Oblikovanje novih prirodnih spojeva u uvjetima *in silico**

KUI – 11/2008
Prispjelo 21. veljače 2007.
Prihvaćeno 18. svibnja 2007.

D. Hranueli,[a] A. Starčević,[a,d] J. Žučko,[b,d] J. Diminić,[a] N. Škunca,[a] V. Željeznak,[a] D. Kovaček,[a] D. Pavlinušić,[b] J. Šimunković,[b] P. F. Long[c] i J. Cullum[d]

[a] Prehrambeno-biotehnološki fakultet, Pierottijeva 6, 10 000 Zagreb, Hrvatska
[b] Novalis d.o.o., Božidara Adžije 17, 10 000 Zagreb, Hrvatska
[c] Sveučilište u Londonu, 29/39 Brunswick Square, London WC1N 1AX, Velika Britanija
[d] Sveučilište u Kaiserslauternu, Poštanski pretinac 3049, D–67653 Kaiserslautern, Njemačka

Poliketidi i neribosomski 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 ili sintetaza neribosomskih peptida. U posljednjem desetljeću 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 te opće nedovoljne identičnosti različitih modula. Osim toga, homolognom bi se rekombinacijom povećala učestalost rekombinacije te potaknula kombinatorna raznolikost rekombinanata. U tijeku je razvoj integralnih računalnih programskih paketa, ***CompGen*** i ***ClustScan***, za modeliranje tih procesa u uvjetima *in silico*. Okosnica je programskog paketa ***CompGen*** specifično strukturirana baza podataka koja povezuje biosintetski put sinteze sa sekvencijama DNA genskih nakupina. Povezanost sekvencija DNA s biosintetskim putem omogućuje njezinu povezanost sa strukturom produkta. Jedna je od funkcija računalnoga programskog paketa, temeljena na toj povezanosti, sposobnost oblikovanja virtualnih rekombinanata između genskih nakupina. To se obavlja pomoću modela rekombinacije da bi se *in silico* predvidjele sekvencije DNA u kojima dolazi do homologne rekombinacije. ***CompGen*** iz tih podataka predviđa kemijsku strukturu nove supstancije i sprema je u bazu podataka virtualnih kemijskih struktura radi daljnjega molekulskog modeliranja. Računalni programski paket omogućuje i analizu tzv. 'obrnutom genetikom'. Naime, ako se pretpostavi poželjna kemijska struktura, program može predvidjeti kako bi trebala izgledati genska nakupina poliketid-sintazâ ili sintetaza neribosomskih peptida, koja bi sintetizirala takvu strukturu na temelju građevnih jedinica genskih nakupina u bazi podataka. U cjelini, ***CompGen*** će omogućiti oblikovanje baze podataka novih prirodnih kemijskih entiteta u uvjetima *in silico*, koja se zatim može upotrijebiti za istraživanja na području tehnologije računalnoga dizajna novih lijekova. Drugi će integralni generički računalni programski paket, ***ClustScan***, moći prepoznati i anotirati nove genske nakupine iz sekvencija cjelovitih mikrobnih genoma ili genskih nakupina u metagenomima mikroorganizama koji žive u tlu ili u simbiozi s morskim organizmima.

Ključne riječi: *Poliketidi i neribosomski sintetizirani peptidi, modularne genske nakupine PKS i NRPS, rekombinacija, baza podataka, računalni programski paket*

Uvod: biološke i kemijske osnove

Mikroorganizmi i biljke sintetiziraju velik 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.[1,2] 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 dvadeset tisuća antibiotički aktivnih supstancija, izoliranih od bakterija do sisavaca, streptomicete sintetiziraju gotovo 40 %. Štoviše, vrste roda *Streptomyces* proizvode 75 % svih klinički važnih antibiotika.[3] Nedavno su *Watve i sur.* pokušali matematički prosuditi broj antibiotika koje ove bakterije mogu sintetizirati. Njihov je model pokazao da je taj broj oko 100 000,[4] od kojih je manje od 10 % do sada otkriveno. Svaka pojedina vrsta roda *Streptomyces* može sintetizirati više od jednog sekundar-

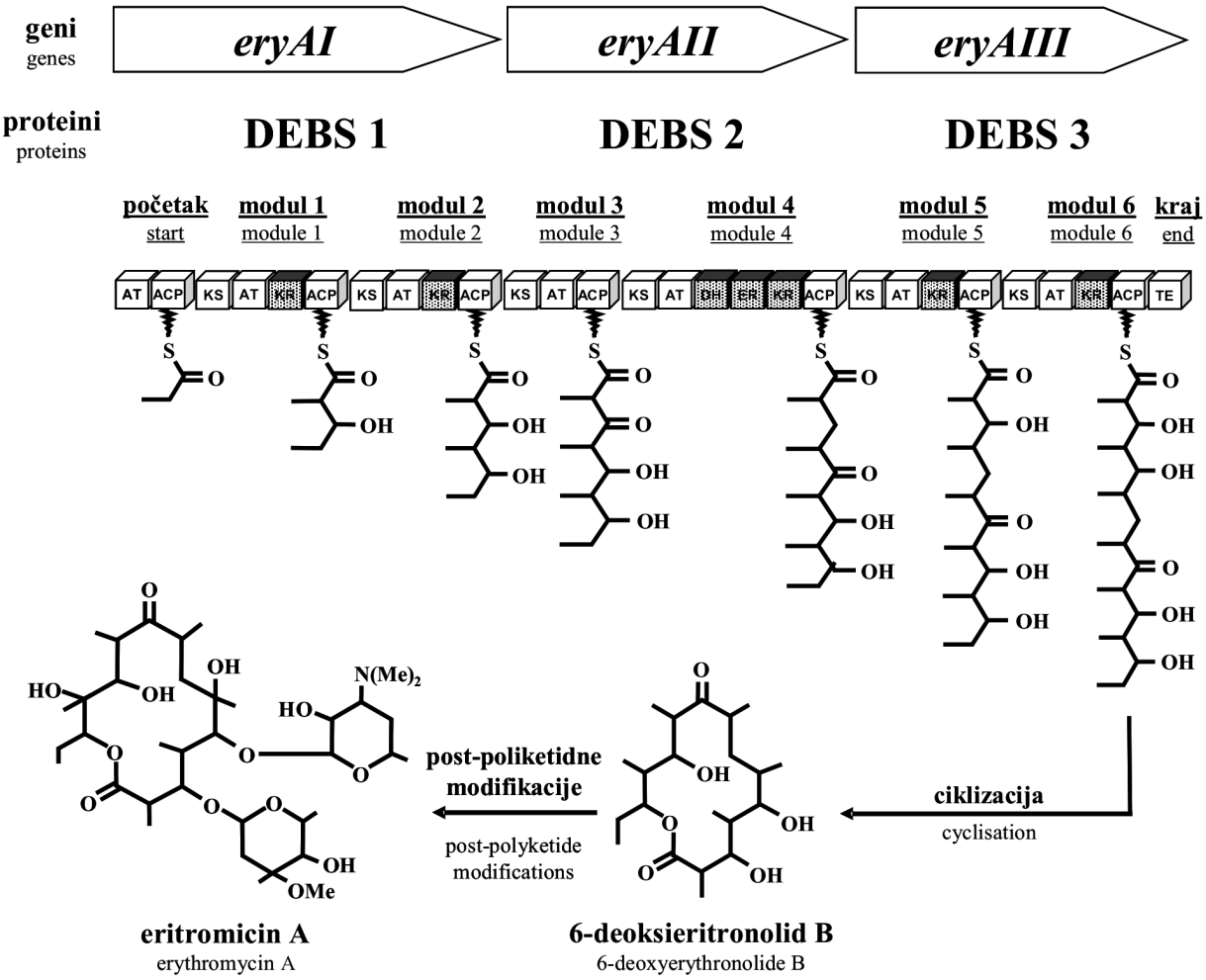
* Podaci navedeni u ovom revijalnom prikazu priopćeni su na Jubilarnom XX. hrvatskom skupu kemičara i kemijskih inženjera koji je održan u Zagrebu od 26. veljače do 1. ožujka 2007. godine.

