotrebljavan program jest računalni program BLAST (eng. "Basic Local Alignment Search Tool"). Program BLAST je prosječno pedeset puta brži od algoritma dinamičkog programiranja. Algoritam koji se upotrebljava u programu BLAST se temelji na statističkim metodama koje su razvili Altschul i suradnici (42). Nepoznatu sekvenciju upit (eng. "query") program BLAST uspoređuje, tj. poravnava, sa svakom sekvencijom koja se nalazi u bazi podataka sekvencija DNA ili proteina. Tijekom usporedbe program BLAST traži dijelove sekvencija, tzv. riječi, koji pokazuju najveći stupanj sličnosti. Svaka riječ predstavlja minimalni dio DNA ili proteina (11 za DNA, 3 za proteine), koji je potreban da se dobije broj riječi koji je dovoljno velik da se prihvati kao značajan, ali ne toliko velik da bi se previdjeli neki kratki, ali značajni dijelovi. Kao rezultat, iz baze se podataka izdvajaju one sekvencije koje sadržavaju dio DNA ili proteina dovoljno sličan nepoznatoj sekvenciji da bi se smatrao statistički značajnim. Tijekom poravnavanja, za bodovanje sličnosti između sekvencija program BLAST upotrebljava jednu od matrica za bodovanje (npr. BLOSUM62 ili PAM250). Matrica PAM250 se temelji na evolucijskom modelu promjene proteina, dok se matrica BLOSSUM62 upotrebljava za identifikaciju članova iste porodice (38).

Algoritam obavlja analizu u nekoliko osnovnih stupnjeva. Program prvo uklanja dijelove niske složenosti, koje nisu značajne, i stvara listu riječi. Nakon toga se za svaku riječ određuje bodovna vrijednost (ovisno o odabranoj matrici), te se određuje T vrijednost (eng. "T Value"). T vrijednost predstavlja minimalnu bodovnu vrijednost potrebnu da bi se ustanovila sličnost. Nakon toga, algoritam traži u sekvencijama iste dužine, tj. dijelove sastavljene od mnogo riječi, koji (poravnani bez praznina) daju najbolji rezultat. Takva se lokalna poravnanja nazivaju parovi s najvećim brojem bodova (eng. "High-Scoring Pairs"; HSP). U slijedećem se stupnju svaki HSP, koji prelazi rezultat S (empirijski određena statistička vrijednost), proširuje, te se dopušta uvođenje praznina (ne više od jednog proširenja na pedeset sekvencija iz baze podataka). Svako uvođenje praznina se boduje negativno (eng. "penalty score"). Konačni cilj je pronalaženje poravnanja, bilo ono sa ili bez praznina, s maksimalnim rezultatom S (eng. "score"). I na kraju, u zadnjem stupnju algoritam procjenjuje statistički značaj dobivenog maksimalnog rezultata. Ako ocjeni da je značajan, sekvencija iz baze podataka biti će izabrana i prikazana na listi koju se dobije kao izvješće (39). Danas postoje mnoge verzije računalnog programa BLAST, no najčešće

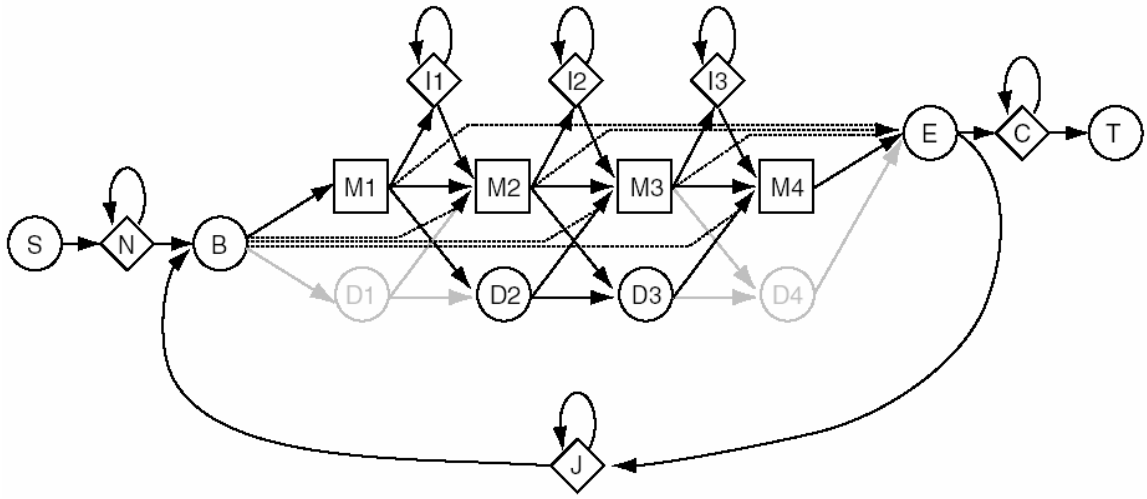
se upotrebljavaju "blastn" (specijaliziran za poravnanje sekvencija DNA), "blastp" (specijaliziran za poravnanje sekvencija proteina) i "blast2" (specijaliziran za neposrednu usporedba dviju sekvencija).

2.2.6. Bioinformatički programski paket HMMER

Skriveni Markovljev model (HMM) je statistički model gdje se za sustav koji se modelira pretpostavlja da je Markovljev proces s nepoznatim parametrima. Pri izradi modela HMM potrebno je odrediti skrivene parametre. Na taj se način izrađeni model i njegovi parametri mogu upotrijebiti za provođenje daljnjih analiza, primjerice za prepoznavanje uzoraka. U uobičajenom modelu HMM stanje je neposredno vidljivo promatraču, te su zbog toga vjerojatnosti prelaska iz jednog stanja u drugo jedini parametri koji se promatraju. U skrivenom modelu HMM stanje nije izričito vidljivo, ali su varijable na koje stanje utječe vidljive. Svako stanje ima distribuciju vjerojatnosti preko mogućeg izlaznog rezultata (engl. "output"). Zbog toga sekvencije reprezentativnih rezultata, koji su generirani modelom HMM, daju informacije o prijelazu stanja. Modeli HMM su posebno poznati u prepoznavanju vremenski promjenjivih uzoraka (engl. "temporal pattern recognition") kao što je govor, rukopis i prepoznavanje gesti. Posebnu primjenu imaju u bioinformatici u analizi linearnih sekvencija aminokiselina.

Krogh i suradnici (43) su uveli arhitekturu modela HMM koja je bila prilagođena za prikazivanje profila višestrukih poravnanja (engl. "multiple alignments"). Za svaki istoznačni stupac (engl. "consensus") višestrukog poravnanja odgovarajuće stanje (engl. "match") opisuje distribuciju aminokiselina koje su dopuštene u stupcu. Stanja umetanja (engl. "insert") i brisanja (engl. "delete") u svakom stupcu omogućavaju uvođenje jedne ili više aminokiseline između tog i slijedećeg stupca ili brisanje određene aminokiseline. Postoji i modificirani profil–HMM čija se arhitektura naziva Plan 7 (Slika 10). Ona je nešto složenija od originalne arhitekture profila–HMM koju su prvi upotrijebili Krogh i suradnici (43). Suština profila–HMM je linearni niz pogodaka (engl. "match states" – M), po jedan za istoznačni (engl. "consensus") stupac u višestrukom poravnanju. Svako stanje M se poravnava s jednom aminokiselinom, s vjerojatnošću koja je određena učestalošću da su aminokiseline bile prisutne u određenom stupcu višestrukog poravnanja. Svako stanje pogotka, dakle, sadržava vektor od 20 vjerojatnosti, za 20 aminokiselina. Profil–HMM može modelirati poravnanja koja sadržavaju praznine, kao što su umetanja (engl. "insert" – I) i delecije (engl. "delete" – D), što računalnom programu omogućuje da opiše potpuno konzerviranu domenu, a ne samo mali motiv bez praznina. Programski paket

HMMER poznaje skupinu od tri stanja (M/D/I) na istom položaju u poravnanju istoznačnog slijeda (engl. "consensus sequence"), takozvani čvor (engl. "node"). Ova stanja su međusobno povezana sa strelicama koje se nazivaju vjerojatnosti prelaska stanja (engl. "state transition probabbilities"). Model započinje i završava s ne-emitirajućim stanjima B (engl. "beginning" – početak) i E (engl. "end" – kraj). Ovaj osnovni odjeljak modela Plan 7, koji se sastoji od čvorova M, D, i I omeđenim stanjima B i E, je u osnovi model Krogh/Haussler (43).



Slika 10. Arhitektura modela Plan 7: Kvadrati prikazuju stanja pogotka (engl. "match states") koja modeliraju istoznačne pozicije u poravnanju. Rombovi prikazuju pozicije umetanja (engl. "insert states") koje modeliraju umetanja prema istoznačnom poravnanju (engl. "consensus sequence") i posebna nasumična stanja koja emitiraju sekvencije. Krugovi prikazuju stanja nedostatka (engl. "delete states") koja modeliraju delecije u odnosu na istoznačno poravnanje, te posebna stanja koja označuju početak ili kraj. Strelice pokazuju prijelaze stanja (engl. "state transitions").

Za razliku od originalnog modela Krogh/Haussler, model Plan 7 nema prelaska D→I i I→D. Ostala stanja jesu: S – početno stanje (engl. "Start state"), N – amino – terminalno stanje neporavnate sekvence (engl. N – "terminal unaligned sequence state"), C – karboksi – terminalno stanje neporavnate sekvencije (engl. C – "terminal unaligned sequence state") i J – povezani segment neporavnatog stanja sekvencija. Ova se stanja nazivaju posebna stanja (engl. "special states"), te u kombinaciji s posebnim vjerojatnostima ulaska B i izlaska E kontroliraju svojstva modela koja nisu ovisna o algoritmu.

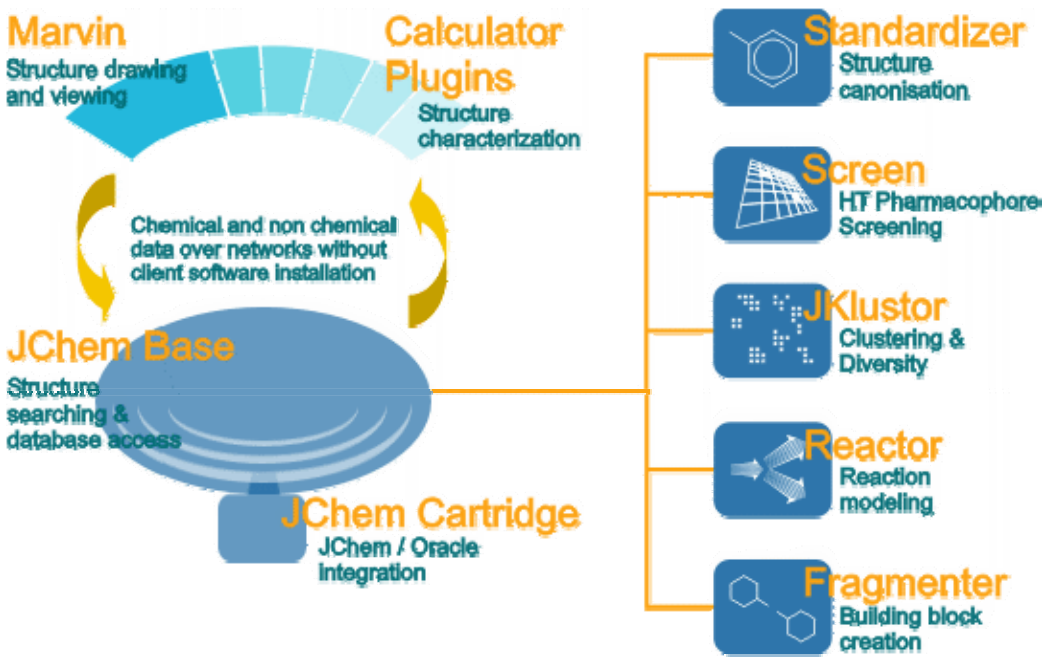
Budući da je način poravnanja u arhitekturi modela Plan 7 određen konfiguracijom modela HMM, a ne algoritmom koji pretražuje, stil poravnanja se odabire upotrebom opcija kada se gradi model, a ne kada se pretražuje sekvencija.

Postoji nekoliko dostupnih računalnih paketa u koje su implementirani profili–HMM a temeljeni na arhitekturi modela Plan 7. Jedan od takvih je programski paket HMMER. Postoji devet programa koji su sastavni dio programskog paketa HMMER: **hmmalign** - poravnava sekvencije prema postojećem modelu, **hmmbuild** - gradi model iz višestrukog poravnanja sekvencija, **hmmcalibrate** - uzima HMM i empirijski određuje parametre koji se koriste da bi se pretrage napravile osjetljivima proračunavajući očekivane vrijednosti (engl. "e – values"), **hmmconvert** - prevodi datoteku modela u različite oblike zapisa, **hmmemit** - prikazuje sekvencije s određenom vjerojatnošću iz profila HMM, **hmmfetch** - izdvaja jedan model iz baze podataka HMM, **hmmindex** - indeksira bazu podataka HMM, **hmmpfam** - pretražuje bazu sekvencija potencijalnim pogocima prema modelu kojim pretražuje i **hmmsearch** - pretražuje sekvenciju potencijalnim pogocima prema modelu kojim pretražuje. Postoje dvije velike zbirke dostupnih profila–HMM: baza podataka Pfam (35) te baza podataka PROSITE (44). U bazi podataka Pfam upotrijebljen je programski paket HMMER za izgradnju modela obitelji Pfam. Za svaku obitelj Pfam dostupni su određeni podaci:

- **Osnovno poravnanje** (engl. "seed alignment"); ručno pregledano višestruko poravnanje koje predstavlja danu obitelj. Model HMM, koji je proizašao iz osnovnog poravnanja, može biti upotrebljen za pronalaženje novih domena kao članova obitelji. Ovakav je model moguće upotrijebiti za pretraživanje skupa značajnih sekvencija. Jedan model HMM je u obliku **ls** (globalno), drugi je u obliku **fs** (lokalno),
- **Potpuno poravnanje** (engl. "full alignment"); automatsko poravnanje svih domena prisutnih u bazi podataka za čije su pretraživanje upotrebljena dva modela HMM (**ls** i **fs**),
- Anotacija sadržava kratak opis domene, poveznice prema drugim bazama podatka i neke specifične podatke za bazu podataka Pfam, koji bilježe način na koji je opisana obitelj konstruirana,
- Distribucija vrsta iz kojih su poravnate domene potekle,
- Arhitektura proteinskih domena,
- Poznate proteinske strukture, i
- Poveznice s drugim bazama podataka.

2.2.7. Kemoinformatički program ChemAxon

Kemoinformatički program ChemAxon jest specijalizirani program koji pruža mogućnost programiranja računalnim jezikom Java i primjenu programa za uređivanje i prikazivanje kemijskih struktura, makromolekula i kemijskih reakcija. Najznačajniji korisnici ovog programa jesu farmaceutske, agrokemijske i biotehnološke kompanije, kao i akademske istraživačke grupe koje žele integrirati kemoinformatiku u svoja istraživanja (45).

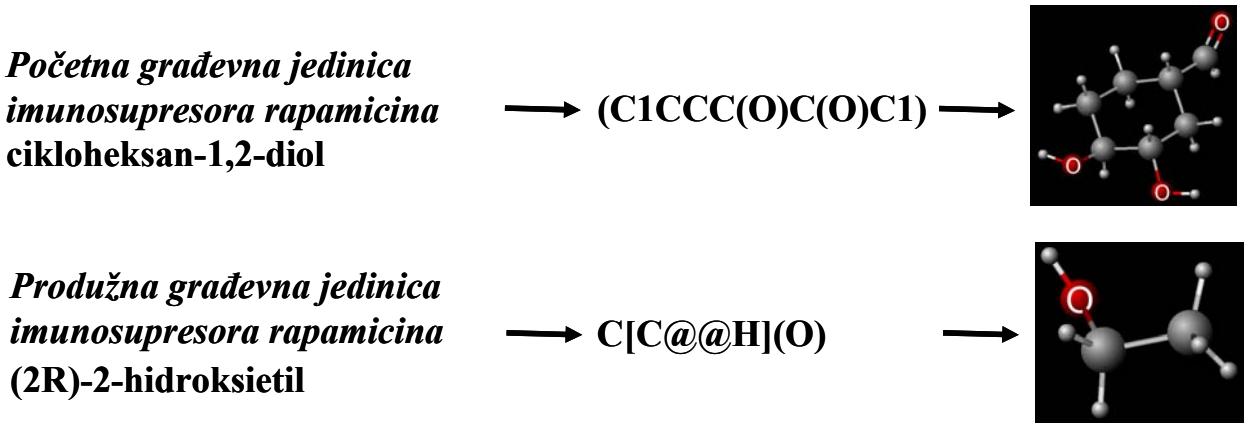


Slika 11. Grafički prikaz različitih alata za analiza kemijskih struktuta koje pruža kemoinformatički program ChemAxon.

Kemoinformatički alati programa ChemAxon jesu: (1) alat Marvinovi appleti (engl. "Marvin Applets": "MarvinSketch", "MarvinView" i "MarvinSpace") za uređivanje i prikazivanje kemijskih struktura, makromolekula i kemijskih reakcija; (2) alat dodaci kalkulatou (engl. "Calculator Plugins") za pretraživanje kemijskih struktura i reakcija različitih svojstava; (3) alat "JChem Base" za pretraživanje i rukovanje kemijskim bazama podataka; (4) alat za brzu aplikacija "Instant JChem" za osobno računalo radi lokalnog i udaljenog upravljanja, pretraživanja i predviđanja struktura u kemijskim bazama podataka; (5) modul "JChem" za

uključivanje podataka u bazu podataka Oracle; (6) alat za standardizaciju (engl. "Standardizer"), tj. za pretvorbu kemijskih struktura iz različitih izvora; (7) reaktor za preobrazbu kemijskih struktura; (8) alat "Screen" za pretraživanja farmakofora i pretraživanja temeljenih na kemijskim strukturama; (9) alat "JKlustor" za analize raznolikosti raspoređivanjem u pojedinačne razrede, i (10) alat "Fragmenter" za raščlanjivanje kemijskih struktura na dijelova i tzv. R-grupe (Slika 11) (45).

Kemoinformatički program ChemAxon pruža mogućnost da se kemijska struktura molekule, izražene u obliku izomeričkih zapisa SMILES (37), iscrta u dvodimenzionalnom (2D) ili trodimenzionalnom (3D) obliku (Slika 12) (45).



Slika 12. Grafički prikaz početne i produžne građevne jedinice poliketidnog imunosupresora rapamicina iscrtanih pomoću kemoinformatičkog programa ChemAxon.

