UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA
ENOTA MEDODDELČNEGA ŠTUDIJA MIKROBIOLOGIJE

Jurij PUHEK

MOZAIČNOST REPLIKACIJSKIH REGIJ PLAZMIDOV SKUPINE IncF KOT JO PRIKAŽE RAČUNALNIŠKI PROGRAM JUP 1.0 ZA ANALIZO NUKLEOTIDNIH ZAPOREDIJ

DIPLOMSKO DELO
Univerzitetni študij

MOSAICISM OF IncF PLASMID REPLICATION REGIONS AS REVEALED BY COMPUTER PROGRAM JUP 1.0 FOR NUCLEOTIDE SEQUENCE ANALYSIS

GRADUATION THESIS
University studies

Ljubljana, 2016

Diplomsko delo je zaključek univerzitetnega mededdelčnega študija mikrobiologije na Biotehniški fakulteti Univerze v Ljubljani.

Za mentorico diplomskega dela je imenovana izr. prof. dr. Marjanca Starčič Erjavec ter za recenzenta doc. dr. Tomaž Accetto.

Mentorica: izr. prof. dr. Marjanca STARČIČ ERJAVEC

Recenzent: doc. dr. Tomaž ACCETTO

Komisija za oceno in zagovor:

Predsednica:	prof. dr. Ines MANDIĆ MULEC Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za živilstvo
Članica:	izr. prof. dr. Marjanca STARČIČ ERJAVEC Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za biologijo
Član:	doc. dr. Tomaž ACCETTO Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za zootehniko

Datum zagovora:

Podpisani izjavljam, da je naloga rezultat lastnega raziskovalnega dela. Izjavljam, da je elektronski izvod identičen tiskanemu. Na univerzo neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravici shranitve avtorskega dela v elektronski obliki in reproduciranja ter pravico omogočanja javnega dostopa do avtorskega dela na svetovnem spletu preko Digitalne knjižnice Biotehniške fakultete.

Jurij Puhek

KLJUČNA DOKUMENTACIJSKA INFORMACIJA (KDI)

ŠD Dn
DK UDK 575.112:004.42:577.21(043)=163.6
KG bioinformatika/plazmidi/mozaičnost/vizualizacija homologije genov/BLAST/ GenBank/BioPython
AV PUHEK, Jurij
SA STARČIČ ERJAVEC, Marjanca (mentorica) / ACCETTO Tomaž (recenzent)
KZ SI-1000 Ljubljana, Jamnikarjeva 101
ZA Univerza v Ljubljani, Biotehniška fakulteta, Enota medoddelčnega študija mikrobiologije
LI 2016
IN MOZAIČNOST REPLIKACIJSKIH REGIJ PLAZMIDOV SKUPINE IncF KOT JO PRIKAŽE RAČUNALNIŠKI PROGRAM JUP 1.0 ZA ANALIZO NUKLEOTIDNIH ZAPOREDIJ
TD Diplomsko delo (univerzitetni študij)
OP XI, 63 str., 4 pregl., 36 sl., 10 pril., 111 vir.
IJ Sl
JI sl/en
AI Plazmidi po Gramu negativnih bakterij imajo pogosto zapise za dejavnike virulence in odpornosti proti protimikrobnim sredstvom. Takšni plazmidi lahko soprispevajo k patogenosti bakterije. Replikacija plazmidov je odvisna od replikacijskih regij, ki jih imenujemo replikoni. Te razvrščamo v t.i. inkompatibilnostne skupine. Ena izmed večjih inkompatibilnostnih skupin je skupina IncF. Replikacijske regije so gensko nestabilne regije z značilnimi mozaičnimi lastnostmi. Fenomen mozaičnosti replikacijskih regij definiramo kot značilnost genskih zaporedij replikacijskih regij, da vsebujejo elemente genoma gostiteljskega organizma ali druge genske elemente sorodnih plazmidov. Tipično se isti geni v različnih plazmidih in organizmih nahajajo na različnih mestih, obdani z različnimi elementi. Naloga je bila razdeljena na dve fazi. Prva ja obsegala razvoj računalniškega diagnostičnega orodja, ki prek analize rezultatov poizvedb v spletne baze podatkov NCBI omogoča analizo mozaičnosti replikacijskih regij. Druga je obsegala analizo replikacijskih regij plazmidov skupine IncF, ki smo jo opravili s pripravljenim orodjem. Rezultati intuitivno in dobro pokažejo mozaičnost replikacijskih regij z jasnimi območji večje in manjše homologije zaporedij. Orodje pokaže neustrezno označevanje genov, saj iste gene raziskovalci večkrat poimenujejo z različnimi imeni, kar povečuje kompleksnost podatkov in otežuje avtomatizirano računalniško obdelavo. Razvita računalniška rešitev je splošno uporabna za analize razvoja in kombiniranja genskih zaporedij genov in ne le zaporedij, kot to nudijo spletne rešitve NCBI.

KEY WORDS DOCUMENTATION (KWD)

ND Dn
DC UDC 575.112:004.42:577.21(043)=163.6
CX bioinformatics/plasmids/mosaicism/gene homology visualization/BLAST/ GenBank/BioPython
AU PUHEK, Jurij
AA STARČIČ ERJAVEC, Marjanca (supervisor) / ACCETTO Tomaž (reviewer)
PP SI-1000 Ljubljana, Jamnikarjeva 101
PB University of Ljubljana, Biotechnical Faculty, Interdepartmental Programme in Microbiology
PY 2016
TI MOSAICISM OF IncF PLASMID REPLICATION REGIONS AS REVEALED BY COMPUTER PROGRAM JUP 1.0 FOR NUCLEOTIDE SEQUENCE ANALYSIS
DT Graduation Thesis (University studies)
NO XI, 63 p, 4 tab., 36 fig., 10 ann., 111 ref.
LA sl
AL sl/en
AB Plasmids, present in Gram-negative bacteria, are often carriers of virulence factors and antimicrobial resistance genes. These plasmids contribute to the pathogenicity of bacteria. Plasmid replication is governed by genetic regions called replicons. They are classified into so-called incompatibility groups of which one of the largest is the IncF group. Replication regions are genetically unstable often with typical mosaic structure. Mosaicism is a genetic trait which denotes multiple origins of genetic elements present in one sequence. Origins can vary from host cell genome to genetic elements from related plasmids. Typically, the same genes can occur in different plasmids and organisms in various places with various adjacent elements. The fist aim of this research was to develop a computer aided analytical tool which queries online gene banks for similar genetic sequences and displays their gene annotations in unified, consolidated comparison result. The second aim was to use the developed tool and analyze the mosaic structure of the IncF replication regions. Results confirm the mosaic structure of analyzed regions with clear and intuitive display of areas of higher and lower sequence homology. The tool also clearly displays the inadequacy in labeling genes as many of them are named with several names. This contributes to the complexity of the data and makes it difficult to process automatically with computer aided tools. Developed software solution is multi-purpose tool that can be used for analysis of genetic evolution and combination which is not available from the current public NCBI online tools.

KAZALO VSEBINE

KAZALO PREGLEDNIC

KAZALO SLIK

KAZALO PRILOG

OKRAJŠAVE IN SIMBOLI

BLAST	orodje za iskanje poravnav zaporedij (angl. **B**asic **L**ocal **A**lignment **S**earch **T**ool)
BLASTN	orodje za iskanje poravnav nukleotidnih zaporedij (angl. **N**ucleotide – nucleotide **B**asic **L**ocal **A**lignment **S**earch **T**ool)
bp	nukleotidni (bazni) par
CDS	označba zaporedja nukleotidov, ki kodira proteinski produkt (angl. **C**o**d**ing **S**equence)
DDBJ	Japonska javna baza nukleotidnih zaporedij (angl. **D**NA **D**ata **B**ank of **J**apan)
DNA	deoksiribonukleinska kislina (angl. **d**eoxyribo**n**ucleic **a**cid)
EMBL	Evropska javna baza nukleotidnih zaporedij (angl. **E**uropean **M**olecular **B**iology **L**aboratory)
ESBL	beta laktamaze z razširjenim spektrom delovanja (angl. **e**xtended-**s**pectrum **β**-**l**actamases)
ExPEC	zunajčrevesni patogeni sevi *E. coli* (angl. **ex**traintestinal **p**athogenic ***E***. ***c****oli*)
GenBank	javna baza nukleotidnih zaporedij, sponzorirana s strani Nacionalnega inštituta zdravja Združenih držav Amerike v v okviru NCBI
HSP	visoko točkovano parno območje (angl. **H**igh-scoring **S**egment **P**air)
HTML	jezik za označevanje nadbesedila (angl. **H**yper**T**ext **M**arkup **L**anguage)
JSON	skriptni javanski zapis objekta (angl. **J**ava**S**cript **O**bject **N**otation)
NCBI	nacionalni biotehnološki center Združenih držav Amerike (angl. **N**ational **C**enter for **B**iotechnology **I**nformation)
RNA	ribonukleinska kislina (angl. **r**ibo**n**ucleic **a**cid)
uORF	zgornji bralni okvir upstream (angl. **u**pstream **ORF**)
XML	razširljiv označevalni jezik (angl. **E**xtensible **M**arkup **L**anguage)

SLOVARČEK

BioPython	odprtokodni projekt z orodji programskega jezika Python za nekomercialno uporabo v bioinformatiki
FASTA	format zapisa zaporedja DNA v obliki niza oznak nukleotidov A, G, C in T ter morebitnih degenerativnih kod in znaka za vrzeli

1 UVOD

Escherichia coli (E. coli) je po Gramu negativna paličasta bakterija. Različni sevi *E. coli* so z gostiteljskim organizmom v različnih simbiontskih odnosih. Tako nekateri sevi s svojim gostiteljem živijo v mutualističnem odnosu. Te lahko najdemo v mikrobioti spodnjega črevesnega trakta organizmov s stalno telesno temperaturo (Eckburg in sod., 2005). Za človeka je pomemben fakultativen anaerob, ki sintetizira vitamin K, s katerim prevzema vitamine B-kompleksa in drugim patogenim bakterijam preprečuje kolonizacijo prebavnega trakta. Drugi sevi iz skupine zunajčrevesnih patogenih sevov *E. coli* (ExPEC) pa lahko ob določenih pogojih povzročijo okužbo praktično kateregakoli zunaj črevesnega anatomskega mesta pri zdravih in imuno kompromitiranih gostiteljih (Russo in Johnson, 2000). V patogenezi *E. coli* igrajo pomembno vlogo različni dejavniki virulence, ki so lahko kodirani v genomu ali plazmidih.

Plazmidi so majhni, krožni, zunaj kromosomski elementi DNA, sposobni avtonomnega podvajanja. Prisotni so v vseh treh domenah živega *Archaea*, *Bacteria* in *Eukarya* (Holmes in sod., 1995; Solar in sod., 1998; Zillig in sod., 1998). Poleg genov, ključnih za lastno podvajanje in uravnavanje števila primerkov v gostiteljski celici, lahko vsebujejo širok in heterogen nabor genov za različne druge, gostitelju koristne lastnosti, kot so razgradnja ksenobiotičnih spojin, virulenca in odpornost proti antibiotikom, odpornosti proti težkim kovinam in zmožnost alternativnih metabolnih poti (Kado, 1998). Velja, da ti dodatni geni niso ključni za rast in razvoj gostiteljske celice, lahko pa doprinesejo k njenemu preživetju v specifičnih razmerah (Thomas in sod., 2005). Izjemna lastnost nekaterih plazmidov je tudi zmožnost horizontalnega prenosa v druge vrste, rodove in celo družine bakterij s procesom konjugacije (Firth in sod., 1996). Plazmidi so zmožni z rekombinacijo in transpozicijo prevzeti in vključiti vase tudi gene iz kromosoma in s tem povišati gensko izmenjavo med bakterijskimi populacijami (Solar in sod., 1998).

Spletna analitična orodja in baze podatkov NCBI nudijo široko paleto načinov analize genskih zaporedij, ki jih želimo primerjati oz. identificirati. Baza podatkov GenBank, ki je nastala v sodelovanju DDBJ, EMBL in GenBank, pri NCBI hrani nukleotidna zaporedja velikega števila organizmov, ki so prek spletnih analitičnih orodij javno na voljo raziskovalcem. Zaradi javnega značaja baz podatkov, v katere lahko raziskovalci donirajo svoje prispevke sekvenciranih genskih zapisov in jih sami označujejo, prihaja ob izostanku jasnih pravil za označevanje do različno poimenovanih enakih zaporedij, kar analizo otežuje.

1.1 NAMEN DELA

V nalogi smo se osredotočili na analizo mozaičnosti replikacijskih regij plazmidov skupine IncF, primarno na replikone RepFIA, RepFIB, RepFIC in RepFIIA, ki jih najdemo v veliko bakterijah družine *Enterobacteriaceae*.

Ker za analizo mozaičnosti doslej ni bilo primernih orodij, je bil velik del naloge namenjen izdelavi spletnega računalniškega programa, ki bo znanemu vhodnemu zaporedju nukleotidov preko zaledne spletne storitve BLAST poiskal podobna zaporedja nukleotidov in zatem preko njihovih akcesijskih številk (unikatni identifikatorji zaporedij) v bazi GenBank pridobil podatke o anotacijah identificiranih regij ujemanja. Zatem bo ujemanja in anotacije skupaj z njihovimi metapodatki prikazal v enotnem prikazu. Področja večjega in manjšega ujemanja bo prikazal z različnimi barvnimi toni in tako označil nivo ujemanja različnih delov nukleotidnega zaporedja. Tak prikaz je za analizo mozaičnosti nekega zaporedja najprimernejši, saj obstoječi grafični prikazi spletnih orodij NCBI ne omogočajo enotnega prikaza več nukleotidnih zaporedij z anotacijami v okviru enega prikaza. Analiza se tako podaljša, prav tako je prikaz mozaičnosti manj učinkovit. Raziskovalci so se zato prisiljeni zatekati h dodatni grafični obdelavi rezultatov za intuitivnejši prikaz.

Cilji naloge:

- Izdelati računalniški program, ki na podlagi vhodnega genskega zaporedja v bazah nukleotidnih zaporedij BLASTN poišče podobna zaporedja in zatem preko njihovih akcesijskih številk v bazi podatkov GenBank pridobi dodatne podatke o genih, ki so kodirani na regijah ujemanja in vse skupaj prikaže tabelarično ter grafično z neposrednim in kontekstnim dostopom do metapodatkov zaporedij in izvornih virov podatkov.
- Z izdelanim programom analizirati replikacijske regije RepFIA, RepFIB in RepFIC in RepFIIA plazmidov skupine IncF, poiskati podobna zaporedja in grafično prikazati mozaičnost teh regij.
- Z rezultati naloge pokazati visoko stopnjo mozaičnosti še drugih replikacijskih regij plazmidov skupine IncF.

2 PREGLED OBJAV

2.1 PLAZMIDI

2.1.1 Splošne značilnosti

Plazmidi so raznoliki po velikosti, številu primerkov v gostiteljski celici in genskem ustroju. Večino predstavljajo kovalentne zaprte dvoverižne molekule DNA (Kado, 1998). Poznamo pa tudi več plazmidov z linearno dvoverižno molekulo DNA. Linearne plazmide najdemo v vrstah rodu *Streptomyces* (Hayakawa in sod., 1979; Kinashi in sod., 1987; Netolitzky in sod., 1995) in vrstah rodu *Borrelia* (Casjens in sod., 1995), kot tudi v kvasovkah in filamentoznih glivah, kjer so primarno prisotni v njihovih mitohondrijih (Fukuhara, 1995; Miyashita in sod., 1990). Velikost plazmidov variira med približno 300 bp in 2,4 milijona bp (Kado, 1998). Manjši plazmidi so v gostiteljski celici prisotni v večjem številu kopij (do 100), medtem ko večji plazmidi nastopajo v eni ali največ dveh kopijah (Madigan in sod., 2014).

Glede na splošni genski ustroj lahko plazmide delimo na dve vrsti (Helinski in sod., 1996):

- nekonjugativne oz. neprenosljive, ki vsebujejo gene za začetek in regulacijo replikacije, vendar jim manjkajo funkcionalni geni za konjugativni prenos v drugo gostiteljsko celico;
- konjugativne oz. samoprenosljive plazmide, ki vsebujejo gene za podvajanje in regulacijo le-tega, imajo pa tudi gene, ki omogočajo konjugativni prenos v drugo gostiteljsko celico.

2.1.2 Replikacija

Neodvisno od velikosti in oblike imajo vsi plazmidi lastnost samostojnega podvajanja. Celotna informacija, potrebna za podvajanje, je tipično zbrana v segmentu, ki je manjši od 3000 bp in ga imenujemo replikacijska regija oz. replikon. Del tega segmenta je tudi zaporedje *ori*, dolgo okoli 250 bp, kjer se podvajanje plazmida začne. Večina plazmidov ima eno samo replikacijsko regijo, ni pa redko, da imajo plazmidi tudi dve ali več replikacijskih regij. Pogosto imajo več kot eno replikacijsko regijo večji plazmidi. Geni, tipično označevani s predpono *rep*, kodirajo za plazmid specifične proteine, ki omogočajo njihovo replikacijo (Helinski in sod., 1996).

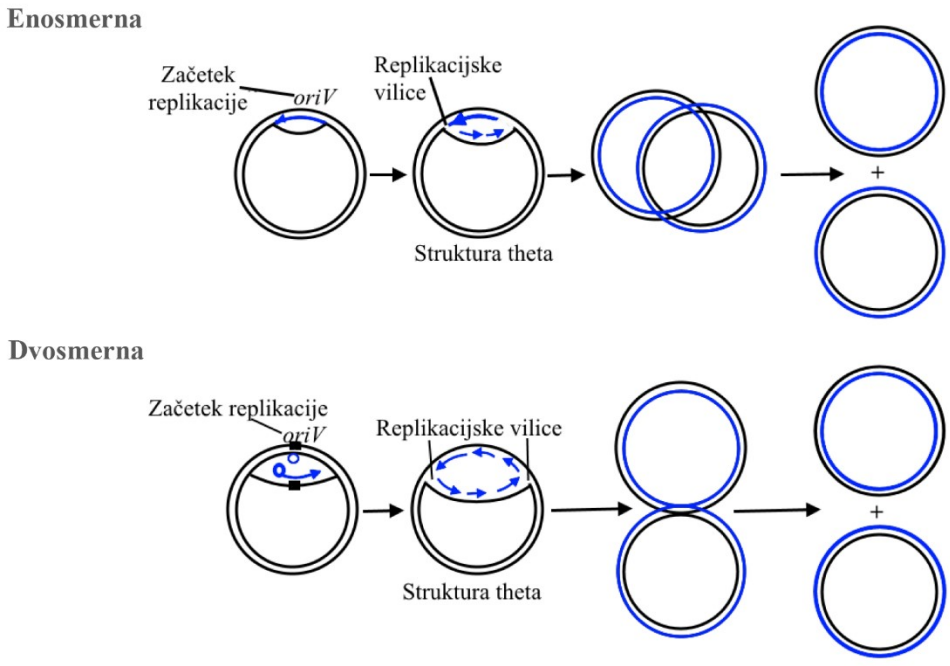
2.1.2.1 Replikacija krožnih plazmidov

Pri krožnih plazmidih poznamo tri mehanizme replikacije: mehanizem theta (θ), mehanizem kotalečega se kroga (angl. rolling circle) in mehanizem premestitve verig (Solar in sod., 1998).

2.1.2.1.1 Mehanizem theta (θ)

Podvajanje DNA z mehanizmom theta (θ) se začne s cepitvijo starševskih verig v zaporedju *ori* s pomočjo iniciatorskega proteina Rep, kodiranega na plazmidu. Nato primaza ali polimeraza RNA sintetizirata začetni oligonukletotid RNA, ki se nato nadaljuje v sintezo DNA s kovalentnim podaljšanjem začetnega oligonukleotida hčerinske verige (Kornberg in Baker, 1992). Sinteza obeh verig DNA je povezana in na vodilni verigi poteka kontinuirano od 5' proti 3', medtem ko sinteza druge verige poteka nezvezno preko Okazakijevih fragmentov. DNA-polimeraza III je nujna za replikacijsko podaljševanje plazmidne DNA. Zaključek podaljševanja kodirajo posebna zaporedja *ter*, ki so mesta vezave signalnih proteinov za terminacijo replikacije plazmidov (Lilly in Camps, 2015)

Mehanizem theta za podvajanje plazmidov je široko razširjen med plazmidi po Gramu negativnih bakterij, vendar je prisoten tudi v nekaterih po Gramu pozitivnih bakterijah. Sinteza DNA se lahko začne na enem koncu ali več delih plazmida, lahko je enosmerna ali dvosmerna. Intermediati pri tem tipu replikacije tvorijo obliko črke theta (θ), od tod tudi ime (Moat in sod., 2002).