nog metabolita. Tako, na primjer, bakterija *S. coelicolor* sintetizira najmanje četiri antibiotika: aktinorodin, undecilprodigiosin, metilenomicin i lipopeptidni antibiotik CDA. Štoviše, sekvenciranje je genomskih DNA bakterija *S. coelicolor*, *S. avermitilis* i *Saccharopolyspora erythraea* pokazalo da svaka od tih bakterija sadržava 18, 35 i 25 genskih nakupina za biosintezu sekundarnih metabolita.[5,6,7]

Ti sekundarni metaboliti nisu samo osebujni 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, katkad izmijenjene polusintetskim ili sintetskim postupcima.[1] Osim antibiotika, fungicida, antivirusnih sredstava i citostatika, u kliničkoj se primjeni nalaze i imunosupresori, antihipertenzivna sredstva, antidijabetici, antimalarici te antikolesterolemici.[8,9] Komercijalno su važni i proizvodi za agroindustriju, kao što su 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. S 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. Njihova se struktura, međutim, ne može predvidjeti iz sekvencije DNA gena koji 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. Unatoč strukturnoj raznolikosti, poliketidi se sintetiziraju pomoću enzimskih kompleksa nazvanih poliketid-sintaze (PKS) (slika 1), 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. Kao produžne se građevne jedinice mogu upotrijebiti samo oni ostaci organskih kiselina koji nakon vezivanja na koenzim A imaju slobodnu karboksilnu skupinu. Nasuprot tome, početne građevne jedinice ne moraju zadovoljavati taj uvjet i kemijski mogu biti vrlo različite.[9]

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.[10,11,12] Plijesni poliketide sintetiziraju pomoću monomodularnih ponavljajućih PKS tipa I i monomodularnih ponavljajućih PKS tipa III.[13,14] Sekvenciranje genskih nakupina ponavljajućih PKS tipa II, kao što je to PKS aromatskoga poliketidnog antibiotika aktinorodina, upućuje na postojanje 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.[9] Isto vrijedi i za monomodularne ponavljajuće PKS tipa I i tipa III.[13,14] 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 *Sacc. erythraea* (slika 1; vidi[9]), pokazalo da su to multifunkcionalni enzimi multimodularne organizacije. Svaki je modul odgovoran za rast poliketidnog lanca za jednu produžnu građevnu jedinicu, a može sadržavati i domene potrebne za redukciju β-ugljika. Moduli su najčešće raspoređeni na nekoliko polipeptida, od kojih svaki sadržava više od jednog modula. Unutar modula PKS genskih nakupina nalazi se niz domena koje provode različite stupnjeve procesa biosinteze. Tako na primjer acil-transferaze (AT) dovode građevne jedinice na male polipeptide nosače acila (ACP). Keto-sintaze (KS) prenose rastuće poliketidne lance 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 postoje keto-reduktaze (KR), keto-skupine se reduciraju u hidroksilne-skupine. Ako uz KR postoje i dehidrogenaze (DH), hidroksilne će skupine biti reducirane u alkene. Prisutnost će enoil-reduktaza (ER), uz KR i DH domene, 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 PKS eritromicina završava domenom tioesteraza (TE) odgovornom za odvajanje poliketidnog ugljikova lanca od PKS i ciklizaciju, pri čemu nastaje poliketidni aglikon 6-deoksieritronolid B. Nakon toga, takozvani post-poliketidni enzimi, kao što su hidroksilaze i glikozilaze, sudjeluju u sintezi makrolidnog antibiotika eritromicina A.[9]

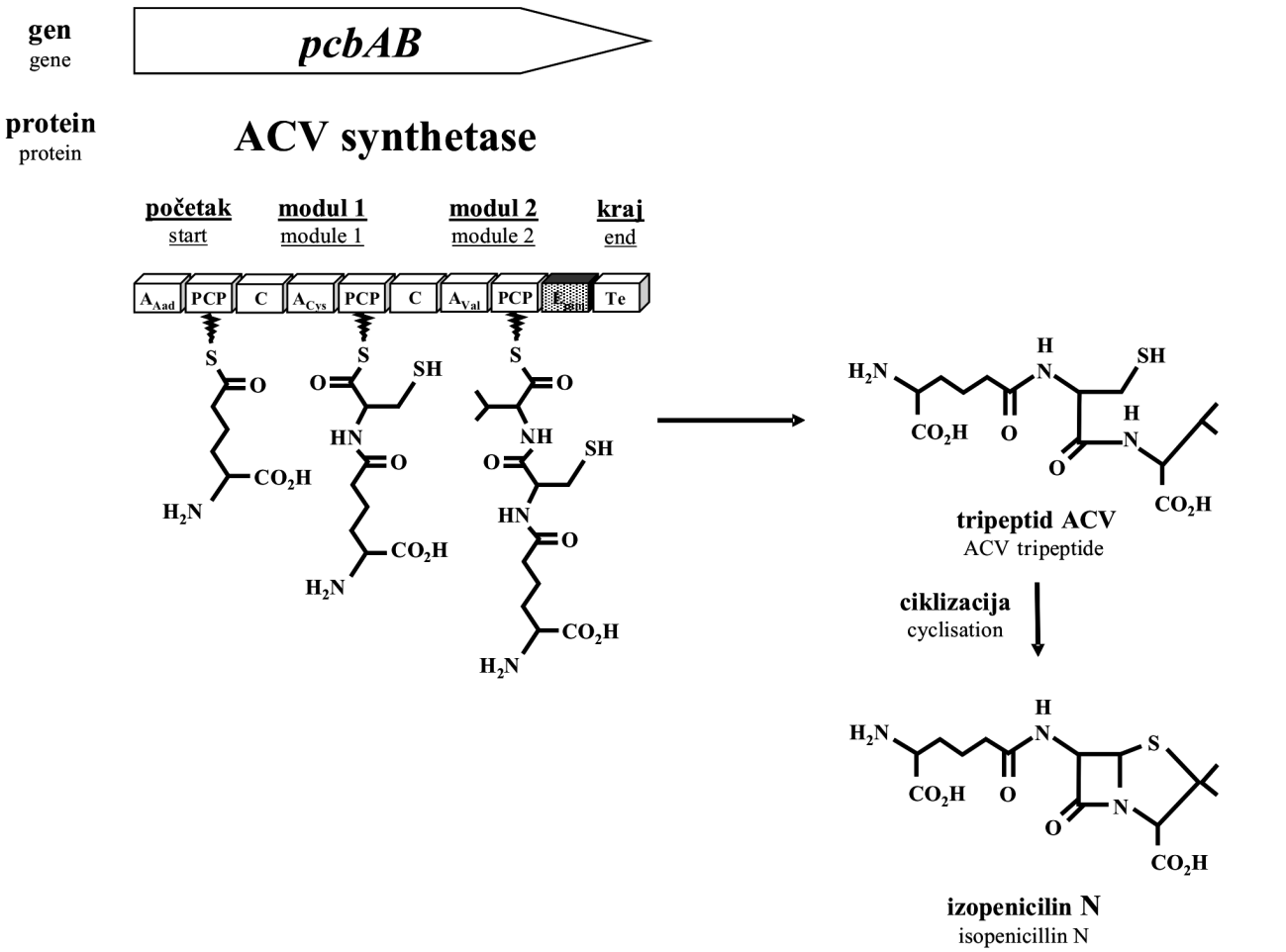
Neribosomski sintetizirani peptidi su, osim poliketida, druga velika skupina sekundarnih metabolita. Među neribosomskim peptidima također se nalaze poznati antibiotici (npr. tripeptid ACV, preteča penicilina), fungicidi (npr. fengicin) i imunosupresori (npr. ciklosporin). Osim toga opisane su i miješane poliketidno/peptidne supstancije poput citostatika bleomicina i epotilona. Oni se sintetiziraju pomoću sintetaza neribosomski sintetiziranih peptida (NRPS), kao što je NRPS peptidnog antibiotika penicilina u plijesni *Penicillium chrysogenum* (slika 2; vidi[15]). 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 kondenzaciju (C), adenilaciju aminokiselina (A) 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 2), redukciju (R), *N*-, *C*- ili *O*-metilaciju (MT) i druge. Na kraju se posljednjeg modula, kao i u enzima PKS, najčešće 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 neribosomski sintetiziranih peptida kao građevne jedinice mogu poslužiti i brojne druge izmijenjene aminokiseline (oko 200 različitih aminokiselina). Prema tome, biosintetski im je potencijal još raznovrsniji od poliketid-sintaza.[16]