2.3. Programski jezici i informatički alati

Posljednjih nekoliko desetljeća informatička je tehnologija doživjela nagli razvoj te je današnje moderno djelovanje čovjeka gotovo nemoguće zamisliti bez pomoći njenih brojnih programskih jezika i informatičkih alata. Programski su jezici i informatički alati toliko brojni da su u ovom poglavlju opisani smo oni upotrijebljeni u izradi ovog diplomskoga rada.

2.3.1. Projekt BioSQL

Projekt BioSQL (46) ima za zadaću izgraditi shemu baze podataka namijenjenu pohrani složenih bioloških podataka. Dio je standarda OBDA (46) koju je razvio Open Bioinformatics Foundation. Nastao je kao suradnja između projekata BioPerl, BioPython, BioJava i BioRuby s ciljem da bi se na standardan način mogli pohranjivati i preuzimati složeni biološki podaci. Tako npr. Perl upotrebljava Perl DBI, a Java upotrebljava JDBC biblioteku (engl. "Java Database Connectivity") za pristup istoj bazi podataka izgrađenoj sa shemom BioSQL.

U shemu BioSQL se mogu pohranjivati biološke sekvencije i anotacije iz različitih izvora kao što su GeneBank ili Swissprot. Shema BioSQL trenutno podržava tri različita sustava za upravljanje bazama podataka DMBS (engl. "Data Base Management System"): MySQL, PostgreSQL i Oracle. Uz male izmjene može podržavati i ostale sustave za upravljanje bazama podataka. Shema BioSQL (verzija 1.5; 46) sadržava 28 tablica. U te se tablice mogu pohranjivati gotovo svi biološki podaci od sekvencija, anotacija, filogenetskih stabla, bibliografskih podataka, komentara, poveznica na različite baze podataka i druge. Najnovija verzija sheme se može preuzeti sa repozitorija CVS (engl. "Concurrent Versioning System"), to jest s adrese dostupne na ovoj Web stranici (46).

2.3.2. Sustav za upravljanje bazama podataka PostgreSQL

Računalni program PostgreSQL (47) je objektno-relacijski sustav za upravljanje bazama podataka. Taj se program aktivno razvija već 15 godina, i izdaje se uz jednu od najslobodnijih licenci BSD (48). Računalni program PostgreSQL je otvorenoga kôda i može se slobodno dograđivati, mijenjati pa čak i prodavati što dopušta licenca BSD. Program je dostupan za više platformi kao što su: Linux, Unix (UNIX AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64) i Windows. Usklađen je sa standardima ANSI SQL92 (engl. "American National Standards Institute"; 49) i SQL99. Mogućnosti, kao što su transakcije, nasljeđivanje tablica (objektne osobine), automatski postupci za pozivanje procedura (engl. "trigger"), funkcije i pohranjene procedure (engl. "stored procedures"), svrstavaju računalni program PostgreSQL u naprednije sustave. On predstavlja izvrsnu zamjenu za skupe komercijalne sustave za upravljanje bazama podataka poput Oracle ili Microsoft SQL Server. Računalni program PostgreSQL podržava velike količine podataka, što vidljivo iz Tablice 1.

Tablica 1. Mogućnosti računalnog programa PostgreSQL s obzirom na količinu podataka.

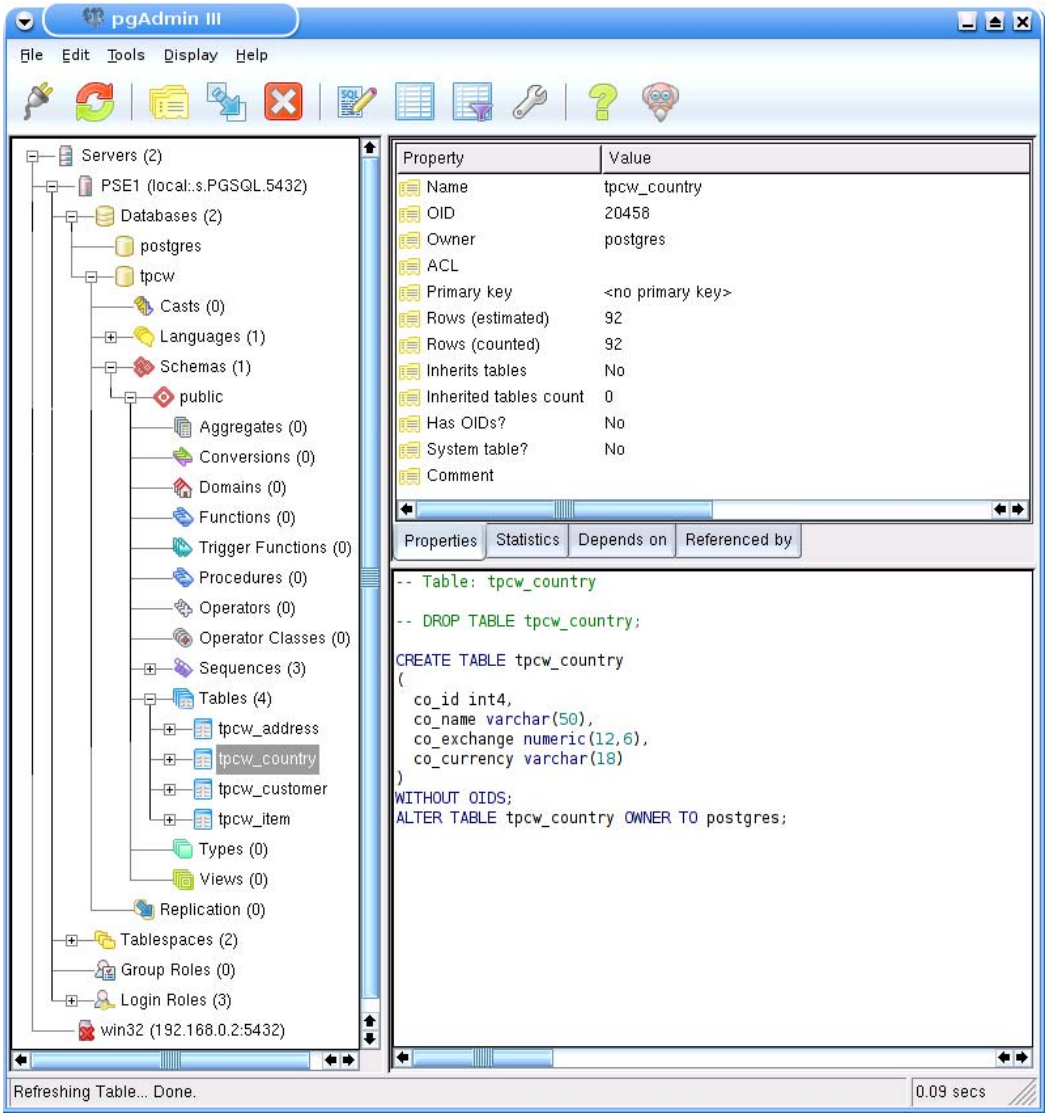
Ograničenje	Veličina
Maksimalna veličina baze	Neograničeno
Maksimalna veličina tablice	32 TB
Maksimalna veličina redova	1.6 TB
Maksimalna veličina polja	1 GB
Maksimalni broj redova po tablici	Neograničeno
Maksimalni broj kolona po tablici	250 - 1600 ovisno o tipu kolone
Maksimalni broj registara po tablici	Neograničeno

Iako se za administriranje i rad s programom PostgreSQL može upotrebljavati računalni jezik SQL, u tu se svrhu najčešće upotrebljava program pgAdmin (verzija 3.; 50). Program pgAdmin ima grafičko sučelje (Slika 13) i može se pokretati na skoro svim platformama kao i sam program PostgreSQL. Program pgAdmin pruža mogućnost da se na vidljiv način jednostavno izvršavaju radnje kao što su to: kreiranje baza i tablica, pregledavanje podataka iz baze, mijenjanje postavke baze, administracija korisnika baze, pisanje i uređivanje kôda SQL itd. Program pgAdmin je također otvorenoga kôda.

2.3.3. Računalni jezik SQL

Računalni jezik SQL (engl. "Structured Query Language") je jezik za postavljanje upita. To je standardni jezik za izradu, pronalaženje, izmjenu i brisanje podataka iz relacijskih baza podataka. Računalni jezik SQL je tijekom vremena standardiziran standardima ANSI SQL 89 (SQL1), ANSI SQL 92 (SQL2) i ANSI SQL 99 (SQL3) (49). Gotovo svaki relacijski sustav DBMS podržava računalni jezik SQL i vrlo ga često proširuje svojim funkcijama i naredbama. Iako je računalni jezik SQL namijenjen za krajnje korisnike, često se upotrebljava od strane programa koji krajnjim korisnicima pruža jednostavnije korisničko sučelje. Standardnim naredbama računalnog jezika SQL, kao što su to: izaberi (engl. "select"), uloži (engl. "insert"), ažuriraj (engl. "update"), izbriši (engl. "delete"), kreiraj (engl. "create") i odbaci (engl. "drop"), te

kombinacije istih, mogu se obaviti gotovo sve akcije koje su potrebne za upravljanje podacima u bazi podataka.



Slika 13. Prikaz grafičkog sučelja programa pgAdmin.

Kombinacije naredbi, naprednije jezične konstrukcije, pružaju mogućnost razvrstavanja podataka po nekom kriteriju, aritmetičke operacije nad podacima, definiranja pogleda (engl. "view") izvedenih iz postojećih relacijskih tablica, kontrolu transakcije (naredbe započni, izvrši i povratak na prethodno stanje engl. "begin", "commit" i "rollback"), definiranje prava pristupa (naredbe dozvoli i poništi engl. "grant" i "revoke") i drugo. Računalni jezik SQL pruža

mogućnost za izvršavanje i najkompliciranijih zadataka. Sve su naredba opisnog tipa preuzete iz govornog engleskog jezika.

2.3.4. Računalni jezik Java

Računalni jezik Java je opće namjenski, objektno orijentirani, programski jezik. Razvija ga tvrtka Sun Microsystems (51) od 1995. godine. Računalni jezik Java je srodnan računalnom jeziku C++, koji je neposredan potomak računalnog jezika C. Veći dio osobina računalnog jezika Java naslijeđen je od ova dva jezika. Za razliku od nekih drugih programskih jezika, izvorni se kôd računalnog jezika Java prevodi u specijalni "bytecode" i tek tokom pokretanja programa taj se kôd prevodi u strojni kôd. Specijalni kôd "bytecode" je optimalan skup instrukcija, koje u trenutku pokretanja programa prevodi Javina virtualna "mašina" JVM (engl. "Java Virtual Machine"). Programi napisani računalnim jezikom Java se pokreću samo pomoću virtualne mašine JVM. Zbog toga je moguće da se isti kôd izvršava na različitim operativnim sustavima kao npr. Windows, Linux, Unix, Mac OS, mobilni uređaji, odnosno na svakoj platformi za koju postoji virtualna mašina JVM. To je moguće bez potrebe za ponovnim prevođenjem programa.

Za pokretanje programa pisanih računalnim jezikom Java, potrebno je na korisničkom računalu instalirati programski paket JRE (engl. "Java Runtime Editon"). To je programski paket koji pokreće programe Java. Takav način pokretanja programa ima više prednosti:

- **prenosivost**: programi pisani na jednoj platformi mogu se izvršavati na drugim platformama što omogućuje virtualna mašina JVM. Većinu razlika raznih platformi svodi na jedinstveno programsko sučelje API (engl. "aplication programming interface"),
- **sigurnost**: programi se izvršavaju unutar unaprijed određenog djela radne memorije dodijeljenog od strane virtualne mašine JVM. Na taj način program ne može vidjeti memorijski prostor koji mu nije dostupan. Isto tako, virtualna mašina JVM provjerava „bytecode“ prije izvršavanja. Osim toga, virtualna mašina JVM osigurava sigurno izvršavanje kôda koji se pokreče na udaljenim računalima,
- **jednostavnost**: kao što se virtualna mašina JVM brine da omogući potrebnu memoriju za izvršavanje programa, tako se brine i da oslobodi memoriju koju su zauzeli objekti Java, vanjski resursi kao skupovi znakova i simbola, boje ili slike. To olakšava samo programiranje jer programer o tome ne mora brinuti. O tome brine virtualna mašina JVM

tijekom izvršavanja programa, i

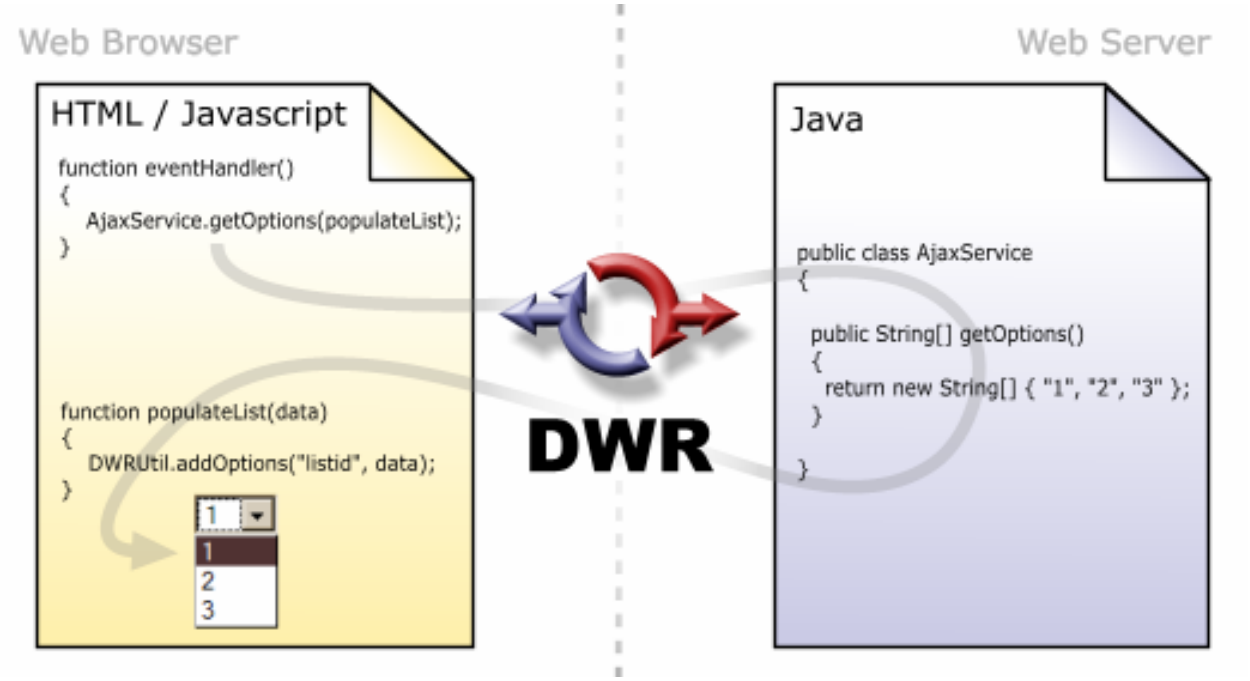
- **optimizacija kôda**: virtualna mašina JVM upotrebljava različite tehnike optimizacije kôda prije izvršavanja programa.

Na Web stranicama tvrtke Sun (51) može se besplatno preuzeti programski paket JRE. Instalirani programski paket JRE sadržava virtualnu mašinu JVM i skup manjih programa da bi korisnik mogao pokretati programe pisane u računalnom jeziku Java.

Računalni jezik Java sastoji se od tri temeljne tehnologije: (1) tehnologija Java ME (engl. "Micro Edition") namijenjenu mobilnim uređajima, (2) tehnologija Java SE (engl. "Standard Editor") namijenjenu za aplikacije klijenta i poslužitelja, te (3) tehnologija Java EE (engl. "Enterprise Edition") namijenjenu za prenosive poslužiteljske aplikacije koje učinkovito upotrebljavaju računalnim sustavima s više procesora. U ovom je radu za izradu Web stranica upotrijebljena tehnologija Java "Standard Edition" 5.0 (51) i dodaci "Java Servleti" i "Java Server Pages" (JSP). Kao "servlet kontejner" je upotrijebljen "Apache Tomcat" (52). Od ostalih tehnologija, upotrijebljena je tehnologija "Hibernate" (53).

Tehnologija Hibernate je vrlo moćna objektno/relacijska tehnologija koja povezuje objekte Java s relacijskim tablicama baze podataka. Ona proširuje računalni jezik SQL u računalni jezik HQL (engl. "hibernate query language") koji ima više objektnih svojstava. Samo je povezivanje izvedeno na taj način da se izradi konfiguracijska datoteka XML koja određuje koje "klase" računalnog jezika Java treba prenijeti u koje relacijske tablice. Na koji način i koje varijable istih klasa odgovaraju kojim poljima u bazi podataka, te kojega su tipa. Tehnologija Hibernate podržava sve osnovne "tipove" objektno orijentiranog računalnog jezika Java, kao i skupove, nizove i tipove koje određuju korisnici. Svi tipovi se automatski pretvaraju u tipove SQL, i obrnuto. Velika prednost programiranja s tehnologijom Hibernate je, nasuprot čistom pristupu biblioteke JDBC, što tehnologija Hibernate izvana radi na transparentan način prema određenoj bazi podataka, odnosno, na gotovo jednak način se upotrebljava sa svakom bazom podataka. Da bi tehnologija Hibernate mogla koristiti specifičnosti određene baze podataka, upotrebljava se specifičan dijalekt Hibernate, koji je zapravo jedna od klasa računalnog jezika Java. Za računalni program PostgreSQL potrebno je, u konfiguracijskoj datoteci, odrediti da upotrebljava dijalekt PostgreSQL. Program se može značajno ubrzati upotrebljavajući primarni i sekundarni "cache" Hibernate.

Biblioteka DWR (engl. "Direct Web Remoting"; 54) je biblioteka otvorenoga kôda koja pruža mogućnost komunikacije između računalnog jezika JavaScript, iz Internet pretraživača, s računalnim jezikom Java na poslužitelju (Slika 14). Za to se upotrebljava popularna metoda AJAX (engl. "Asynchronous JavaScript and XML"). Princip rada biblioteke DWR se temelji na udaljenom pozivanju metoda objekata Java na poslužitelju iz objekata JavaScript u Web pregledniku. Biblioteka DWR dinamički kreira klase JavaScript koje odgovaraju klasama Java na poslužitelju.



Slika 14. Grafički prikaz rada biblioteke DWR. Na poslužitelju su objekti Java vidljivi iz objekata JavaScript u Web pregledniku. Princip rada je sličan poznatoj Java tehnologiji RMI (engl. "Remote Method Invocation").