Slika 1: Podvajanje plazmida z mehanizmom theta in vmesno obliko (Chaudhari, 2014)

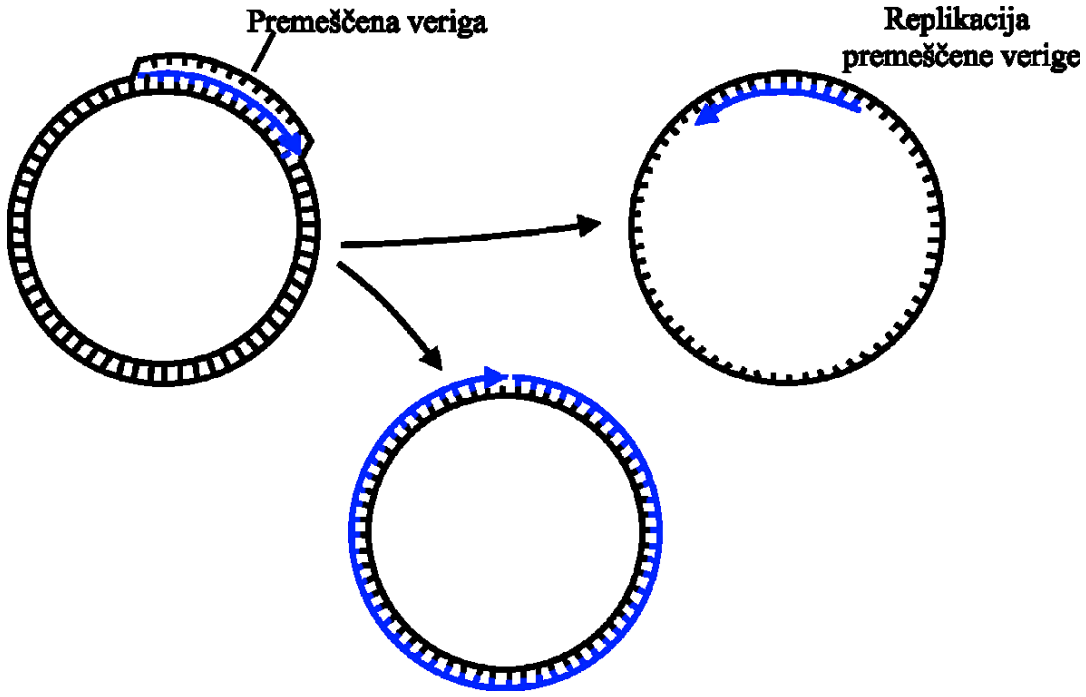
2.1.2.1.2 Mehanizem premestitve verig

Za iniciacijo replikacije oz. sinteze DNA po mehanizmu premestitve verig je ključna združitev treh na plazmidu kodiranih proteinov RepA, RepB in RepC. Replikacija je dvosmerna in se začne na mestu *ori*. V primeru plazmidov s tem tipom replikacije izvorna regija za začetek replikacije vsebuje iterone, bogate z nukleotidi GC, en segment AT in dve

palindromski zaporedji *ssiA* in *ssiB* (angl. small palindromic sequences), locirani nasproti regije *ori* (Solar in sod., 1998). Iteroni so mesta vezave RepC (Gruss in Ehrlich, 1988) medtem ko zaporedji *ssiA* in *ssiB* prepozna RepB (primaza) (Ingmer in Cohen, 1993).

Replikacija se začne, ko sta regiji *ssiA* in *ssiB* izpostavljeni na enoverižni obliki DNA. Razdružitev dvoverižne DNA, ki vodi do enoverižne oblike DNA, je odvisna od RepC in RepA (helikaza DNA). RepB katalizira pripravo začetnega oligonukleotida, ki sproži podvajanje DNA in poteka na eni verigi do konca. Replikacija druge verige se začne na mestu *ssi*. Zankam podobne strukture (angl. steam-loop), ki nastanejo s pomočjo zaporedij *ssi*, so potrebne za sestavljanje primaze RepB (Moat in sod., 2002).

Zaradi lastno kodiranih proteinov RepA, RepB in RepC je replikacija neodvisna od gostiteljskih faktorjev replikacije kot so polimeraza RNA, DnaA in DnaB. To je lahko vzrok, da se ti plazmidi lahko podvajajo v široki paleti gostiteljev (Wilson, 2006).



Slika 2: Podvajanje plazmida po mehanizmu premestitve verig (Chaudhari, 2014)

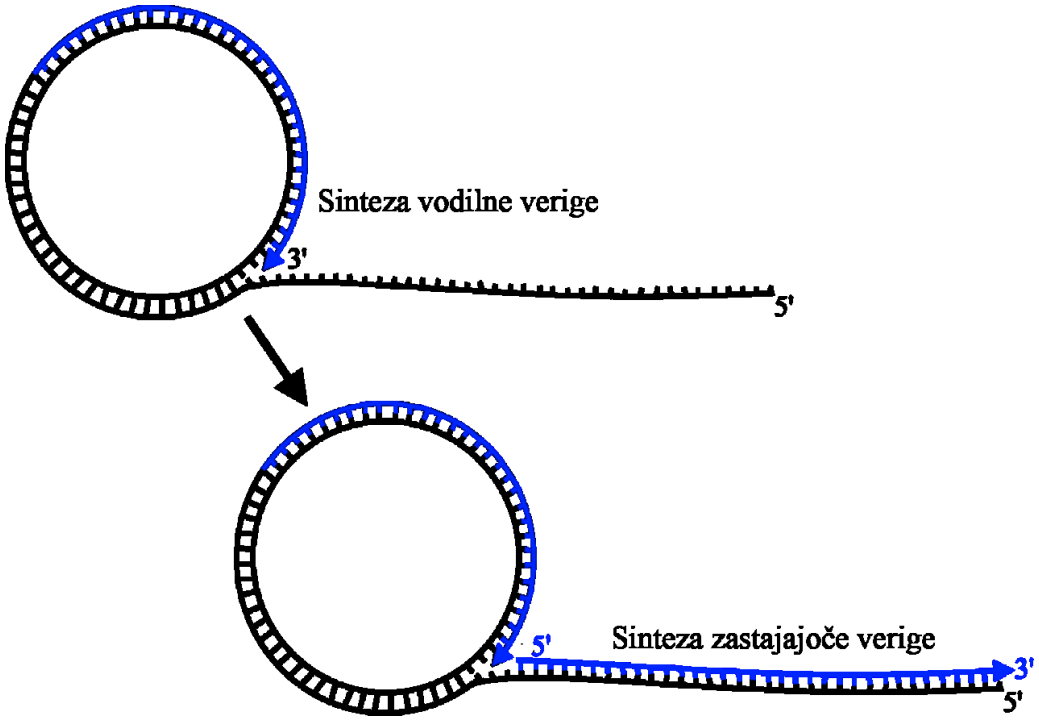
2.1.2.1.3 Mehanizem kotalečega se kroga (angl. rolling circle)

Mehanizem je enosmeren, saj je sinteza obeh verig razdružena, asimetrična. Ena ključnih posebnosti tega mehanizma podvajanja je lastnost, da vodeča veriga plus pri podvajanju ostane kovalentno vezana na njeno starševsko verigo plus (Solar in sod., 1998).

Začne se z vezavo na plazmidu kodiranega iniciatorskega proteina Rep na verigo plus plazmida, ki povzroči prekinitev verige plus na področju imenovanem *dso* (angl. double-stranded origin). Protein Rep ostane kovalentno vezan na 5'-fosfatu na mestu

prekinitve. Konec 3'-OH se uporabi kot začetno zaporedje sinteze vodilne verige. Za replikacijo so potrebni replikacijski proteini gostitelja (polimeraza DNA III, SSB in helikaza) (Khan, 2000). Elongacija konca 3'-OH poteka kot neprekinjena replikacija vodilne verige, dokler se ne konča v terminalnem delu. Podvajanje vodilne verige poteka ločeno od zastajajoče (Moat in sod., 2002).

Ta vrsta replikacije je razširjena med plazmidi, ki v gostiteljskih celicah praviloma nastopajo v več kopijah in so navadno manjši od 10 kb. (Solar in sod., 1998).



Slika 3: Podvajanje plazmida po mehanizmu kotalečega se kroga (Chaudhari, 2014)

2.1.2.2 Replikacija lineranih plazmidov

Replikacija linearnih plazmidov navadno poteka prek mehanizma vezave replikacijskega proteina na 5'-konec vsake verige (Madigan in sod., 2014). Ti predstavljajo začetne oligonukleotide za sintezo DNA. Linearni plazmidi z lasnicami se replicirajo s pomočjo konkatemernih intermediatov (Moat in sod., 2002).

2.1.3 Inkompatibilnost

Klasifikacija in identifikacija plazmidov mora temeljiti na njihovih lastnostih, ki so vedno prisotne in konstantne. Izkazalo se je, da so za razvrščanje plazmidov najbolj primerne lastnosti, povezane z replikacijo plazmidov (DeNap in Hergenrother, 2005). Leta 1971 sta Hedges in Datta predlagala shemo za klasifikacijo plazmidov, ki temelji na fenomenu imenovanem »inkompatibilnost« (Datta in Hedges, 1971; Hedges in Datta, 1971).

Inkompatibilnost izvira iz enakih ali zelo podobnih mehanizmov podvajanja plazmidov (Datta in Hedges, 1971; Novick in Hoppensteadt, 1978; Novick, 1987). Definirana je kot nezmožnost sobivanja dveh plazmidov iste inkompatibilnostne skupine (Inc) v isti gostiteljski celici v odsotnosti zunanje selekcije. Plazmida pripadata isti inkompatibilnostni skupini, če na stabilnost enega vpliva prisotnost drugega plazmida.

Nezmožnost sobivanja pripisujejo v glavnem tekmovanju za pomanjkljive vire v gostiteljski celici, ki so ključni za vzdrževanje plazmida (Yarmolinsky, 2000). Velja, da soroden replikacijski mehanizem dveh plazmidov navadno povzroča njuno inkompatibilnost (Helinski in sod., 1996). Dobro znani elementi inkompatibilnosti so protiprepisna RNA (ctRNA), ki poleg delovanja na sam plazmid, deluje tudi na sosednji plazmid z enakim tipom replikona (Tamm in Polisky, 1983) in iteroni, ki regulirajo število primerkov plazmida v gostiteljski celici (Tolun in Helinski, 1981).

Sekcija »Plasmid« pri National Collection of Type Cultures (London, Velika Britanija) trenutno vodi 27 inkompatibilnostnih skupin v družini *Enterobacteriaceae* (Couturier in sod., 1988; Carattoli in sod., 2005; Villa in sod., 2010) vključno s sedmimi skupinami IncF (FI do FVII) in tremi IncI skupinami (I1, Iγ, I2).

2.2 INKOMPATIBILNOSTNA SKUPINA IncF

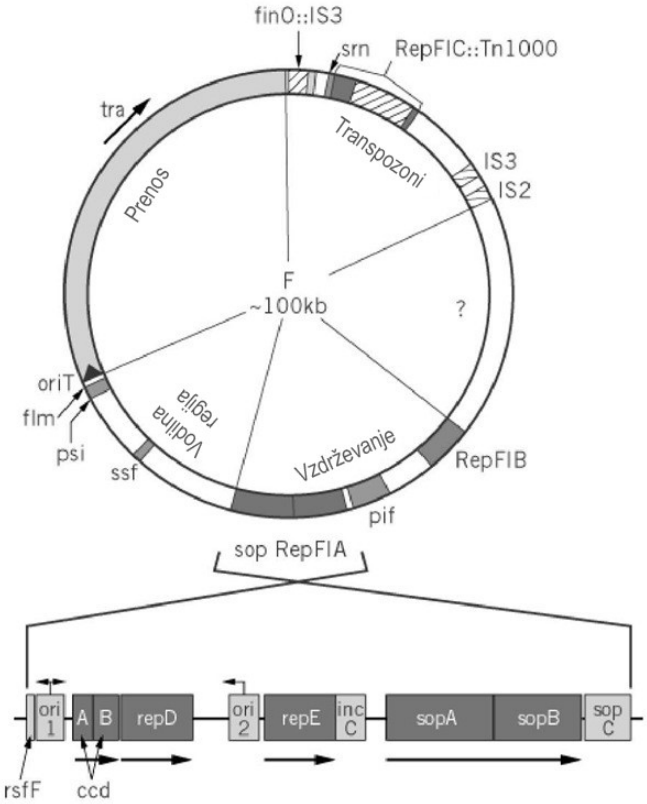
Plazmidi inkompatibilnostne skupine IncF predstavljajo eno prevladujočih inkompatibilnostnih skupin pri *Enterobacteriaceae* (Carattoli, 2009; Mathers in sod., 2015). Od skupaj 924 plazmidov, katerih genomi so deponirani v nukleotidni bazi pMLST (http://pubmlst.org/plasmid/, 30. januar 2015), jih 214 pripada inkompatibilnostni skupini IncF. Od teh jih je bilo iz *Escherichia coli* izoliranih 158 oz. 74 %.

Plazmidi IncF imajo mehanizme, ki zagotavljajo njihovo avtonomno podvajanje in gene, ki uravnavajo število kopij plazmida v gostiteljski celici ter zagotavljajo stabilno dedovanje med celično delitvijo (Cohen, 1976; Kado, 1998; Hayes, 2003). Poleg tega izkazujejo sposobnost integracije širokega spektra genov, ki omogočajo odpornost proti antibiotikom, vključno z β-laktami, aminoglikozidi, tetraciklini, kloramfenikolom in kinoloni (Liao in sod., 2013; Liu in sod., 2013).

Na splošno plazmidi skupine IncF niso homogena skupina, njihova velikost variira med 50 in 200 kb. Vsebujejo različne tipe replikonov (Garcillán-Barcia in sod., 2009; de Been in sod., 2014; Lanza in sod., 2014). Virulentne lastnosti, kodirane v plazmidih, so praktično ekskluzivno povezane s plazmidi skupine IncF (Johnson in Nolan, 2009), zato so tarče mnogih raziskav, ki bi omogočile učinkovito klinično obvladovanje razširjanja protimikrobne rezistence različnih enterobakterij (Osborn in sod., 2000; DeNap in Hergenrother, 2005; Baquero in sod., 2011).

2.2.1 Anatomija IncF plazmidov

Kljub heterogeni sestavi IncF plazmidov se je kot tipični predstavnik te skupine uveljavil samo en plazmid, in sicer plazmid F. Njegova sestava je prikazana na sliki 4.



Slika 4. Karta plazmida F (F Plasmid – Molecular Biology, 2016)
Krožni plazmid F je glede na funkcijo razdeljen na pet sektorjev (nakazano z dolgimi črtami, ki segajo iz središča karte). Razlaga simbolov genov je podana v besedilu. Puščica nakazuje smer prenosa plazmida F v konjugaciji. *oriT* je mesto, kjer se konjugacijski prenos začne.

Vodilna regija, sicer slabo raziskana in brez ugotovljene vloge, leži med RepFIA in *oriT*. Ta regija prva vstopa v gostiteljsko celico v procesu konjugacijskega prenosa (Ray in Skurray, 1983).

Replikacijska regija RepFIA je primarno odgovorna za tipične replikacijske lastnosti plazmida F, vsebuje enosmeren (*oriS*) in dvosmeren (*oriV*) začetek replikacije (Lane, 1981). V tej regiji so zapisi za pomembne determinante vzdrževanja F-plazmida v celici. Če replikacijsko regijo RepFIA izoliramo in pridružimo genu z zapisom za odpornost proti kakšnemu antibiotiku za potrebe selekcije, ima tako pripravljen miniplazmid enake replikacijske, delitvene in stabilizacijske lastnosti, kot jih ima izvorni F-plazmid (Lane, 1981).

Sekundarna replikacijska regija RepFIB je neodvisna od RepFIA in omogoča podvajanje plazmida tudi v odsotnosti regije RepFIA (Lane, 1981).

Odsek na koncu operona *tra* vsebuje delne ostanke replikacijske regije RepFIC. Ta replikacijska regija je nefunkcionalna zaradi vstavitve transpozicijskega elementa Tn*1000* (Saadi in sod., 1987). Vse kaže, da je odsek na koncu operona *tra* in predela RepFIC past za transpozicijske elemente, saj so v tem odseku ob Tn*1000* vključeni še dodatni transpozicijski elementi, IS2 in dve kopiji IS3 (Guyer, 1978). Ena kopija IS3 je vstavljena v regulatorni gen *finO* operona *tra*. Zaradi vstavitve IS3 v gen *finO*, je le-ta gen inaktiviran. Posledično se konstantno izražajo geni operona *tra*, kar vodi v višje frekvence konjugacijskega prenosa (Cheah in Skurray, 1986).

Operon *tra* obsega približno 33 kb in kodira komponente, ki F-plazmidu omogočajo konjugacijo, horizontalni prenos DNA iz donorske in recipientske celice ob neposrednem kontaktu obeh celic (Kokate in sod., 2011).

Plazmidi skupine IncF za svoje vzdrževanje in replikacijo uporabljajo tako produkte genov, ki so zapisani na plazmidu, kot produkte genov, ki so zapisani v kromosomu gostiteljske celice. Plazmidi IncF so za podvajanje odvisni od giraze DNA, DnaB, DnaC, DnaG, SSB (proteini, ki se vežejo na enoverižno DNA) in DNA-polimeraze III, ki so vsi zapisani v kromosomu gostiteljske celice (Toukdarian, 2004).

2.3 POMEN PLAZMIDOV SKUPINE IncF

Povečana odpornost proti protimikrobnim sredstvom ima lahko širše posledice za zdravje ljudi in živali. Zadnje študije kažejo, da so plazmidi še učinkovitejši medij za širjenje genov protimikrobnih faktorjev, kot je bilo to do sedaj znano (Taylor in sod., 2004; García-Fernández in sod., 2009; Dolejska in sod., 2011; Accogli in sod., 2013; Dahmen in sod., 2013). Plazmidi inkompatibilnostne skupine IncF lahko vsebujejo širok diapazon genov, katerih produkti nudijo gostiteljski celici faktorje za odpornost proti večini razredov protimikrobnih učinkovin. Širjenje takih plazmidov med sevi *Enterobacteriaceae* lahko vplivajo na klinično obvladovanje okužb z gostiteljskimi organizmi.

Prisotnost plazmidov s faktorjem IncF-*bla*$_{\text{CTX-M}}$ v sevih *E. coli*, izoliranih iz ljudi in živali (npr. R100), je ključna pri širitvi *bla*$_{\text{CTX-M-14}}$ v Hong Kongu, Združenih državah Amerike in Franciji (Woodford in sod., 2009; Dahmen in sod., 2013). V plazmidih IncF so nedavno identificirali gene *rmtB*, *qepA*, *qnr*, *fosA3* in *oqxAB* iz sevov *E. coli* na Kitajskem, Koreji in Španiji (Tamang in sod., 2008; Li in sod., 2012; Ruiz in sod., 2012; Ho in sod., 2013). Kaže, da imajo plazmidi iz skupine IncF potencial prevladujoče vplivati na razširitev genov za različne protimikrobne faktorje. Geni, ki bakterijam *E. coli* dajejo virulentne lastnosti in ki izvirajo iz plazmidov, skoraj ekskluzivno pripadajo plazmidom inkompatibilnostne skupine IncF (Johnson in Nolan, 2009).

Popolno sekvenciran plazmid IncF pIP1206, identificiran v *E. coli* v Franciji, je nosilec genov *rmtB* in *qepA*, slednji odgovoren za odpornost proti hidrofilnim fluorokinolonom. pIP1206 nosi dve kopiji replikona RepFII in dva dodatna replikona tipa RepFIA ter RepFIB. Ta zanimiv multireplikonski plazmid poleg tega nosi tudi zapise za sistem toksin-antitoksin in virulentne faktorje (Perichon in sod., 2008).

Gena *qnrB4* in *qnrB6*, povezana z geni *armA* in geni β-laktamaz z razširjenim spektrom delovanja (angl. extended-spectrum beta-lactamases ali ESBL), najdena v *E. coli*, *K. pneumoniae* in *E. cloacae* v Koreji, so identificirali v plazmidih IncF z replikoni IncFIIA. Ti plazmidi so zelo podobni virulentnim plazmidom bakterij sevov *Salmonella* (Tamang in sod., 2008).

V zaporedjih DNA več plazmidov IncF so zaznali prisotnost gruče genov, ki potencialno prispevajo k virulenci gostiteljskih bakterij, kot je to primer z aerobaktinskim sistemom za prevzem železa pri pRSB107 (Szczepanowski in sod., 2005) in ABC prenašalcev ter operoni za deaminazo rafinoze in arginina pri pIP1206 (Perichon in sod., 2008).