S l i k a 1 – *Shematski prikaz organizacije gena i proteina multimodularne 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 aciltransferaze (AT), male polipeptide, nosače acila (ACP), β-ketoacil-sintaze (KS), keto-reduktaze (KR), dehidratazu (DH) enoil-reduktazu (ER) i tioesterazu (TE)] te poliketidnog aglikona 6-deoksieritronolida B (6-dEB) i makrolidnog antibiotika eritromicina A. Osjenčane su domene koje sudjeluju u redukciji β-ugljika (KR, DH i ER). (Preinačeno prema citatu[9] s dopuštenjem izdavača)*

F i g. 1 – *Schematic representation of genes and proteins of the type I multi-modular polyketide synthase responsible for the biosynthesis of erythromycin A. The linear arrangement of genes (*ery*AI, *ery*AII and* ery*AIII) corresponds to the order of proteins (DEBS1, DEBS2 and DEBS3) in biosynthesis. Domains for acyltransferases (AT), acyl carrier proteins (ACP), β-ketoacyl synthases (KS), ketoreductases (KR), dehydrase (DH), enoyl reductase (ER) and thioesterase (TE) are shown as well as the polyketide aglycon 6-deoxyerythronolide B (6-dEB) and the final product, the macrolide antibiotic erythromycin A. Domains that are involved in β-carbon reduction are shadowed. (Modified from reference[9] with permission of the publisher)*

U prirodi su tijekom evolucije već upotrijebljena svojstva kombinatorne biosinteze i sintetizirano je preko 170 000 poznatih prirodnih spojeva (vidi: "Chapman & Hall dictionary of natural products"[17]). Do danas izolirani prirodni poliketidi/peptidi čine, međutim, samo mali dio kombinatornog potencijala koji se može postići permutacijama gena PKS/NRPS. Na primjer, ako se u šest modula PKS eritromicina permutira svih sedam acil-transferaza i svih pet 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 (vidi sliku 4) dovela do sinteze nevjerojatnog broja od 10^{14} poliketida.[9,18] Zbog toga su znanstvenici u složenim poliketidima, 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 uklanjanja cijelih modula multimodularnih gena PKS ili se upotrebljava takozvana kemijska biosinteza, radi sinteze novih makrolida s različitom veličinom ugljikovih kostura.[9,19,20] 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.[21] 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[22] rekombinacijom (slika 3) koja nastaje nakon prirodnog prijenosa genetičkoga materijala konjugacijom ili transformacijom u bakterija. Vjeruje se će taj pristup omogućiti rekombinaciju između sekvencija DNA, koje se i ina-



S l i k a 2 – *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 peptidne preteče ACV i peptidnog antibiotika izopenicilina N. Osjenčana je domena koja sudjeluje u epimerizaciji građevne jedinice (Epim)*

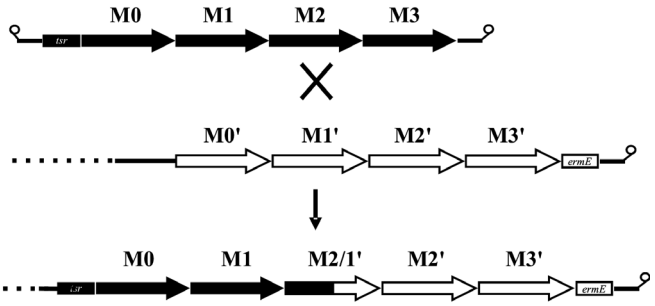
F i g. 2 – *Schematic representation of the gene and protein of the multi-modular peptide synthetase responsible for the biosynthesis of the penicillin precursor (ACV). Linear arrangement of the gene* (pcb*AB*) *and protein (AVC synthetase), that contains catalytic domains for adenylations (A), peptidile carrier proteins (PCP), condensations (C) and thioesterase (Te) are shown as well as the peptide precursor ACV and the final product, the peptide antibiotic isopenicillin N. The domain that is involved in the epimerisation of the building block (Epim) is shadowed*

če događaju u prirodi. Tako bi se smanjila mogućnost da nastanu nefunkcionalna neprirodna čvorišta, kao tijekom rekombinacije u uvjetima *in vitro*.[11]

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 detaljno opisati (anotirati) genske nakupine PKS/NRPS, smještene u genomima nedovoljno dobro opisanih organizama (npr. u genomu amebe *Dictyostelium discoideum*[23]), kao i genske nakupine PKS/NRPS u metagenomima mikroorganizama koji žive u tlu ili u simbiozi s morskim organizmima.[24,25]

Računalna platforma za oblikovanje u uvjetima *in silico*

Uz podršku tvrtke Novalis d. o. o., u tijeku je razvoj generičkih računalnih programskih paketa ***CompGen*** i ***ClustScan*** (akronimi za "***Com****pound* ***Gen****erator*" i "***Clust****er* ***Scan****ner"), napisanih objektno orijentiranim računalnim jezikom (poput npr. Java[26,27]). Ti će programski paketi (koje će biti dostupni posredstvom tvrtke Novalis d. o. o., Zagreb, Hrvatska) omogućiti da se do sada opisani procesi modeliraju u uvjetima *in silico*. Temeljni cilj programskog paketa ***CompGen*** jest strukturiranje baze podataka potpuno novih, u prirodi nepoznatih, virtualnih kemijskih entiteta, počevši od sekvencija DNA poliketidnih/peptidnih genskih nakupina koje se nalaze u javnim bazama poput baze podataka GenBank.[28] U javnim bazama podataka[29,30] 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 bazi podataka GenBank već danas ima nekoliko stotina i njihov broj eksponencijalno se povećava sa svakim novim sekvenciranim bakterijskim genomom.[28]



S l i k a 3 – *Predložena strategija za homolognu rekombinaciju između dvaju poliketidnih genskih nakupina u vrsta roda Streptomyces. Jedna bi genska nakupina (⟹) trebala biti klonirana na linearni plazmidni vektor koji sadržava genetički biljeg za otpornost prema antibiotiku tiostreptonu (tsr). Druga bi genska nakupina (⟹) trebala biti klonirana u blizini kraja linearnog kromosoma koji sadržava genetički biljeg za otpornost prema antibiotiku eritromicinu (ermE). Nova bi poliketidna genska nakupina (⟹) nastala jednostrukim premošćenjem između dviju genskih nakupina. Ona bi se selekcionirala prijenosom rekombinantnoga plazmida, koji sadržava oba genetička biljega, u drugu vrstu roda Streptomyces (Preinačeno prema citatu[11] dopuštenjem izdavača.)*

F i g. 3 – *Suggested strategy to obtain homologous recombination between polyketide gene-clusters in Streptomyces. One gene-cluster (⟹) should be cloned on a linear plasmid with an adjacent thiostrepton resistance gene (tsr). The other gene-cluster (⟹) should be cloned near one end of the linear chromosome with an adjacent erythromycin resistance gene (ermE). A single cross over between clusters to generate the novel polyketide gene-cluster (⟹) can be selected by transfer of the recombinant plasmid carrying both resistance markers to a second Streptomyces strain (Modified from reference[11] with permission of the publisher.)*

Struktura baze podataka

Upotrebom notacijskog jezika UML (Universal Markup Language[31]) u tijeku je izrada, o platformi neovisnih, generičkih programskih paketa ***CompGen*** i ***ClustScan*** za automatsko prikupljanje, pohranu i obradu podataka o poliketidima, neribosomskim peptidima i ostalim produktima modularnih biosintetskih putova iz javno dostupnog izvora baze podataka GenBank.[28] 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.[32] 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 (akronim za "**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem"[33]), dvodimenzionalne (2D) i trodimenzionalne (3D) strukture poliketidnih lanaca te konačne cikličke kemijske strukture poliketidnih/peptidnih okosnica, tj. aglikona.