Na samom se početku mora odrediti koji se objekti i koje metode mogu pozivati iz Web preglednika. Određivanje se vrši u specifičnoj konfiguracijskoj datoteci XML, "dwr.xml", na poslužitelju i postavlja se u standardni direktorij WEB-INF. U standardnoj servlet konfiguracijskoj datoteci "web.xml" određuje se servlet biblioteke DWR koji će rukovati svim pozivima iz Web preglednika. Servlet biblioteke DWR se poziva i za određivanje klasa JavaScript, kao i za razmjenu podataka s poslužiteljem, odnosno udaljenim pozivanjem metoda.

Svi pozivi prema poslužitelju su asinkroni što i slijedi iz definicije metode AJAX. Nasuprot tome, oblik podataka koji se razmjenjuju između Web preglednika i poslužitelja nije oblik zapisa XML već oblik zapisa JSON (engl. "JavaScript Object Notation"). Oblik zapisa JSON se upotrebljava jer ima nekoliko prednosti naspram oblika zapisa XML. Oblik zapisa JSON je puno jednostavniji od oblika zapisa XML. Podaci spremljeni u obliku zapisa JSON su manji nego isti spremljeni u obliku zapisa XML. Budući da je oblik zapisa JSON prirodni oblik računalnog jezika JavaScript, nije ga potrebno posebno raščlanjivati (engl. "parse"). Prevodilac računalnog jezika JavaScript u Web pregledniku ga odmah može izvršavati čime se dobije na brzini izvođenja.

2.3.5. Računalni jezik JavaScript

Računalni jezik JavaScript (55) je skriptni programski jezik, koji se izvršava u Web pregledniku (na strani korisnika). Razvijen je da bude sličan računalnom jeziku Java, da bi bio lakši za upotrebu. Međutim, računalni jezik JavaScript nije objektno orijentirani jezik kao što je to računalni jezik Java. On se temelji na prototipu i tu prestaje svaka veza sa programskim jezikom Java. Razvila ga je tvrtka Netscape (56) 1996. godine i uključen je u Web preglednik "Netscape Navigator" 2.0 (NN). Računalni jezik JavaScript je implementacija standarda "ECMAScript" i primarno se upotrebljava unutar Web stranica na strani korisnika u Web pregledniku. Taj jezik omogućuje rukovanje kôdom HTML, modelom DOM (engl. "Dokument Object Model") i kôdom CSS (engl. "Cascading Style Sheets"), što se tada zajedničkim imenom naziva DHTML (engl. "Dynamic HTML").

Prilikom poziva Web stranice, poslužitelj šalje korisniku (odnosno Web pregledniku) izvorni kôd računalnog jezika JavaScript zajedno sa kôdovima HTML i CSS. Pored toga, mogu se slati slike, "flash" sadržaji i Java appleti. Nakon što je Web preglednik obradio tekstove HTML i CSS, započinje prevoditi izvorni kôd računalnog jezika JavaScript i odmah ga izvršava. Računalni jezik JavaScript se najčešće upotrebljava za provjeravanje podataka iz Web formulara, stvaranje vizualnih efekata, a u novije vrijeme za neposrednu komunikaciju s programom na poslužitelju s kojega je preuzet cjelovit kôd računalnog jezika JavaScript. Računalni jezik JavaScript metodom AJAX (engl. "Asynchronous JavaScript and XML") omogućuje Web stranicama da asinkrono komuniciraju s programom na poslužitelju. To čini Web aplikaciju više interaktivnom i lakšom za upotrebu.

Jedan od problema s koji se danas programeri susreću jest činjenica da svaki Web preglednik ima različitu implementaciju računalnog jezika JavaScript. Program JavaScript koji dobro radi na jednom Web pregledniku najčešće ne radi na drugim Web preglednicima. Stoga je odlučeno da se generički računalni programski paket ***CompGen*** razvije za Web preglednik Firefox (57).

2.3.6. Računalni jezici HTML i CSS

Računalni jezik HTML (engl. "HyperText Markup Language") (58) je jezik za prezentaciju i kreiranje Web stranica. Njime se određuje izgled Web stranice, skupovi znakova i simbola, boje, slike i druga svojstva. Računalni jezik HTML se piše u obliku ključnih riječi svojstvenih za opisivanje sadržaja određene Web stranice, tj. u obliku takozvanih "tagova", i strukturom je sličan računalnom jeziku XML (engl. "Extensible Markup Language"). Kasnije je razvijen računalni jezik XHTML. On predstavlja proširenje računalnog jezika HTML koji je potpuno usklađen s računalnim jezikom XML. Dok se računalnim jezikom HTML definira sadržaj i strukturu dokumenta, stilski jezik CSS (engl. "Cascading Style Sheets") (58) opisuje izgled dokumenta za prezentaciju. Računalni jezik CSS dodatno opisuje dokument HTML, tako da se svaka klasa CSS veže za jedan ili više tagova HTML. Računalni jezik CSS se može pisati unutar dokumenta HTML odnosno unutar "<style>" taga ili može biti u zasebnoj datoteci koja se veže na stranicu HTML. Dokument HTML se može podijeliti u dva glavna dijela: zaglavlje (engl. "header") i tijelo (engl. "body"). U zaglavlju se definiraju elementi koji se ne prikazuju neposredno u Web pregledniku. Podaci potrebni za cjelovitu Web stranicu kao jesu:

- adresa datoteka CSS na poslužitelju, koje Web preglednih dohvaća i kombinira zajedno s računalnim jezikom HTML,
- adrese datoteka JavaScript na poslužitelju,
- naslov (engl. "title") Web stranice, i
- razni "meta" podaci određene Web stranice, kao npr. ključne riječi, opis stranice i dr.

U tijelo se postavljaju svi tagovi koju su zaslužni za prikaz stranice.

Računalni jezik HTML je na početku bio zamišljen kao jezik za prikaz dokumenata. Naknadno je popularnost Interneta dovela i do velike popularnosti Web stranica. Danas se s računalnim jezicima JavaScript i HTML izrađuju bogate i interaktivne Internet aplikacije, koje

sučeljem skoro nimalo ne zaostaju za standardnim aplikacijama grafičkog sučelja. Danas su Web stranice (za razliku od grafičkog sučelja) najčešći izbor za korisničko sučelje čemu pridonosi nekoliko faktora:

- laka izmjena aplikacije koja se odmah odražava u svih korisnika,
- svi popularni operativni sustavi imaju Web preglednik, pa korisnik ne mora dodatno instalirati program, i
- jednostavnost programiranja i bogatstvo opcija za definiranje izgleda Web stranica.

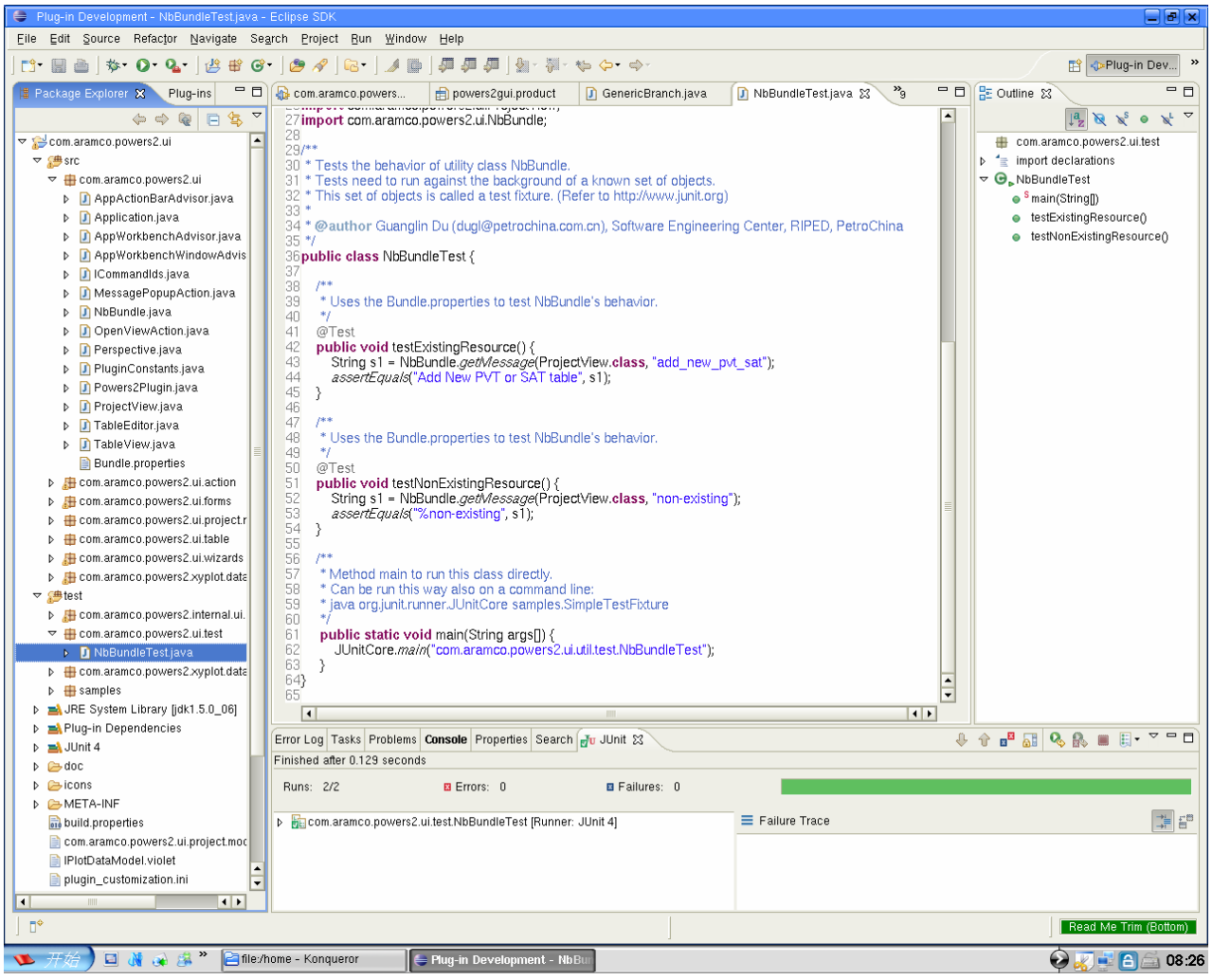
2.3.7. Pomoćni program Eclipse

Pomoćni program Eclipse je razvojno okruženje otvorenog kôda IDE (eng. "Integrated Develpment Environment") (59). Prvenstveno je namijenjen za razvijanje programa pisanih računalnim jezikom Java (Slika 15). I sam je pomoćni program Eclipse napisan računalnim jezikom Java. Pomoćni je program Eclipse u početku razvila tvrtka IBM. Tvrtka je IBM kasnije ustupila cjelovit kôd zajednici "*open source*". Taj je pomoćni program jedan od najboljih programa za uređivanje kôda računalnog jezika Java. Prednost programa Eclipse nad ostalim alatima sličnih mogućnosti jest što se može lako proširiti potrebnim dodacima (eng. "plugin") i time proširiti mogućnosti samog programa. Za ovaj su rad upotrijebljeni dodaci za razvoj Web aplikacija (eng. "Web tools platform, WTP") i dodatak za razvoj aplikacija baza podataka s tehnologijom Hibernate (eng. "Hibernate synchronizer"). Web aplikacija WTP je proširenje koje između ostalog pruža mogućnost izrade dinamičkih Web stranica tehnologijom JSP i dodatkom Java Servlet.

2.3.8. Programski paket Apache Tomcat

Programski paket Apache Tomcat je poslužitelj Web aplikacija otvorenoga kôda kojeg je razvila organizacija Apache Software Foundation (ASF) (60). Tomcat Web kontejner je službena implementacija JSR 154: Java Servlet 2.4 (61) i JSR 152: JavaServer Pages specifikacije 2.0 (62). Taj se programski paket može izvršavati na više platformi: Linux, Windows, Unix i drugim. Programski paket Apache Tomcat smatra se vrlo stabilnim Web poslužiteljem koji se upotrebljava za najzahtjevnije projekte. Sam je programski paket napisan računalnim jezikom Java. Pomoćni se program Eclipse može jednostavno integrirati s poslužiteljem Apache Tomcat. Naime, pomoćni program Eclipse pruža mogućnost pristupa funkcionalnostima preko

korisničkog sučelja (npr. pokretanje, zaustavljanje, mijenjanje postavki rada Web poslužitelja i dr.). Sva se konfiguracija Web poslužitelja Apache Tomcat obavlja u datotekama koje se nalaze u direktoriju "conf", unutar samog programa. Važna konfiguracijska datoteka je "server.xml", a upotrebljava se oblik zapisa XML. Određivanje korisnika (engl. "user") i pripadajućih zaporki (engl. "password"), kao i imena prava (engl. "roles"), definiraju se u konfiguracijkoj datoteci "tomcat-users.xml" koja je također u obliku zapisa XML.



Slika 15. Prikaz pomoćnog programa Eclipse IDE, najpopularnijeg pomoćnog programa za uređivanje kôda računalnog jezika Java.

3. Eksperimentalni dio

3.1. Materijal

3.1.1. Sklopovlje i operativni sustav

Ovaj je diplomski rad izrađen na računalima slijedećih sklopovlja: (**a**) osobno računalo s dva diska kapaciteta 120 GB, procesor AMD Athlon (tm) 3800XP+, radna memorija 512 DDR RAM, te grafika kartica ASUS 9600SE AGP 8X i (**b**) prijenosno računalo IBM, model T41, s diskom kapaciteta 80 GB, procesor Intel Mobile (tm) 2,6 GHz, radna memorija 768 RAM i grafika kartica 32MB ATI Radeon 7521.

Na čvrstom se disku osobnog računala nalazi operativni sustav Windows XP Professional. Na prijenosnom računalu je instaliran operativni sustav Linux Debian verzija 3.0. Ovaj diplomski rad u potpunosti je izrađen na operativnom sustavu Linux Debian. Upotrijebljen je grafički sustav X.Org, verzija 6.7.0. (63), operativnog sustava Linux Debian s grafičkim sučeljem Gnome, verzija 2.8.1. (64).

3.1.2. Informatička programska podrška

U ovom je radu upotrijebljeno nekoliko programskih jezika i tehnologija da bi generički programski paket ***CompGen*** mogao u cijelosti funkcionirati kao Web aplikacija. Za razvoj generičkog programskog paketa ***CompGen*** potrebno je na računalu instalirati programski paket JRE (51), verziju 1.5 ili višu, te Mozilla kompitablni Web preglednik Firefox (57). Glavni je dio izrađen na poslužitelju (engl. "Server side") kao Java Web aplikacija koju pokreće program Jakarta Tomcat (60). Kao operativni sustav se upotrebljava Linux Debian. Korisničko sučelje je temeljeno na Web sučelju upotrebom računalnog jezika DHTML (engl. "Dynamic HTML") (55) i tehnologije Java appleta (51). Za upravljanje bazama podataka upotrijebljen je računalni program PostgreSQL (47). Generički programski paket ***CompGen*** je izrađen kao klasična troslojna arhitektura, odnosno program je organiziran u tri djela:

1. Korisničko sučelje, Web preglednik (sloj za prezentaciju),
2. Poslovna logika na strani poslužitelja kao takozvani Java "servleti" (srednji sloj), i
3. Baza podataka, potrebni podaci smješteni su u bazi podataka (sloj podataka).

3.1.2.1. Shema baze podataka BioSQL

Projekt BioSQL (46) služi za izgradnju sheme baze podataka namijenjene pohrani složenih bioloških podataka. Shema BioSQL (verzija 1.5) trenutno podržava 28 tablica. U ovom je radu shema BioSQL poslužila za dizajniranje vlastite sheme baze bioloških i kemijskih podataka za pohranjivanje svih podataka od sekvencija DNA genskih nakupina poliketid sintaza i sintetaza neribosomalno sintetiziranih peptida, primarnih sekvencija aminokiselina proteina, modula, domena i njihovih poveznica, kemijskih struktura početnih i produžnih građevnih jedinica u obliku izomeričkih zapisa SMILES, dvodimenzionalnih (2D) i trodimenzionalnih (3D) struktura poliketidnih lanaca sve do konačnih cikličkih kemijskih struktura poliketidnih okosnica, tj. do aglikona.

3.1.2.2. Sustav za upravljanje bazama podataka PostgreSQL

Računalni program PostgreSQL (47) je objektno-relacijski sustav za upravljanje bazama podataka. U ovom je radu s tim programom podržano administriranje baze podataka, to jest pohrana podataka, brzo pronalaženje podataka i ažuriranje podataka.

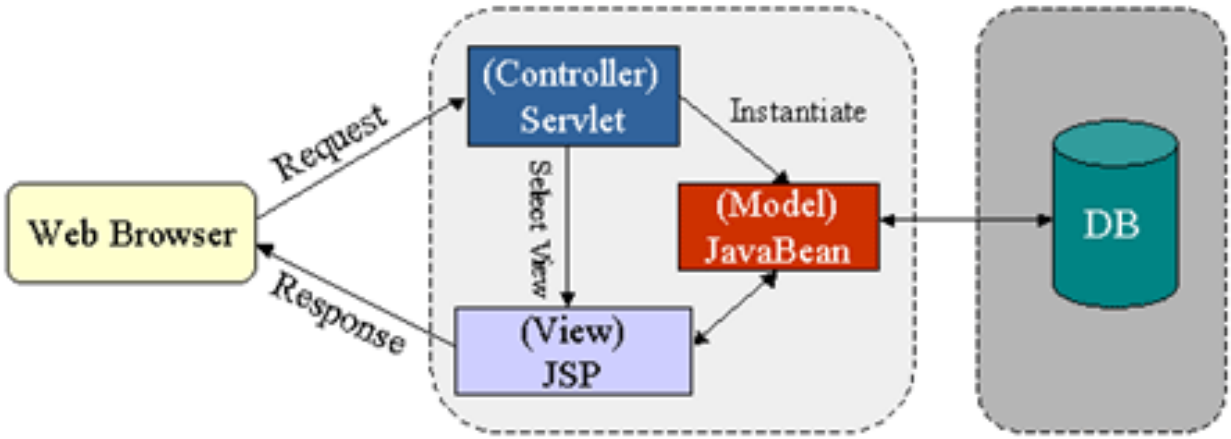
3.1.2.3. Računalni jezik SQL

Računalni jezik SQL (engl. "Structured Query Language", 49) je jezik za postavljanje upita bazi podataka. U ovom je radu taj računalni jezik upotrijebljen za komunikaciju između računalnog programa PostgreSQL i objektno orijentiranog računalnog jezika Java u kojem je razvijen generički programski paket ***CompGen***.