Pregledni članek (Carattoli, 2009) je zbral podatke različnih raziskav in identificiral 331 plazmidov skupine IncF s prisotnostjo genov *aac(6')-Ib-cr*, *bla*$_{\text{CMY-2}}$, *bla*$_{\text{CTX-M-1-2-3-9-14-15-24-27}}$, *bla*$_{\text{DHA-1}}$, *bla*$_{\text{SHV-2-5-12}}$, *bla*$_{\text{TEM-1}}$, *armA*, *rmtB*, *qepA*, *qepA2*, *qnrA1*, *qnrB2*, *qnrB4*, *qnrB6*, *qnrB19* in *qnrS1* za odpornost proti antibiotikom v bakterijah vrste *Enterobacter aerogenes (E. aerogenes)*, *Enterobacter cloacae (E. cloacae)*, *E. coli*, *Klebsiella pneumoniae (K. pneumoniae)*, *Salmonella enterica (S. enterica)*, *Serratia marcescens (S. marcescens)* in *Shigella sonnei (S. sonnei)*. Prisotnost omenjenih plazmidov so dokazali v 54 % sevov *E. coli*, izoliranih iz fecesa zdravih ljudi, ki niso jemali antibiotikov ter v 67 % sevov *E. coli*, izoliranih iz fecesa testnih subjektov iz različnih družin ptic.

2.4 REPLIKONI PLAZMIDOV SKUPINE IncF

2.4.1 RepFIA

Replikon RepFIA je približno 6.500 bp dolg odsek DNA, ki je razdeljen na tri dele. Prvi je *ori2*, drugi je esencialni replikacijski gen *repE*, katerega produkt je odgovoren za iniciacijo replikacije in *incC*, ki regulira vzdrževanje oz. število kopij plazmida v gostiteljski celici. Plazmid, sestavljen iz teh delov, se v gostiteljski celici normalno podvaja, vendar je vseeno nestabilen. Stabilnost zahteva mehanizem, ki usmerja kopije v vsako izmed novih gostiteljskih celic. Ta mehanizem je zapisan v pridruženem lokusu *sop*. Lokus *sop* v bistvu predstavlja delitveni modul, ki lahko deluje neodvisno od lokusa *rep*, na katerega je sicer običajno vezan. Lokusa *rep* in *sop* konstituirata osnovni vzdrževalni sistem plazmida F. Ta ureditev je presenetljivo podobna ureditvi profaga P1. Če zaporedje replikona RepFIA izoliramo v samostojen mini-F plazmid, se lahko ta samostojno vzdržuje in podvaja (Murakami in sod., 1987).

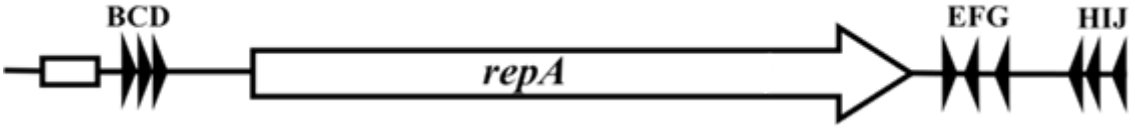
Slika 5: Karta replikona RepFIA (F Plasmid – Molecular Biology, 2016)
Označeni so geni/regije tipičnega replikona RepFIA. Razlaga posamičnih genov je v tekstu.

2.4.2 RepFIB

Replikacijsko regijo RepFIB so prvič odkrili pri plazmidu F, na 7,5 kb dolgem fragmentu *Eco*RI f7. Replikon RepFIB, ki sta ga prvič izolirala Lane in Gardner (Lane in Gardner, 1979), so kasneje našli še na drugih plazmidih IncF, npr. pCG86, R386, pHH507, R453 in ColV3-K30. Preučevanje pri ostalih plazmidih, ki vsebujejo RepFIB, je pokazalo visoko homologijo DNA med vsemi temi plazmidi v približno 2 kb dolgi regiji. Replikacijska regija RepFIB je zelo majhna regija, saj vsebuje samo en bralni okvir. Ta kodira odločilen protein za replikacijo plazmida, RepA, velik od 28,9 do 34,7 kDa (Picken in sod., 1984; Bergquist in sod., 1985; Bergquist in sod., 1986; Perez-Casal in Crosa, 1984)

Replikacijsko regijo RepFIB sestavlja gen *repA*, ki je navzgor obdan z direktnimi trojnimi ponovitvami (B, C in D) oziroma iteroni in navzdol s šestimi kompleksnejšimi ponovitvami (E, F, G, H, I in J). Zaporedje, ki se ponavlja, je 5'-ANATAAGCTTAGNNNGYAAA-3'. Pri ponovitvah E, F, G, H, I in J je odgovorno za inkompatibilnost replikona RepFIB s skupino IncE. Zaporedje *repA* je pri RepFIB homologno z ostalimi plazmidi, ki vsebujejo *repA* (R6K, RK2, pSC101), čeprav ni nobene podobnosti med ponavljajočimi se zaporedji (slika 6) (Gibbs in sod., 1993).

Uravnavanje podvajanja poteka preko vezave proteina RepA na iteronska zaporedja (Gibbs in sod., 1993).



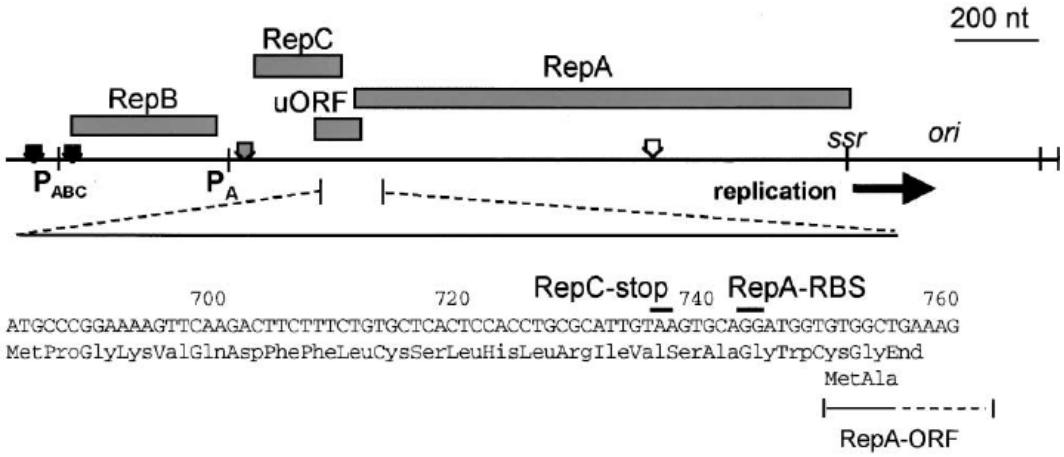
Slika 6: Karta replikona RepFIB (Gibbs in sod., 1993)
Minimalni replikon RepFIB je sestavljen iz samo enega gena, ki je nujen regulator za replikacijo, gen repA. repA je obdan z obeh strani s ponavljajočimi se zaporedji DNA (t.i. iteroni): BCD, EFG in HIJ. BCD ponavljajoča se regija s pripadajočim predzaporedjem označenim kot bel pravokotnik tvorita regijo začetka podvajanja.

2.4.3 RepFIC

Z razliko od RepFIB in RepFIA, ki sta iteronska tipa replikonov, je RepFIC tip protiprepisne replikacijske regije, katere kontrola je podobna replikonom plazmidov R1 in R100 v *E. coli*

in ColIb-P9 pri bakteriji *Shigella*. V svoji divji obliki vsebujejo promotorsko regijo, ki ji sledijo tri genske kasete. Prva je opcijska, drugi dve pa ključni za replikacijo. Opcijska kaseta I vsebuje gen za represorski protein RepA2 in njegovo tarčo, promotor Pa. Kaseta II vsebuje informacijo za protiprepisno lastnost replikona in uORF (angl. upstream ORF). Kaseta III vsebuje strukturne gene proteina RepA1 in minimalno mesto *ori*, zato jo imenujemo tudi kaseta RepA-*ori*. Proteini RepA1 različnih replikonov tega tipa imajo vsak svoja specifična mesta *ori*, vselej v smeri 5' proti 3' (Couturier in sod., 1988).

Prepis genov replikona RepFIC rezultira v dveh proteinih, uORF in RepA1. RepA2, ki zaradi svoje neključne narave ni vedno izražen, je klasični represor transkripcije. Njegova funkcija je utišanje promotorja Pa (Maas in Wang, 1997).



Slika 7: Karta replikona RepFIC (Maas, 2001)
Shema je v merilu. Puščice nakazujejo smer prepisa in translacije. Sivi okvirji kažejo izražene proteine. Prepis, združen z zamikom okvirja cistein-metionin je prikazan v povečani podrobnosti.

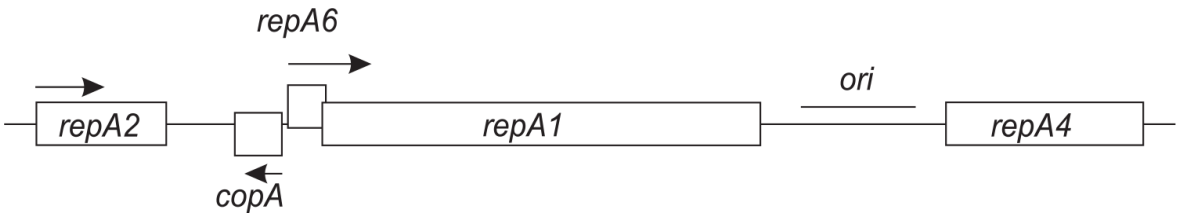
Replikona RepFIC in RepFIIA si delita tri močno ohranjene regije homologije, kar kaže na sorodne mehanizme regulacije, zato jih oba lahko uvrščamo tudi v razširjeno družino replikonov RepFIIA (Saadi in sod., 1987).

2.4.4 RepFIIA

Replikoni družine RepFIIA tipično vsebujejo 5 genov. Prvi je gen *repA2*, ki vsebuje zapis za represorski protein, ki s svojo vezavo na promotor gena *repA1* zavira sintezo mRNA s tega gena in tako preprečuje nastanek replikacijskega proteina Rep (Vanooteghem in Cornelis, 1990). Drugi gen *copA* kodira protiprepisno RNA molekulo, ki regulira prevajanje mRNA gena *repA1* (Vanooteghem in Cornelis, 1990). Gen *copA* je tipično dolg 90 bp, izjemoma je lahko tudi daljši. Tretji gen v RepFIIA je *repA6*, ki kodira krajši začetni peptid. Njegovo izražanje inhibira vezava CopA ctRNA, ki tako prepreči replikacijo plazmida (Blomberg in sod., 1992). Regija je dolga le 75 bp. Četrti gen je *repA1*, ki kodira protein RepA, ključen za replikacijo plazmida (Helinski in sod., 1996). Domneva se, da represor RepA2 regulira prepis mRNA gena *repA1*, medtem ko protiprepisna RNA gena *copA*, ki je

komplementarna začetnemu zaporedju mRNA gena *repA1*, regulira translacijo (Starčič Erjavec in Žgur-Bertok, 2006). Peti in zadnji gen replikona RepFIIA je regija gena *repA4*, ki vpliva na stabilnost plazmida v gostiteljski celici (Jiang in sod., 1993).

Replikoni v tej družini so mozaični (Osborn in sod., 2000), posamični geni, prisotni v replikonu izvirajo iz različnih virov.



Slika 8: Karta replikona RepFIIA (Starčič Erjavec in Žgur-Bertok, 2006)
Označeni so geni/regije tipičnega replikona RepFIIA. Zaradi razvidnosti lege preučevanih zaporedij, so nekateri okvirčki, ki prikazujejo gene/regije, premaknjeni. Označena je tudi smer prepisa mRNA iz posameznega gena. *ori* je regija, kjer se veže replikatorski protein RepA1 in prične s podvajanjem plazmida.

Preglednica 1: Predvideni geni/regije replikona RepFIIA (AY234377) plazmida pRK100 (Starčič Erjavec in Žgur-Bertok, 2006)

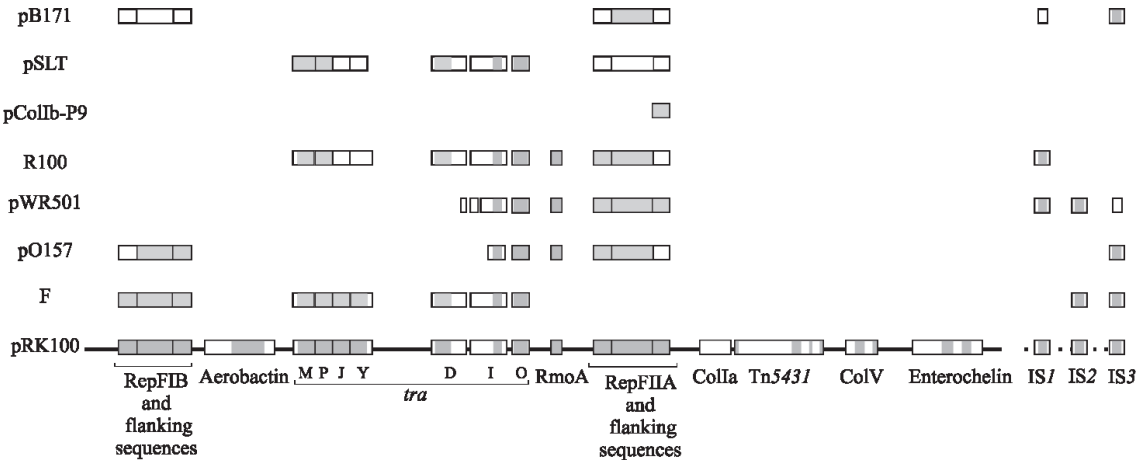
Predviden gen/regija	Okvir	Začetek (bp)	Konec (bp)	Dolžina (bp)
repA2	+3	1401	1661	261
copA	+1 C	1874	1782	93
repA6	+2	1886	1960	75
repA1	+3	1953	2810	858
repA4	+2	3173	3556	384

2.5 MOZAIČNOST

Mozaična narava plazmidov je široko sprejeto dejstvo. Pojavi se zaradi različnih genskih mobilnih elementov kot so transpozicijski elementi (transpozoni in insercijska zaporedja), integroni, genske kasete (Hall in Vockler, 1987; Pansegrau in sod., 1994).

Mozaičnost velikih naravnih plazmidov skupine IncF (npr pRK100) se kaže v njihovi sestavi, saj vsebujejo elemente, ki so sicer prisotni v kromosomih gostiteljskih celic in drugih naravnih velikih plazmidih. Pojavi se zaradi različnih rekombinacijskih dogodkov, ki so pogosto opaženi na regijah s transpozicijskimi elementi (npr. IS) (Starčič Erjavec in sod., 2003).

Mozaičnost pRK100 kaže na to, da je plazmid himera, sestavljen iz elementov več plazmidov in kromosomov gostiteljskih celic (Starčič Erjavec in sod., 2003). Analiza nukleotidnih zaporedij plazmida pRK100 je pokazala, da ima zapise, ki se nahajajo tudi na drugih plazmidih (slika 9).

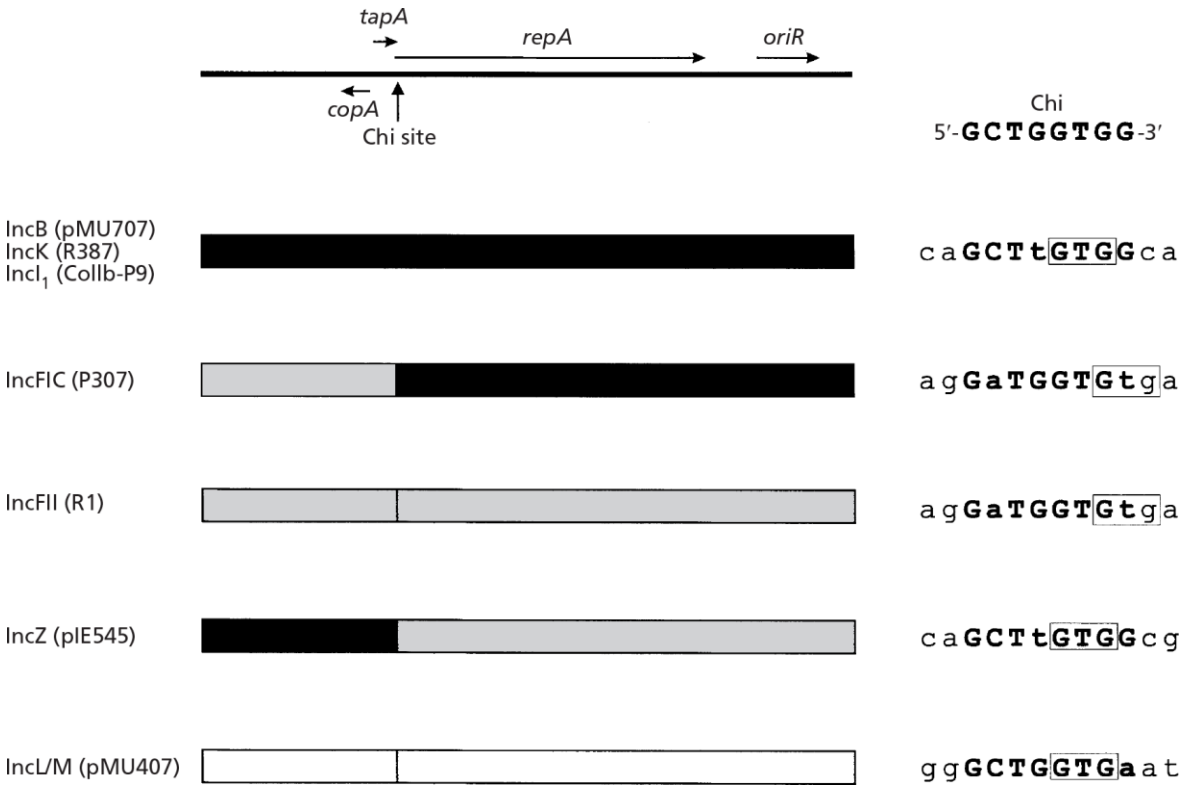


Slika 9: Mozaičnost plazmida pRK100 (Starčič Erjavec in sod., 2003)

S pomočjo analize sekvenciranih regij plazmida pRK100 BLAST se mozaičnost določenih regij genoma lepo izrazi. Osenčene regije slike 9 kažejo na veliko podobnost (>95 %) s fragmenti drugih plazmidov. Prekinitve sicer zveznih regij na drugih plazmidih kažejo na izgubo določenih elementov regije (Starčič Erjavec in sod., 2003).

Tipičen primer mozaičnosti pRK100 je regija *tra*, ki omogoča konjugacijski prenos plazmida in RepFIB, za katero kaže, da jo je pridobil iz drugih F-sorodnih plazmidov. RepFIIA po vsej verjetnosti izhaja iz R1-podobnih plazmidov, sistem za privzem železovih ionov (aerobaktin) in kolicin V verjetno izvirajo iz pColV-podobnih plazmidov (Starčič Erjavec in sod., 2003).

Povišani nivoji rekombinacije so povezani s prisotnostjo mest *Chi*, sestavljenih iz nukleotidnih oktamer 5'-GCTGGTGG-3' (Smith in sod., 1981). Mesta *Chi* prepoznava protein RecBCD (*E. coli*), ki omogoča homologno rekombinacijo (Smith, 1987; Taylor in Smith, 1995). Raziskava replikonov skupin IncB, FII, FIC, I, K, L/M in Z pokaže prisotnost *Chi*-podobnih zaporedij na začetku genov, ki kodirajo replikacijske proteine, kar razlaga mozaičnost (slika 10) (Osborn in sod., 2000).



Slika 10: Mozaična struktura replikonov z visoko stopnjo homologije z replikoni IncFII in zaporedja potencialnih mest *Chi* (Osborn in sod., 2000)
Relacija med replikoni je pokazana z nivojem osivljenosti:
črna – podobna IncB/I/K zaporedjem,
siva – podobna IncFII zaporedjem,
bela – podobna IncL/M zaporedjem.
Pozicija genov, udeleženih pri replikaciji R1 in plazmidov IncFII, je prikazana na sliki. Pozicija mest *Chi* je prikazana odebeljeno. Začetni kodoni genov, ki kodirajo replikacijske proteine, so prikazani v okvirjih.

3 MATERIAL IN METODE

3.1 MATERIALI

3.1.1 Strojna oprema - platforma

3.1.1.1 Razvojno okolje

Za razvoj analitičnega orodja JuP 1.0 in njegovo uporabo za namen pridobitve rezultatov tega diplomskega dela smo uporabili prenosni računalnik Lenovo T450s, ki temelji na 64 bitnem procesorju Intel Core i7 5600U pri 2,60GHz, 12GB delovnega spomina, 512GB SSD trdi disk in operacijskem sistemu Microsoft Windows 10 Professional v sl_SI lokalizaciji.

3.1.1.2 Izvajalno okolje

Za izvajalno okolje analitičnega orodja JuP 1.0 smo uporabili virtualno okolje VMware vSphere 6 Enterprise Plus, delujoče na fizični strojni opremi HP ProLiant DL380 Gen9, 2x Intel Xeon E5-2640 v3 pri 2,60GHz, 128GB delovnega spomina in 1,02TB SSD lokalnemu diskovnemu polju.