Računalni program za prikupljanje, anotaciju i pohranu bioloških podataka

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.[34] Te su primarne sekvencije aminokiselina modelnih domena pripremljene upotrebom baze podataka Pfam[35] i funkcije 'hmmemit' bioinformatičkoga programskog paketa HMMER.[36] One se upotrebljavaju kao sekvencije-upiti za pretraživanje baze podataka GenBank bioinformatičkim programom BLAST.[37] BLAST je akronim za "**B**asic **L**ocal **A**lignment **S**earch **T**ool" i upotrebljava se za brzo pretraživanje mogućih lokalnih poravnanja sekvencija-upita sa sekvencijama iz javnih baza podataka. Pri stupnjevanju pogodaka rabi se statistička analiza za prosudbu važnosti svakog pogotka koja se izražava kao 'očekivana vrijednost', to jest vrijednost E ('expected value'). ***CompGen*** zatim provodi pretragu na *web*-poslužitelju Nacionalnog centra za biotehnološke informacije (NCBI), koji održava bazu podataka GenBank.[28]

Izvještaj programa BLAST, oblikovan programskim paketom ***CompGen***, sadržava prikaz 100 zapisa baze podataka GenBank. Identičnost, sličnost ili različitost podataka izražena je vrijednošću E i bodovnim stanjem poravnanja ('score'). Što je vrijednost E manja, a bodovno stanje poravnanja veće, to je pronađena sekvencija proteina 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 udovoljava, 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'.[28] Prije nego što ga spremi, korisnik može još jednom pogledati detalje zapisa na *web*-poslužitelju Nacionalnog centra za biotehnološke informacije.

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[35] i bioinformatičkoga programskog paketa HMMER.[36] Na temelju vlastitih proteinskih profila domena PKS/NRPS i programskog paketa HMMER ***CompGen*** prepoznaje i grafički prikazuje domene u sekvenciji aminokiselina u proteinima. U pozadini toga grafičkog prikaza nalaze se podaci gdje svaka točka na zaslonu (piksel) predočuje 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. Nakon toga, pomoću opcije 'kreiraj biosintetski put', ***CompGen*** poreda proteine, module i domene u njihov biosintetski redoslijed.

Katalitički je aktivne domene potrebno posebno podrobno anotirati radi predviđanja njihove aktivnosti/neaktivnosti i specifičnosti. Programski paket ***CompGen*** mora moći prepoznati da li s domene PKS/NRPS sustava, prisutne unutar genskih nakupina, aktivne ili neaktivne. Naime, iz podrobno 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).[30] Zbog toga programski paket ***CompGen*** mora prvo moći analizirati i razlikovati sekvencije aminokiselina različitih domena po njihovoj duljini, 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 'hmmemit' bioinformatičkoga programskog paketa HMMER.[36]

Predviđanje specifičnosti domena AT/A za supstrate (građevne jedinice) koji se ugrađuju u poliketidne/peptidne lance te specifičnosti domena KR za modifikaciju (stereospecifičnost skupina –OH i $–CH_3$) 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[38–41] te funkcije 'hmmbuild' bioinformatičkoga programskog paketa HMMER.[36] Ti će profili pripremljeni odvojeno za C2, C3 i druge poliketidne građevne jedinice, a i za različite aminokiseline kao građevne jedinice neribosomski sintetiziranih peptida, poslužiti za poravnanje nepoznatih domena AT/A i KR upotrebom funkcije 'hmmalign' bioinformatičkoga programskog paketa HMMER.[36] 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.[38–41]

Računalni programski paket ***CompGen*** sadržava i bazu podataka početnih i produžnih građevnih jedinica u obliku izomeričkih zapisa SMILES.[33] 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 program ChemAxon,[42] za pripremu novog zapisa SMILES. 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 (2D) ili trodimenzionalnom (3D) obliku.