3.1.2.4. Računalni jezik Java

Računalni jezik Java je opće namjenski, objektno orijentirani, programski jezik (51). U ovom je radu objektno orijentirani računalni jezik Java upotrijebljen za razvoj velikog dijela generičkog programskog paketa ***CompGen***. Poslovna logika programskog paketa ***CompGen***, koji se uglavnom nalazi na strani poslužitelja (engl. "server side"), izrađen je Java Servletima. Java Servleti primaju HTTP zahtjeve klijenta (engl. "client"). Web poslužitelj Apache Tomcat je zadužen za pozivanje odgovarajućeg Java Servleta. Java Servleti primaju podatke poslane od Web preglednika, obrađuju ih, ažuriraju bazu podataka i obavljaju logiku programskog paketa

CompGen. Nakon toga, tok programa se preusmjerava na tehnologiju JSP. Tehnologija JSP je upotrijebljena za dinamičko generiranje dokumenta HTML, koji predstavlja korisničko sučelje (Slika 16).



Slika 16. Shematski prikaz obrade zahtjeva HTTP. Java Servleti zaprimaju zahtjev, obrađuju ga i obavljaju poslovnu logiku programskog paketa ***CompGen***. Nakon toga, tehnologija JSP dinamički generira dokument HTML.

3.1.2.5. Računalni jezik JavaScript

Računalni jezik HTML jest "statični" jezik. Naime, s računalnim se jezikom HTML može samo odrediti kako će Web stranica izgledati (vidi: *3.1.2.6. Računalni jezik HTML*). Da bi se Web stranicu učinilo više interaktivnom, upotrebljava se računalni jezik JavaScript (55). U ovom je radu računalni jezik JavaScript upotrijebljen za kreiranje interaktivnih Web stranica. Na primjer, Web stranica "createModule.jsp" prikazuje jedan protein te, pomoću programa "hmmpfam" programskog paketa HMMER koji se nalazi na poslužitelju, pronalazi sve domene u proteinskom produktu toga gena. Na tim se Web stranicama mogu kreirati moduli koji sadržavaju jednu ili više domena.

3.1.2.6. Računalni jezik HTML

Računalni jezik HTML (engl. "HyperText Markup Language") (58) je jezik za prezentaciju i kreiranje Web stranica. U ovom je radu taj jezik upotrijebljen za izradu korisničkog sučelja.

Korisničko sučelje se sastoji od Web formulara za unos podataka, tablični prikaz rezultata pretraživanja pomoću bioinformatičkog računalnog programa BLAST (42), grafički prikaz proteina, modula, domena i njihovih poveznica nakon analize s programskim paketom HMMER (43) i grafički prikaz puta biosinteze.

3.1.2.7. Pomoćni program Eclipse

Pomoćni program Eclipse (59) je razvojno okruženje otvorenog kôda IDE (eng. "Integrated Develpment Environment"). Prvenstveno je namijenjen za razvijanje programa pisanih računalnim jezikom Java. U ovom je radu pomoćni program Eclipse upotrijebljen za uređivanje kôda računalnih jezika Java i JavaScript, prevođenje kôda računalnog jezika Java, i uređivanje svih datoteka potrebnih za cjelokupan programski paket. Pomoćni program Eclipse pruža mogućnost za cjelokupno upravljanje projektom, od pisanja kôda do isporuke gotovog programa.

3.1.2.8. Programski paket Apache Tomcat

Programski paket Apache Tomcat je Web poslužitelj otvorenoga kôda kojeg je razvila organizacija Apache Software Foundation (ASF) (60). U ovom je radu taj programski paket upotrijebljen kao poslužitelj za isporučivanje dinamičkih Web stranica klijentu. Pri svakom zahtjevu klijenta za Web stranicom, programski paket Apache Tomcat osigurava da se izvrši potrebni kôd računalnog programa Java (Java Servleti i JSP) koji dinamički kreira stranicu HTML. Pored toga, programski paket Apache Tomcat osigurava autentičnost i autorizaciju korisnika generičkog programskog paketa ***CompGen***.

3.2. Metode

3.2.1. Strukturiranje scheme baze podataka po uzoru na BioSQL

Shema baze podataka poliketidnih/peptidnih genskih nakupina izrađena je po uzoru na Projekt BioSQL (46), pri čemu su iskorištena sva prijašnja iskustva biologa i informatičara koji su razvijali shemu baze podataka BioSQL. Bazom podataka se upravlja sustavom za upravljanje bazama podataka PostgreSQL (47), dok je za postavljanje upita upotrijebljen računalni jezik SQL (49).

U bazu podataka se spremaju svi podaci potrebni za kreiranje poliketidnih/peptidnih genskih nakupina. Temeljna tablica je tablica gena (engl. "gene"), u koju se spremaju podaci o sekvencijama DNA (broj nukleinskih baza) i proteina (broj aminokiselina) toga gena. Tablica "gene" je roditeljska tablica tablici modula (engl. "module") što znači da jedan gen ima niti jedan ili više modula. Tablica "module" je roditeljska tablica tablici domena (engl. "domain") što isto tako znači da jedan modul ima niti jednu ili više povezanih domena. Da se podaci ne bi udvostručavali, samo tablica "gene" sadržava same sekvencijje DNA i proteina i čuva ih u koloni sekvencija (engl. "sequence"). Nasuprot tome, tablice "module" i "domain" sadržavaju kolone "pos_start" i "pos_stop" koje ukazuju na njihovu poziciju, tj. koordinate, unutar proteina iz tablice "gene" (na koju su vezani pomoću "stranog ključa"), to jest gdje se točno nalazi pojedini modul i domena. Tablica genske nakupine (engl. "clusters") sadržava samo sekvencije DNA i predstavlja roditeljsku tablicu tablici "gene". Ona je roditeljska tablica zbog toga što su sekvencije DNA gena proizašle iz sekvencije DNA genske nakupine. Naime, bioinformatičkim su alatima ti geni otkriveni unutar genske nakupine i prevedeni u sekvenciju aminokiselina proteina. U tablici "clusters" spremljen je i podatak o dužini sekvencije DNA kao broj znakova nukleinskih baza koje genska nakupina sadržava.

U bazi podataka čuvaju se i literaturni podaci o sekvencijama DNA i proteina. Oni se spremaju u tablicu literatura (engl. "reference"). Tablica "reference" je povezana s tablicom "clusters" (koja čuva sekvencije DNA) s nula ili više naprema jedan. To znači da jedna sekvencija DNA ili proteina može imati niti jedan ili više literaturnih podataka. Od literaturnih podataka se čuva jedinstveni broj, "ID", koji potječe iz baze podataka PubMed koju održava centar NCBI. Na taj je način korisniku uvijek moguće pružiti sve dodatne informacije o sekvencijijama DNA i proteina. U programu su, kod prikaza sekvencija DNA i proteina, uspostavljene hiperveze (engl. "hyperlink") prema "online" bazi podataka PubMed. Neki su literaturni podaci o sekvencijama DNA i proteina unaprijed automatski izvučeni iz baze podataka PubMed i spremljeni u vlastitu bazu podataka. Ti se podaci odnose na naslov rada (engl. "title"), naziv časopisa (engl. "journal") i imena autora radova (engl. "author"), koji se svi čuvaju u tablici "reference". Razlog tome jest pretpostavka da korisnik želi u programskom paketu ***CompGen*** odmah vidjeti osnovne literaturne podatke o sekvencijama DNA i proteina, dok se za detalje upućuje na "online" bazu podataka PubMed.

3.2.2. Prikupljanje i strukturiranje bioloških podataka

Za prikupljanja primarnih sekvencija DNA i proteina, pristupa se već postojećim javno dostupnim bazama podataka na udaljenim poslužiteljima. Kao temeljni izvor sekvencija DNA i proteina upotrebljava se baze podataka GenBank (29). Prikupljanje je obavljeno pomoću bioinformatičkog programa BLAST (42). Unutar baze podataka GenBank, Web stranice bioinformatičkog programa BLAST (65) pružaju mogućnost pretraživanja nekoliko različitih baza podataka. Među njima se nalazi i baza podataka sekvencija proteina "nr" (66) koja se sastoji od svih nepreklapajućih sekvencija proteina iz baza podataka proteina: "GenBank CDS translations", "RefSeq Proteins", "PDB", "SwissProt", "PIR" i "PRF".

Generički programski paket ***CompGen*** pruža mogućnost da se za pretraživanje pomoću bioinformatičkog programa BLAST (42) podese slijedeći parametri:

- **Izbor matrice**: BLOSUM62, BLOSUM80 ili BLOSUM45, te PAM30 ili PAM60. Te matrice sadržavaju vrijednosti proporcionalne s vrijednošću vjerojatnosti da aminokiselina **i** mutira u aminokiselinu **j** za sve parove te aminokiseline. Matrice se izrađuju upotrebom velikih i raznolikih uzoraka potvrđenih poravnanja parova aminokiselina. Ako je uzorak dovoljno velik da bi bio statistički reprezentativan, dobivene matrice bi trebale biti stvaran pokazatelj vjerojatnosti pojave mutacije tijekom evolucije.

- **Izbor vrijednosti E**: Identičnost, sličnost ili različitost podataka izražena je vrijednošću E (engl. "E value") i bodovnim stanjem poravnanja S (engl. "Score"). Vrijednost E je mjera pouzdanosti bodovnog stanja poravnanja S. To je vjerojatnost da se pukom slučajnošću naiđe na neko drugo poravnanje koje ima sličnost veću od one bodovnog stanja poravnanja S.

- **Podešavanje pod-sekvencija**: Podešavanjem se pod-sekvencije (engl. "Set subsequence") određuje koliki dio sekvencije, kojom se pretražuje, uzima bioinformatički program BLAST i prosljeđuje ga dalje kao upit. Vrijednosti koje se unose mogu biti bilo koja 2 cijela prirodna broja koji se nalaze u rasponu od 1 do "l", gdje je "l" stvarna dužina sekvencije, tj. ukupan broj znakova koji je sačinjavaju.

- **Podešavanje cijene praznina**: "Cijena praznina" (engl. "Gap Costs") je dužina praznine u poravnanju koja se uvodi kao kompenzacija za delecije i insercije u sekvencijama DNA/proteina u odnosu na neku drugu sekvenciju DNA/proteina. Da bi se spriječilo uvođenje

previše praznina u poravnanju dvaju sekvencija, određuje se cijena praznina kao negativno bodovanje praznina u poravnanju.

- **Podešavanje pretraživanja CD**: Servis "pretraživanje CD" (engl. "CD-Search") je servis koji pruža centar NCBI. On se može upotrijebiti za prepoznavanje konzerviranih domena prisutnih u sekvenciji proteina koja se proslijeđuje kao upit. Većina poznatih konzerviranih domena nalazi se u bazi podataka "Conserved Domain Database" koju održava centar NCBI. Servis pretraživanje CD pruža mogućnost pretraživanja te baze podataka.

- **Ograničavanje upita**: Htio to netko ili ne, pretragom s postojećim "standardnim" parametrima bioinformatičkog programa BLAST automatski će se pretražiti sve sekvencije koje pripadaju mnogim različitim organizmima. Ograničenje upita (engl. "Limit by Entrez query") pruža mogućnost sužavanja kriterija pretrage određivanjem specifične grupe organizama čije će se sekvencije pretraživati.

Pretraživanje pomoću bioinformatičkog programa BLAST se može obaviti pomoću Web sučelja na Web stranicama centra NCBI, ili je moguće izraditi vlastiti računalni program. U ovom je radu razvijen vlastiti računalni program za pretraživanje baze podataka "nr" pomoću bioinformatičkog programa BLAST. Slijedi objašnjenje dijela kôda računalnog jezika Java kako programski pristupiti pretraživanje baze podataka "nr" pomoću bioinformatičkog programa BLAST. Cjeloviti se kôd generičkog računalnog paketa ***CompGen*** nalazi u Prilogu 3 na kompaktnom disku (CD-R). Na Web stranici "formBlastSearch.jsp" generičkog programskog paketa ***CompGen*** moguće je, za pretraživanje pomoću bioinformatičkog programa BLAST, kao upit upotrijebiti sekvenciju jedne ili više domene u obliku zapisa FastA (33). Ukoliko korisnik kao upit odabere sekvenciju samo jedne domene, programa BLAST obavi samo jedno pretraživanje i u izvješću prikaže sve moguće pogotke (engl. "hits"). Nasuprot tome, ukoliko se kao upit odabere *n* sekvencija u obliku zapisa FastA, programa BLAST obavi *n* pretraživanja i prikaže njihov presjek. Na taj se način dobiju precizniji rezultati.

Pretraživanje pomoću mrežnog servisa BLAST (65) programski je moguće isključivo upotrebom protokola HTTP (engl. "Hyper Text Transfer Protocol") (58). Centar NCBI pruža programerima mogućnost da izrade vlastiti program za pretraživanje njihove baze podataka pomoću servisa BLAST URLAPI (67). Standardno aplikacijsko sučelje API (eng. "Application Program Interface"), BLAST URLAPI, omogućava pristupanje na servis centra NCBI. To sučelje

upotrebljava zahtjeve HTTP prema Web poslužitelju centra NCBI. Za taj su dio programskog paketa ***CompGen*** upotrebljene Java biblioteke Jakarta Commons HttpClient (68) i Castor (69).

Računalni program, razvijen u ovom diplomskom radu, najprije mrežnom servisu NCBI BLAST šalje upit HTTTP GET (engl. "query string") na Web adresu http://www.ncbi.nlm.nih.gov/blast/Blast.cgi. Za svaki upit mora se navesti vrijednost *CMD* koja mrežnom servisu "govori" da upit šalje (*CMD=Put*) ili da se traže rezultati upita (*CMD=Get*). Upit se šalje prema sljedećim pravilima. Najprije se kreira "query string" koji mora sadržavati:

- CMD=*Put* [baza podataka odabire se pomoću parametra *DATABASE* (npr. *DATABASE*=nr)],
- *QUERY* [sekvencija se odabire parametrom *QUERY* (npr. *QUERY*=SSWWHVMGPP …)], i
- *PROGRAM* [baza podataka odabire se parametrom *PROGRAM* (npr. *PROGRAM*=blastp).

Ostale parametre nije nužno navesti. Upotrebljavat će se standardne vrijednosti. "Jakarta Commons HttpClient" biblioteka sadržava klasu za sastavljanje upita HTTP:

```
UrlQuery putQuery = new UrlQuery();
putQuery.setNameValue("CMD", "Put");
putQuery.setNameValue("DATABASE", "nr");
putQuery.setNameValue("QUERY", "SSWWAHVEMGPPDPILGVTEAY");
putQuery.setNameValue("PROGRAM", "blastp");
```

Nakon toga se kreira nova klasa HttpClient:

```
HttpClient http = new HttpClient();
GetMethod getMethod = new GetMethod("http://www.ncbi.nlm.nih.gov/blast/Blast.cgi")
getMethod.setQueryString(putQuery);
```

Zahtjev se šalje pomoću slijedećeg kôda:

```
int statusCode = http.executeMethod(getMethod);
```

Status rezultata provjerava se pomoću kôda:

```
if (statusCode != HttpStatus.SC_OK)
```

Dobiveni rezultat jest HTML Web stranica koja se sprema u "*String*" objekt:

```
 InputStreamReader inputStream = new InputStreamReader(
```

```
    getMethod.getResponseBodyAsStream(),
    getMethod.getResponseCharSet());
 BufferedReader buffer = new BufferedReader(inputStream);
 StringBuilder resultBuffer = new StringBuilder(2048);
 String line;
 while ((line = buffer.readLine()) != null) {
    result.append(line);
 }
 Strigin result = resultBuffer.toString();
```

Na kraju se zatvaraju svi otvoreni priključci (engl. "connection") programa:

```
getMethod.releaseConnection();
inputStream.close();
```

Varijabla "*result*" sadržava sadržaj HTML. Taj sadržaj nije konačno izvješće programa BLAST koje se želi dobiti. U tom se sadržaju HTML nalaze dvije vrijednosti omeđene komentarima unutar kôda. Skraćeni primjer jednog takvoga kôda jest:

```
    <!–
    QBlastInfoBegin
                RID = 954517067-8610-1647
                RTOE = 207
                QBlastInfoEnd
     -->
```

Kada se želi dobiti konačno izvješće programa BLAST, vrijednost se *RID* (engl. "Request Identifier"; koji predstavlja identifikacijski broj) treba vratiti mrežnom servisu BLAST. Vrijednost *RTOE* (engl. "Request Time of Execution"; u sekundama) je vrijeme potrebno za provođenje cjelokupnog pretraživanja. Vrijednosti *RID* i *RTOE* se mogu rasčlaniti pomoću slijedećih "regularnih izraza" (engl. "regular expression"):

```
public static final String RTOE_RID_PATTERN =
"<!--\\s*QBlastInfoBegin\\s*RID = (.*)\\s*RTOE = (.*)\\s*QBlastInfoEnd\\s*-->";
Pattern pattern = Pattern.compile(RTOE_RID_PATTERN);
Matcher matcher = pattern.matcher(result);

String resultRID;
String resultRTOE;
if (matcher.find()) {
  if (matcher.groupCount() != 2) {
        return false;  }
```

```
}
resultRID = matcher.group(1);
resultRTOE = matcher.group(2);
```

Nakon što mrežni servis BLAST obavi pretraživanje, program provjerava da li je zahtjev izvršen:

```
UrlQuery getQuery = new UrlQuery();
getQuery.setNameValue("RID", rid);
getQuery.setNameValue("CMD", "GET");
```

Potrebno je rasčlaniti vrijednost "*Status*" na slijedeće vrijednosti: *WAITING* (označava da je potrebno čekati još nekoliko sekundi), *READY* (označava da je zahtjev obrađen), *UNKNOWN* (označava da je upit izgubljen) ili *ERROR* (označava grešku prilikom upita):

```
<!--
  QBlastInfoBegin
  Status=READY
  QBlastInfoEnd
  -->
```

Ukoliko je dobiven "*Status*"=READY može se kreirati upit GET da bi se dobilo konačno izvješće u obliku zapisa XML. Pogodci se iz zapisa XML rasčlane pomoću Castor biblioteke. Od tih se podataka kreira konačno izvješće bioinformatičkog programa BLAST i prikazuje se kao Web stranica (Primjer cjelovitog izvješća bioinformatičkog brograma BLAST oblikovanog računalnim programom za prikupljanje bioloških podataka prikazan je kao Prilog 1).

Korisnik iz konačnog izvješća bioinformatičkog programa BLAST odabire jedan ili više pogodaka. Program će iz baze podataka GenBank preuzeti sve podatke, svakog obilježenog pogotka, u obliku zapisa "DBsource" (primjer oblika zapisa "DBsource" za poliketidni antibiotik eritromicin prikazan je kao Prilog 2). Iz oblika zapisa "DBsource", program pojedinačne podatke unosi u tablice vlastite baze podataka pri čemu je upotrijebljen objektno orijentirani računalni jezik Java (51) i tehnologija Hibernate (53).