Virtualni strežnik za izvajanje aplikacije smo definirali z 8 vCPU, 4GB delovnega spomina, 128GB SSD diskovne kapacitete, ki je temeljil na verziji 11 virtualne strojne opreme VMware in deloval v neredundančni konfiguraciji. Na virtualni strežnik smo namestili operacijski sistem CentOS Linux release 7.2.1511 (Core) z zadnjimi popravki na dan 15. 5. 2016 (jedro: 3.10.0-327.18.2.el7.x86_64).

3.1.2 Programska oprema

3.1.2.1 Razvojno okolje:

Za razvoj zaledne skripte za pridobivanje in analizo rezultatov BLAST (Altschul in sod., 1990) in GenBank (Sayers in sod., 2009), ročno analitiko različnih rezultatov funkcij BioPython (Cock in sod., 2009), funkcijske teste posameznih sklopov aplikacije ter razvoj in testiranje spletne aplikacije smo uporabili prenosni računalnik z Microsoft Windows 10 in naslednjo programsko opremo:

- Notepad++ v6.8.8 (Ho, 2016) z dodatki:
  - XML Tools 2.6.8,
  - JSON Viewer 1.22,
  - JSTool 1.16.10
- Python 2.7.6 32 bit (Python Software Foundation, 2016) z dodatnimi moduli:
  - Python 2.7 biotypthon-1.66 (Cock in sod., 2009)

- Spletna komponenta aplikacije JuP 1.0:
  - Bootstrap v3.3.6 (HTML, CSS in JS ogrodje) (Twitter, 2016)
  - Font Awesome v4.6.3 (CSS ogrodje in grafični elementi) (Gandy, 2016)
  - bootstrap-slider v7.1.1 (dodatek za prikaz drsnikov) (Kemp in Kalkur, 2016)
  - Bootstrap Colorpicker v2.3.3 (dodatek za prikaz drsnikov) (Petre in Aguiar, 2016)
  - jQuery (JS ogrodje, kompresirana produkcijska verzija) v2.2.4 (The jQuery Foundation, 2016)
- Firefox brskalnik v46.0.1 v en_US lokalizaciji

3.1.2.2 Izvajalno okolje

- CentOS Linux 7.2.1511 (Core) (popravki na dan 15.5.2016) z naslednjimi dodatki:
  - httpd-2.4.6-40.el7.centos.1.x86_64 z nujnimi dodatki, ki omogočijo njegovo namestitev
  - mod_python 3.5.0 (izvorna koda, prevedena v izvajalnem okolju)
  - python-2.7.5-34.el7.x86_64 z nujnimi dodatki, ki omogočijo njegovo namestitev
  - numpy-1.7.1-11.el7.x86_64 z nujnimi dodatki, ki omogočijo njegovo namestitev
  - python-devel-2.7.5-34.el7.x86_64 z nujnimi dodatki, ki omogočijo njegovo namestitev
  - httpd-devel-2.4.6-40.el7.centos.1.x86_64 z nujnimi dodatki, ki omogočijo njegovo namestitev
- Python dodatki nameščeni prek rpm ali pip orodja:
  - numpy (1.7.1)
  - mod-python (3.5.0)
  - biopython (1.66)

3.2 METODE

3.2.1 Priprava rezultatov BLAST za vizualizacijo in dodatno obdelavo

Glavno orodje za iskanje ujemajočih se genskih zaporedij je spletna aplikacija nukleotid-nukleotid BLAST oz. *blastn* ameriškega nacionalnega inštituta za zdravje NIH oziroma njihovega oddelka za biotehnološko informatiko (National Center for Biotechnology Information – angl. NCBI) (Altschul in sod., 1990). Aplikacija poleg uporabe prek spletne strani http://blast.ncbi.nlm.nih.gov/ omogoča njeno uporabo tudi prek programskih vmesnikov, ki omogočajo integracijo v programske rešitve tretjih ponudnikov, kot je to primer z JuP.

Na voljo je več javno dostopnih in brezplačnih ogrodij za zaledno delo z aplikacijo BLAST, ki temeljijo na njegovem spletnem programskem vmesniku in uporabnikom omogočajo integracijo aplikacije BLAST v njihove programske rešitve. Razlika med njimi je v glavnem

v uporabljeni tehnologiji izvedbe in odločitev katero izbrati navadno temelji na tehnologiji in arhitekturi končne programske rešitve.

BioPython (Cock in sod., 2009) je ogrodje, ki temelji na tehnologiji in programskem jeziku Python (Python Software Foundation, 2016), ki je široko uporabljan jezik v sodobnih, zalednih sistemih. Zaradi možnosti uporabe vhodnih parametrov omogoča izvajanje natančnih poizvedb BLAST, ki izboljšajo kakovost rezultatov. Modul *Bio.Blast.NCBIWWW* ogrodja BioPython sproži nukleotid-nukleotid primerjavo BLAST v oddaljeni spletni aplikaciji in rezultate klicatelju vrne oblikovane v razširljivem označevalnem jeziku ali krajše XML. Modul *Bio.Blast.NCBIXML* ogrodja BioPython te rezultate prek transformacije obdela s pomočjo shem in jih klicatelju vrne v obliki znanih konstruktov objektnega modela BioPython, ki jih lahko aplikacija Python uporabi brez dodatne obdelave (nizi objektov z znanimi lastnostmi, atributi in metodami). Tako pridobljeni podatki so sorodni tistim iz spletne različice aplikacije, vendar so z razliko od različice HTML primernejši za računalniško obdelavo.

V okviru analize rezultatov BLAST je bilo potrebno poiskati podatke, ki so potrebni za želeno funkcijo aplikacije JuP, jih ovrednotiti, razvozlati format, ki je v dokumentaciji ogrodja BioPython precej slabo opisan in preveriti stabilnost transformacije na več testnih vektorjih.

Koncept iskanja podobnih nukleotidnih zaporedij preko nukleotid-nukleotid analize BLAST temelji na predpostavki, da obstajajo statistično pomembnejša območja ujemanja nukleotidnih zaporedij, ki jim pravimo visoko točkovana parna območja (angl. High scoring Segment Pairs ali HSP). Te aplikacija BLAST išče prek hevristične metode, ki je sorodna Smith-Waterman algoritmu. Rezultat poizvedbe je niz identificiranih genskih zaporedij v okviru katerih je vsakemu pripisano eno ali več visoko točkovanih parnih območij.

Rezultat poizvedbe komponenta filtrira glede na vhodne parametre filtriranja (vključna in izključna gesla). Če so ti podani, jih pretvori v prenosljivo obliko za vizualizacijo v spletnem delu aplikacije in en del uporabi za vhodne podatke pridobivanja z rezultati povezanih podatkov o genih iz aplikacije GenBank, ki jih analiza BLAST ne vrača.

3.2.2 Povezava rezultatov BLAST in GenBank

GenBank (Sayers in sod., 2009) je baza genskih podatkov ameriškega nacionalnega inštituta za zdravje NIH, ki na enem mestu nudi javni dostop do informacij nukleotidnih zaporedij iz baz DDBJ, EMBL in GenBank pri NCBI, ki dnevno izmenjujejo podatke o novo deponiranih genskih zaporedjih.

Primerjava BLAST omogoča iskanje ujemajočih se nizov nukleotidov in njihovo poravnavo, ne pa tudi dostopa in analize do označb zaporedij, kar je glavni namen razvoja JuP. S podatkom o identifikatorju organizma, v katerem je identificirano ujemajoče se zaporedje in

nukleotidnim razponom ujemanja, je prek vpogleda v bazo podatkov GenBank možno pridobiti informacijo o označbah regije genskega zaporedja, ki je za zaporedja proteinskih produktov označena s kratico CDS.

Vhodni parameter drugemu delu zaledne komponente JuP so podatki o identificiranem ujemajočem zaporedju nukleotidov ter niz z njim povezanih visoko točkovanih parnih območij. Ta del aplikacije za vsako visoko točkovano parno območje prek modula *Bio.Entrez* ogrodja BioPython izvede spletno poizvedbo v bazo podatkov GenBank in pridobi obstoječe označbe z njihovimi podrobnostmi na razponu dotičnega območja HSP. Rezultate poizvedbe *Bio.Entrez* s pomočjo modula *Bio.Entrez.Parser* pretvori v objekte ogrodja BioPython, ki so primerni za računalniški dostop do rezultatov poizvedbe. Visoko točkovana parna območja brez označb CDS zavrže, saj za tip primerjave, ki ga implementira JuP, niso relevatna.

Neučinkovitost in dolgotrajnost zaporedne obdelave rezultatov BLAST, ki so sicer med seboj neodvisni, nas je vodila v večnitno izvedbo teh poizvedb. Uporabili smo koncept skupine niti (angl. thread pool) s privzetimi 5 nitmi v skupini (število je sicer nastavljivo), kjer vsaki prosti niti vzporedno dodelimo obdelavo enega elementa rezultata poizvedbe BLAST in njegovih visoko točkovanih parnih območij. Ko nit obdela dodeljen zahtevek, se vrne v skupino prostih niti in v obdelavo vzame naslednji zahtevek. Tako smo analizo časovno in performančno bistveno izboljšali.

Tako pridobljene podatke o kodnih zaporedjih proteinov (CDS) pripišemo v niz objektov CDS vsakemu objektu HSP in jih pripravimo za vizualizacijo, normalizacijo ter kasnejšo navzkrižno analizo s podatki ujemanja BLAST.

3.2.3 Normalizacija podatkov BLAST in GenBank

Za pravilno poravnavo regij CDS visoko točkovanih parnih območij z poravnavo BLAST in iskanje stopnje ujemanja je ključna normalizacija pridobljenih podatkov. Analiza BLAST kompenzira za vse vrzeli neujemanja vhodnega in tarčnega zaporedja. Izhodna podatka o začetku in koncu visoko točkovanega parnega območja sta torej neposredno primerljiva z odsekom tarčnega nukleotidnega zaporedja iz baze podatkov, medtem ko je dolžina baznih parov primerjave BLAST podaljšana za število vrzeli v primerjavi.

V kolikor izvedemo poizvedbo v bazo GenBank za območje med začetkom in koncem ujemanja s tarčnim zaporedjem, bomo pridobili vse zapise CDS in njihove odseke. Njihovi začetki in konci niso neposredno primerljivi z odseki na poravnavi, saj ne vsebujejo kompenzacije vrzeli. Tako smo morali za pravilno poravnavo regije CDS nekega visoko točkovanega parnega območja in poravnave BLAST kompenzirati podatka o začetku in koncu regije CDS za število vrzeli v poravnavi na tistem odseku. Število vrzeli na nekem odseku primerjave nukleotidnega zaporedja predstavlja seštevek znakov '−' v vrnjenem zaporedju primerjave.

Normalizacija je ključna za pravilno poravnavo elementov pri vizualizacji (CDS[QUERY_START] in CDS[QUERY_END]), izračunih in statistike stopnje ujemanja poravnave (niz poravnave HSP na odseku od CDS[ALIGN_START] do CDS[ALIGN_END] in izračuna statistike odseka CDS (število vrzeli v nizu nukleotidov QUERY in SUBJECT poravnave HSP od CDS[ALIGN_START] do CDS[ALIGN_END]).

Surovi podatki, pomembni za izračun pozicije elementa CDS v poravnavi, ki jih pridobimo iz baz podatkov, so:

HSP[QUERY_START]	začetek HSP v vhodnem zaporedju
HSP[QUERY_END]	konec HSP v vhodnem zaporedju
HSP[SUBJ_START]	začetek HSP v najdenem zaporedju
HSP[SUBJ_END]	konec HSP v najdenem zaporedju
HSP[ALIGN_LEN]	dolžina ujemanja HSP
CDS[HSP_SUBJ_START]	začetek regije CDS v najdenem zaporedju relativno na HSP
CDS[HSP_SUBJ_END]	konec regije CDS v najdenem zaporedju relativno na HSP

3.2.3.1 Normalizacija izhodnih podatkov v primeru verig plus/plus

Formula za izračun odseka elementa CDS v poravnavi zaporedju ALIGN_xxx in QUERY_xxx:

CDS[ALIGN_START] = CDS[HSP_SUBJ_START] + število vrzeli v zaporedju SUBJECT zaporedju od začetka do nukleotida na poziciji CDS[HSP_SUBJ_START]

CDS[ALIGN_END] = CDS[HSP_SUBJ_START] + število vrzeli v zaporedju SUBJECT od začetka do nukleotida na poziciji CDS[HSP_SUBJ_END]

CDS[QUERY_START] = HSP[QUERY_START] + CDS[ALIGN_START] – število vrzeli v zaporedju QUERY do CDS[ALIGN_START] – 1

CDS[QUERY_END] = HSP[QUERY_START] + CDS[ALIGN_END] – število vrzeli v zaporedju QUERY do CDS[ALIGN_END] – 1

Postopek normalizacije lahko prikažemo na primeru analize ujemanja plus/plus zaporedij pod akcesijskima številkama AY234375 (vhodno zaporedje) in KF719970 (najdeno zaporedje), ki jih vrne orodje BLAST (E=10.0, št. BLAST zadetkov 500) v spletnem načinu in načinu dostopa s programskim vmesnikom:

```
                |======> smer iskanja vrzeli v zaporedju QUERY
Query  1685     TGCTAGATTACTGATCGTTTAAGGAATTTTGTGGCTG-GCCACGCCGTAAGGTGGCAAGG  1743
                ||||||||  |||| || | |  ||||||| || || ||||||||||||||||||| ||
Sbjct  174332   TGCTAGATTGTTGATGGTCTGAATAATTTTG-GG-TGTGCCACGCCGTAAGGTGGCAGGG  174389
                |======> smer iskanja vrzeli v zaporedju SUBJECT
Query  1744     AACTGGTTCTGATGTGG-ATTTA-CAGGAGCCAGAAAAGCAAAAACCCCGATAATCTTCT  1801
                |||||||||||||| || || || | ||  |||| ||||||||||||||||||||||||
Sbjct  174390   AACTGGTTCTGATGAGGTATCTACCCGGGACCAG-AAAGCAAAAACCCCGATAATCTTCT  174448

Query  1802     TCAACTTTGGCGAGTACGAAAAGATTACCGGGGCCCACTTAAACCGTATAGCCAAC-AAT  1860
                ||||  |||||| | | |||||||||| |||||||  | ||||| | ||||  |||   |
Sbjct  174449   TCAATCTTGGCG-GAAGGAAAAGATTAACGGGGCCTTCATAAACTGCATAG--AACGTGT  174505
                                          |===> začetek regije CDS
Query  1861     TCAGCTATGCGGGGAGTATAGTTATATGCCCGGAAAAGTTCAAGACTTC-TTTCTGTG-C  1918
                |   |||||| |||||||||  || |||| | ||||| |||||| ||   ||||||||  |
Sbjct  174506   TGCTCTATGCAGGGAGTATATGTACATGCTCAGAAAACTTCAAG-CTCAGTTTCTGTGTC  174564
                                    konec regije CDS  ===>|
Query  1919     --TCGCTCCTTCTGCGCATTGTAAGTGCAGGATGGTGTGACTGATCTTC-ACCAAA-C-G  1973
                  ||||||||||||| |||   |||| |||||| |  ||||||||||| |  | ||| |
Sbjct  174565   ATTCGCTCCTTCTGTGCAACATAAGCGCAGGAAGCGGTGACTGATCTCCTTCAAAATCAC  174624

<nadaljevanje segmenta odstranjeno zaradi primernejšega prikaza in nerelevantnosti>
```

Slika 11: Prikaz plus/plus ujemanja vhodnega zaporedja AY234375 in najdenega ujemajočega se zaporedja v KF719970, kot ga prikaže spletni vmesnik BLAST z ročno označeno regijo gena *tapA*

```
                <odstranjen začetni segment informacije>

     gene            200..277
                     /gene="tapA"
     CDS             200..277
                     /gene="tapA"
                     /codon_start=1
                     /transl_table=11
                     /product="Leader peptide, replication control"
                     /protein_id="AHG55683.1"
                     /db_xref="GI:575010271"
                     /translation="MLRKLQAQFLCHSLLLCNISAGSGD"

        <odstranjen segment zaradi nerelevatnosti informacije>

ORIGIN
        1 tgctagattg ttgatggtct gaataatttt gggtgtgcca cgccgtaagg tggcagggaa
       61 ctggttctga tgaggtatct acccgggacc agaaagcaaa aaccccgata atcttcttca
      121 atcttggcgg aaggaaaaga ttaacggggc cttcataaac tgcatagaac gtgttgctct
      181 atgcagggag tatatgtaca tgctcagaaa acttcaagct cagtttctgt gtcattcgct
      241 ccttctgtgc aacataagcg caggaagcgg tgactgatct ccttcaaaat cactattcac
      301 aggttaaaaa cccgaatccg gtattcacgc cgcgtgaagg gaaaaagacc ctgccgttct
      361 gccgtaagct gatggcgaaa gccgaaggct tcacgtcccg ttttgatttt tccatccatg

                <odstranjen segment informacije>
```

Slika 12: Prikaz izhodnih podatkov iz baze podatkov GenBank za odsek visoko točkovanega parnega območja z ročno označeno regijo gena *tapA*

Surovi podatki analize iz baz podatkov:

HSP[QUERY_START] = 1685
HSP[QUERY_END] = 3479
HSP[SUBJ_START] = 174332
HSP[SUBJ_END] = 176127
HSP[ALIGN_LEN] = 1810
CDS[HSP_SUBJ_START] = 200
CDS[HSP_SUBJ_END] = 277

Normalizirani podatki, ki jih zaledna komponenta kot svoje izhodne podatke preda vizualizacijski komponenti:

CDS[ALIGN_START] = 200 + 6 = 206
CDS[ALIGN_END] = 277 + 7 = 284
CDS[ALIGN_LEN] = 284 – 206 + 1 = 79
CDS[QUERY_START] = 1685 + 206 – 4 – 1 = 1886
CDS[QUERY_END] = 1685 + 284 – 8 – 1 = 1960

3.2.3.2 Normalizacija izhodnih podatkov v primeru verig plus/minus

Normalizacijo pridobljenih podatkov iz visoko točkovanih parnih območij na komplementarnih verigah smo izvedli invertno. Poizvedba po označbah CDS na komplementarnih verigah vrne podatke obrnjene, zato je potrebno podatke o začetku in koncu regije CDS glede na poravnavo obrniti, število vrzeli pa določiti od konca pokrivanja v smer začetka, saj gre za komplement.

Formula za izračun odseka CDS elementa v poravnavi zaporedja ALIGN_xxx in QUERY_xxx:

CDS[ALIGN_START] = HSP[ALIGN_LEN] – CDS[HSP_SUBJ_START] – število vrzeli v zaporedju SUBJECT od konca do nukleotida na poziciji CDS[HSP_SUBJ_START] + 1

CDS[ALIGN_END] = HSP[ALIGN_LEN] – CDS[HSP_SUBJ_END] – število vrzeli v zaporedju SUBJECT od konca do nukleotida na poziciji CDS[HSP_SUBJ_END] + 1

CDS[QUERY_START] = HSP[QUERY_START] + CDS[ALIGN_START] – število vrzeli v zaporedju QUERY od začetka do nukleotida na poziciji CDS[ALIGN_START] – 1

CDS[QUERY_END] = HSP[QUERY_START] + CDS[ALIGN_END] – število vrzeli v zaporedju QUERY od začetka do nukleotida na poziciji CDS[ALIGN_END] – 1

Postopek normalizacije lahko prikažemo na primeru analize ujemanja plus/minus zaporedij pod akcesijskima številkama AY234375 (vhodno zaporedje) in AY091607.1 (najdeno zaporedje), ki jih vrne orodje BLAST (E=10.0, št. BLAST zadetkov 500) v spletnem načinu in načinu dostopa z programskim vmesnikom:

se nadaljuje.

Slika 13: Prikaz plus/minus ujemanja vhodnega zaporedja AY234375 in najdenega ujemajočega se zaporedja v AY091607.1, kot ga prikaže spletni vmesnik BLAST z ročno označeno regijo gena *repA3*

nadaljevanje slike 13.