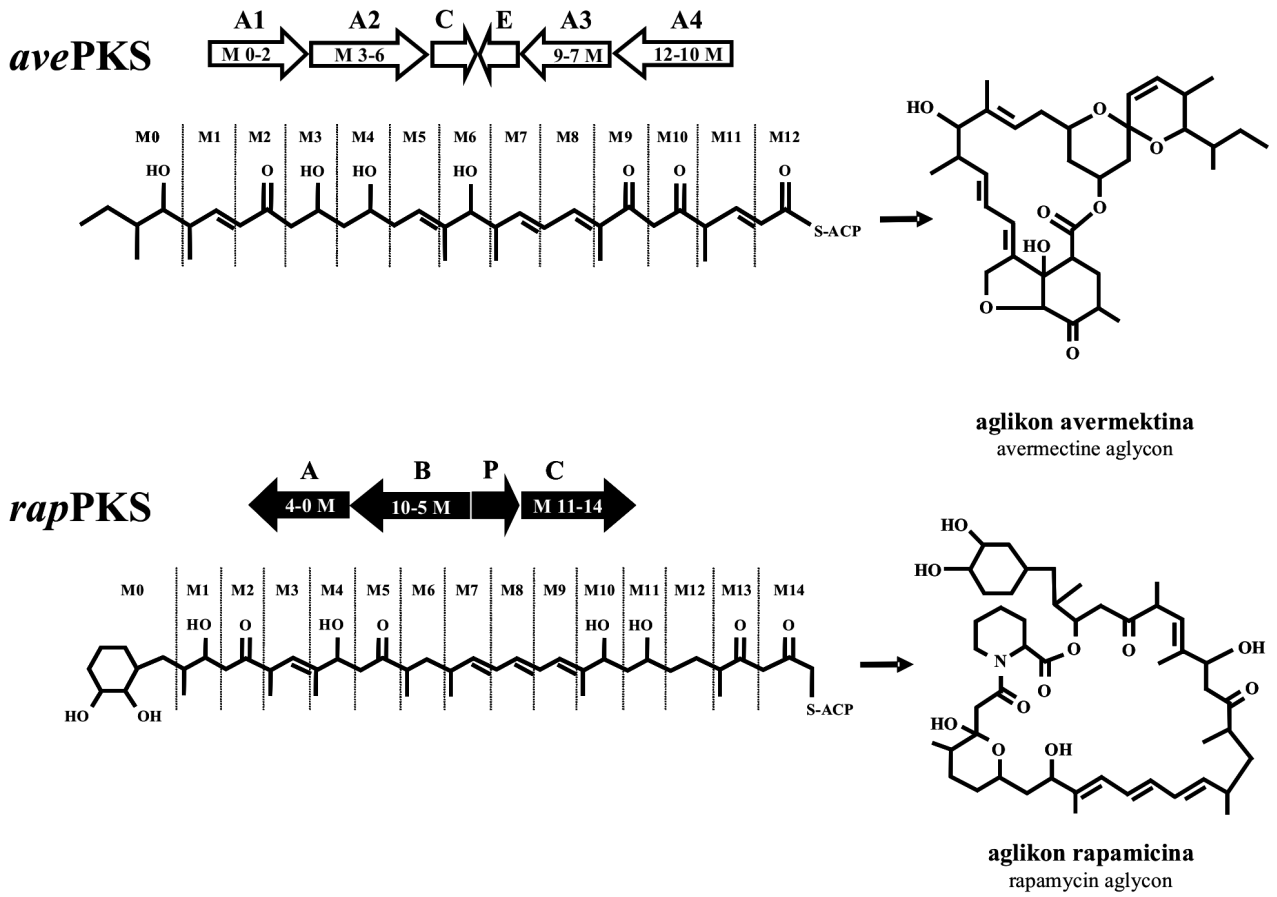
Specijalizirani računalni programi

Sljedeći važan računalni izazov zastupa molekulsko modeliranje procesa ciklizacije linearnih novih nepoznatih poliketidnih/peptidnih lanaca u cikličke strukture. Za to modeliranje upotrebljavamo poznate poliketidne/peptidne lance i poznate aglikone.[30] U postupku molekulskoga modeliranja procesa ciklizacije linearnih poliketidnih lanaca u prvom je stupnju potrebno generirati sve potencijalne stereoizomere linearnih poliketidnih lanaca. Različiti stereoizomeri nastaju kao posljedica nedovoljnog poznavanja stereospecifičnosti biosintetskoga puta. Sljedeći stupanj je pretvorba 2D u 3D kemijsku strukturu.[42] Tek nakon toga slijedit će opsežna konformacijska analiza primjenom metoda molekulske mehanike i kvantne mehanike. Konformacijska će analiza biti izrađena primjenom programskoga paketa Spartan,[43] upotrebom metode MonteCarlo. Do sada se postupak sustavnog pretraživanja konformacijskog prostora pokazao neprikladnim s obzirom na utrošeno računalno vrijeme (1000 puta veći utrošak računalnog vremena). Potrebno je istaknuti da su se rezultati konformacijske analize, dobiveni primjenom molekulske mehanike, pokazali neprikladnim u usporedbi s konformacijskom analizom temeljenom na semiempirijskom kvantno-kemijskom pristupu AM1.[44] Taj pristup, za razliku od molekulske mehanike, predviđa konformere s manjim energijama, koji su se nakon analize geometrijskih parametara pokazali kao potencijalni kandidati za reakcije ciklizacije. Pri tome treba biti svjestan ograničenja semiempirijskih postupaka. Zbog toga je prijeko potrebna daljnja analiza reakcija ciklizacije na kvantno-kemijskoj razini *ab initio*.[45] Tek će se nakon rezultata dobivenih na razini *ab initio* pristupiti razvoju modela[46] čija će se svojstva moći opisati jednadžbom: $A \cdot S \approx$ konstanta.[47] Ta jednadžba podsjeća na Heisenbergov princip neodređenosti,[48] gdje je S veličina molekule, dok A ukazuje na točnost računa. Ako je vrijednost S veća, manja je točnost i vjerodostojnost rezultata. U teorijskom pristupu velikih molekulskih sustava nastoji se pronaći najbolja aproksimativna metoda koja će biti najbolji mogući kompromis između jednostavnosti i primjenjivosti s jedne te kvalitete rezultata s druge strane. Zbog toga se odabiru metode poklanja velika pozornost. Kao što je to već istaknuto, prikladnost metode provjeravat će se na nizu dobro opisanih poliketida/peptida[30] za koje, radi usporedbe, postoje odgovarajući eksperimentalni podaci.

Također je u tijeku i modeliranje procesa homologne rekombinacije[22] između sekvencija DNA dviju poliketidnih/peptidnih genskih nakupina. U prirodi proces rekombinacije zahtijeva minimalnu duljinu apsolutne identičnosti dviju sekvencija DNA. Minimalna se duljina apsolutne identičnosti, MEPS ('**M**inimal **E**ssential **P**airing **S**equence')[49] 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. Specijalizirani računalni program koji pronalazi identičnost i sličnost sekvencija DNA već je napisan računalnim jezikom Perl.[50] U tijeku je 'prevođenje' toga programa u računalni jezik Java,[26,27] da bi ***CompGen*** mogao grafički prikazati roditeljske genske nakupine i nove genske nakupine dobivene rekombinacijom.

Kao primjer prikazani su rezultati programskog paketa ***CompGen*** dobiveni upotrebom modeliranja procesa rekombinacije između genskih nakupina koje sadržavaju genetičku uputu za biosintezu poliketidnog antiparazitika avermektina[51] i poliketidnog imunosupresora rapamicina.[52] Genska nakupina za biosintezu avermektina sadržava četiri gena s 12 modula i u poliketidni lanac ugrađuje 13 građevnih jedinica. Nasuprot tome, genska nakupina za biosintezu rapamicina sadržava tri gena s 14 modula i u poliketidni lanac ugrađuje 15 građevnih jedinica (slika 4).

Specijalizirani računalni program za modeliranje procesa rekombinacije, napisan računalnim jezikom Perl,[50] predviđa 30 mogućih rekombinanata, to jest 30 novih genskih



S l i k a 4 – *Shematski prikaz organizacije gena multimodularnih poliketid sintaza avermektina (ave*PKS*) i rapamicina (rap*PKS*). Prikazan je linearni raspored gena ave*PKS *(A1-A2-C-E-A3-A4) i rap*PKS *(A-B-P-C), modula ave*PKS *i rap*PKS *(M0-M12 i M0-M14), linearnih poliketidnih lanaca te aglikona avermektina i rapamicina*

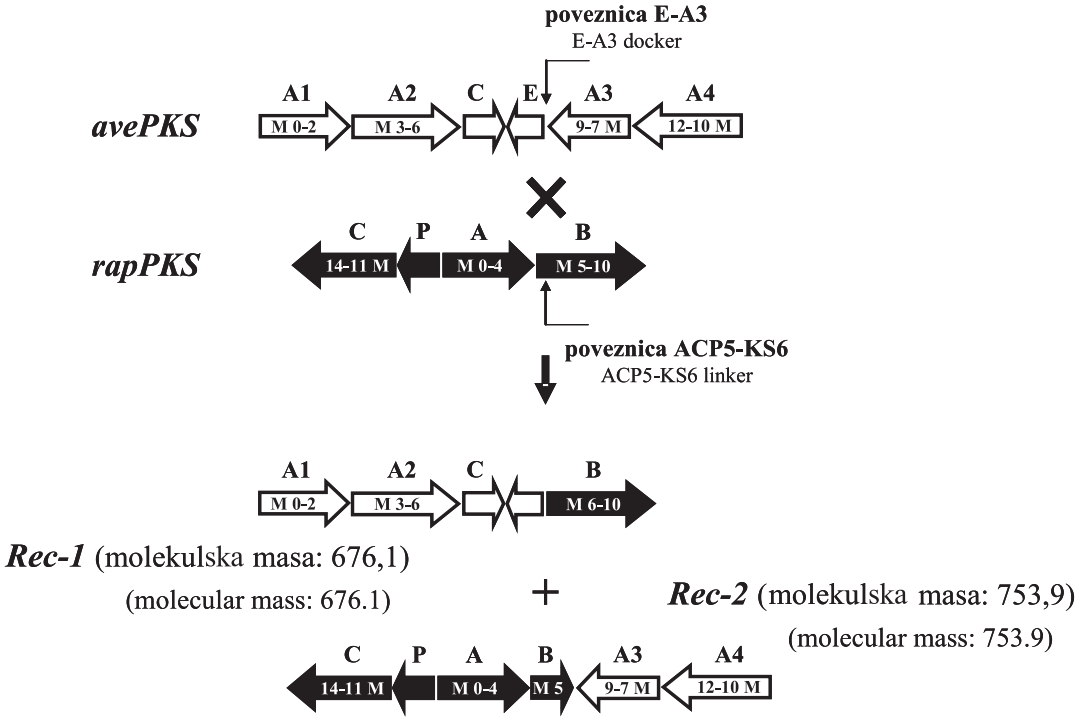
F i g. 4 – *Schematic representation of genes of the multi-modular avermectin (ave*PKS*) and rapamycin (rap*PKS*) polyketide synthases. The linear arrangement of ave*PKS *(A1-A2-C-E-A3-A4) and rap*PKS *(A-B-P-C) genes, ave*PKS *and rap*PKS *modules (M0-M12 and M0-M14), linear polyketide chains, as well as avermectin and rapamycin aglycones are shown*

nakupina. Jedan od 30 mogućih rekombinanata mogao bi se dogoditi u području DNA između gena *ave*E i *ave*A3, te petog i šestog modula (M5 i M6) gena *rap*B (slika 5). U tom se području nalazi apsolutno identična sekvencija DNA (MEPS) od 54 para nukleotida. Sekvencija se nalazi unutar poveznice između gena *ave*E i *ave*A3 te poveznice između domena ACP5 i KS6 petog i šestog modula (M5 i M6) gena *rap*B. Računalni program predviđa koordinate početka i kraja apsolutno identične sekvencije DNA (*ave*E i *ave*A3: 49545-49598 te ACP5 i KS6 (M5-M6) *rap*B: 15450--15505). Specijalizirani računalni program te koordinate povezuje sa specifično strukturiranom vlastitom bazom podataka da bi se prepoznali geni, moduli, domene ili poveznice koje sudjeluju u rekombinaciji. Taj se rezultat može i nacrtati (slika 5). Rekombinacijom između poveznice gena *ave*E-*ave*A3 i poveznice domena ACP5-KS6 gena *rap*B nastaju dvije nove genske nakupine označene kao Rec–1 i Rec-2. U ovoj fazi razvoja programa, ***CompGen*** može predvidjeti molekulske mase njihovih produkata koje iznose 676,1 i 753,9.