3.2.3. Anotacija modula, domena i njihovih poveznica

Iz vlastite baze podataka program uzima sekvenciju proteina svakog pojedinoga gena i u njoj pronalazi domene i njihove poveznice pomoću bioinformatičkog programskog paketa

HMMER (43). Nakon što je završeno pretraživanje s bioinformatičkim programskim paketom HMMER, program prikazuje korisniku, na interaktivnoj Web stranici za anotaciju gena "createDomains.jsp", gen po gen s na njima pronađenim domenama. Web stranica za anotaciju prikazuje sve važnije podatke gena i proteina: ime gena, položaj proteina unutar sekvencije DNA iz koje je preveden translacijom, dužinu proteina i druge. Korisnik ovdje može kreirati module koji sadržavaju jednu ili više domena.

Na primjer, Web stranica "createModule.jsp" prikazuje jedan protein te, pomoću programa "hmmpfam" programskog paketa HMMER, koji se nalazi na poslužitelju, pronalazi sve domene u proteinskom produktu toga gena. Na toj je Web stranici protein prikazan u obliku vertikalne crvene linije, gdje svaki piksel predstavlja jedno slovo koje obilježava svaku pojedinu aminokiselinu u proteinu. Ispod vertikalne crvene linije nacrtane su proteinske domene u obliku pravokutnika različitih boja. Zbroj dužina domena i njihovih poveznica je proporcionalan dužini proteina u kojemu se domene nalaze. Da bi korisnik od pripadajućih domena mogao kreirati module, upotrijebljen je računalni jezik JavaScript za međusobno interaktivno djelovanje vertikalne crvene linije proteina s pravokutnicima domena. To se postiže na taj način da se mišem označe odabrane domene, pri čemu kôd računalnog jezika JavaScript upotrebljava "onClick" i "onMouseOver" događaje.

Za izradu dinamičkih Web stranica ("createDomains.jsp" i "createModule.jsp") upotrijebljeni su slijedeći računalni jezici i tehnologije: računalni jezici Java (51), JavaScript (55), HTML i CSS (58), te tehnologije JSP (62), Hibernate (53) i DWR (54).

3.2.4. Anotacija biosintetskih puteva poliketida/peptida

Za izradu dinamičkih Web stranica "createCluster.jsp" i "orderGene.jsp" upotrijebljeni su isti računalni jezici i tehnologije kao i za izradu dinamičkih Web stranica "createDomains.jsp" i "createModule.jsp". Naime, nakon kreiranja modula program vodi korisnika na sljedeću Web stranicu, "createCluster.jsp". Na toj Web stranici korisnik ima pregled svih gena s pripadajućim modulima, domenama i njihovim poveznicama. U ovome se stupnju kreiraju genske nakupine odnosno, korisnik određuje koji geni pripadaju kojoj genskoj nakupini, te im se pridodaje ime genske nakupine za kasnije lakše pretraživanje baze podataka.

Nakon kreiranja genske nakupine program dinamički kreira hipervezu prema sljedećoj Web stranici, "orderGene.jsp". Na toj Web stranici korisnik određuje biosintetički redoslijed gena u genskoj nakupini. Prvi gen obično započinje takozvanom početnom domenom, "Loading domain" (LD), a zadnji gen obično završava domenom tioesteraza (TE). Ukoliko se između prvog i zadnjeg gena nalazi više od jednog gena, njihov redoslijed u biosintetskom putu odgovara njihovom redoslijedu u sekvenciji DNA. Osim toga, na Web stranici "orderGene.jsp", korisnik za svaki pojedini modul odabire i kemijske strukture građevnih jedinica. Te se građevne jedinice nalazi unutar programskog paketa ***CompGen*** u obliku izomeričkih zapisa SMILES. Korisnik može vidjeti kemijske strukture građevnih jedinica u 2D ili 3D obliku. Taj vizualni prikaz omogućuje alat kemoinformatičkog programa ChemAxon (45) koji je ugrađena u Web stranicu "orderGene.jsp" kao Java aplet. Izborom određenog modula, programski paket ***CompGen*** prikazuje građevnu jedinicu koja je uz taj modul vezana. Pri vrhu prikaza Web stranice "orderGene.jsp" nalazi se padajuća lista, na kojoj se mogu odabrati druge građevne jedinice te pridodati pojedinom modulu. Sve građevne jedinice se mogu prikazati u 2D i 3D obliku, te detaljnije pregledati pomoću alata kemoinformatičkog programa ChemAxon.

3.2.5. Grafičko prikazivanje kemijskih struktura poliketida/peptida

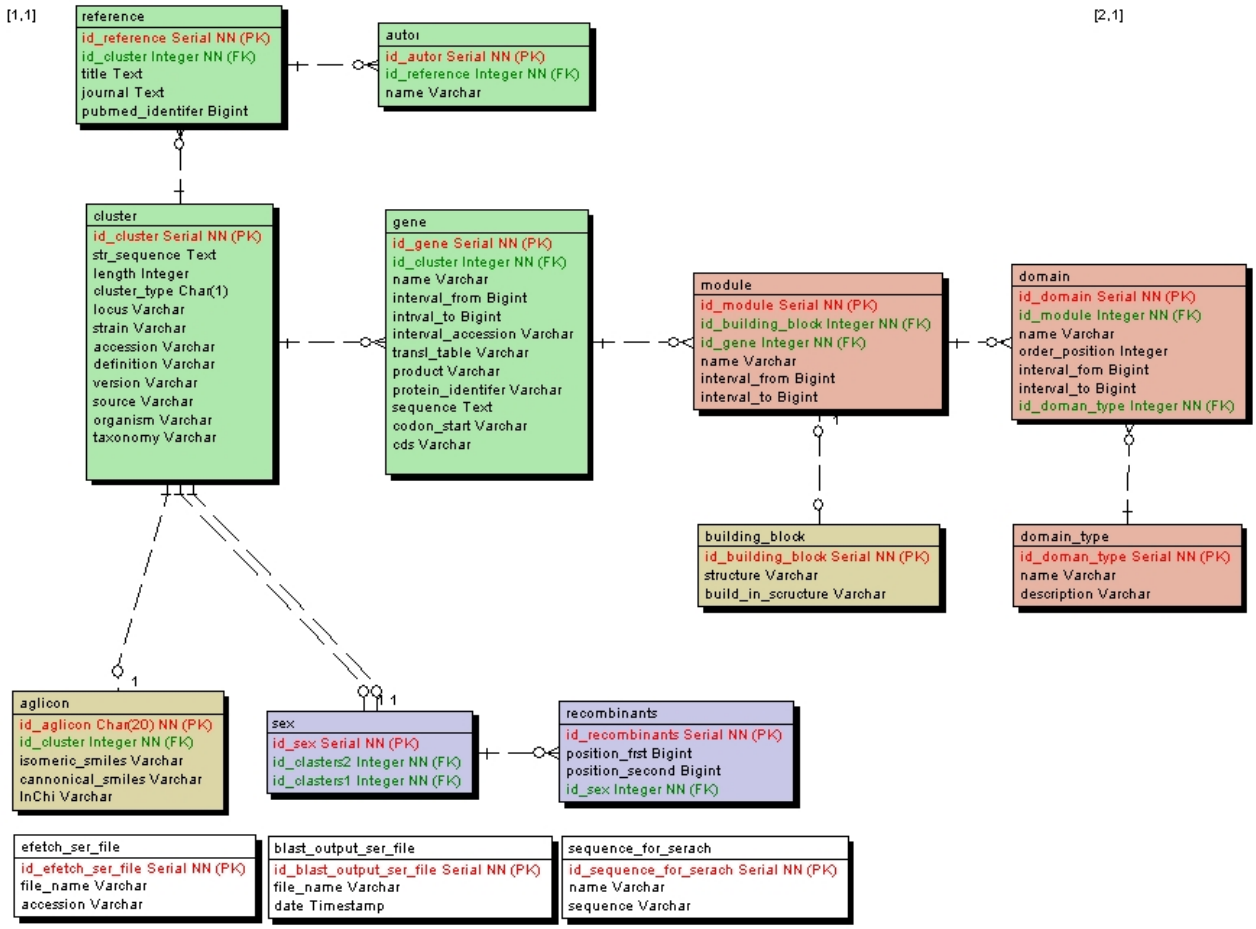
Na temelju biosintetičkog redoslijeda pojedinih gena u genskoj nakupini i građevnih jedinica u obliku zapisa SMILES, generički programski paket ***CompGen*** može na jednostavan način kreirati cjelovitu kemijsku strukturu linearnoga lanca poliketida/peptida.

Na dinamičkoj se Web stranici "viewStructure.jsp" korisniku pruža mogućnost vizualnog prikaza produkta biosintetskog puta genske nakupine Za izradu se dinamičke Web stranice "viewStructure.jsp" također upotrjebljavaju računalni jezici Java (51), JavaScript (55), HTML i CSS (58), te tehnologije JSP (62), Hibernate (53) i DWR (54). Kemijska se struktura linearnog poliketidnog/peptidnog lanca može prikazati u obliku 2D ili 3D. To omogućavaju ChemAxon alpeti Java (45). Java apleti (51) su aplikacije s grafičkim korisničkim sučeljem koje se mogu izvršavati u Web pregledniku na gotovo svim važnijim platformama. Na dinamičkoj se Web stranici "viewStructure.jsp" nalazi i tablica kreiranih genskih nakupina. U tablici se nalaze hiperveze koji pružaju korisniku mogućnost da pregleda svaku novokreiranu gensku nakupinu i kemijsku strukturu linearnog poliketidnog/peptidnog lanca u 2D ili 3D obliku. Hiperveza u koloni "Order and structure" vodi korisnika na prethodni stupanj ukoliko naknadno želi promijeniti kemijsku strukturu modula ili redoslijed gena u biosintetičkom putu.

4. Rezultati

4.1. Baza podataka poliketida/peptida

Okosnica generičkog programskog paketa ***CompGen*** jest specifično strukturirana baza podataka poliketidnih/peptidnih genskih nakupina izrađena po uzoru na BioSQL, koji predstavlja standardan model za prikazivanje bioloških podataka. Specifično strukturirana baza podataka sadržava sve podatke od sekvencija DNA genskih nakupina poliketid sintaza i sintetaza neribosomalno sintetiziranih peptida, primarnih sekvencija proteina gena, modula i domena, kemijskih struktura početnih i produžnih građevnih jedinica u obliku izomeričkih zapisa SMILES, dvodimenzionalnih (2D) i trodimenzionalnih (3D) struktura poliketidnih lanaca do konačnih cikličkih kemijskih struktura poliketidnih okosnica, tj. do aglikona. Struktura vlastite baze podataka prikazana je na Slici 17.



Slika 17. Shematski prikaz strukture vlastite baze podataka poliketida/peptida izrađene po uzoru na BioSQL.

4.2. Prikupljanje podataka

Generički programski paket ***CompGen*** zahtijeva prikupljanje bioloških i kemijskih podataka iz javnih baza podataka.

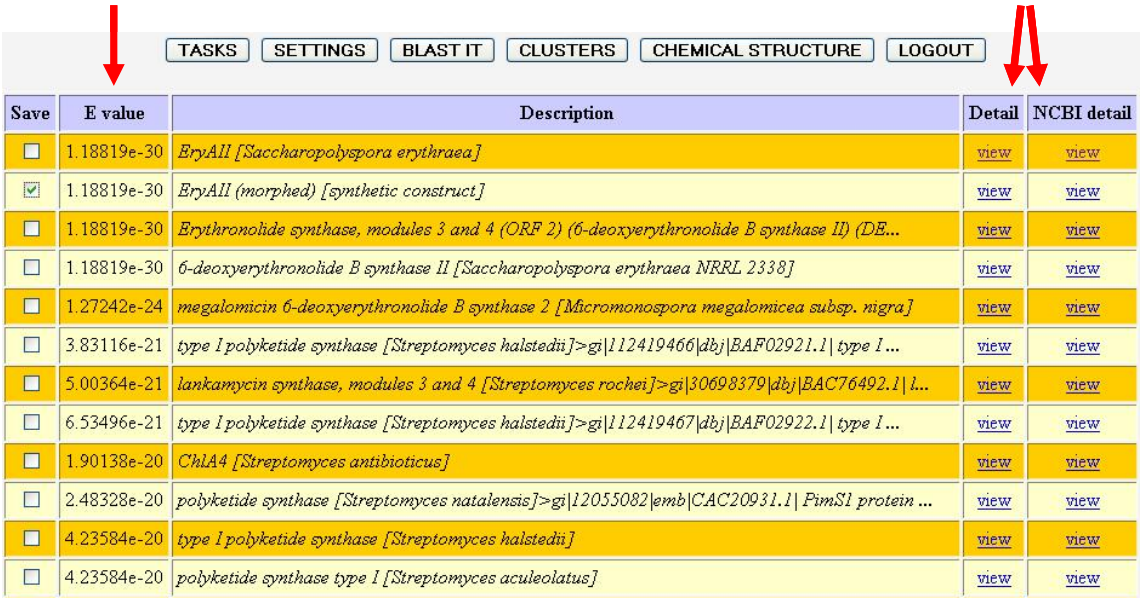
Da bi se ti biološki podaci prikupili iz javnih baza podataka i specifično strukturirali u vlastitoj bazi podataka, dio računalnog programa generičkog programskog paketa ***CompGen*** sadržava konsensus (tj. 'idealne') sekvencije aminokiselina domena poliketid/peptid sintaza/sintetaza u obliku zapisa FastA (Antonio Starčević, neobjavljeni rezultati) (Slika 18), koje se upotrebljavaju kao sekvencije, to jest upiti, za pretraživanje baze podataka GenBank bioinformatičkim programom BLAST.

```
>Consensus|KS|422 bp
epiaivgigcrfpggigspeqlwrvlaeganlttgfpadrgwdigrlyhpdpdnpgtsyv
dkggfltdaadfdpgffgitprealamdpqqrlmletaweaveragidpdalrgtdtgvf
vgmngqsymqllageaervdgyqglgnsasvlsgriaytfgwegpaltvdtacssslvgi
hlamqalrrgecslalaggvtvmsdpytfvdfstqrglasdgrckafsaradgfalsegv
aalvleplsraranghqvlavlrgsavnqdgasnglaapngpsqervirqalaasgvpaa
dvdvveahgtgtelgdpieagaliatygqdrdrplrlgsvktnightqaaagaagvikvv
lamrhgmlprslhadelsphidwesgavevlreevpwpagerprragvssfgvsgtnahv
iv
```

Slika 18. Konsenzus primarne sekvencije aminokiselina domene ketosintaza (KS) u obliku zapisa FastA.

Generički programski paket ***CompGen*** provodi pretragu na Web poslužitelju, to jest serveru, Nacionalnog centra za biotehnološke informacije (NCBI), koji održava bazu podataka GenBank. Dio računalnog programa generičkog programskog paketa ***CompGen*** prikuplja sve do sada sekvencirane genske nakupine poliketida/peptida i grafički ih prikazuje. Izvještaj programa BLAST, oblikovan generičkim programskim paketom ***CompGen***, sadržava prikaz od prvih 100 zapisa baze podataka GenBank (Slika 19 A i B). Identičnost, sličnost ili različitost podataka izražena je vrijednošću E (engl. "E value") i bodovnim stanjem poravnanja (engl. "Score"). Što je vrijednost E manja, a bodovno stanje poravnanja veće, to je pronađena sekvencija DNA sličnija sekvenciji upotrijebljenoj kao sekvencija-upit.

A



Save	E value	Description	Detail	NCBI detail
☐	1.18819e-30	*EryAII [Saccharopolyspora erythraea]*	view	view
☑	1.18819e-30	*EryAII (morphed) [synthetic construct]*	view	view
☐	1.18819e-30	*Erythronolide synthase, modules 3 and 4 (ORF 2) (6-deoxyerythronolide B synthase II) (DE...*	view	view
☐	1.18819e-30	*6-deoxyerythronolide B synthase II [Saccharopolyspora erythraea NRRL 2338]*	view	view
☐	1.27242e-24	*megalomicin 6-deoxyerythronolide B synthase 2 [Micromonospora megalomicea subsp. nigra]*	view	view
☐	3.83116e-21	*type I polyketide synthase [Streptomyces halstedii]>gi|112419466|dbj|BAF02921.1| type 1...*	view	view
☐	5.00364e-21	*lankamycin synthase, modules 3 and 4 [Streptomyces rochei]>gi|30698379|dbj|BAC76492.1| l...*	view	view
☐	6.53496e-21	*type I polyketide synthase [Streptomyces halstedii]>gi|112419467|dbj|BAF02922.1| type 1...*	view	view
☐	1.90138e-20	*ChlA4 [Streptomyces antibioticus]*	view	view
☐	2.48328e-20	*polyketide synthase [Streptomyces natalensis]>gi|12055082|emb|CAC20931.1| PimS1 protein ...*	view	view
☐	4.23584e-20	*type I polyketide synthase [Streptomyces halstedii]*	view	view
☐	4.23584e-20	*polyketide synthase type 1 [Streptomyces aculeolatus]*	view	view

B

☐	3.2221e-17	*beta-ketoacyl synthase [Salinispora tropica CNB-440]>gi|113904267|gb|EAU23104.1| beta-ke...*	view	view
☐	4.2082e-17	*Plm2-3 [Streptomyces sp. HK803]*	view	view
☐	4.2082e-17	*modular polyketide synthase [Streptomyces avermitilis]*	view	view
☐	4.2082e-17	*borrelidin polyketide synthase, type I [Streptomyces parvulus]*	view	view
☐	4.2082e-17	*modular polyketide synthase [Streptomyces avermitilis MA-4680]>gi|29604057|dbj|BAC68129....*	view	view
☐	4.2082e-17	*NapA [Streptomyces hygroscopicus subsp. duamyceticus]*	view	view
☐	4.2082e-17	*Beta-ketoacyl synthase:Acyl transferase domain:Phosphopantetheine-binding domain [Franki...*	view	view
☐	4.2082e-17	*beta-ketoacyl synthase [Salinispora tropica CNB-440]>gi|113901348|gb|EAU20216.1| beta-ke...*	view	view

Save — Number of results: 100

Detailed Analysis of Results

☐ Accession: AAV51821 NCBI detail Top
Hit definition: *EryAII [Saccharopolyspora erythraea]*
HSP Information: **Score** = 337 **E-value** = 1.1407e-30

```
Query:  EPIAIVGIGCRFPGGIGSPEQLWRVLAEGAMLTTGFPADRGWDIGRLYHPDPDNPGTSYV
        EPIAIVGIGCRFPGGIGSPEQLWRVLAEGAMLTTGFPADRGWDIGRLYHPDPDNPGTSYV
Hit:    EPIAIVGIGCRFPGGIGSPEQLWRVLAEGAMLTTGFPADRGWDIGRLYHPDPDNPGTSYV
```