Slika 13: Prikaz plus/minus ujemanja vhodnega zaporedja AY234375 in najdenega ujemajočega se zaporedja v AY091607.1, kot ga prikaže spletni vmesnik BLAST z ročno označeno regijo gena *repA3*

```
        <odstranjen začetni segment informacije>

     gene            complement(1819..1947)
                     /gene="repA3"
     CDS             complement(1819..1944)
                     /gene="repA3"
                     /note="similar to Shigella flexneri plasmid R100
                     replication-associated protein A3"
                     /codon_start=1
                     /transl_table=11
                     /product="RepA3"
                     /protein_id="AAM14716.1"
                     /db_xref="GI:22035193"
                     /translation="MWIYRSQKSENPDNLLQVWRLERLPGPSKNRIANNSAMRGV"
ORIGIN
        1 tatgcagagt atataacctt tcagacacag ccaccaacat atgccgcgcc agcggcatgg
       61 aaggcggcac tcagttgttg catatgatac cggagtaaaa ccgccgaagc ccggcgtcag
      121 ccggtactga ttgatagatt tcaccttacc catccccagc cctgccagac catacccgct
      181 ttcagccatg agagagcttc tgtgcgcggt cggagtggtc ccgacgaggg tttacccgaa
      241 gtcggggcgt atctccgcgt tagcgggccg tgagggccgc ttacgagcgt gtactgagga
      301 cttccagcga gaagactgac agcgatgatg atgtagttac aacattcata attaaaagcg
      361 actctgttcc ggccctttgg gccggggcgg ggccgctttt cagttatggg agaggggctt
      421 tgtggtttcg gttctgcgct ggactggggt tttcctggtg gttgtatttg agtgttgtaa
      481 ctaaagtggc tccggttggg gcccgccgct tgcggtggga ggtgcatatc tgtctgtcca
      541 caggacaggc agtgaatagg ttttcttttt aaatgaatgt aattaagtag tttaaaggag
      601 atataaacag gtgtttaaaa gatacattgc accctgtaac gctggcggct ggcgctttat
      661 gaggtgaacg gttgtaactt tatcgggaag tcccttgcag tttaatgtgg ataagcaaaa
      721 ttacccgtct gtgaggcgta ttttgtatta aaaacagggg gaaccggatg caccagaagg
      781 tggatggtga ggttatttt tgtatgtggt gctgattttt tgtgcgctga cggacttcag
      841 gcgtgcggat gcctccggcg cgggccggat tattctgagg aggtcacttt caggggaag
      901 ctgtggccag ccgactgtaa ttgcggttac gtgacagaat catgcgctcc ttcacacgac
      961 gctccacttc acgttttacc gcctcacgac tggcagtgaa gcgcccttcc gcgatttcac
     1021 gcgtcagctg tcgtttcacc agggtgacga tatcctgacg ttccctgcct gcatcacgac
     1081 gcgcacgggc acgctttatt ccccgggact taagctctgt ctggtaactg cggaaacgct
     1141 cacgcacaaa acgccaggct ttcgctatca gttcatccat acccagggta tccagcccct
     1201 gctttttgcg ctgtctgttt tcccattcca cacggctgcg gcgcgcggcg gccactgcct
     1261 cctctgatac atcgagggca gcaaacagtg cagatgtgaa cgtgatatcg gtcggaatgt
     1321 agcacccgat aagcgggtca tattccgtct ggtaggtaat cagtcccagc tctgacagga
     1381 acgtcagggc acgggtggca cgggtgatgg agagttttcc tgcaccggac tctgtcgcca
     1441 gtccgcactc aatggccagc gtggtgatgg agcactggac gcggttggcc agcggatcat
     1501 agtgaaaaca gagtccctgc agcagcgcat caatagcccg tcgacgcagc accggtggca
     1561 tgcgccgacg caaaccacgg gaacgggcat gcgccacatg aatggcgaaa tcaaaacggg
     1621 aggtgaaacc caccgccttt tccatcagtt tttcgcagaa cttcagcgtt ccggcacctt
     1681 tacggggtgt aaacaccgga ttcgggttct ttacctggcg gtaatacgtt tgttgaagat
     1741 cagtcacacc atcctgcact tacaatgcgc agaaggagtg agcacagaaa gaagtcttga
     1801 acttttccgg gcatataact atactccccg catagctgaa ttgttggcta tacggttttt
     1861 agatggcccc ggtaatcttt ctagtcgcca aacttgaaga agattatcgg ggttttcact
     1921 tttctggctc ctgtaaatcc acatcagaac cagttccctg ccaccttacg gcgtggccag
     1981 ccacaaaatt ccttaaacga tcagtaatct agca
```

Slika 14: Prikaz izhodnih podatkov iz baze podatkov GenBank za odsek visoko točkovanega parnega območja z ročno označeno regijo gena *repA3*

Surovi podatki analize iz baz podatkov:

HSP[QUERY_START] = 1685
HSP[QUERY_END] = 3702
HSP[SUBJ_START] = 5939
HSP[SUBJ_END] = 3926
HSP[ALIGN_LEN] = 2019
CDS[HSP_SUBJ_START] = 1944
CDS[HSP_SUBJ_END] = 1819

Normalizirani podatki, ki jih zaledna komponenta kot svoje izhodne podatke preda vizualizacijski komponenti:

CDS[ALIGN_START] = 2019 – 1944 – 5 + 1 = 71
CDS[ALIGN_END] = 2019 – 1819 – 3 + 1 = 198
CDS[ALIGN_LEN] = 198 – 71 + 1 = 128
CDS[QUERY_START] = 1685 + 71 – 0 – 1 = 1755
CDS[QUERY_END] = 1685 + 284 – 8 – 2 = 1960

3.2.4 Vizualizacija rezultatov v spletni komponenti

Podatki analize ujemanj regij CDS so po normalizaciji pripravljeni za vizualizacijo. Podatke zaledna skripta vrne spletni komponenti v JSON obliki po shemi iz priloge A, ki je glede na obliko zapisa najprimernejša kot vir podatkov za vizualizacijo v spletnih pregledovalnikih. Spletna komponenta JSON podatke najprej v prvi zavihek rezultatov transformira v preglednico podatkov o regijah poravnav CDS, organiziranih po organizmu in visoko točkovanem parnem območju, kjer je lociran in sortiranih po skupnem številu točk vseh visoko točkovanih parnih območij. Podatke regij HSP in označb CDS vsake regije HSP enega organizma nato združi v niz struktur podatkov, jih sortira po začetnem položaju glede na vhodno zaporedje nukleotidov in jih na podlagi dolžine vhodnega zaporedja, normalizirane lege oznake CDS in glede na poravnavo normalizirane dolžine označbe CDS v drugi zavihek izriše na pasovih tako, da se zaradi preglednosti ne prekrivajo. Vsaki oznaki dodeli tudi informacijski oblak, ki ga lahko uporabnik prikliče z pritiskom na oznako HSP ali CDS. Vsako oznako HSP in CDS obarva začetno z tonom za oznako določene barve glede na stopnjo ujemanja vhodnega in najdenega zaporedja, pri čemer je 80 % stopnja ujemanja 0 % osnovna barva (bela) in 100 % stopnja ujemanja 100 % osnovna barva. Izris lahko uporabnik aplikacije nadzoruje z oznakami na preglednici rezultatov pred imenom organizma.

3.2.5 Podrobni prikaz ujemanja s segmentacijo področij HSP in CDS

Obarvanje področja oznake HSP ali CDS glede na povprečno stopnjo ujemanja daje splošno informacijo o ujemanju celotnega območja oznake HSP ali CDS. Ta je navadno številčno prikazana tudi v statističnih podatkih spletnega in programskega vmesnika BLAST kot število identitet glede na dolžino elementa HSP. Ker so področja ujemanja v različnih odsekih ujemanja večja in manjša, je ta področja smiselno prikazati tudi podrobneje in tako natančneje označiti področja večjega in manjšega ujemanja znotraj enega področja HSP ali CDS.

Aplikacija JuP daje uporabniku možnosti barvanja stopnje ujemanja različnih področij oznak HSP in CDS tudi v t.i. segmentiranem načinu. V tem načinu vizualizacijska komponenta glede na nastavitev velikosti segmenta razdeli oznako HSP ali CDS na segmente in za vsak segment izračuna njegovo relativno stopnjo ujemanja. Uporabnik se lahko odloči in prikaže segmente:

- diskretno – vsak segment je obarvan s svojim tonom osnovne barve glede na stopnjo ujemanja segmenta po njegovi celotni površini segmenta, robovi med segmenti so ostri
- zvezno – vsakemu segmentu vizualizacijska komponenta izračuna stopnjo ujemanja in njegovo sredino ter jo obarva s tonom osnovne barve glede na izračunano stopnjo ujemanja. Barva segmenta levo in desno od sredine gradientno prehaja v ton osnovne barve, ki kaže povprečno stopnjo ujemanja s sosednjim segmentom. Robovi segmentov so tako zvezni oziroma se gradientno prelivajo glede na izračun stopnje ujemanja segmentov.

3.2.6 Analizirane replikacijske regije plazmidov skupine IncF

Pripravljeno orodje JuP smo uporabili za analizo mozaičnosti replikacijskih regij plazmidov skupine IncF. Zaporedja replikonov smo pridobili z analizo člankov, ki poročajo o opaženi inkompatibilnosti pri različnih plazmidih. V kolikor zaporedje posamičnega replikona iz člankov ni bilo samostojno deponirano v bazi nukleotidnih zaporedij NCBI, smo uporabili podatke o replikacijski regiji, ki naj bi bila povezana z določeno inkompatibilnostjo in pripravili zaporedja, izrezana iz genoma posameznega plazmida. Kjer zaporedje replikona plazmida, ki izraža določeno inkompatibilnost, ni bilo znano, smo analizirali zaporedje nuleotidov v okolici replikacijskega gena. Analizirali smo replikone, ki so prikazani v preglednici 2. Zaporedja analiziranih replikonov so prikazna v prilogah od B do I v obliki FASTA.

Preglednica 2: Analizirani replikoni inkompatibilnostnih skupin IncF

Replikon	Izvoren plazmid	Akcesijska številka	Odsek na deponiranem zaporedju	Referenca
RepFIA	plazmid F, *E. coli* K-12	AP001918	od 44600 do 53300 bp	Gubbins in sod., 2005
RepFIB	plazmid F, *E. coli* K-12	AP001918	od 36000 do 40000 bp	Gubbins in sod., 2005
RepFIC	p307, *E. coli*	M16167	celotno zaporedje	Saadi in sod., 1987
RepFIIA	pR100, *S. flexneri* 2b	AP000342	od 88200 do 90500 bp	Villa in sod., 2010
RepFIII	pSU316, *E. coli*	M26937	celotno zaporedje	López in sod., 1989b
RepFIV	pMP-R124, *P. fluorescens* R124	CM001562	od 1 do 1272 bp (*repA*) in od 43607 do 43794 bp (neanotiran konec zaporedja)	Campbell in sod., 1987
RepFVI	pSU212, *E. coli*	X55895	celotno zaporedje	López in sod., 1991
RepFVII	pSU316, *E. coli*	M28097	zaporedje *incFVII* determinante brez replikacijskih genov	López in sod., 1989

4 REZULTATI

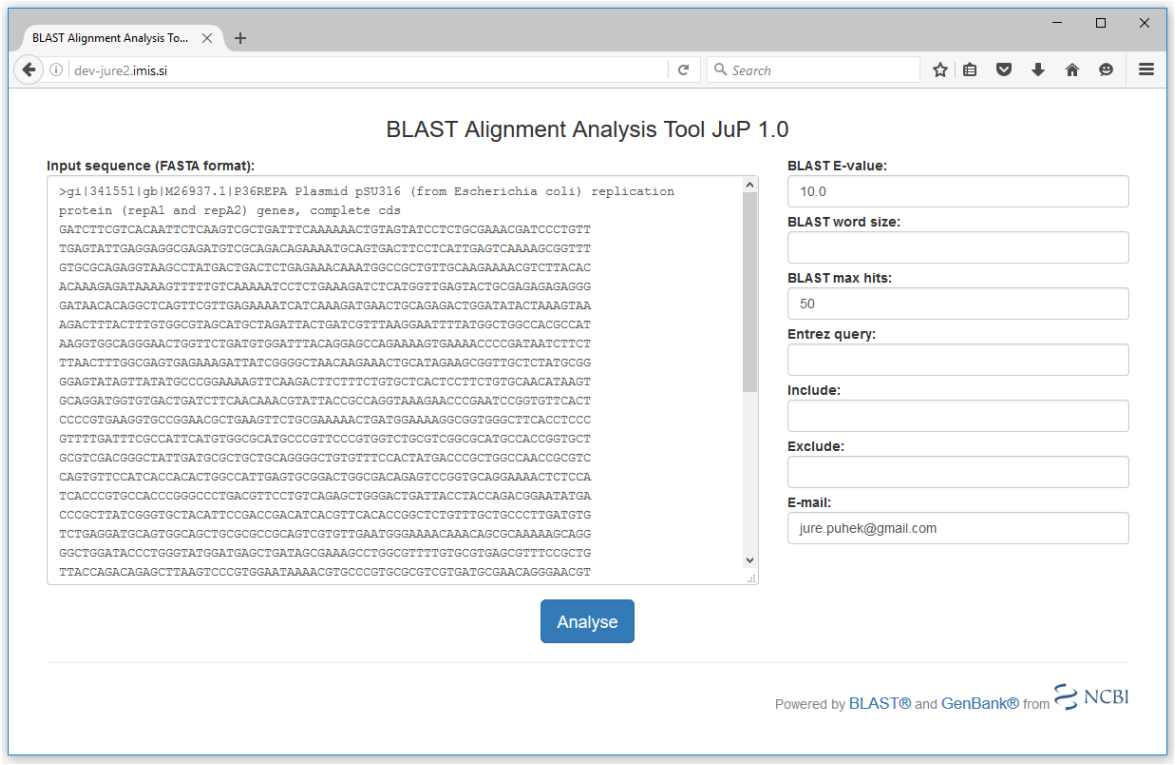
4.1 RAČUNALNIŠKA APLIKACIJA JuP

4.1.1 Vnosna maska

Na zaželeno in primerno delovanje aplikacije vplivajo posredovani vhodni parametri, ki določajo obseg in način analize vhodnega nukleotidnega zaporedja. Aplikacija JuP podpira naslednje vhodne parametre:

- »Input sequence«: niz alfa numeričnih znakov; vhodno zaporedje nukleotidov v obliki FASTA; edini obvezen vhodni parameter;
- »BLAST E value«: pozitivno decimalno število; privzeta vrednost je 10,0; vrednost E (angl. Expect value) oz. »pričakovana« vrednost, določa pričakovano število naključnih zadetkov. Manjša kot je E vrednost, manj naključnih ujemanj lahko pričakujemo v rezultatu in bolj značilna je poravnava. E vrednost ponazarja »naključen zaledni šum«;
- »BLAST word size«: pozitivna celoštevilska vrednost, večja od 7; privzeta vrednost 28; BLAST identificira homologijo zaporedja z razdelitvijo v t.i. »besede«, katere predstavljajo osnovno enoto primerjanja pri iskanju homolognih zaporedij. Osnovno pravilo določa, da naj bo iskano zaporedje najmanj 2-krat večje od velikosti »besede«. Pri krajših, gensko nestabilnih zaporedjih (npr. replikacijske regije plazmidov) in analizi mozaičnosti lahko zmanjšanje osnovne velikosti »besede« pripomore k širšemu prvotnemu obsegu najdenih homolognih zaporedij in s tem omogoči kvalitetnejši pregled ter izbiro relevantnih najdenih homolognih zaporedij. Manjše velikosti »besede« podaljšajo analizo BLAST, saj parameter vpliva na fragmentacijo iskalnih nizov;
- »BLAST max hits«: pozitivna celoštevilska vrednost, večja od 0; privzeta vrednost 50; predstavlja zgornjo omejitev števila ujemanj, ki naj jih analiza BLAST vrne zaledni komponenti v nadaljno obdelavo;
- »Entrez query«: niz alfa numeričnih znakov; neobvezen parameter; nabor nukleotidnih zaporedij, ki so predmet analize, lahko omejimo z poizvedbo Entrez. Podrobni opis sintakse poizvedbe je dostopen na spletu na naslovu http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html;
- »Include«: niz alfa numeričnih znakov; neobvezen parameter; seznam besed, ki predstavljajo seznam regularnih izrazov (angl. regular expression) ujemanja z nazivi organizmov, ki naj bodo vključeni v rezultat poizvedbe;

- »Exclude«: niz alfa numeričnih znakov; neobvezen parameter; seznam besed, ki predstavljajo seznam regularnih izrazov (angl. regular expression) ujemanja z nazivi organizmov, ki naj bodo izključeni iz rezultata poizvedbe;
- »email«: niz alfa numeričnih znakov; privzeta vrednost »noreply@gmail.com«; parameter predstavlja elektronski naslov uporabnika aplikacije; NCBI za delo z njihovimi storitvami priporoča identifikacijo uporabnika.



Slika 15: Ekranska slika vnosne maske orodja za analizo nukleotidnih zaporedij JuP 1.0

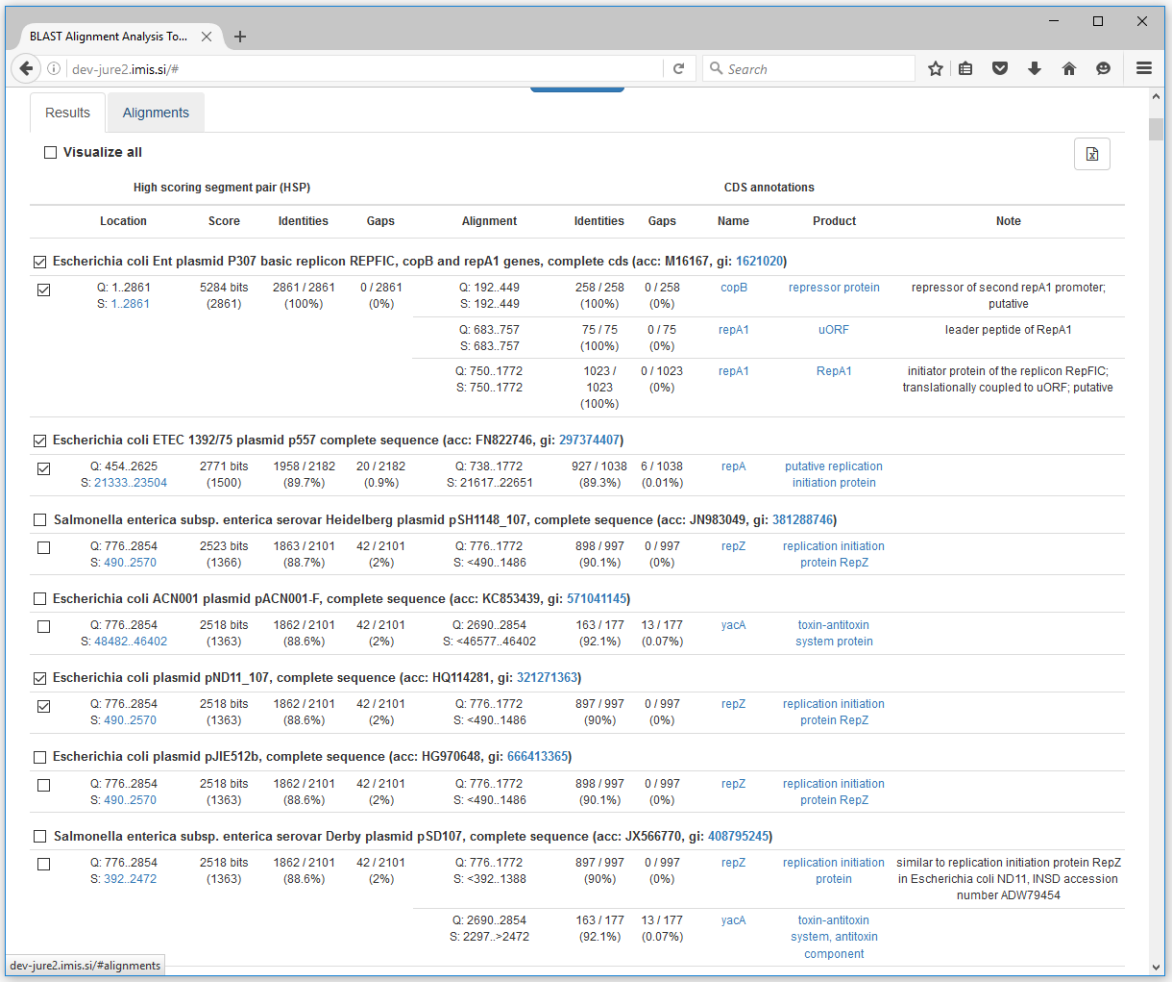
4.1.2 Tabelarični prikaz rezultatov analize

Rezultat analize zaledne komponente vizualizacijska komponenta transformira v preglednico rezultatov, sortiranih po »Score« rezultatu visoko točkovanega parnega območja padajoče, ki je prikazana na sliki 16. Preglednica vsebuje niz organizmov, ki jih spletna storitev BLAST identificira kot ujemajoča se zaporedja nukleotidov, organizirana v seznam visoko točkovanih parnih območij, ki kažejo visok nivo ujemanja z vhodnim zaporedjem nukleotidov glede na določene vhodne parametre analize. Vsako ujemajoče območje v preglednici obsega osnovne statistične podatke o področju in stopnji ujemanja ter številu identitet in vrzeli z njihovimi odstotnimi stopnjami. Preglednica vsebuje tudi hiperpovezave do podatkov GenBank in grafičnih prikazov visoko točkovanih parnih območjih.