Kada specijalizirani računalni program, napisan računalnim jezikom Perl,[50] bude 'preveden' u računalni jezik Java,[26,27] ***CompGen*** će proces rekombinacije i grafički prikazati. 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 '**P**rediction of **A**ctivity **S**pectra for **S**ubstances')[53] ili upotrebom računalom vođenog dizajna lijekova, tehnologijom CDD.[54]

Konkurencija u svijetu

Koliko nam je poznato, u svijetu postoje četiri moguća konkurenta. Kao prvo, tvrtka Ecopia BioSciences Inc.[55] razvila je bioinformatički sustav ***DecipherIT***™ koji automatski anotira genske nakupine mikroorganizama iz tla. Svrha je sustava predviđanje kemijske strukture molekula koje sintetiziraju enzimi s određenom genetičkom uputom. Drugo, istraživači iz Nacionalnog instituta za imunologiju, New Delhi, Indija, razvili su računalni program ***SEARCHPKS*** za anotiranje 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.[38,56] Treće, postoji i računalni program ***NRPSpredictor***,[39,57] koji predviđa specifičnosti domena adenilacije (domena A), genskih nakupina NRPS, za odabir građevnih jedinica tijekom biosinteze neribosomski sinteti-



S l i k a 5 – *Shematski prikaz procesa homologne rekombinacije između multimodularnih poliketid-sintaza avermektina (ave*PKS*) i rapamicina (rap*PKS*). Rekombinacijom između poveznice gena ave*E*-ave*A3 *i poveznice domena ACP5-KS6 gena rap*B *nastaju dvije nove genske nakupine označene kao* Rec-*1 *i* Rec-*2. Prikazane su i pretpostavljene molekulske mase njihovih produkata*

F i g. 5 – *Schematic representation of homologous recombination process between the avermectin (ave*PKS*) and rapamycin (rap*PKS*) polyketide synthases. The recombination occurs between the ave*E*-ave*A3 *docker and ACP5-KS6 linker of the rap*B *gene producing two recombinant gene clusters* Rec-*1 *and* Rec-*2. Predicted molecular masses of their potential products are also shown*

ziranih peptida. Konačno, nedavno je objavljen računalni program ***Biogenerator***,[58] koji sadržava program PASS,[59] koji će 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 bazu podataka virtualnih poliketida koja će se moći pretraživati u uvjetima *in silico* da bi se predvidjela biološka aktivnost tih kemijskih supstancija.[53] 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***.

Zaključci i perspektive

Prema tome, do sada je strukturirana specifična baza podataka poliketidnih/peptidnih genskih nakupina po uzoru na BioSQL. Izrađen je računalni program za prikupljanje, strukturiranje, anotaciju i grafičko prikazivanje podataka poliketidnih/peptidnih biosintetskih puteva objektno orijentiranim računalnim jezikom poput jezika Java. U tijeku je modeliranje procesa ciklizacije linearnih poliketidnih lanaca upotrebom semiempirijskih kvantno-kemijskih metoda i modela. Napisan je računalni program za modeliranje procesa homologne rekombinacije računalnim jezikom Perl. U tijeku je prevođenje napisanog računalnog programa u objektno orijentirani računalni jezik Java. Integralni generički računalni programski paket ***CompGen*** poslužit će za oblikovanje baze podataka potpuno novih, u prirodi nepoznatih, virtualnih kemijskih entiteta na temelju sekvencija DNA genskih nakupina koje sadrže genetičku uputu za biosintezu poliketida i neribosomalno sintetiziranih peptida, koje su anotirane pomoću integralnoga računalnog programskog paketa ***ClustScan***. Inovativnost programskog paketa ***CompGen*** sastojat će se u sposobnosti računalnoga programskog paketa da oblikuje nove genske nakupine (tj. nove biosintetske procese) virtualnom homolognom rekombinacijom. Program predviđa kemijske strukture na temelju novih biosintetskih procesa koji se zatim pohranjuju u bazu podataka radi daljnjeg molekulskog modeliranja. Računalni programski paket omogućit će i analizu takozvanom 'obrnutom genetikom'. Naime, ako se zamisli poželjna kemijska struktura, program može predvidjeti koja bi genska nakupina PKS/NRPS sintetizirala takvu strukturu na temelju građevnih jedinica genskih nakupina u bazi podataka. Budućnost će računalnoga programskog paketa ***CompGen*** biti predviđanje biološke aktivnosti novih, *in silico* oblikovanih, kemijskih entiteta određivanjem aktivnosti kemijskih supstancija računalnim alatom PASS ili upotrebom računalom vođenog dizajna lijekova, CDD tehnologijom.

ZAHVALA

Prikazani rezultati su nastali unutar programa TEST – Tehnologijski istraživačko-razvojni projekt (TP-05/0058-23) uz potporu Ministarstva znanosti, obrazovanja i športa Republike Hrvatske.

Literatura
References

1. *A. Demain*, Pharmaceutically active secondary metabolites of microorganisms. Appl. Microbiol. Biotechnol. **52** (1999) 455.
2. *T. M. Kutchan*, Ecological arsenal and developmental dispatcher. The paradigm of secondary metabolism. Plant Physiology **125** (2001) 58.
3. *J. Bérdy*, Bioactive microbial metabolites. J. Antibiot. (Tokyo) **58** (2005) 1.
4. *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.
5. *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.
6. *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.
7. *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. (2007).
8. *S. Grabley, R. Thiericke*, The impact of natural products on drug discovery, u *S. Grabley, R. Thiericke* (ur.), Drug Discovery from Nature, Springer, Berlin, 1999, str. 3–37.
9. *D. Hranueli, J. Cullum*, Novi hibridni poliketidi dobiveni kombinatornom biosintezom. Kem. Ind. **50** (2001) 381.
10. *K. J. Weissman*, Polyketide biosynthesis: understanding and exploiting modularity. Philos. Transact. A Math. Phys. Eng. Sci. **362** (2004) 2671.
11. *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.
12. *H. B. Bode, R. Muller*, Possibility of bacterial recruitment of plant genes associated with the biosynthesis of secondary metabolites. Plant Physiol. **132** (2003) 1153.
13. *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. Antonie van Leeuwenhoek **78** (2000) 287.
14. *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.
15. *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.
16. *S. A. Sieber, M. A. Marahiel*, Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. Chem. Rev. **105** (2005) 715.
17. *Dictionary of Natural Products*, http://www.chemnetbase.com/scripts/dnpweb.exe
18. *J. Gonzalez-Lergier, L. J. Broadbelt, V. Hatzimanikatis*, Theoretical considerations and computational analysis of the complexity in polyketide synthesis pathways. (2005) 9930.
19. *J. Staunton, B. Wilkinson*, Combinatorial biosynthesis of polyketides and nonribosomal peptides. Curr. Opin. Chem. Biol. **5** (2001) 159.
20. *C. Khosla, J. D. Keasling*, Metabolic engineering for drug discovery and development. Nat. Rev. Drug Discov. **2** (2003) 1019.
21. *K. J. Weissman, P. F. Leadlay*, Combinatorial biosynthesis of reduced polyketides. Nat. Rev. Microbiol. **3** (2005) 925.
22. *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.
23. *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.
24. *B. F. Milne, P. F. Long, A. Starcevic, D. Hranueli, M. Jaspars,* Spontaneity in the patellamide biosynthetic pathway. Org. Biomol. Chem. **4** (2006) 631.
25. *W. C. Dunlap, M. Jaspars, D. Hranueli, C. N. Battershill, N. Perić-Concha, J. Zucko, 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.
26. Java Reference Documentation, 1994–2006, Sun Microsystems, Inc., http://java.sun.com/ reference/docs/index.html
27. *B. Eckel*, Thinking in Java, 4 edition. Prentice Hall PTR; 2006, str. 1–1150.
28. *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** (Data base issue) (2006) D173.
29. A knowledge based resource for analysis of Non-ribosomal Peptide Synthetases and Polyketide Synthases, http://www.nii.res.in/nrps-pks.html
30. *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.
31. *J. Rumbaugh, I. Jacobson, G. Booch*, The Unified Modeling Language Reference Manual (Second Edition). Addison-Wesley, Boston, 2005, 1–496.
32. Open Bioinformatics Foundation, http://obda.open-bio.org/