HSP Information: **Score** = 241 **E-value** = 1.54528e-19

```
Query:  EPIAIVGIGCRFPGGIGSPEQLWRVLAEGAMLTTGFPADRGWDIGRLYHPDPDNPGTSYV
        +PIAIV + CR PGG+ +P++LW +L EG     +GFP DRGWD+ RL+HPDPDNPGTSYV
Hit:    DPIAIVSMACRLPGGVNTPQRLWELLREGGETLSGFPTDRGWDLARLHHPDPDNPGTSYV
```

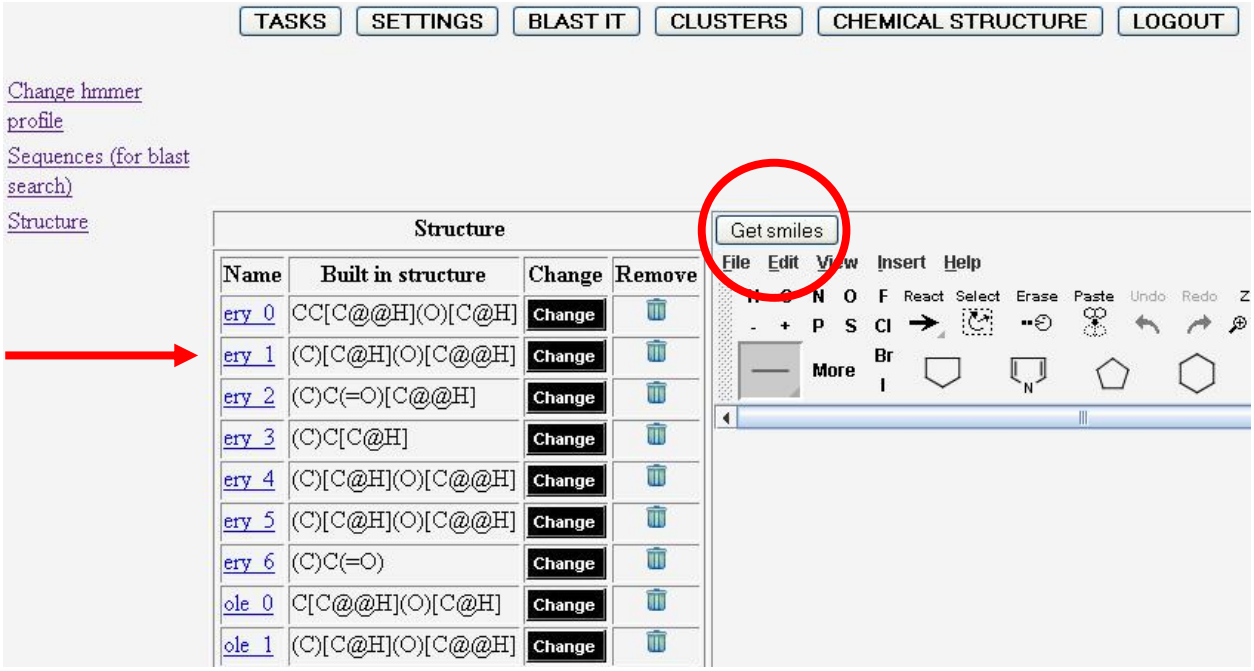
Slika 19. Grafički prikaz početnog (**A**) i završnog (**B**) dijela izvješća programa BLAST, oblikovanog dijelom programa generičkog programskog paketa ***CompGen***.

Detalji se svakog zapisa, prije pohranjivanja u vlastitu bazu podataka, mogu provjeriti u izvornoj javnoj bazi podataka GenBank (Slika 20). Osim toga, iz samog se izvješća programa BLAST može uočiti identičnost, sličnost ili različitost redoslijeda aminokiselina u izabranoj domeni (Slika 19; vidi. "Detailed Analysis of Results"). Ako dobiveni rezultat zadovoljava, korisnik programskog paketa ***CompGen*** sprema ga u vlastitu bazu podataka pomoću opcije "spremi" (engl. "Save"). Računalni program pronalazi i sprema zapis iz baze podataka GenBank, oblikovan programskim paketom ***CompGen***, u obliku zapisa "DBsource". Prije nego što ga spremi, korisnik može još jednom pogledati detalje zapisa na Web poslužitelju Nacionalnog centra za biotehnološke informacije (NCBI).

Slika 20. Grafički prikaz poveznice generičkog programskog paketa ***CompGen*** s bazom podataka GenBank iz koje su podaci prikupljeni.

Za prikupljanje kemijskih podataka, to jest početnih i produžnih građevnih jedinica u obliku izomeričkih zapisa SMILES, iz javnih baza podataka, programski paket ***CompGen*** koristi bazu podataka PubChem Compound. Izrađena je vlastita baza podataka svih do sada poznatih početnih i produžnih građevnih jedinica (Nives Škunca, neobjavljeni rezultati). Dio računalnog

programa generičkog programskog paketa ***CompGen*** dohvaća prikupljene podatke iz vlastite baze podataka u obliku izomeričkih zapisa SMILES, grafički ih prikazuje, i ugrađuje ih u linearni poliketidni lanac (Slike 21, 24 i 25).



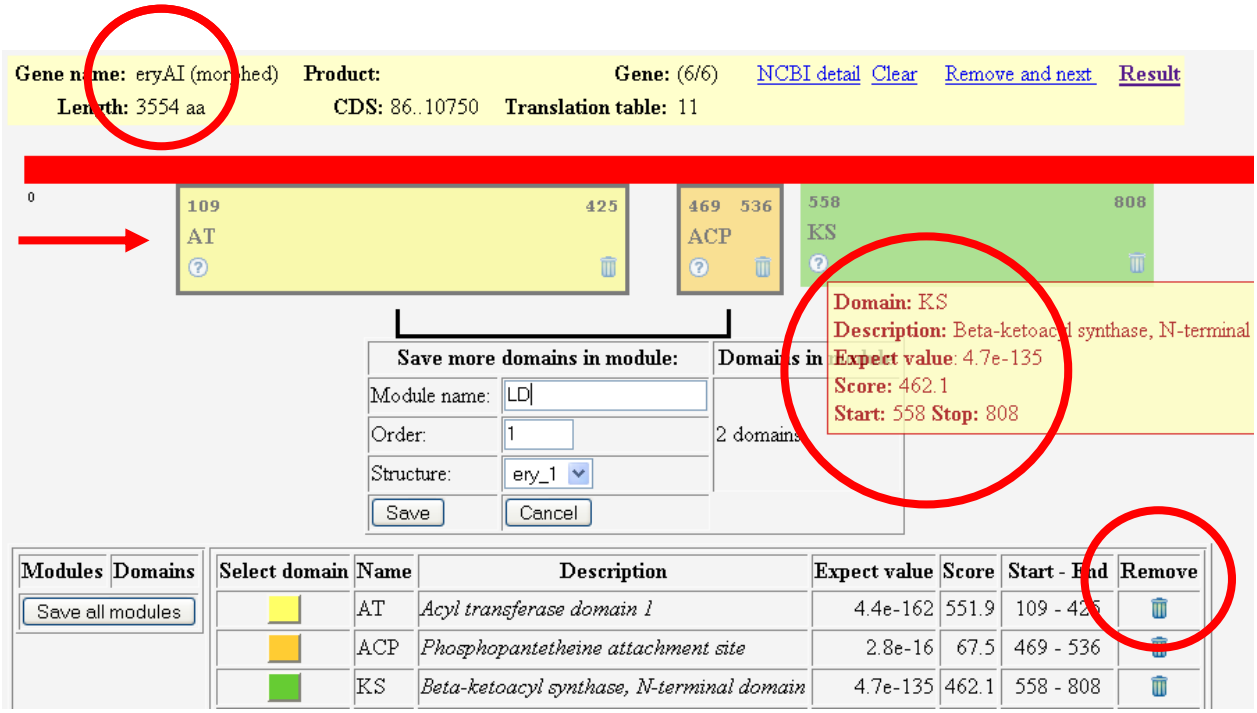
Slika 21. Grafički prikaz baze podataka početnih i produžnih građevnih jedinica u obliku izomeričkih zapisa SMILES i dijela program ChemAxon pomoću kojeg se može pripremiti novi zapis.

Kada se u znanstvenoj literaturi pojavi nova poliketidna genska nakupina, to jest novi biosintetski put poliketida koji upotrebljava do sada nepoznatu građevnu jedinicu, dio računalnog programa generičkog programskog paketa ***CompGen*** podržava program ChemAxon za pripremu novog zapisa SMILES (Slika 21).

4.3. Obrada podataka

Nakon prikupljanja bioloških podataka u obliku zapisa "DBsource" slijedi anotacija proteina poliketidnih gena i prepoznavanje njegovih modula, domena i poveznica. Generički programski paket ***CompGen*** sadržava vlastite proteinske profile domena gena PKS/NRPS

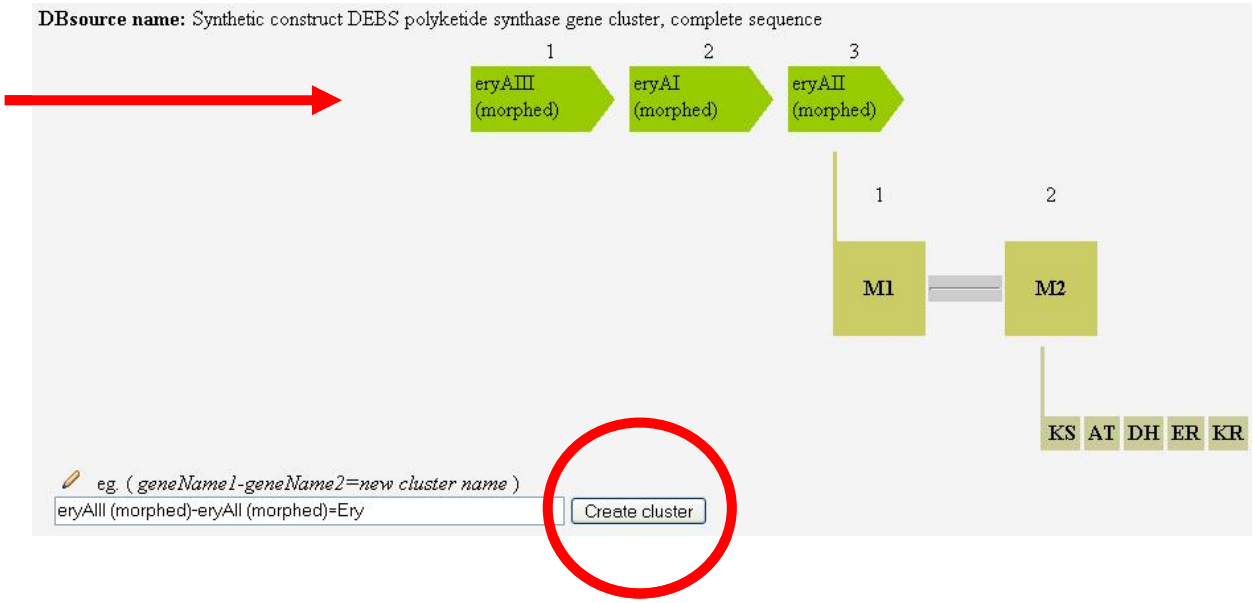
pripremljene upotrebom baze podataka Pfam i bioinformatičkoga programskog paketa HMMER (Antonio Starčević, neobjavljeni rezultati).



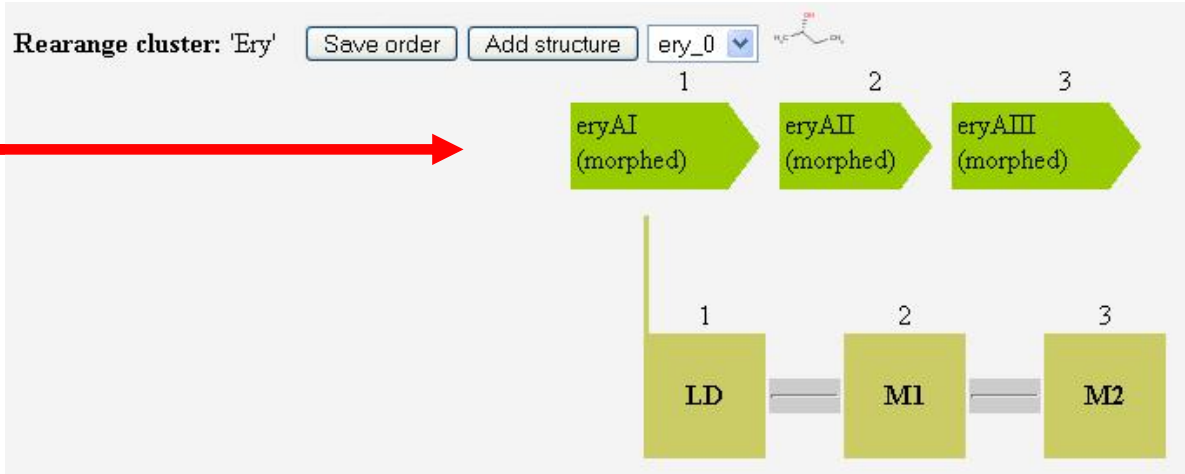
Slika 22. Grafički prikaz modula, domena i njihovih poveznica koje je dio programa gereričkog programskog paketa ***CompGen*** pronašao u sekvenciji aminokiselina proteina poliketidnog gena *ery*AI.

Na temelju vlastitih proteinskih profila domena poliketid sintaza i programskog paketa HMMER, dio računalnog programa generičkog programskog paketa ***CompGen*** prepoznaje domene u sekvenciji aminokiselina proteina, koju grafički prikazuje vertikalnom crvenom linijom (Slika 22). U pozadini se te crvene linije nalaze izvorni podaci, gdje svaki piksel predstavlja jedno slovo, kojim je obilježena svaka pojedina aminokiselina. U prikazanom primjeru crvena linija predstavlja 3554 aminokiselina proteina gena *ery*AI. Domene su grafički prikazane pravokutnicima različitih boja, i sadržavaju podatke o početku i kraju domene, te identičnosti, sličnosti ili različitosti sekvencije aminokiselina s upitom u obliku vrijednosti E i bodovnog stanja poravnanja S. Tijekom anotacije, korisnik programskog paketa ***CompGen*** može prihvatiti ili odbaciti domenu koju mu računalni programski paket predlaže. Na taj se način anotiraju svi geni i njihovi proteinski produkti svake pojedine poliketid/peptid sintaze/sintetaze koja se sprema u vlastitu bazu podataka.

A



B

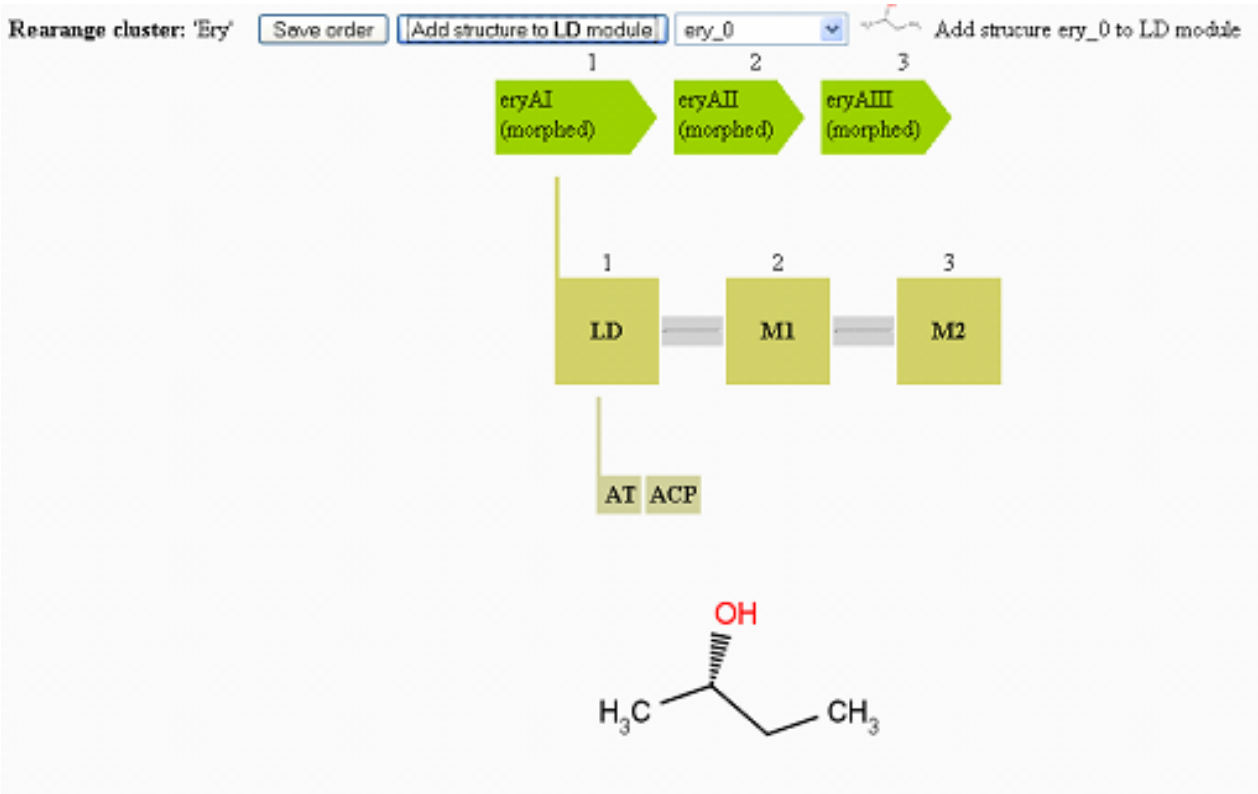


Slika 23. Grafički prikaz anotacija biosintetskoga puta analizirane poliketidne genske nakupine koja sadržava genetičku uputu za biosintezu eritromicina (*ery*AI, *ery*AII i *ery*AIII). Redoslijed proteina, modula, domena i poveznica zapisan u sekvenciji DNA (**A**) i njihov redoslijed nakon redanja u njihov biosintetski redoslijed (**B**).