Visoko točkovana parna območja so naprej razdeljena na pripadajoče oznake CDS. Vsaka oznaka CDS v preglednici obsega osnovne statistične podatke o področju in stopnji

ujemanja, številu identitet in vrzeli z njihovimi odstotnimi stopnjami in podatkih o genu in produktu, ki ga območje CDS kodira. Vsebuje tudi hiperpovezave do podatkov GenBank o elementu CDS, njegovem grafičnem prikazu in produktu (proteinu), ki ga kodira. Preglednico je prek gumba 'X' v zavihku »Rezultati« (angl. Results) zgoraj desno možno izvoziti v obliki Microsoft Excel za nadaljno obdelavo.

Funkcija preglednice je tudi izbor visoko točkovanih parnih območij, ki so predmet grafičnega prikaza v sosednjem zavihku »Poravnave« (angl. Alignments).



Slika 16: Ekranska slika metapodatkovnega rezultata analize homolognih zaporedij v JuP v obliki preglednice

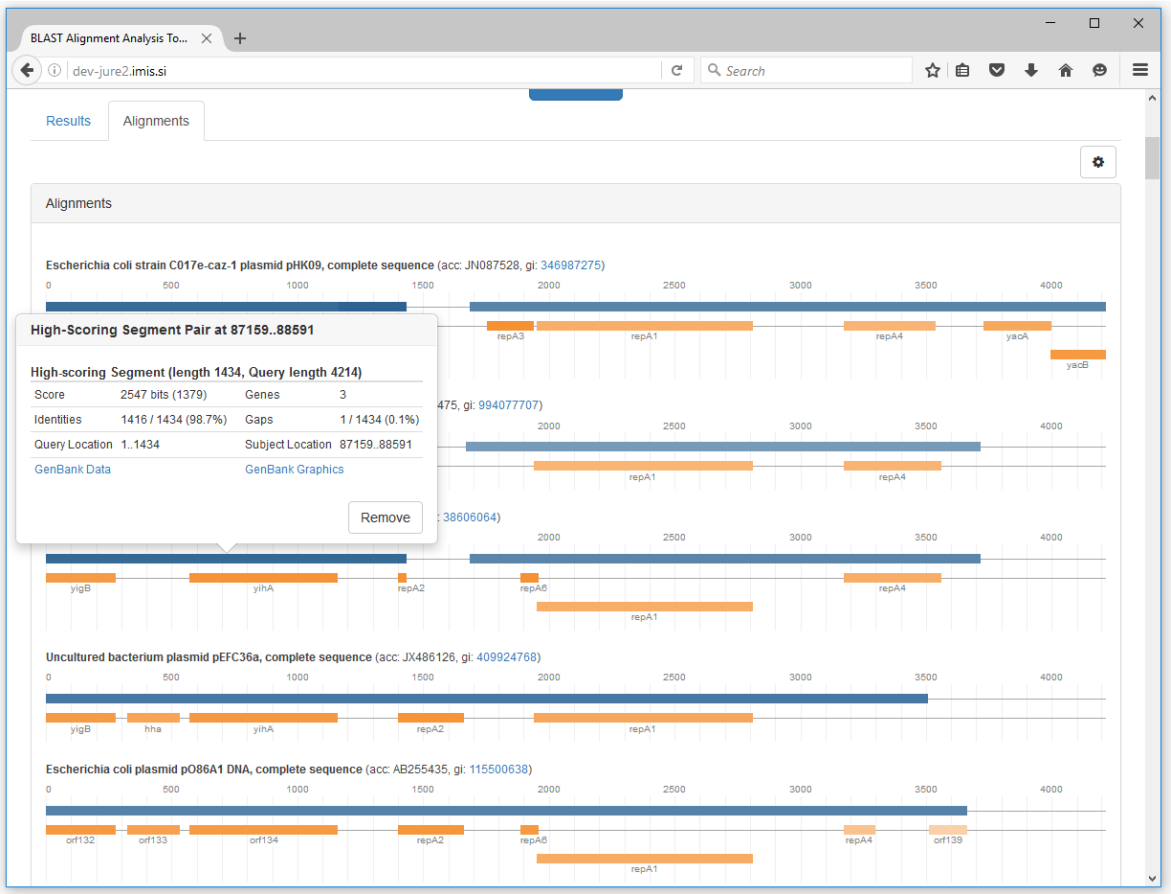
4.1.3 Grafični prikaz rezultatov analize

V preglednici rezultatov izbrana visoko točkovana parna območja v zavihku »Poravnave« (angl. Alignments) aplikacija grafično prikaže enega nad drugim z nazivom organizma, kateremu pripadajo in nadaljuje z barvnimi pasovi elementov. Elementi se začnejo na točki začetka ujemanja z vhodnim nukleotidnim zaporedjem in končajo s točko konca ujemanja z vhodnim nukleotidnim zaporedjem ne glede na dolžino poravnave, ki je lahko zaradi vrzeli

daljša. Privzet način prikaza obarva pasove s tonom izbrane osnovne barve na lestvici od 0 % do 100 % polnila. Polnilo 0 % predstavlja ujemanje s stopnjo nastavljenega praga prikaza ujemanja (privzeto 80 %), 100 % polnilo pa popolno ujemanje.

Pasovi so postavljeni na vertikalno mrežo oznak dolžine vhodnega nukleotidnega zaporedja, katero aplikacija dinamično določa glede na dolžino vhodnega zaporedja. Elementi CDS so označeni z imeni za lažjo in hitro identifikacijo.

Elementi se odzivajo na izbiro z miško ali navigacijskimi tipkami tipkovnice in prikažejo informacijski oblak izbranega elementa z njegovimi podrobnimi informacijami. Elemente HSP oz. visoko točkovana parna območja je mogoče iz prikaza izključiti tudi prek dejanja »Odstrani« (angl. Remove) v informacijskem oblaku, če uporabnik meni, da njegov prikaz ni relevanten. Dejanje ima enako funkcijo kot izključitev prikaza v preglednici rezultatov.

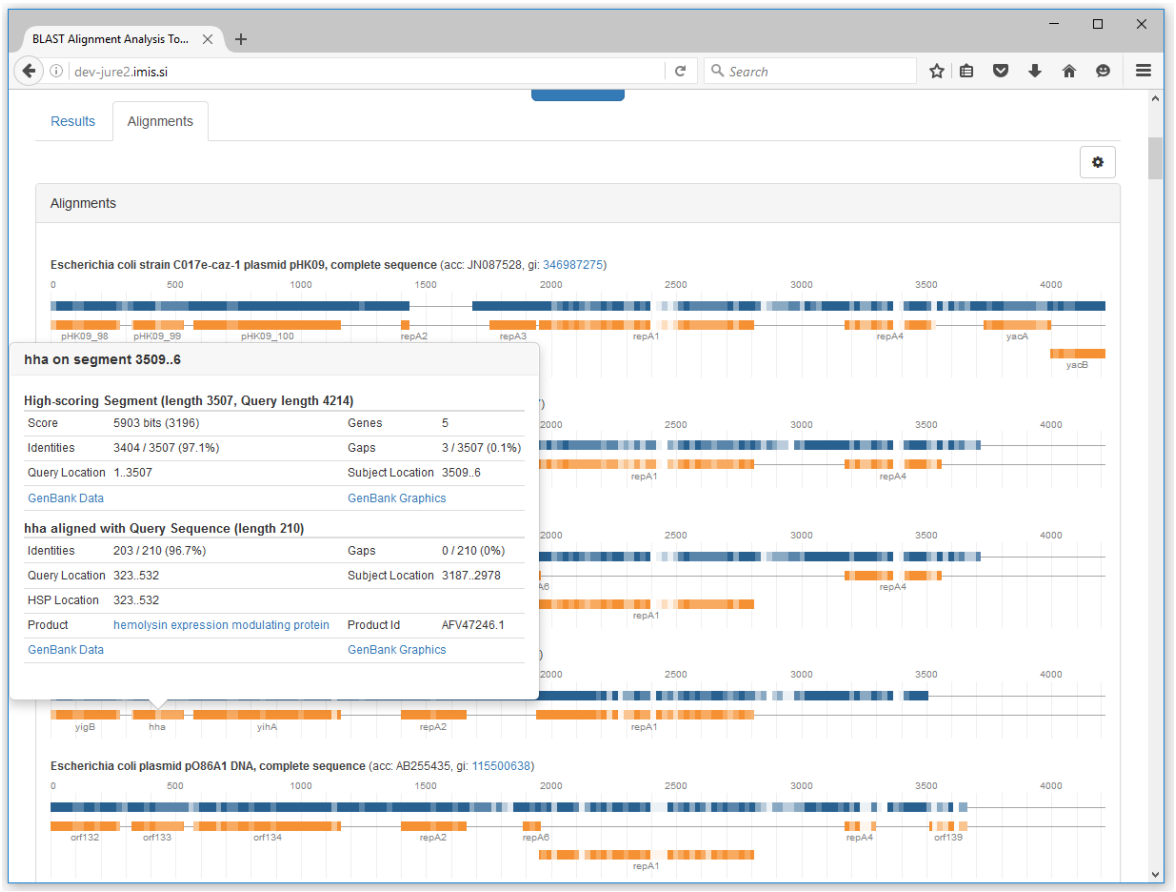


Slika 17: Ekranska slika prikaza homolognih zaporedij s prikazom povprečne stopnje homologije visoko točkovanih parnih območij in posameznih elementov CDS s primerom prikaza podrobnih podatkov o visoko točkovanem parnem območju

V primeru izbire prikaza podrobnega ujemanja s segmentacijo elementov HSP in CDS aplikacija pasove razdeli v segmente, katerih velikost je odvisna od uporabniške nastavitve. Privzeta velikost segmenta je sicer 1 % dolžine vhodnega zaporedja. Za vsak segment izračuna relativno stopnjo ujemanja in ga obarva s tonom izbrane osnove barve na lestvici

od 0 % do 100 % polnila. Polnilo 0 % predstavlja ujemanje s stopnjo nastavljenega praga prikaza ujemanja (privzeto 80 %), 100 % polnilo pa popolno ujemanje.

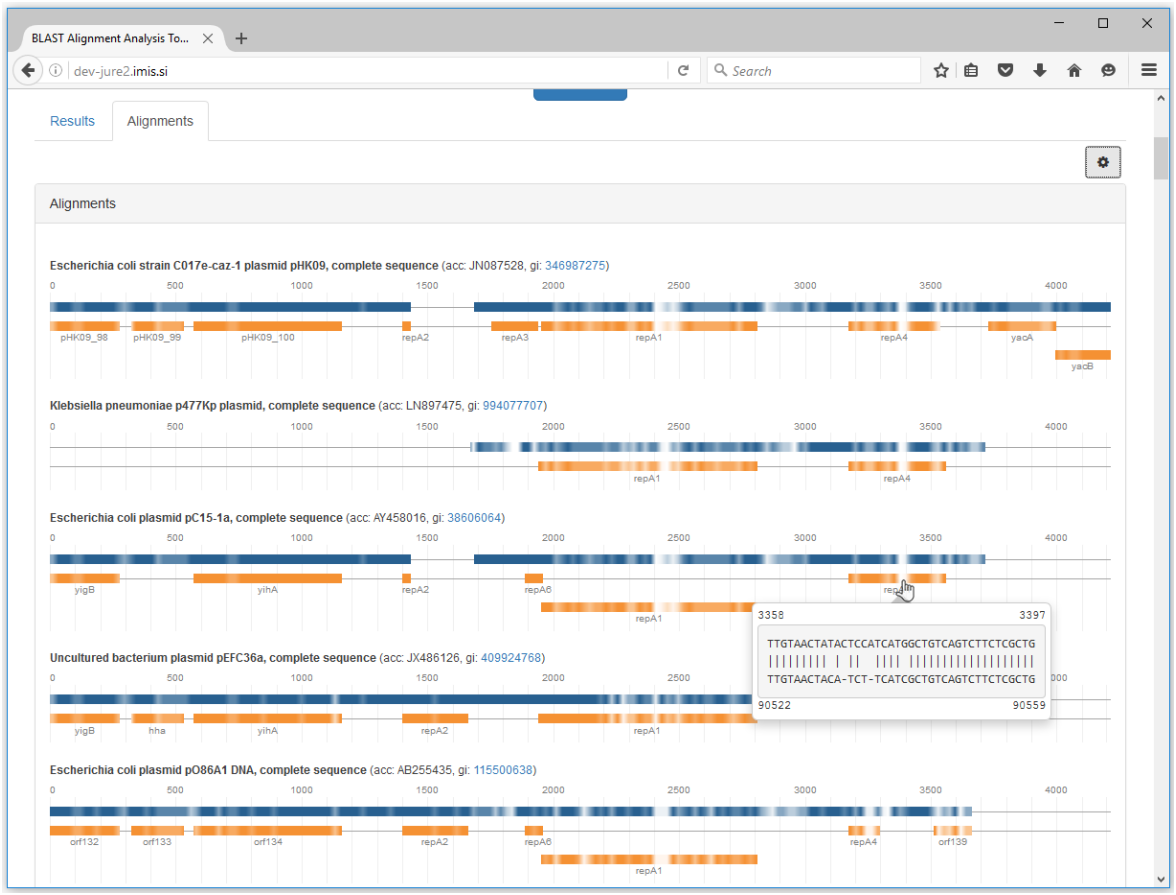
Prikaz podrobnega ujemanja s segmentacijo dodatno označi področja večjega in manjšega ujemanja z vhodnim zaporedjem in tako olajša identifikacijo lokacij večjih genskih nestabilnosti.



Slika 18: Ekranska slika prikaza homolognih zaporedij z diskretno obarvanim segmentiranim prikazom stopnje homologije visoko točkovanih parnih območij in posameznih elementov CDS s primerom prikaza podrobnih podatkov o elementu CDS

Segmentirana območja HSP in CDS je možno prikazati tudi zvezno. Pri tem prikazu aplikacija z izračunom povprečne stopnje ujemanja med segmenti in uporabo barvnih gradientov ustvari zvezno prehanjanje med segmenti.

Aplikacija omogoča vpogled v podrobnosti poravnave s prehodom miške preko področja pasu elementa HSP ali CDS. Informacijski oblak, ki se pri tem pojavi, podrobnost poravnave prikaže z začetekom in koncem vhodnega in najdenega zaporedja v obsegu nastavljenega območja podrobnosti ter med njima uporabnikom BLAST dobro znan grafični način prikaza ujemanja z znaki '|' za ujemanje nukleotidnega para in ' ' za njuno neujemanje. Vrzeli v poravnavi vhodnega in najdenega zaporedja prikaže z znakom '−'.



Slika 19: Ekranska slika prikaza homolognih zaporedij z zvezno obarvanim segmentiranim prikazom stopnje homologije visoko točkovanih parnih območij in posameznih elementov CDS s primerom prikaza poravnave iskanega in najdenega zaporedja elementa CDS

4.1.4 Nastavitve parametrov grafičnega prikaza poravnav

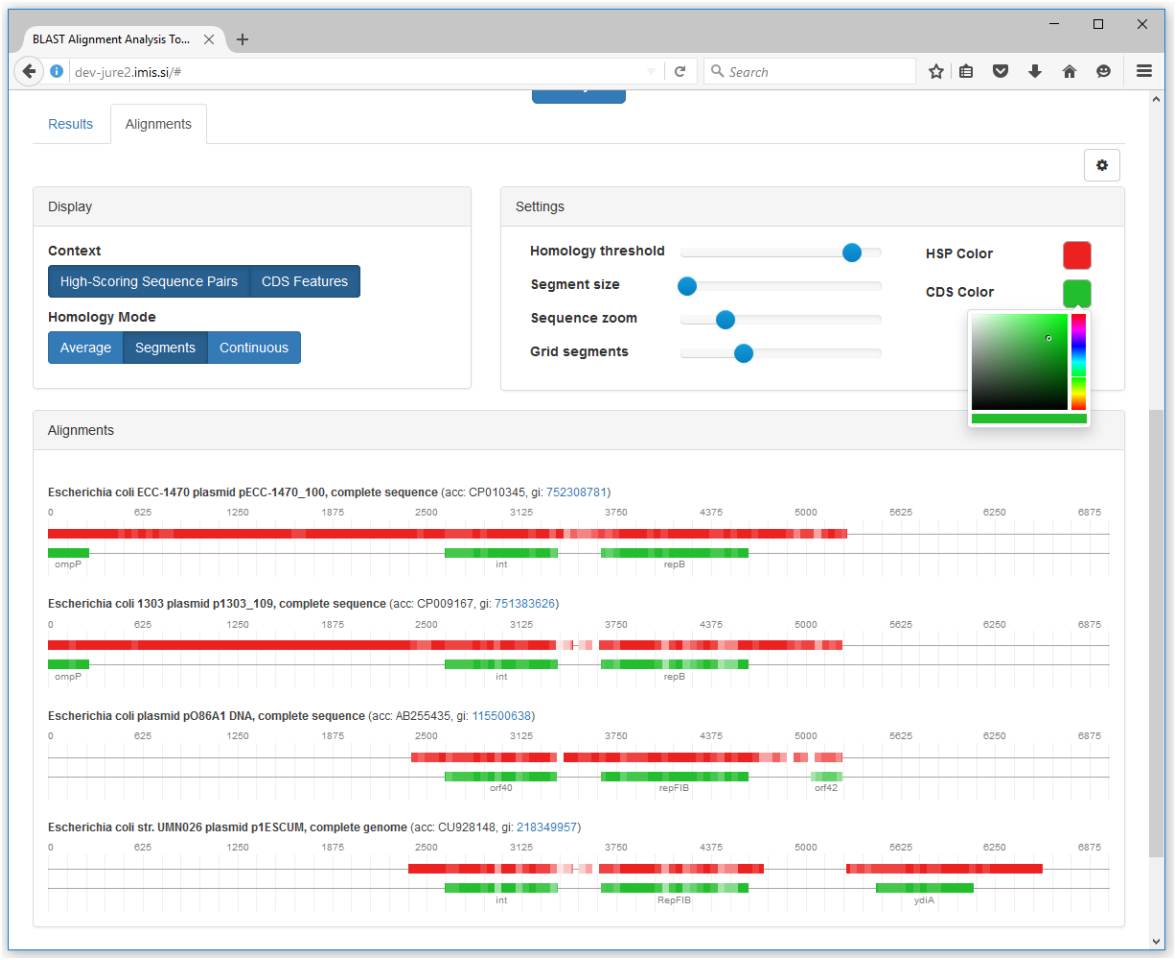
Uporabnik lahko dodatno prilagodi prikaz poravnav s prilagoditvijo parametrov vizualizacije, ki so dostopni prek dejanja pod gumbom z zobatim kolesom zgoraj desno v zavihku »Poravnave« (angl. Alignments). Področje nastavitev se po prilagoditvi skrije in tako omogoči večjo površino prikaza poravnav.

- »Display« / »Context« – gumba »High-Scoring Sequence Pairs« in »CDS Features« vsak zase vključujeta in izključujeta prikaz označenih visoko točkovanih parnih območij in oznak CDS. Izbrati je mogoče eno, drugo ali obe možnosti.
- »Display« / »Homology Mode« – možno je izbrati eno izmed naslednjih možnosti prikaza homologije poravnave:
  a) »Average« – dejanje povzroči obarvanje pasov poravnav s tonom izbrane osnovne barve, ki ga določa povprečno ujemanje poravnave celotnega elementa na lestvici od 0 % do 100 % polnila. Polnilo 0 % predstavlja ujemanje s stopnjo

nastavljenega praga prikaza ujemanja »Homology threshold«, 100 % polnilo pa popolno ujemanje;

b) »Segments« – dejanje povzroči diskretno segmentirano obarvanje pasov poravnav. Za vsak segment aplikacija izračuna relativno stopnjo ujemanja in ga obarva s tonom izbrane osnove barve na lestvici od 0 % do 100 % polnila, kjer 0 % polnilo predstavlja ujemanje s stopnjo nastavljenega pragu prikaza ujemanja »Homology threshold«, 100 % polnilo pa popolno ujemanje.
c) »Continuous« – dejanje povzroči zvezno segmentirano obarvanje pasov poravnav. Aplikacija vsakemu segmentu izračuna stopnjo ujemanja in njegovo sredino obarva s tonom osnovne barve glede na izračunano stopnjo ujemanja z upoštevanjem pragu prikaza ujemanja »Homology threshold«. Barva segmenta levo in desno od sredine gradientno prehaja v ton osnovne barve, ki kaže povprečno stopnjo ujemanja s sosednjim segmentom.

- »Settings« / »Homology threshold« – vrednost določa prag prikaza stopnje ujemanja. Področja ujemanja poravnave elementov ali njihovih segmentov pod določenim pragom bodo obarvana z belo barvo, ton med 0 % in 100 % pa predstavlja stopnjo ujemanja med določenim pragom in popolnim ujemanjem poravnave. Prilagajanje vrednosti povzroči ojačanje razlik področij z podobno homologijo ujemanja poravnave, ko so ta podobno homologna in s tem določa občutljivost prikaza ujemanja.
- »Settings« / »Segment size« – vrednost določa velikost (širino) segmenta v primeru segmentiranega prikaza homologije poravnav (»Segments«, »Continuous«). Z prilagoditvijo nastavitve višamo natančnost prikaza področij višje in nižje homologije poravnave znotraj elementov HSP in CDS.
- »Settings« / »Sequence zoom« – vrednost določa širino informacijskega oblaka prikaza poravnave, ko uporabnik z miško prehaja po elementu HSP ali CDS.
- »Settings« / »Grid segments« – vrednost določa število razdelkov v vertikalni mreži oznak dolžine vhodnega nukleotidnega zaporedja. Z večanjem števila razdelkov se natančnost mreže povečuje.
- »Settings« / »HSP Color« in »Settings« / »CDS Color« – barva pod nastavitvama določa osnovno barvo za barvanje visoko točkovanih parnih območij in elementov CDS. Z intuitivnim načinom izbira uporabniku prilaganjanje prikaza poravnav.