33. Daylight, Chemical Information Systems Inc., http://www.daylight.com/smiles/f_smiles. html
34. *S. Markel, D. Leon*, Sequence Analysis in a Nutshell: O'Reilly, Sebastopol, CA, 2003, 1–302.
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. *S. R. Eddy*, Profile hidden Markov models. Bioinformatics **14** (1998) 755.
37. *S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman,* Basic local alignment search tool. J. Mol. Biol. **215** (1990) 403.
38. *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.
39. *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.
40. *P. Caffrey*, Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. ChemBioChem **4** (2003) 654.
41. *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.
42. *F. Csizmadia*, JChem: Java applets and modules supporting chemical database handling from web browsers. J. Chem. Inf. Comput. Sci. **40** (2000) 323.
43. *Y. Shao, L. F. Molnar, Y. Jung, J. Kussmann, C. Ochsenfeld, S. T. Brown, A. T. B. Gilbert, L. V. Slipchenko, S. V. Levchenko, D. P. O'Neill, R. A. DiStasio Jr., R. C. Lochan, T. Wang, G. J. O. Beran, N. A. Besley, J. M. Herbert, C. Y. Lin, T. Van Voorhis, S. H. Chien, A. Sodt, R. P. Steele, V. A. Rassolov, P. E. Maslen, P. P. Korambath, R. D. Adamson, B. Austin, J. Baker, E. F. C. Byrd, H. Dachsel, R. J. Doerksen, A. Dreuw, B. D. Dunietz, A. D. Dutoi, T. R. Furlani, S. R. Gwaltney, A. Heyden, S. Hirata, C.–P. Hsu, G. Kedziora, R. Z. Khalliulin, P. Klunzinger, A. M. Lee, M. S. Lee, W. Z. Liang, I. Lotan, N. Nair, B. Peters, E. I. Proynov, P. A. Pieniazek, Y. M. Rhee, J. Ritchie, E. Rosta, C. D. Sherrill, A. C. Simmonett, J. E. Subotnik, H. L. Woodcock III, W. Zhang, A. T. Bell, A. K. Chakraborty, D. M. Chipman, F. J. Keil, A. Warshel, W. J. Hehre, H. F. Schaefer, J. Kong, A. I. Krylov, P. M. W. Gill, M. Head-Gordon*, Advances in methods and algorithms in a modern quantum chemistry program package. Phys. Chem. Chem. Phys. **8** (2006) 3172.
44. *M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart*, The development and use of quantum mechanical molecular models. J. Amer. Chem. Soc. **107** (1985) 3902.
45. *M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B .B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, J. A. Pople*, Gaussian 03, Revision C.02, Gaussian, Inc., Wallingford CT, 2004, str. 1- 330.
46. *Z. B. Maksić (ur.),* Theoretical Models of Chemical Bonding, Vol. 1–4, Springer, Berlin, 1990–1991, Vol. 1: 1–342, Vol. 2: 1–643, Vol. 3: 1–638, Vol. 4: 1–458.
47. *D. Kovaček,* Disertacija, Institut Ruđer Bošković, 1994.
48. http://www.aip.org/history/heisenberg/
49. *P. Shen, H. V. Huang*, Homologous recombination in Escherichia coli: dependence on substrate length and homology. (1986) 441.
50. PERL, http://www.perl.org/
51. *H. Ikeda, T. Nonomiya, M. Usami, T. Ohta, S. Ōmura*, Organization of the biosynthetic gene cluster for the polyketide anthelmintic macrolide avermectin in *Streptomyces avermitilis*. Proc. Natl. Acad. Sci. USA **96** (1999) 9509.
52. *T. Schwecke, J. F. Aparicio, I. Molnar, A. Konig, L. E. Khaw, S. F. Haydock, M. Oliynyk, P. Caffrey, J. Cortés, J. B. Lester, G. A. Bohm, J. Staunton, P. F. Leadlay*, The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. Proc. Natl. Acad. Sci. USA **92** (1995) 7839.
53. *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.
54. *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.
55. Ecopia BioSciences Inc., http://www.ecopiabio.com/
56. SEARCHPKS, A Program for Detection and Analysis of Polyketide Synthase Domains, http://www.nii.res.in/searchpks. html
57. NRPSpredictor, http://www-ab.informatik.uni tuebingen.de/software
58. *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.
59. Prediction of Activity Spectra for Substances (PASS), http://195.178.207.233/PASS/

Kratice
Abbreviations

A – domena za adenilaciju aminokiselina (NRPS)
– **A**denylation domain (NRPS)

ACP – mali polipeptid nosač acila (PKS)
– **A**cyl **C**arrier **P**rotein (PKS)

ACV – L-alfa-aminoadipil-L-cisteinil-D-valin – preteča penicilina
– L-alpha-amino**A**dipyl-L-**C**ysteinyl-D-**V**aline – penicillin precursor

AM1 – semiempirijski kvantno kemijski pristup, Austin Model 1
– semiempirical quantum chemical approach, **A**ustin **M**odel **1**

AT – domena acil-transferaze (PKS)
– **A**cyl-**T**ransferase domain (PKS)

ave – gen čiji proteinski produkt sudjeluje u biosintezi poliketida avermektina
– gene whose protein product is involved in **aver**mectin biosynthesis

BLAST – bioinformatički računalni program (akronim za "Basic Local Alignment Search Tool")
– bioinformatics computer program (acronim for "**B**asic **L**ocal **A**lignment **S**earch **T**ool")

C – domena za kondenzaciju (NRPS)
– **C**ondensation domain (NRPS)

CDA – antibiotik ovisan o kalciju
– **C**alcium **D**ependent **A**ntibiotic

CDD – računalom vođen dizajn lijekova
– **C**omputer-aided **D**rug **D**esign

Cy – domena za formiranje heterocikličkih prstenova (NRPS)
– domain for hetero**Cy**clic ring formation (NRPS)

DH – domena dehidrataze (PKS)
– **D**e**H**ydratase domain (PKS)

DNA – deoksiribonukleinska kiselina
– **D**eoxyribo**N**ucleic **A**cid

E – očekivana vrijednost
– **E**xpected value

Epim – domena za epimerizaciju (NRPS)
– **Epim**erisation domain (NRPS)

ER – domena enoil-reduktaze (PKS)
– **E**noyl **R**eductase domain (PKS)

KR – domena keto-reduktaze (PKS)
– **K**eto **R**eductase domain (PKS)

KS – domena β-ketosintaze (PKS)
– β-**K**etoacyl**S**ynthase domain (PKS)