Sljedeći je stupanj anotacija biosintetskoga puta analiziranog poliketida/peptida. U početku dio programa generičkog programskog paketa ***CompGen***, na temelju podataka pohranjenih u

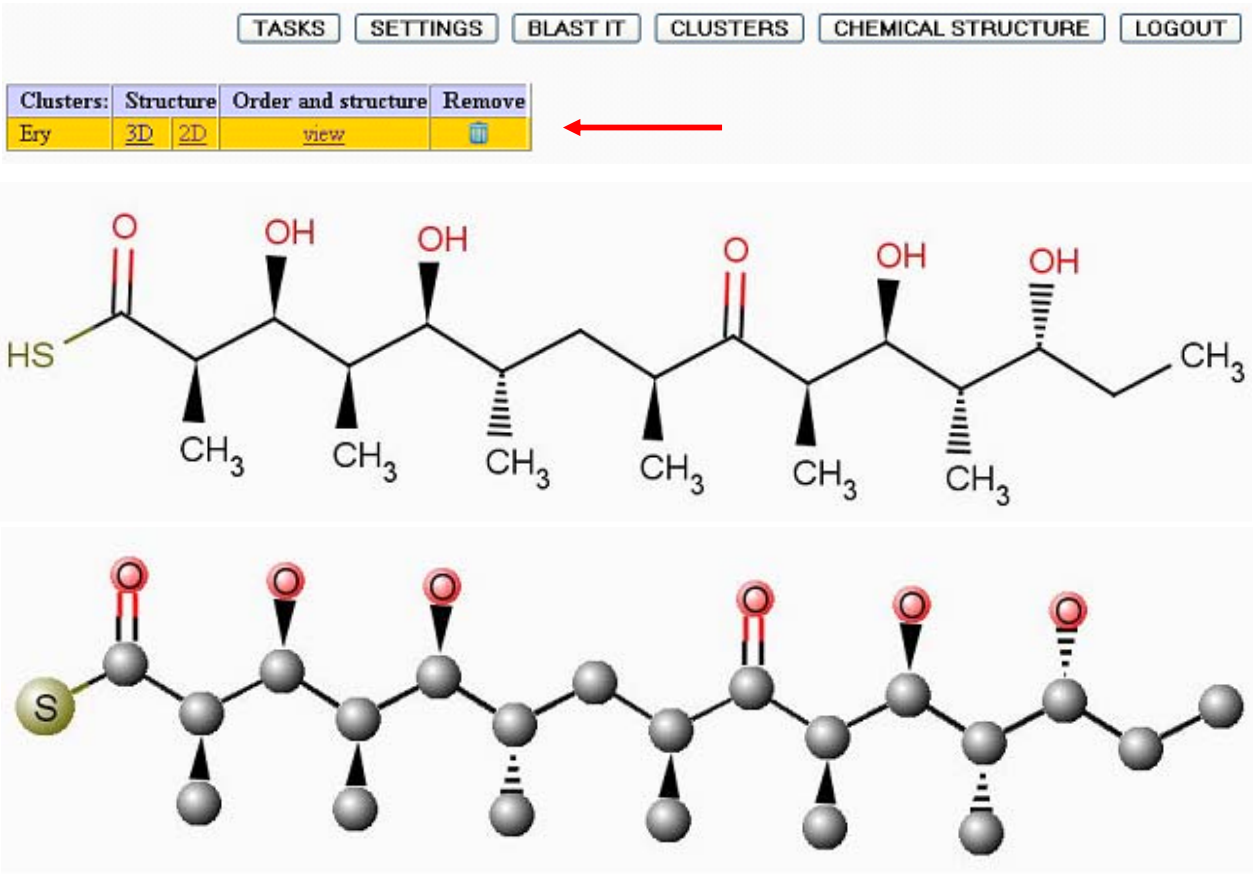
obliku zapisa "DBsource", grafički prikazuje redoslijed proteina, modula, domena i poveznica zapisan u sekvenciji. Na primjer, *ery*AIII, *ery*AI i *ery*AII (Slika 23 A). Nakon toga, slijedeći dio programa generičkog programskog paketa ***CompGen*** poreda, pomoću opcije "kreiraj biosintetski put" (engl. "Create cluster"), proteine, module i domene u njihov biosintetski redoslijed, to jest *ery*AI, *ery*AII i *ery*AIII, odnosno početna (tzv. "Loading"; LD) domena, Modul-1, Modul-2 i tako dalje. (Slika 23 B).

Iz do sada navedenih podataka, koji se nalaze u vlastitoj bazi podataka, dio programa generičkog programskog paketa ***CompGen***, počevši od sekvencije DNA genske nakupine, te sekvencija aminokiselina proteina, modula, domena te građevnih jedinica u obliku izomeričkih zapisa SMILES (Slika 24), predviđa linearni poliketidni lanac koji može prikazati u dvodimenzionalnom (2D) ili trodimenzionalnom (3D) obliku (Slika 25) prikladnim za daljnju računalnu obradu.



Slika 24. Grafički prikaz odabira građevnih jedinica poliketidnog antibiotika eritromicina, koje programski paket ***CompGen*** odabire iz baze podataka građevnih jedinica i povezuje u kemijsku strukturu linearnog lanca poliketida.

Na Slici 25 grafički su prikazane dvodimenzionalne (2D) i trodimenzionalne (3D) kemijske strukture linearnog lanca poliketidnog antibiotika eritromicina.



Slika 25. Grafički prikaz dvodimenzionalne (2D) i trodimenzionalne (3D) strukture linearnog poliketidnog lanca antibiotika eritromicina.

5. Rasprava

U tijeku je razvoj generičkih računalnih programskih paketa, ***CompGen*** i ***ClustScan*** (akronimi za "***Comp***ound ***Gen***erator" i "***Clust***er ***Scan***ner"), napisanih objektno orijentiranim računalnim jezikom Java (51, 70) i skriptnim računalnim jezikom JavaScript (55). Temeljni cilj programskog paketa ***CompGen*** jest strukturiranje virtualne baze podataka potpuno novih, u prirodi nepoznatih, kemijskih entiteta, počevši od sekvencija DNA poliketidnih/peptidnih genskih nakupina koje se nalaze u općim javnim bazama poput baze podataka GenBank (30). U specijaliziranim javnim bazama podataka (31, 32) ima nekoliko desetaka detaljno opisanih, to jest anotiranih, genskih nakupina. Nasuprot tome, programski bi paket ***ClustScan*** trebao pomoći pri anotaciji sekvenciranih poliketidnih/peptidnih genskih nakupina čija su funkcija i konačni kemijski produkt još nepoznati. Takvih sekvenciranih genskih nakupina u općoj javnoj bazi podataka GenBank već danas ima nekoliko stotina i njihov broj eksponencijalno raste (30).

Upotrebom računalnog jezika UML ("Universal Markup Language"; 71) u tijeku je izrada, o platformi neovisnih, generičkih programskih paketa ***CompGen*** i ***ClustScan*** za automatsko prikupljanje, pohranu i obradu podataka o poliketidima, neribosomalnim peptidima i ostalim produktima modularnih biosintetskih putova iz javno dostupnog izvora baze podataka GenBank (30). Okosnica programskog paketa ***CompGen*** specifično je strukturirana baza podataka poliketidnih/peptidnih genskih nakupina izrađena po uzoru na BioSQL v1.29, koji predstavlja standardni model za prikazivanje bioloških podataka (46). Baza podataka sadržava sve podatke od sekvencija DNA genskih nakupina PKS/NRPS do sekvencija DNA i proteina gena, modula, domena i njihovih poveznica. Sadržava i kemijske strukture početnih i produžnih građevnih jedinica u obliku izomeričkih zapisa SMILES (37), dvodimenzionalne (2D) i trodimenzionalne (3D) strukture poliketidnih lanaca, te konačne cikličke kemijske strukture poliketidnih/peptidnih okosnica, tj. aglikona (Slika 17).

Da bi se ti biološki i kemijski podaci prikupili iz javnih baza podataka i pohranili u vlastitoj bazi, programski paket ***CompGen*** sadržava konsenzus (tj. "idealne") primarne sekvencije aminokiselina domena proteina PKS/NRPS u obliku zapisa FastA (33) (Slika 18). Te su primarne sekvencije aminokiselina modelnih domena pripremljene upotrebom baze podataka Pfam (35) i funkcije programa "hmmemit" bioinformatičkoga programskog paketa HMMER (43). One se upotrebljavaju kao sekvencije-upiti za pretraživanje baze podataka GenBank bioinformatičkim programom BLAST (42). BLAST je akronim za "Basic Local Alignment Search Tool" i upotrebljava se za brzo pretraživanje mogućih lokalnih poravnanja sekvencija-upita sa sekvencijama iz javnih baza podataka. Pri stupnjevanju pogodaka koristi se statistička analiza za

procjenu važnosti svakog pogotka koja se izražava kao "očekivana vrijednost", to jest vrijednost E (engl. "E value"). ***CompGen*** zatim provodi pretragu na Web poslužitelju Nacionalnog centra za biotehnološke informacije (NCBI), koji održava bazu podataka GenBank (30).

Izvješće programa BLAST, oblikovano programskim paketom ***CompGen***, sadržava prikaz 100 zapisa baze podataka GenBank (Slika 19 A i B). Identičnost, sličnost ili različitost podataka izražena je vrijednošću E i bodovnim stanjem poravnanja S (engl. "Score"). Što je vrijednost E manja, a bodovno stanje poravnanja veće, to je pronađena sekvencija DNA sličnija sekvenciji upotrijebljenoj kao sekvencija-upit. Detalji se svakog zapisa, prije pohranjivanja u vlastitu bazu podataka, mogu provjeriti u izvornoj javnoj bazi podataka GenBank. Osim toga, iz samog se izvješća programa BLAST može uočiti identičnost, sličnost ili različitost redoslijeda aminokiselina u izabranoj domeni. Ako dobiveni rezultat zadovoljava, korisnik programskog paketa ***CompGen*** sprema ga u vlastitu bazu podataka pomoću opcije "spremi". Računalni program pronalazi i sprema zapis iz baze podataka GenBank, oblikovan programskim paketom ***CompGen***, u obliku zapisa "DBsource" (30). Prije nego što ga spremi, korisnik može još jednom pogledati detalje zapisa na Web poslužitelju Nacionalnog centra za biotehnološke informacije (Slika 20).

Nakon toga slijedi anotacija proteina poliketidnih/peptidnih gena i prepoznavanje njegovih modula, domena i poveznica. ***CompGen*** sadržava vlastite proteinske profile domena gena PKS/NRPS pripremljene upotrebom baze podataka Pfam (Bateman *i sur.*, 2002.) i bioinformatičkoga programskog paketa HMMER (43). Na temelju vlastitih proteinskih profila domena PKS/NRPS i programskog paketa HMMER, ***CompGen*** prepoznaje i grafički prikazuje domene u sekvenciji aminokiselina u proteinima (Slika 21). U pozadini toga grafičkog prikaza nalaze se podaci, gdje svaki piksel predstavlja jedno slovo koje obilježava svaku pojedinu aminokiselinu. Domene su također grafički prikazane i sadržavaju podatke o početku i kraju domene, te identičnosti, sličnosti ili različitosti sekvencije aminokiselina sa sekvencijom-upitom u obliku vrijednosti E i bodovnog stanja poravnanja. Tijekom anotacije, korisnik programskog paketa ***CompGen*** može prihvatiti ili odbaciti domenu koju mu predlaže računalni program. Tako se anotiraju svi geni i njihovi proteinski produkti svake pojedine PKS/NRPS koja se sprema u vlastitu bazu podataka.

Sljedeći je stupanj anotacija biosintetskoga puta analiziranog poliketida/peptida. U početku ***CompGen***, na temelju podataka pohranjenih u obliku zapisa "DBsource", grafički prikazuje

redoslijed proteina, modula, domena i poveznica zapisan u sekvenciji (Slika 23 A). Nakon toga, pomoću opcije "kreiraj biosintetski put", ***CompGen*** poreda proteine, module i domene u njihov biosintetski redoslijed (Slika 23 B).

Katalitički je aktivne domene potrebno posebno detaljno anotirati radi predviđanja njihove aktivnosti/neaktivnosti i specifičnosti. Programski paket ***CompGen*** mora moći prepoznati jesu li domene PKS/NRPS sustava, prisutne unutar genskih nakupina, aktivne ili neaktivne. Naime, iz detaljno anotiranih genskih nakupina poliketida i peptida poznato je da se unutar genskih nakupina nalaze i katalitički neaktivne domene. To je posebno zapaženo u domena dehidrogenaza (DH) (32). Zbog toga programski paket ***CompGen*** mora prvo moći analizirati i razlikovati sekvencije aminokiselina različitih domena po njihovoj dužini, to jest broju aminokiselina. Nakon toga mora moći analizirati i razlikovati sekvencije aminokiselina različitih domena s obzirom na prisutnost/odsutnost aminokiselina važnih za katalitičko svojstvo na određenim pozicijama u njihovu redoslijedu. U tu će svrhu također biti pripremljeni profili pojedinih domena upotrebom baze podataka Pfam (35) i funkcije programa "hmmemit" bioinformatičkoga programskog paketa HMMER (43).

Predviđanje specifičnosti domena AT/A za supstrate (građevne jedinice) koji se ugrađuju u poliketidne/peptidne lance, te specifičnosti domena KR za modifikaciju (stereospecifičnost –OH i $–CH_3$ skupina) građevnih jedinica nakon ugradnje u poliketidni/peptidni lanac, drugi je bitan čimbenik koji će programski paket ***CompGen*** morati razriješiti. I u tu će svrhu biti pripremljeni vlastiti specifični profili domena AT/A i KR upotrebom baze podataka Pfam (35) i podataka iz znanstvene literature (72, 73, 74, 75, 76), te funkcije programa "hmmbuild" bioinformatičkoga programskog paketa HMMER (43). Ti će profili pripremljeni odvojeno za C2, C3 i druge poliketidne građevne jedinice, a i za različite aminokiseline kao građevne jedinice neribosomalno sintetiziranih peptida, poslužiti za poravnanje nepoznatih domena AT/A i KR upotrebom funkcije programa "hmmalign" bioinformatičkoga programskog paketa HMMER. Programski će paket ***CompGen*** moći prepoznati ključne aminokiseline s obzirom na prisutnost/odsutnost aminokiselina važnih za svojstvo specifičnosti na određenim položajima u redoslijedu domena AT/A i KR (72, 73, 74).

Računalni programski paket ***CompGen*** sadržava i bazu podataka početnih i produžnih građevnih jedinica u obliku izomeričkih zapisa SMILES (37). Kada se u znanstvenoj literaturi pojavi nova poliketidna/peptidna genska nakupina, to jest novi biosintetski put poliketida/peptida

koji upotrebljava do sada nepoznatu građevnu jedinicu, ***CompGen*** podržava i kemoinformatički program ChemAxon (77), za pripremu novog zapisa SMILES (Slika 21). Iz do sada navedenih podataka, što se nalaze u specifično strukturiranoj bazi podataka, ***CompGen***, počevši od sekvencije DNA genske nakupine, predviđa linearni poliketidni/peptidni lanac koji se može prikazati u dvodimenzionalnom (2 D) ili trodimenzionalnom (3 D) obliku.

U tijeku je molekularno modeliranje procesa ciklizacije linearnih novih nepoznatih poliketidnih/peptidnih lanaca u cikličke strukture. Za to modeliranje upotrebljavamo poznate poliketidne/peptidne lance i poznate aglikone (32). Također je u tijeku i modeliranje procesa homologne rekombinacije (23) između sekvencija DNA dviju poliketidnih/peptidnih genskih nakupina. U prirodi proces rekombinacije zahtijeva minimalnu dužinu apsolutne identičnosti dviju sekvencija DNA. Minimalna se dužina apsolutne identičnosti, MEPS (engl. "Minimal Essential Pairing Sequence"; 78) mora nalaziti u području DNA razmjerno velike sličnosti sekvencija. Koordinate identičnosti i sličnosti dviju sekvencija DNA povezane su s bazom podataka kako bi se mogle prepoznati domene i poveznice koje sudjeluju u rekombinaciji. Kada računalni programi za modeliranje procesa ciklizacije i homologne rekombinacije budu napisani, budućnost će računalnih programskih paketa ***CompGen*** i ***ClustScan*** biti predviđanje biološke aktivnosti novih prirodnih kemijskih entiteta oblikovanih u uvjetima *in silico*, određivanjem aktivnosti kemijskih supstancija računalnim alatom PASS (akronim za "Prediction of Activity Spectra for Substances"; 79) ili upotrebom računalnom vođenog dizajna lijekova, CDD tehnologijom (80).

Koliko nam je poznato, u svijetu postoje četiri moguća konkurenta. Kao prvo, tvrtka Ecopia BioSciences Inc. (81) razvila je bioinformatički sustav ***DecipherIT™*** koji automatski anotira genske nakupine mikroorganizama iz tla. Svrha je sustava da predvidi kemijske strukture molekula koje sintetiziraju enzimi čiju genetičku uputu sadržavaju. Drugo, istraživači su iz Nacionalnog instituta za imunologiju, New Delhi, Indija, razvili računalni program ***SEARCHPKS*** za anotaciju domena poliketid sintaza u njihovim sekvencijama proteina. Njihov se računalni pristup temelji na iscrpnoj analizi dobro opisanih modularnih PKS sadržanih u bazi podataka PKSDB. Računalnom se programu ***SEARCHPKS*** može pristupiti na Web stranicama http://www.nii.res.in/searchpks.html (72). Treće, postoji i računalni program ***NRPSpredictor*** (73, 82) koji predviđa specifičnosti domena adenilacije (domena A), genskih nakupina NRPS, za odabir građevnih jedinica tijekom biosinteze neribosomalno sintetiziranih peptida. Konačno, nedavno je objavljen računalni program ***Biogenerator*** (83) koji sadržava program PASS (engl.

"Prediction of Activity Spectra for Substances"; 84), što će ga na tržište staviti norveška tvrtka Biosergen AS iz Trondheima. Taj je računalni program sličan sustavu ***DecipherIT™*** po tome što također automatski anotira sekvencije genoma radi predviđanja mogućih poliketidnih produkata sintetiziranih poliketid sintazama. Osim toga, taj će računalni program generirati i virtualnu bazu podataka poliketida koja će se moći pretraživati u uvjetima *in silico* da bi se predvidjela biološka aktivnost tih kemijskih supstancija (79). Međutim, ni jedan od navedenih računalnih programa ne predviđa procese homologne rekombinacije između dviju poliketidnih/peptidnih genskih nakupina radi predviđanja novih prirodnih kemijskih entiteta poput generičkoga računalnog programskog paketa ***CompGen***.

6. Zaključci

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

1. Za generički računalni programski paket ***CompGen*** specifično je strukturirana baza podataka poliketidnih/peptidnih genskih nakupina po uzoru na BioSQL.

2. Upotrebom objektno orijentiranog računalnog jezika Java i skriptnog računalnog jezika JavaScript razvijeni su dijelovi generičkog računalnog programskog paketa ***CompGen*** za:

   - prikupljanje i strukturiranje bioloških i kemijskih podataka,

   - detaljan opis, to jest anotaciju, modula, domena i njihovih poveznica,

   - anotaciju biosintetskih puteva poliketida/peptida, i

   - grafičko prikazivanje dvodimenzionalnih i trodimenzionalnih kemisjkih struktuta linearnih lanaca poliketida/peptida.