Slika 20: Ekranska slika nastavitvenih elementov prikaza homolognih zaporedij v JuP 1.0

4.2 MOZAIČNOST REPLIKACIJSKIH REGIJ PLAZMIDOV SKUPINE IncF

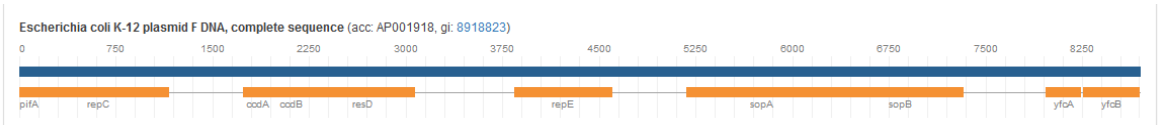
Podatke za analizo smo črpali iz baze podatkov »nt« v okviru orodja BLAST. Primarna analiza ujemanja vhodnih zaporedij je pokazala, da vpliv spreminjanja vrednost E ni signifikanten in je pri vrednostih 10^{-8}, 1, 5 in 10 rezultiral v enakem nefiltriranem rezultatu. Drugačen vpliv je imel parameter dolžine »besede« (angl. BLAST Word Size). Pri vrednosti 11, kar je priporočena vrednost za analize, optimirane za nehomologna zaporedja (v spletni različici orodja BLAST gre za optimizacijo »Optimize for More dissimilar sequences (discontiguous megablast)«), JuP v nefiltriranem rezultatu vrne več ujemajočih se zaporedij kot pri vrednosti 28, kar je priporočena vrednost za analize, optimirane za visoko ujemajoča se zaporedja (v spletni različici orodja BLAST gre za optimizacijo »Highly similar sequences (megablast)«).

V analizi mozaičnosti smo želeli objeti kar se da širok nabor ujemajočih se zaporedij, saj genski procesi kot so insercije, delecije, ipd, vplivajo na stopnjo ujemanja zaporedij in s tem

na mozaičnost. Po nekaj testnih analizah smo se odločili za naslednje vhodne parametre, ki zajamejo primerno širok nabor zaporedij:

- E vrednost: 1
  Vrednost zagotavlja primerno širok začetni obseg najdenih ujemanj BLAST upoštevajoč dolžino iskanih zaporedij. Predvideva, da bo analiza v bazi podatkov naletela na naključno ujemanje zaporedja s podobnim rezultatom le enkrat.
- BLAST word size: 11, 18, 28
  Rezultat analize z manjšo vrednostjo zajamejo fragmentirano ujemajoča se zaporedja, ki so pri mozaičnosti prav tako signifikantni.
- Najvišje število zadekov BLAST: 1000
  maksimalno število rezultatov BLAST, ki so predmet dodatnega filtriranja (izločitev vseh ujemanj, ki niso anotirana in ki v nazivu vsebujejo izločitvena gesla).
- Izločitvena gesla: »vector«, »assembly«
  zaporedja, ki v svojem nazivu vsebujejo izločitvena gesla, so umetno ustvarjena zaporedja za potrebe genskih manipulativnih procesov kot je kloniranje. Analiza se omejuje na naravno prisotne plazmide ali mini plazmide, ki so sicer umetno ustvarjeni, vendar niso namenjeni kloniranju.

4.2.1 RepFIA



Slika 21: Prikaz anotiranega vhodnega replikona RepFIA

Območje gena *repE* je dobro ohranjeno in visoko homologno pri večini pojavitev. Prav tako ni moč izpostaviti regije gena, ki je gensko bolj nestabilna od ostalih delov. Pri plazmidu pVir68 (*E. coli* Vir68) lahko opazimo insercijo daljšega območja nukleotidov, saj je *repE* prekinjen na približno polovici svojega zaporedja. Program JuP ni pokazal ujemanja po bp 33209 kljub temu, da se v bazi GenBank gen nadaljuje do bp 33230, vendar je z dolžino 477 bp krajši, kot pri plazmidu F z dolžino 756 bp. Zaporedje se nato od 35340 bp nadaljuje s praktično popolnoma homolognim zaporedjem v lokus *sop*, ki je dobro ohranjen.

Pri plazmidih p557, pO157, pETEC_74 (*E. coli*) program JuP pokaže na izgubo prvega dela *repE*. Zaključni del v bazi GenBank večkrat ni anotiran, sodeč po homologiji območja na visoko točkovanem parnem območju pa lahko trdimo, da je dobro ohranjen, kar lahko kaže na ostanek nekoč funkcionalnega gena ali na tako obsežno spremembo genskega zaporedja, da ni mogoče pokazati ujemanja z izvornim zaporedjem gena *repE*.

Slika 22: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIA – visoko homologna zaporedja

Plazmidi p1658/97, pAPEC-O1-ColBM, pAPEC-1, pETN48, pAPEC-O78-ColV, pSMS35_130 in drugi (*E. coli*) ter pCVM29188_146 in pSSAP03302A (*S. enterica* serovar Kentucky) kažejo dobro ohranjen lokus *sop* ob popolni izgubi gena *repE*.

Slika 23: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIA – izguba replikacijske regije ob dobro ohranjeni regiji lokusa *sop*

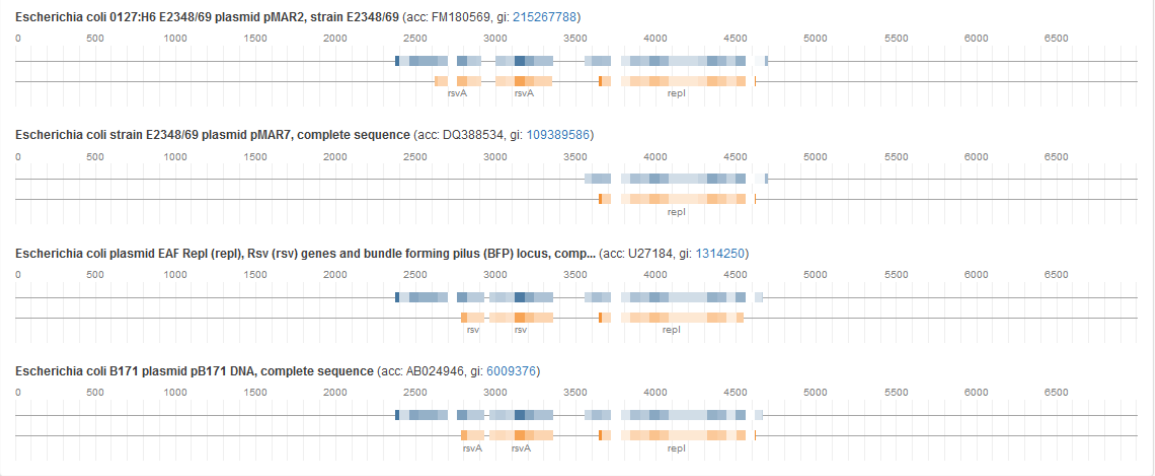
4.2.2 RepFIB

Gen *repA* s sinonimi *repB*, *repA_4*, *repFIB*, *repI* in *repA2* v bazi GenBank je dobro ohranjen replikon, ki je večinoma visoko homologen z iskanim zaporedjem. Njegova dolžina je pri ujemajočih se zaporedjih podobno anotirana, program JuP ne pokaže signifikantnih procesov insercije ali podvajanja replikona. Visoka homologija replikona se ohranja tudi v primeru horizontalnih prenosov na druge družine bakterij, kot je to primer s plazmidi pSH696_117 (*S. enterica* serovar Heidelberg), p33673_IncF, pYT3 in pU302L (*Salmonella typhimurium (S. typhimurium*)), kjer ostane nad 98 %.

Glede na to, da v bazi GenBank ni ločenih anotacij za inkompatibilnostne elemente (iteroni), ki obdajajo *repE* ob dejstvu, da je regija levo in desno od gena iz vhodnega zaporedja navadno dobro ohranjena, kaže na nizko stopnjo mozaičnosti replikona.

Slika 24: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIB – visoko homologna zaporedja

Pri horizontalnem prenosu v druge družine bakterij, kjer stopnja ujemanja pade v območje 80 %, kot je to primer pri plazmidih pP10164-NDM (*Leclercia adecarboxylata (L. adecarboxylata*), pNDM1_EC14653 in pECL3-NDM-1 (*E. cloacae*), pMAR2, pMAR7, pB171 in pK351 (*E. coli*), pride do izraza začetno ožje in osrednje širše območje slabše genske stabilnosti replikona *repE*. Glede na ohranitev funkcionalnosti kaže, da območja ne kodirajo ključnih elementov replikona.



Slika 25: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIB – prikaz izrazitejše regije nizke homologije gena *repE*

4.2.3 RepFIC

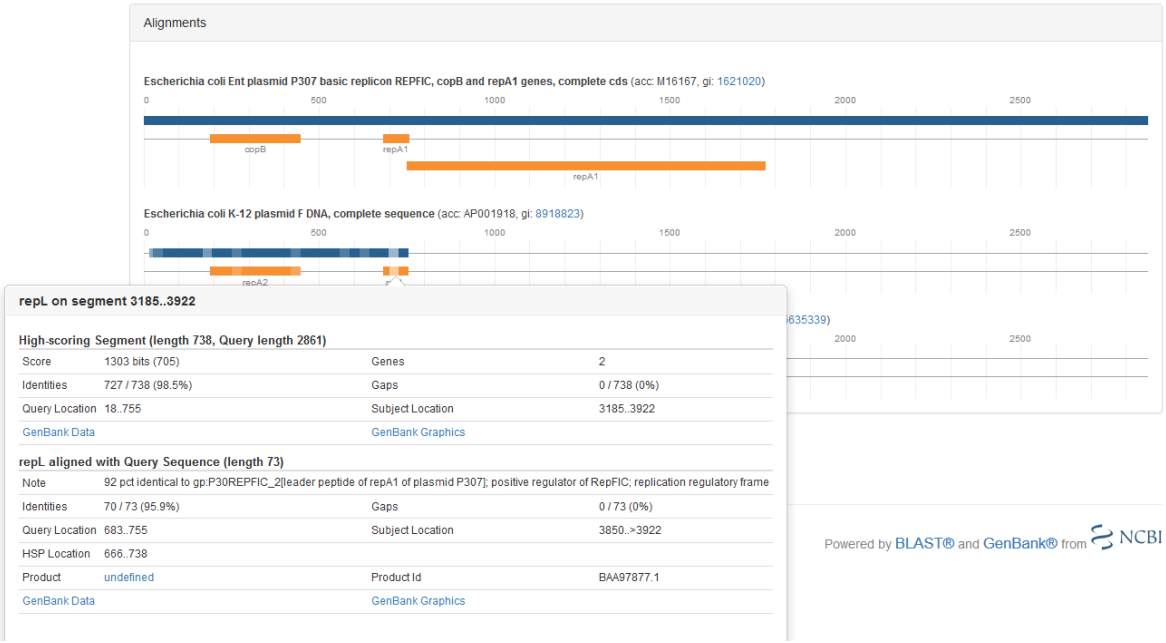
Plazmid p307 kot osnovni replikon RepFIC v bazi podatkov nima visoko homolognih zaporedij. Stopnje ujemanj najpodobnejših zaporedij so vse pod 90 % ne glede na optimizacjo iskanja. Vsa najdena zaporedja izhajajo iz plazmidov, ki izražajo IncFII inkompatibilnost, kar kaže na tesno sorodnost inkompatibilnostnih faktorjev IncFIC in IncFII. Na splošno so vsa signifikantno ujemajoča se zaporedja bistveno manj homologna, kot pri ujemanju replikona RepFIIA.

Gen *repA1* je v 90 % homologen z genom *repA* oz. *repZ* iz plazmidov p557, pND11_107, pJIE512b, pO104_H7, pHNAH4-1 in drugimi (*E. coli*) ter pSH1148_107, pSD107, TY474p2, pCVM29188_101, pSTM709 (*S. enterica*). Slika 26 pokaže območja večje in manjše stopnje ujemanja zaporedja, ki so pri večini omenjenih plazmidov na istih mestih, kar bi lahko kazalo na skupen izvor *repA*.

Primerjava zaporedja s sorodnim replikonom RepFIC iz plazmida F v *E. coli* K-12 pokaže visoko homologijo začetnega dela replikona z 99 % stopnjo homologije visoko točkovanega parnega območja do točke zaporedja vstavljenega transpozona Tn*1000*, kar potrjuje prej objavljene izsledke (Saadi in sod., 1987). Gen *copB* na zaporedju p307 je 99 % homologen z genom *repA2* na zaporedju plazmida F. Podobno stopnjo homologije kaže gen *repA1* na zaporedju p307, ki je 96 % homologen z genom *repL* na zaporedju plazmida F (slika 27). Zaporedje plazmida F *E. coli* K-12 se od bp 3917 podobno kot pri p307 nadaljuje z genom replikacijskega proteina *repA1*. Primerjava zaporedij gena *repA1* iz obeh plazmidov pokaže nezadostno homologije, da bi jo orodje BLAST prek JuP identificiralo v okviru visoko točkovanega parnega območja. Podoben rezultat pokaže primerjava zaporedja p307 z drugimi plazmidi, kot so pAPEC-1 v *E. coli* chi7122, pAPEC-O1-ColBM v *E. coli* APEC O1, pCoo v *E. coli*. Pri plazmidu pHK17a in pIP 1206 v *E. coli* je z razliko od plazmida F

seva K-12 opazna ohranitev homologije s koncem zaporedja plazmida p307, kjer lahko identificiramo gen *repA4* (slika 28).

Slika 26: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIC – visoko homologna zaporedja



Slika 27: Prikaz homolognosti z inaktiviranim replikonom RepFIC v plazmidu F *E. coli* K-12 Inaktivacija je posledica vstavitve transpozona Tn*1000* v zaporedje replikacijskega proteina *repA1*.

Slika 28: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIC – visoko homologna zaporedja ob izgubi homologije gena *repA1*

4.2.4 RepFIIA

Gen *repA2* je po svojem celotnem zaporedju glede na ostale gene v replikonu visoko homologen pri večini pojavitvah, kjer nastopa v svojem celotnem obsegu. To ne drži v primerih, kjer je ohranjena začetna tretjina gena. Dejstvo, da orodje BLAST ne najde ujemanja v nadaljevanju zaporedja, kaže na popoln izbris dveh tretjin gena ali na večje mutacije, ki stopnjo ujemanja zaporedja spustijo do te mere, da jo primerjava BLAST ne identificira kot ujemajočo se in je ne doda v rezultat. To je možno opaziti na primerih plazmidov, kjer najdemo delne ponovitve replikona. V bazah je sicer anotiran tudi kot *copB, repB in cpb2*.

Gen *repA6* je večkrat anotiran tudi kot zaključni del gena *repA3*, nekajkrat je anotiran kot del gena *copA*, pojavi se z anotacijo *repL*, *tapA*, *tap*. Gen ohranja visoko stopnjo homologije v večini pojavitev. Kljub temu, da v bazi GenBank večkrat ni anotiran, je homologija na njegovi lokaciji v visoko točkovanih parnih območj visoka.

Gen *repA1*, v bazi GenBank anotiran tudi kot *repA*, *repB*, *repAFII* in *rep2*, je po svojem celotnem zaporedju slabo homologen. Izstopa njegova osrednja regija z izrazito gensko nestabilnostjo. Pomembna je njegova ohranjenost v smislu obsega. Analiza zaporedij ne pokaže prekinitve zaradi procesov insercije ali delecije ali krajšanja gena od začetka ali

konca, kar kaže na njegovo funkcionalno ključnost v procesu podvajanja in ohranjanja replikona.

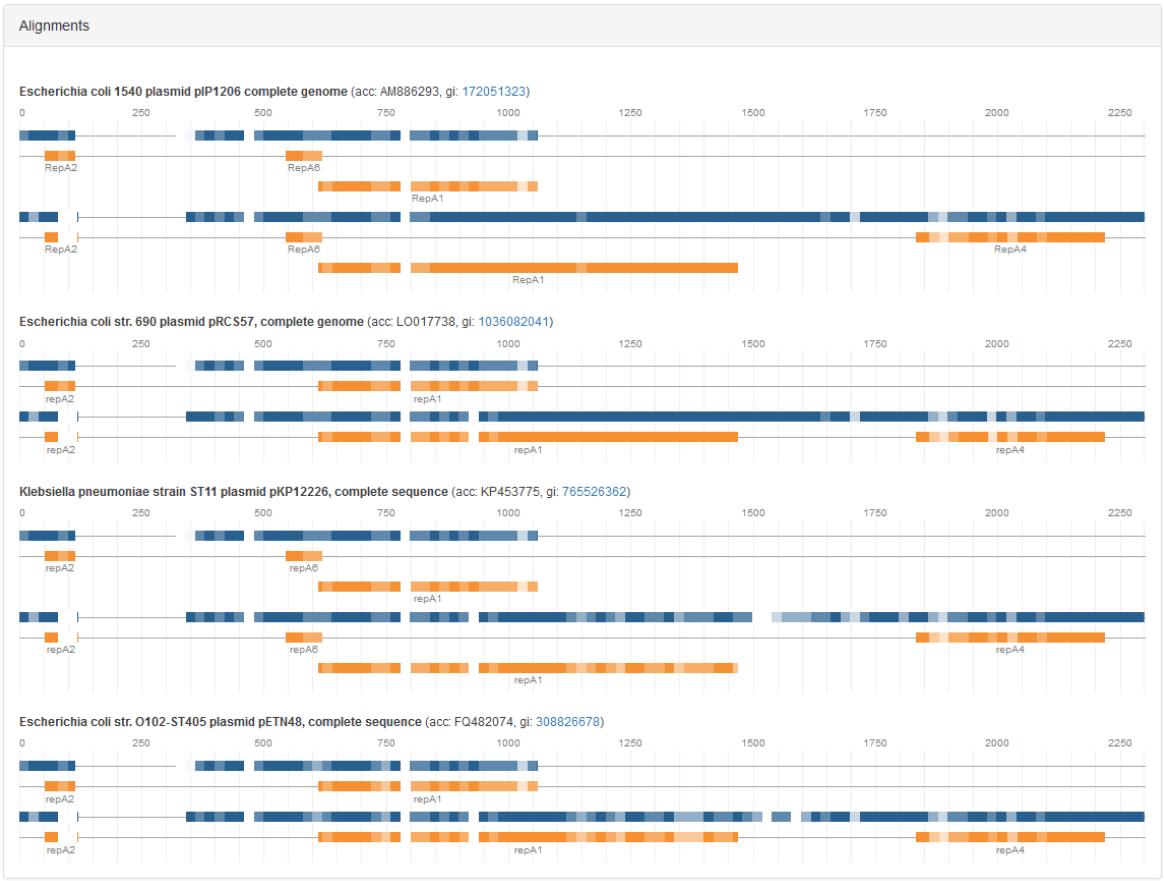
Gen *repA4* je v redkih primerih visoko homologen. Začne se z območjem nizke stopnje homologije. Srednja stopnja se pokaže v njegovem osrednjem delu, medtem ko je njegova zadnja tretjina močno ohranjena. V bazi je njegova dolžina večkrat nepopolno določena, vendar nima sinonimov.

Slika 29: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIIA – visoko homologna zaporedja z jasnim prikazom območij višje in nižje homologije

Pri plazmidih *E. coli* pIP1206, pRCS57, pRCS52, pETN48, pKP12226, RCS105_pl, pAA in plazmidu bakterije *Shigella dysenteriae* (*S. dysenteriae*) pSD1_197 je prisotna delna podvojitev elementov replikona. V najdenem zaporedju v dvojni ponovitvi nastopata skrajšana gena *repA2* in *repA1*. Gen *repA6* pri več zaporedjih ni anotiran, vendar je glede na stopnjo ujemanja njegovega območja v visoko točkovanem parnem območju mogoče trditi, da je prav tako podvojen. Rezultat kaže na delno podvojitev replikona, pri čemer nizka stopnja ujemanja zaporedja podvojitve gena *repA2* in njegovo skrajšanje kaže na nepopolno

podvojitev. Kljub skrajšanju v najdenem zaporedju analiza ne pokaže odseka, kjer bi se gen nadaljeval, kar kaže na delno izgube informacije gena glede na iskano zaporedje.