MEPS – minimalna duljina apsolutne identičnosti u kojoj dolazi do homologne rekombinacije (akronim za 'Minimal Essential Pairing Sequence')
– minimal length of apsolute identity in which the homologous recombination occurs (acronym for '**M**inimal **E**ssential **P**airing **S**equence')

MT – domena za *N*-, *C*- ili *O*-metilaciju (NRPS)
– *N*-, *C*- or *O*-**Me**T**h**ilation domain (NRPS)

NCBI – Nacionalni Centar za Biotehnološke Informacije, Washington, SAD
– **N**ational **C**enter for **B**iotechnology **I**nformations, Washington, USA

NRPS – sintetaza neribosomski sintetiziranih peptida
– **N**on**R**ibosomal **P**eptide **S**ynthetase

PASS – akronim za 'Prediction of Activity Spectra for Substances'
– acronym for '**P**rediction of **A**ctivity **S**pectra for **S**ubstances'

PCP – mali polipeptid nosač peptidila (NRPS)
– **P**eptidyl **C**arrier **P**rotein (NRPS)

PKS – poliketid-sintaza
– **P**oly**K**etide **S**ynthase

PKS/NRPS – mješovita poliketid-sintaza i sintetaza neribosomski sintetiziranih peptida
– mixed **P**oly**K**etide **S**ynthase/**N**on**R**ibosomal **P**eptide **S**ynthetase

R – domena reduktaze (NRPS)
– **R**eductase domain (NRPS)

rap – gen čiji proteinski produkt sudjeluje u biosintezi poliketida rapamicina
– gene whose protein product is involved in ***rap****amycin* biosynthesis

SMILES – akronim za "Simplified Molecular Input Line Entry System"
– acronym for "**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem"

Te – domena tioesteraze (NRPS)
– **T**hio**E**sterase domain (NRPS)

TE – domena tioesteraze (PKS)
– **T**hio**E**sterase domain (PKS)

UML – akronim za "Universal Markup Language"
– acronym for "**U**niversal **M**arkup **L**anguage"

Simboli
Symbols

A – točnost računa (u jednadžbi $A \cdot S \approx$ konstanta)
– calculation accuracy (in equation $A \cdot S \approx$ constant)

S – veličina molekule (u jednadžbi $A \cdot S \approx$ konstanta)
– size of molecule (in equation $A \cdot S \approx$ constant)

SUMMARY

In silico Design of 'Un-Natural' Natural Products

D. Hranueli,[a] A. Starčević,[a,d] J. Žučko,[b,d] J. Diminić,[a] N. Škunca,[a] V. Željeznak,[a] D. Kovaček,[a] D. Pavlinušić,[b] J. Šimunković,[b] P. F. Long,[c] and J. Cullum[d]

Polyketides and non-ribosomal peptides represent a large class of structurally diverse natural products much studied over recent years because the enzymes that synthesise them, the modular polyketide synthases (PKSs) and the non-ribosomal peptide synthetases (NRPSs), share striking architectural similarities that can be exploited to generate 'un-natural' natural products. PKS and NRPS proteins are multifunctional, composed of a co-linear arrangement of discrete protein domains representing each enzymic activity needed for chain elongation using either carboxylic acid or amino acid building blocks. Each domain is housed within larger modules which form the complex. Polyketide and peptide antibiotics, antifungals, antivirals, cytostatics, immunosuppressants, antihypertensives, antidiabetics, antimalarials and anticholesterolemics are in clinical use. Of commercial importance are also polyketide and peptide antiparasitics, coccidiostatics, animal growth promoters and natural insecticides.

Polyketides are assembled through serial condensations of activated coenzyme-A thioester monomers derived from simple organic acids such as acetate, propionate and butyrate. The choice

of organic acid allows the introduction of different chiral centres into the polyketide backbone. The active sites required for condensation include an acyltransferase (AT), an acyl carrier protein (ACP) and a β-ketoacylsynthase (KS). Each condensation results in a β-keto group that undergoes all, some or none of a series of processing steps. Active sites that perform these reactions are contained within the following domains; ketoreductase (KR), dehydratase (DH) and an enoylreductase (ER). The absence of any β-keto processing results in the incorporation of a ketone group into the growing polyketide chain, a KR alone gives rise to a hydroxyl moiety, a KR and DH produce an alkene, while the combination of KR, DH and ER domains lead to complete reduction to an alkane. Most often, the last module contains the thioesterase domain (TE) responsible for the release of linear polyketide chain from the enzyme and final cyclisation. After assembly, the polyketide backbone typically undergoes post-PKS modifications such as hydroxylation(s), methylation(s) and glycosylation(s) to give the final active compound.

Non-ribosomal peptides are assembled by the so-called "multiple carrier thio-template mechanism". Three domains are necessary for an elongation module: an adenylation (A) domain that selects the substrate amino acid, analogous to a polyketide AT domain, and activates it as an amino acyl adenylate; a peptidyl carrier protein (PCP) that binds the co-factor 4-phosphopantetheine to which the activated amino acid is covalently attached, analogous to the ACP of a PKS; and a condensation (C) domain that catalyzes peptide bond formation, again analogous to the KS in modular PKSs. The NRPSs also contain a (Te) domain located at the C-terminal of the protein which is essential for release of linear, cyclic or branched cyclic peptides. Auxiliary activities can further enlarge the structural diversity of the peptide especially common are epimerization domains (E_{pim}) that convert the thioester-bound amino acid from an L- to D- configuration.

There has been a lot of interest in the last few years in generating new compounds for the production of novel drugs by manipulating the programming of such clusters *in vitro* (e.g. the idea of combinatorial biosynthesis). However, an important barrier to the progress is the fact that most changes made by *in vitro* methods result in very low yields or no detectable product. A possible solution to the yield problem would be the generation of novel clusters by homologous recombination *in vivo*, because this would favour more closely related sequences and should reduce problems caused by non-functional incompatible junctions.

The Unified Modeling Language (UML) was used to define the platform independent integral generic program packages, ***CompGen*** and ***ClustScan***, which are under development to model these processes *in silico*. The heart of ***CompGen*** is a specially structured database, based on BioSQL v1.29, which connects the biosynthetic order of synthase/synthetase enzymes to the sequences of the component polypeptides. The additional linkage to the gene sequences allows the integration of DNA sequence with product structure. The database contains sequences of the well-characterised PKS/NRPS clusters, and non-annotated sequenced clusters whose structure and function is yet unknown, to act as building blocks for the production of novel products. It is easy to add custom sequences to the database and to annotate them by the use of propriety protein profiles designed by Pfam database and HMMER. One function of the program is the ability to generate virtual recombinants between clusters. This can be done using a recombination model (with optional parameters) to predict sites for homologous recombination or by user defined recombination sites (e.g. to model *in vitro* genetic manipulation such as module replacement). The program predicts the linear polyketide structure of the resulting 'un-natural' natural products with a chemical description using isomeric SMILES. Molecular modelling of the subsequent spontaneous cyclisation process produces structures for a virtual compound database for further molecular modelling studies using PASS and CDD technology. An optional 'reverse genetics' module analyses a given chemical structure to see if it could be produced by a novel PKS/NRPS synthesis cluster and suggests the DNA sequence of a suitable cluster based on building blocks derived from clusters contained in the database.

Overall, the ***CompGen*** allows *in silico* generation of the database of novel 'un-natural' natural chemical compounds that can be used for *in silico* screening using PASS or CDD technology. The other integral generic program package, ***ClustScan***, will recognise and annotate new gene clusters from microbial genome sequencing projects or in metagenomes of soil and/or marine microorganisms.

[a] *Faculty of Food Technology and Biotechnology, Pierottijeva 6, 10 000 Zagreb, Croatia*

[b] *Novalis Ltd., Božidara Adžije 17, 10 000 Zagreb, Croatia*

[c] *University of London, 29/39 Brunswick Square, London WC1N 1AX, United Kingdom*

[d] *University of Kaiserslautern, Postfach 3049, D–67653 Kaiserslautern, Germany*

Received February 21, 2007
Accepted May 18, 2007