7. Literatura

1. *D. Hranueli, J. Cullum, B. Basrak, P. Goldstein, P.F. Long*, Plasticity of the *Streptomyces* genome - evolution and engineering of new antibiotics. Curr. Med. Chem. **12** (2005) 1697.
2. *D. Hranueli, A. Starcević, J. Žučko, J. Diminić, N. Škunca, V. Željeznak, D. Kovaček, D. Pavlinušić, J. Šimunković, P.F. Long, J. Cullum*. Oblikovanje novih prirodnih spojeva u uvjetima *in silico*. Kem. Ind. (Zagreb), (2007). (u tisku).
3. *A. Demain*, Pharmaceutically active secondary metabolites of microorganisms. Appl. Microbiol. Biotechnol. **52** (1999) 455.
4. *T.M. Kutchan*, Ecological arsenal and developmental dispatcher. The paradigm of secondary metabolism. Plant Physiol. **125** (2001) 58.
5. *J. Bérdy*, Bioactive microbial metabolites. J. Antibiot. (Tokyo) **58** (2005) 1.
6. *M.G. Watve, R. Tickoo, M.M. Jog, B.D. Bhole*, How many antibiotics are produced by the genus *Streptomyces*? Arch. Microbiol. **176** (2001) 386.
7. *S.D. Bentley, K.F. Chater, A.M. Cerdeno-Tarraga, G.L. Challis, N.R. Thomson, K.D. James, D.E. Harris, M.A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C.W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M.A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B.G. Barrell, J. Parkhill, D.A. Hopwood*, Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature **417** (2002) 141.
8. *H. Ikeda, J. Ishikawa, A. Hanamoto, M. Shinose, H. Kikuchi, T. Shiba, Y. Sakaki, M. Hattori, S. Omura*, Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat. Biotechnol. **21** (2003) 526.
9. *M. Oliynyk, M. Samborskyy, J.B. Lester, T. Mironenko, N. Scott, S. Dickens, S.F. Haydock, P.F. Leadlay*, Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. Nat. Biotechnol. **25** (2007) 447.

10. *S. Grabley, R. Thiericke*, The impact of natural products on drug discovery, str. 3-37. *U:* S. Grabley, R. Thiericke *i sur*. (Ured.), Drug Discovery from Nature, Springer, Berlin, 1999.
11. *D. Hranueli, J. Cullum*, Novi hibridni poliketidi dobiveni kombinatornom biosintezom. Kem. Ind. (Zagreb) **50** (2001) 381.
12. *K.J. Weissman*, Polyketide biosynthesis: understanding and exploiting modularity. Philos. Transact. A Math. Phys. Eng. Sci. **362** (2004) 2671.
13. *H.B. Bode, R. Muller*, Possibility of bacterial recruitment of plant genes associated with the biosynthesis of secondary metabolites. Plant Physiol. **132** (2003) 1153.
14. *C.R. Hutchinson, J. Kennedy, C. Park, S. Kendrew, K. Auclair, J. Vederas*, Aspects of the biosynthesis of non-aromatic fungal polyketides by iterative polyketide synthases. Anton. Leeuw. Int. J. G. **78** (2000) 287.
15. *Y. Seshime, P.R. Juvvadi, I. Fujii, K. Kitamoto*, Discovery of a novel superfamily of type III polyketide synthases in *Aspergillus oryzae*. Biochem. Biophys. Res. Commun. **331** (2005) 253.
16. *J.F. Martin*, Molecular control of expression of penicillin biosynthesis genes in fungi: regulatory proteins interact with a bidirectional promoter region. J. Bacteriol. **182** (2000) 2355.
17. *S.A. Sieber, M.A. Marahiel*, Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. Chem. Rev. **105** (2005) 715.
18. Chapman & Hall dictionary of natural products, Dictionary of Natural Products, http://www.chemnetbase.com/, pristupljeno: 7. 5. 2006.
19. *J. Gonzalez-Lergier, L.J. Broadbelt, V. Hatzimanikatis*, Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways. J. Am. Chem. Soc. **127** (2005) 9930.
20. *J. Staunton, B. Wilkinson*, Combinatorial biosynthesis of polyketides and nonribosomal peptides. Curr. Opin. Chem. Biol. **5** (2001) 159.

21. *C. Khosla, J.D. Keasling*, Metabolic engineering for drug discovery and development. Nat. Rev. Drug Discov. **2** (2003) 1019.

22. *K.J. Weissman, P.F. Leadlay*, Combinatorial biosynthesis of reduced polyketides. Nat. Rev. Microbiol. **3** (2005) 925.

23. *C. Rayssiguier, D.S. Thaler, M. Radman*, The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. Nature **342** (1989) 396.

24. *L. Eichinger, J.A. Pachebat, G. Glockner, M.A. Rajandream, R. Sucgang, M. Berriman, J. Song, R. Olsen, K. Szafranski, Q. Xu, B. Tunggal, S. Kummerfeld, M. Madera, B.A. Konfortov, F. Rivero, A.T. Bankier, R. Lehmann, N. Hamlin, R. Davies, P. Gaudet, P. Fey, K. Pilcher, G. Chen, D. Saunders, E. Sodergren, P. Davis, A. Kerhornou, X. Nie, N. Hall, C. Anjard, L. Hemphill, N. Bason, P. Farbrother, B. Desany, E. Just, T. Morio, R. Rost, C. Churcher, J. Cooper, S. Haydock, N. van Driessche, A. Cronin, I. Goodhead, D. Muzny, T. Mourier, A. Pain, M. Lu, D. Harper, R. Lindsay, H. Hauser, K. James, M. Quiles, M. Madan Babu, T. Saito, C. Buchrieser, A. Wardroper, M. Felder, M. Thangavelu, D. Johnson, A. Knights, H. Loulseged, K. Mungall, K. Oliver, C. Price, M.A. Quail, H. Urushihara, J. Hernandez, E. Rabbinowitsch, D. Steffen, M. Sanders, J. Ma, Y. Kohara, S. Sharp, M. Simmonds, S. Spiegler, A. Tivey, S. Sugano, B. White, D. Walker, J. Woodward, T. Winckler, Y. Tanaka, G. Shaulsky, M. Schleicher, G. Weinstock, A. Rosenthal, E.C. Cox, R.L. Chisholm, R. Gibbs, W.F. Loomis, M. Platzer, R.R. Kay, J. Williams, P.H. Dear, A.A. Noegel, B. Barrell, A. Kuspa*, The genome of the social amoeba *Dictyostelium discoideum*. Nature **435** (2005) 43.

25. *J. Žučko, N. Škunca, T. Curk, B. Zupan, P.F. Long, J. Cullum, R. Kessin, D. Hranueli*. Polyketide synthase genes and the natural products potential of *Dictyostelium discoideum*. Bioinformatics, **23** (2007) 2543.

26. *B.F. Milne, P.F. Long, A. Starcevic, D. Hranueli, M. Jaspars*, Spontaneity in the patellamide biosynthetic pathway. Org. Biomol. Chem. **4** (2006) 631.

27. *W.C. Dunlap, M. Jaspars, D. Hranueli, C.N. Battershill, N. Perić-Concha, J. Žučko, S.H. Wright, P.F. Long*, New methods for medicinal chemistry – universal gene cloning and expression systems for production of marine bioactive metabolites. Curr. Med. Chem. **13** (2006) 697.

28. *D. Hranueli, J. Cullum*, From genes to genomes and beyond, str. 278-284. *U:* V. Lelas *i sur*. (Ured.), Proceedings of the 4th Croatian Congress of Food Technologists, Biotechnologists and Nutritionists (Central European Meeting), Zagreb, 2002.

29. GenBank, http://www.ncbi.nlm.nih.gov/Genbank/, pristupljeno: 23. 6. 2006.

30. *D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M., DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, W. Helmberg, Y. Kapustin, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko*, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 34 (Database issue) (2006) D173.

31. National Institute of Immunology, http://www.nii.res.in/nrps-pks.html, pristupljeno: 23. 6. 2006.

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

33. *S. Markel, D. Leon*, Sequence Analysis in a Nutshell: A Guide to Common Tools and Databases, O'Reilly, Sebastopol, CA, 2003, str 1–302.

34. Pfam, http://www.sanger.ac.uk/Software/Pfam/, pristupljeno: 26. 6. 2006.

35. *A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S.R. Eddy, S. Griffiths-Jones, K.L. Howe, M. Marshall, E.L. Sonnhammer*, The Pfam protein families database. Nucleic Acids Res. **30** (2002) 276.

36. PubChem, http://pubchem.ncbi.nlm.nih.gov/, pristupljeno: 27. 6. 2006.

37. Daylight, Chemical Information Systems Inc., http://www.daylight.com/smiles/, pristupljeno: 20. 7. 2006.

38. *D.R. Westhead, J.H. Parish, R.M. Twyman*, Instant Notes: Bioinformatics. BIOS Sci. Pub. Ltd., Oxford, U.K., (2002) str. 1- 272.

39. *D.W. Mount*, Bioinformatics: sequence and genome analysis. Cold Spring Harbor Laboratory Press, New York, (2001) str. 1- 732.

40. *S.B. Needleman, C.D. Wunsch*, A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48** (1970) 443.

41. *T.F. Smith, M.S. Waterman*, Identification of common molecular subsequences. J. Mol. Biol. **147** (1981) 195.

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

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

44. *N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P.S. Langendijk-Genevaux, M. Pagni, C.J. Sigrist*, The PROSITE database. Nucleic Acids Res. **34** (Database issue) (2006) D227.

45. ChemAxon, http://www.chemaxon.com/, pristupljeno: 5. 7. 2006.

46. OBDA project (BioSQL), http://obda.open-bio.org/, pristupljeno: 5. 7. 2006.

47. PostgreSQL, http://postgresql.org/, pristupljeno: 21. 5. 2006.

48. Licenca BSD, http://opensource.org/search/node/licenses+bsd+license, pristupljeno: 22. 5. 2006.

49. ANSI, Američki nacionalni institut za standardizaciju, http://www.ansi.org/, pristupljeno: 1.5. 2006.
50. pgAdmin, http://www.pgadmin.org/, pristupljeno: 1. 5. 2006.
51. Java Sun, http://www.java.sun.com/, pristupljeno: 12. 7. 2006.
52. Apache Tomcat, http://tomcat.apache.org/, pristupljeno: 2. 5. 2006.
53. Hibernate, www.hibernate.org/, pristupljeno: 2. 5. 2006.
54. DWR, Direct Web Remoting, http://getahead.org/dwr/, pristupljeno: 13. 8. 2006.
55. Javascript.com, http://www.javascript.com/, pristupljeno: 23. 8. 2006.
56. Netscape, www.netscape.com/, pristupljeno: 12. 7. 2006.
57. Firefox Web preglednik, http://www.mozilla.com/, pristupljeno: 12. 11. 2006.
58. World Wide Web Consortium (W3C), http://www.w3.org/, pristupljeno: 1. 12. 2006.
59. Eclipse - an open development platform, http://www.eclipse.org/, pristupljeno: 6. 5. 2006.
60. Jakarta Tomcat web server, http://tomcat.apache.org/, pristupljeno: 7. 5. 2006.
61. JSR 154: Java™ Servlet 2.4 Specification, http://jcp.org/en/jsr/detail?id=154, pristupljeno: 8. 8. 2006.
62. JSR 152: JavaServer Pages™ 2.0 Specification, http://jcp.org/en/jsr/detail?id=152, pristupljeno: 8. 8. 2006.
63. The X.Org Foundation, http://www.x.org/wiki/, pristupljeno: 5.7. 2006.
64. GNOME: The Free Software Desktop Project, www.gnome.org, pristupljeno: 2. 3. 2007.
65. NCBI/BLAST Home, http://www.ncbi.nlm.nih.gov/BLAST/, pristupljeno: 7. 7. 2006.
66. Peptide Sequence Databases, http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#protein_databases, pristupljeno: 3. 7. 2006.
67. BLAST URLAPI, http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/new/, pristupljeno: 21. 9. 2006.
68. The Apache Jakarta Project: HttpClient, http://jakarta.apache.org/httpcomponents/httpclient-3.x/, pristupljeno: 3. 2. 2007.

69. The Castor Project, http://www.castor.org/, pristupljeno: 3. 2. 2007.

70. *B. Eckel*, Thinking in Java, 4 edition. Prentice Hall PTR; 2006, str 1–1150.

71. *J. Rumbaugh, I. Jacobson, G. Booch*, The Unified Modeling Language Reference Manual (Second Edition). Addison-Wesley, Boston, 2005, str. 1-496.

72. *G. Yadav, R.S. Gokhale, D. Mohanty*, Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. J. Mol. Biol. **328** (2003) 335.

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

74. *P. Caffrey*, Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. ChemBioChem **4** (2003) 654.

75. *A. Starcevic, J. Cullum, M. Jaspars, D. Hranueli, P.F. Long*. Predicting the nature and timing of epimerisation on a modular polyketide synthase. ChemBioChem, **8** (2007) 28.

76. *A.T. Keatinge-Clay*. A tylosin ketoreductase reveals how chirality is determined in polyketides. Chemistry & Biology, **14** (2007) 898.

77. *F. Csizmadia*, JChem: Java applets and modules supporting chemical database handling from web browsers. J. Chem. Inf. Comput. Sci. **40** (2000) 323.

78. *P. Shen, H.V. Huang*, Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics **112** (1986) 441.

79. *V.V. Poroikov, D.A. Filimonov*, How to acquire new biological activities in old compounds by computer prediction. J. Comput. Aid. Mol. Des. **16** (2002) 819.

80. *R.J. Zauhar, G. Moyna, L. Tian L, Z. Li, W.J. Welsh*, Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. J. Med. Chem. **46** (2003) 5674.

81. Ecopia BioSciences Inc., http://www.ecopiabio.com/, pristupljeno 1. 7. 2006.

82. NRPSpredictor, http://www-ab.informatik.unituebingen.de/software, pristupljeno 24. 5. 2007.

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

84. Prediction of Activity Spectra for Substances, http://195.178.207.233/PASS/, pristupljeno: 24. 5. 2007.

8.1. Popis korištenih kratica u radu

A - domena za adenilaciju aminokiselina (NRPS)

ACP - mali polipeptid nosač acila (PKS)

ACV - L-alfa-amino adipil-L-cisteinil-D-valin - preteča penicilina

AJAX - metoda AJAX (akronim za "Asynchronous JavaScript and XML")

AT - domena aciltransferaze (PKS)

ave - gen čiji proteinski produkt sudjeliju u biosintezi poliketida avermektina

BLAST - bioinformatički računalni program BLAST (akronim za "Basic Local Alignment Search Tool")

C - domena za kondenzaciju (NRPS)

CDA - lipopeptidni, o kalciju ovisan, antibiotik

CDD - računalom vođen dizajn lijekova (akronim za "Computer added Drug Design")

Cy - domena za formiranje heterocikličkih prstenova (NRPS)

CSS - računalni jezik CSS (akronim za "Cascading Style Sheets")

DDBJ - Japanska baza podataka sekvencija DNA (akronim za "DNA Database of Japan")

DEBS - sintaza deoksieritronolida B (akronim za "DeoxyErythronolide B Synthase")

DH - domena dehidrataze (PKS)

DHTML - dinamički HTML (akronim za "Dynamic HTML")

DMBS - sustav za upravljanje bazama podataka DMBS (akronim za "DataBase Management System")

DNA - deoksiribonukleinska kiselina

DOM - model DOM (akronim za "Dokument Object Model")

DWR - biblioteka otvorenog kôda DWR (akronim za "Direct Web Remoting")

E - očekivana vrijednost

E_{pim} - domena za epimerizaciju (NRPS)

EMBL - Europska baza podataka sekvencija DNA (akronim za "EMBL Nucleotide Sequence Database")

ER - domena enoilreduktaze (PKS)

HTTP - protokol HTTP (akronim za "Hypertext Transfer Protocol")

IDE - integrirana razvojna okolina (akronim za "Integrated Develpment Environment")

JSON - oblik zapisa JSON (akronim za "JavaScript Object Notation")

JSP - Java tehnologija JSP (akronim za "Java Server Page")

KR - domena ketoreduktaze (PKS)

KS - domena ketosintaze (PKS)

HTML - računalni jezik HTML (akronim za "Hypertext Markup Language")

JDBC - Java tehnologija JDBC (akronim za "Java Database Connectivity")

MEPS - minimalna dužina apsolutne identičnosti u kojoj dolazi do homologne rekombinacije (akronim za "Minimal Essential Pairing Sequence")

MT - domena za *N*-, *C*- ili *O*-metilaciju (NRPS)

NCBI - Nacionalni Centar za Biotehnološke Informacije, Washington, SAD

NRPS - sintetaza neribosomalno sintetiziranih peptida

PASS - računalni alat PASS (akronim za "Prediction of Activity Spectra for Substances")

PCP - mali polipeptid nosač peptidila (NRPS)

PKS - poliketid sintaza

PKSDB - baza podataka modularnih poliketid sintaza (akronim za "Database of Modular Polyketide Synthases")

PKS/NRPS - mješovita poliketid sintaza i sintetaza neribosomalno sintetiziranih peptide

R - domena reduktaze (NRPS)

rap - gen čiji proteinski produkt sudjeliju u biosintezi poliketida rapamicina

SMILES - oblik izomeričkog zapisa (akronim za "Simplified Molecular Input Line Entry System")

Te - domena tioesteraze (NRPS)

TE - domena tioesteraze (PKS)

UML - računalni jezik UML (akronim za "Universal Markup Language"

WTP - dodaci za razvoj Web aplikacija (akronim za "Web Tools Platform)

XHTML - računalni jezik XHTML (akronim za "Extensible HyperText Markup Language")

XML - računalni jezik XML (akronim za "Extensible Markup Language")

8.2. Sadržaj kompakt diska

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

1. Primjer zapisa u obliku "DBsource" iz baze podataka GenBank.

2. Primjer cijelovitog izvješća bioinformatičkog brograma BLAST oblikovanog računalnim programom za prikupljanje bioloških podataka.

3. Programski kôd cjelovitog računalnog programa za prikupljanje, strukturiranje, pohranu i grafički prikaz podataka.

Gotovi seminarski, maturski, maturalni i diplomski radovi iz raznih oblasti, lektire , puškice, tutorijali, referati - specijalizovan tim za usluge visokokvalitetnog pisanja, istraživanja i obradu teksta za kompletan region Balkana.

Posetite nas na sajtovima ispod:

WWW.MATURSKIRADOVI.NET

WWW.SEMINARSKIRAD.ORG

WWW.MATURSKI.NET

WWW.MATURSKI.ORG

WWW.SEMINARSKIRAD.INFO

Dostupni smo Vam 24h 365 dana u godini.

Za gotove verzije rada obratiti se na mail:

maturskiradovi.net@gmail.com

061/ 11-00-105

Seminarski, diplomski, maturski radovi, prevodi na engleski i eseji…

.