Visoka stopnja homologije zaporedja s plazmidi *E. coli* pKF3-70, pCC1409-1, pCC1410-1, pKP12226 bakterije *K. pneumoniae*, pCROD1 bakterije *Citrobacter rodentium* (*C. rodentium*) ter pCS0010A in pSSAP03302A bakterije *S. enterica* serovar Kentucky poleg horizontalnega prenosa replikonov znotraj vrste kaže na uspešen horizontalen prenos med družinami bakterij.



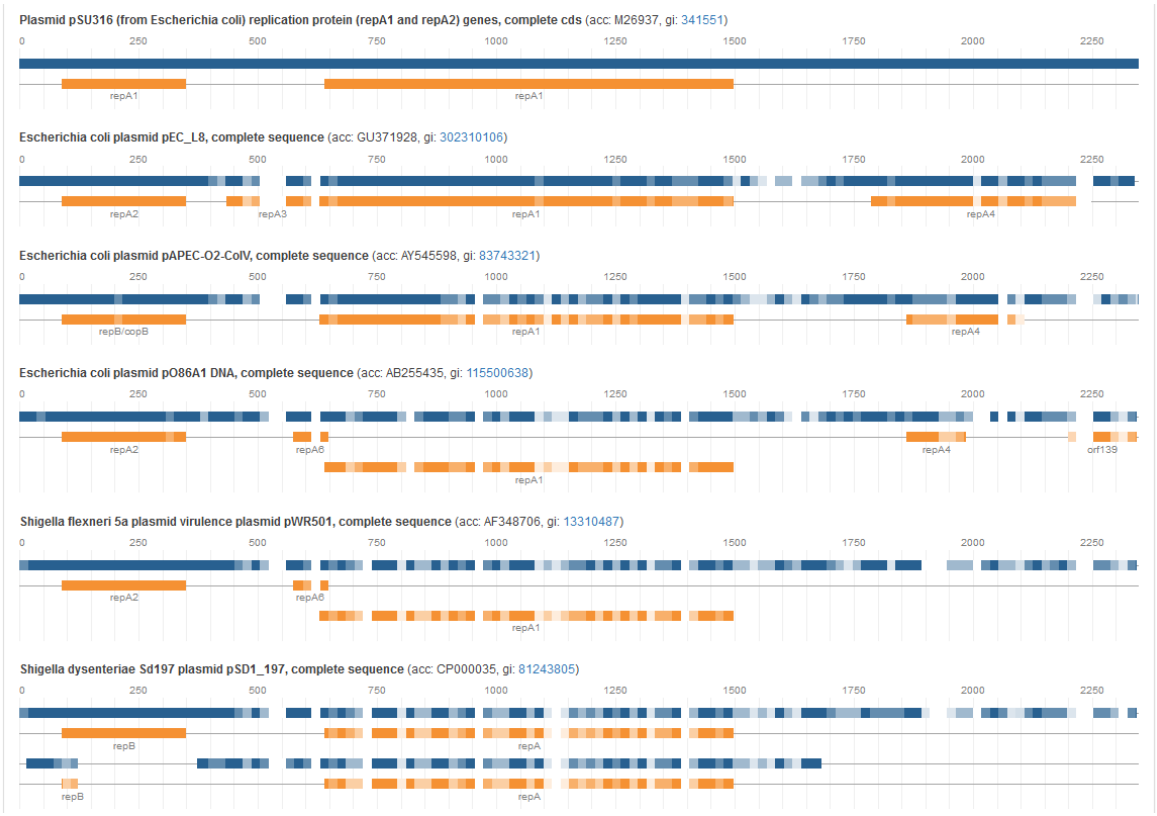
Slika 30: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIIA – podvojena homologna zaporedja

4.2.5 RepFIII

Podobno kot replikon RepFIC tudi RepFIII pripada družini replikonov RepFIIA, kar jasno pokaže visoka stopnja ujemanja plazmida pSU316 s plazmidi pARS3, pEC_L8, pAPEC-O2-ColV, R100 in drugimi, vsi prisotni v *E. coli*, ki kažejo visoko stopnjo ujemanja tudi s plazmidom R100 *S. flexneri* 2b. Ta je dobro znan po svoji inkompatibilnosti IncFIIA. Poleg tega sta p316 (*E. coli*) in R100 (*S. flexneri* 2b) v prisotna v njunih navzkrižnih rezultatih z visokim rezultatom, kar kaže na njuno veliko podobnost.

Pri vseh zaporedjih, ki kažejo homologijo s pSU316, je očitno področje šibke 50 % homologije pred genom *repA6*. V primerjavi z analizo RepFIIA je mogoče šibkejšo homologijo zaznati tudi pri *repA1* in *repA4,* medtem ko je *repA2* močno homologen, kar potrjuje rezultate analize homologije *repA2* pri RepFIIA.

Pri primerjavi s plazmidom pSD1_197 (*S. dysenteriae*) je zaznati ujemanje s podvojenim zaporedjem, ki je homologno vhodnemu replikonu. V prvi ponovitvi je zaporedje gena *repA2* 100 % homologno, medtem ko zaporedje gena *repA1* pokaže 93 % stopnjo ujemanja. V drugi ponovitvi je *repA2* tako nehomologen, da orodje BLAST prek JuP ne pokaže celotnega gena, ampak kot signifikatnen del pokaže le odsek z dolžino 33 bp. Vpogled v zapis najdenega zaporedja v bazi GenBank pokaže, da je gen anotiran v pravi dolžini 261 bp, vendar nehomolognost območja po 33 bp povzroči nevključenost v eno izmed visoko točkovanih parnih območij. Gen *repA1* je podobno kot v prvi ponovitvi tudi v drugi podobno homologen. Podobne primere podvojenih replikonov je možno opaziti tudi pri pRCS57 in pIP1206 (*E. coli*), kjer je izrazita tudi regija z *repA4*, ki je pri pSD1_197 odsotna.



Slika 31: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIII – visoko homologna zaporedja

4.2.6 RepFIV

Primerjava vhodnega zaporedja z velikostjo besede BLAST pri vrednosti 28 ne poda smiselnega rezultata, kar kaže na šibko in fragmentirano homologijo z zaporedji v bazi.

Nastavitev velikosti besede smo zato znižali na 11, kar je priporočena vrednost za iskanje šibko homolognih zaporedij. V tem primeru orodje BLAST prek JuP identificira in razvrsti najdena zaporedja s stopnjo homologije od 75 % navzdol. Temu primerno smo znižali prag prikaza homologije zaporedja (angl. Homology threshold) na vrednost 55.

Vrnjena zaporedja so omejena na plazmide iz sevov vrste *Pseudomonas* in v manjšem obsegu *Aeromonas*, *Xanthomonas* in *Serratia*. Vse primerjave pokažejo enakomerno porazdeljeno homologijo po celotnem območju *repA* z nekoliko opaznejšim področjem šibkejše homologije v osrednjem in drugem delu gena. Analiza ne pokaže znakov genskih insercij ali delecij, saj je ujemanje zvezno.

Slika 32: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFIV – visoko homologna zaporedja

4.2.7 RepFVI

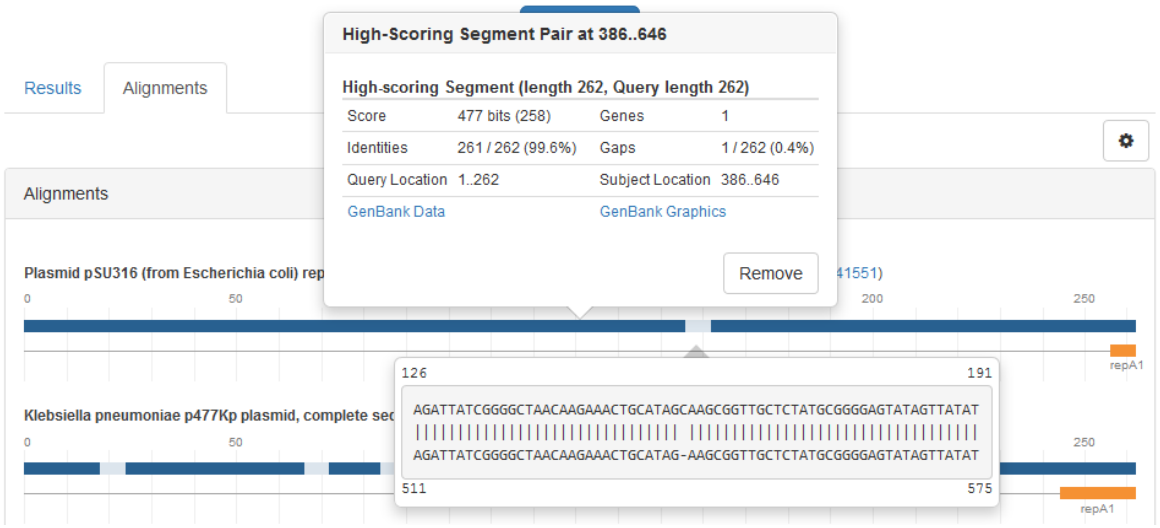
Gre za replikon, ki podobno kot RepFIC in RepFIII, spada v družino replikonov RepFIIA. Gen *repA2* kaže visoko stopnjo homologije z istimi geni v najdenih zaporedijih, lociranih na odsekih plazmidov, ki so značilni za replikon RepFIIA. Zaporedje *incFVI* kaže določeno stopnjo ujemanja z odsekom zaporedja gena *repA3*, vendar je to hkrati tudi najmanj homologna regija v primerjavi s preostankom zaporedja. Homologija zaporedja na mestu genov *repA6* in *repA1* kaže močno ohranjenost z le nekaj spremenjenimi nukleotidi.

Rezultat primerjave vrne tudi plazmide, kjer so deli replikona delno ali v celoti podvojeni, kar kaže na njihovo mozaično naravo. Fenomen je možno zaznati tudi pri replikonih drugih inkompatibilnostnih skupin.

Slika 33: Ekranska slika rezultata analize vhodnega zaporedja replikona RepFVI – visoko homologna zaporedja

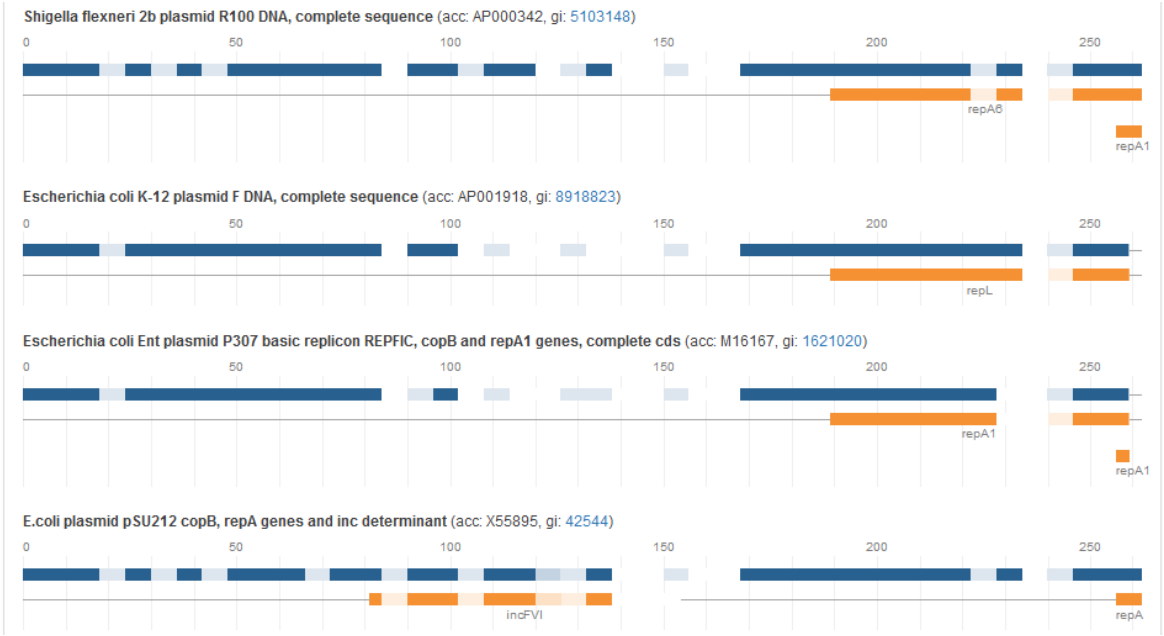
4.2.8 RepFVII

V bazi podatkov GenBank ni deponiranega celotnega zaporedja replikona RepFVII, dostopno je le zaporedje determinante *incFVII*. Zaporedje determinante *incFVII* kaže 99,6 % homologijo z osrednjo regijo zaporedja replikona RepFIII v območju od 386 do 646 bp oz. razliko v enem baznem paru (slika 34). Področje replikona sicer kodira protiprepisno RNA inkompatibilnostno determinanto zato z veliko gotovostjo lahko trdimo, da sta inkompatibilnostni determinanti *incFIII* in *incFVII* visoko homologni oz. se razlikujeta v enem baznem paru. To potrjuje rezultate iz poročila López in sod. (1989), ki omenja visoko stopnjo homologije replikona RepFIII in RepFVII in uvršča pSU233 med plazmide nadskupine RepFIIA.



Slika 34: Ekranska slika rezultata primerjave (neanotirane) inkompatibilnostne determinante *incFIII* in vhodnega zaporedja inkompatibilnostne determinante *incFVII*

Navzkrižna primerjava podobnih zaporedij, ki kodirajo inkompatibilnostne determinante drugih skupin (slika 35), pokaže 89 % homologijo za ekvivalentno regijo na plazmidu R100 (*incFIIA*), 90 % homologijo z zaporedjem ekvivalentne regije na plazmidu pSU212 (*incFVI*), skupaj z pSU316 (*incFIII*) predstavniki nadskupine RepFIIA. Nekoliko nižjo, 88 % stopnjo homologije, pokaže primerjava z inkompatibilnostno regijo plazmida F (*incFIA* in *incFIB*) in še nekoliko nižjo, 87 % stopnjo homologije primerjava z inkompatibilnostno regijo plazmida P307 (*incFIC*). Plazmida R124 (*incFIV*) ni med rezultati homolognih zaporedij.



Slika 35: Navzkrižna primerjava inkompatibilnostnih determinant *incFIIA*, *incFVI*, *incFIII*, *incFIA*, *incFIB* in *incFIC*

Zaporedje inkompatibilnostne determinante *incFVII* je v več primerih anotirano kot del genov *repA6*, *repA3* in *copA*, ki so sami različno veliki (npr *repA3* pri pHNFP460-1, pTUC100 in pAPEC-1), kar spet pokaže precejšnjo nedeterminiranost označevanja genov.

Osrednja regija *incFVII* v območju od 100 bp do 170 bp je gensko nestabilna in kaže manjšo ohranjenost od ostalih delov zaporedja.

Slika 36: Prikaz homologije inkompatibilnostne determinante *incFVII* z zaporediji replikacijskih genov homolognih zaporedij iz različnih plazmidov *Enterobacteriaceae*

5 RAZPRAVA

Plazmidni replikoni so ključni za vzdrževanje plazmidov v gostiteljskih celicah. Plazmidi se lahko razlikujejo po mehanizmih podvajanja, izvoru replikacijskih zaporedij in proteinih, ki sodelujejo pri podvajanju plazmida. Plazmide s podobnimi zaporedji začetka podvajanja in mehanizmi uravnavanja podvajanja uvrščamo v inkompatibilnostne skupine.

5.1 PREVALENCA REPLIKONOV SKUPINE RepFIIA

Analiza BLAST (preglednica 3) zaporedij replikonov RepFIC, RepFIIA, RepFIII in RepFVI v bazi podatkov »nt« pokaže prevelenco replikonov nadskupine RepFIIA v primerjavi s plazmidi drugih F-inkompatibilnostnih skupin. Rezultat potrjuje prej objavljene študije, ki postavljajo nadskupino replikonov RepFIIA kot najbolj razširjeno med izoliranimi sevi *E. coli* (Osborn in sod., 2000; Villa in sod., 2010; Moran in sod., 2015).

Največjo razširjenost in prisotnost replikona RepFIIA potrdi tudi rezultat mojega dela s 475 primeri popolnoma ujemajočih se zaporedij v 458 različnih deponiranih zaporedijh. To je 4 × več od naslednje skupine RepFIA s 140 zaporedji in RepFVI iz skupine RepFIIA s 132 popolnoma ujemajočimi se zaporedji. Rezultat je sicer potrebno kritično presojati, saj ni normaliziran na gostiteljski sev in vpliva večje oziroma manjše raziskanosti nekega organizma ter s tem povezano število deponiranih in ujemajočih se zaporedij ne upošteva.

V tej raziskavi se replikon RepFIB ne pokaže kot zelo razširjen, kar sicer poročajo študije Johnson in sod. (2007 in 2012) ter Moran in sod. (2015). V slednji so objavili podatek o 59 % prisotnosti RepFIB v plazmidih komenzalnih sevov enterobakterij v Avstraliji. Na rezultat lahko vpliva manjša prisotnost deponiranih zaporedij, kjer je replikon prisoten, zato ni mogoče absolutno sklepati na njegovo razširjenost v mikrobni združbi.

Preglednica 3: Prevalenca replikonov prikazana prek števila najdenih, ujemajočih se deponiranih zaporedij

Replikon	**Št. HSP s stopnjo ujemanja nad 90 % / št. različnih zaporedij**	**Št. HSP s 100 % stopnjo ujemanja / št. različnih zaporedij**
RepFIIA	1015 / 831	475 / 458
RepFIA	2547 / 333	140 / 131
RepFIB	74 / 36	23 / 23
RepFIC	783 / 607	4 / 4
RepFIII	1043 / 834	0 / 0
RepFIV	8 / 8	0 / 0
RepFVI	401 / 382	132 / 130

Analiza RepFIIA pokaže divergenco ohranjenosti med različimi geni v replikonu. Pokaže se močna ohranjenost zaporedja gena *repA2* s praktično nespremenjenim zaporedjem v večini relevantnih najdenih zaporedjih, kjer je gen ohranjen v svojem celotnem obsegu 255 bp. Gen

je v več primerih skrajšan na 62 bp, vendar kljub temu ohranja visoko stopnjo homologije. Takrat replikon navadno nastopa v dveh ponovitvah, pri čemer je druga ponovitev gena *repA2* bistveno manj homologna. Podobno stopnjo visoke ohranjenosti pokaže zaporedje gena *repA6*, medtem ko sta gena *repA1* in *repA4* manj ohranjena. To ni nepričakovano, saj *repA4* nima funkcionalnega produkta, ki bi vplival na selekcijske procese.

Mozaičnost RepFIIA ni edini primer mozaičnosti zaporedij v plazmidih skupine IncFIIA. Raziskava Starčič Erjavec in sod. (2002) pokaže, da mozaično strukturo izkazujejo tudi regije *tra* plazmidov, in ne navsezadnje tudi celotna sestava velikih plazmidov. Lep primer je plazmid pRK100, za katerega velja, da so geni *tra* pRK100 homologni genom *tra* plazmida F, replikon RepFIIA plazmida pRK100 pa replikonu RepFIIA pWR501. Vse to nakazuje, da je tak plazmid z mozaično naravo plazmidnih genov nastal z več rekombinacijskimi dogodki med različnimi izvornimi plazmidi (Boyd in sod., 1996).

Primerjava zaporedja replikona RepFIC plazmida P307 z replikonom IncFII plazmida R100 pokaže regije visoke homologije in nehomolognih odsekov, kar naj bi bil rezultat rekombinantnih procesov in potrdi izsledke raziskave Saadi in sod. (1987). Prisotnost zaporedij, podobnih zaporedjem *Chi* na stikih posamičnih mozaičnih elementov kažejo na ključno vlogo zaporedij *Chi* v rekombinacijskih dogodkih in evoluciji replikacijskih družin (Boyd in sod., 1996).

Mozaičnost replikonov otežuje kvalitetno klasifikacijo bakterijskih plazmidov in evolucijske študije sorodnosti med plazmidi in njihovimi replikoni. Za klasifikacijo v inkompatibilnostne skupine uporabljamo specializirane sonde *rep* za razvrščanje novih plazmidov prek postopka hibridizacijske analize. Vendar pa posamezne sonde *rep* za protiprepisne replikone inkompatibilnostne skupine RepFIIA kažejo značilno navzkrižno hibridizacijo (Couturier in sod., 1988), kar omejuje njihovo učinkovito uporabo. Postopke klasifikacije prek sond dodatno otežuje dejstvo, da inkompatibilnost med plazmidi lahko povzroči tudi razlika v enem samem baznem paru zaporedja. Plazmide z več replikoni (t.i. multi-replicon plazmide) je večkrat težko klasificirati v eno inkompatibilnostno skupino, saj lahko izražajo inkompatibilnost z večimi skupinami. V sodobnem času se je tako uveljavila metoda tipiziranja replikonov s pomočjo PCR in PCR v realnem času (Carattoli in sod., 2005; Boot in sod., 2013).

5.2 PROGRAM JuP JE JASNO POKAZAL POMANJKLJIVOSTI V ANOTIRANJU ZAPOREDIJ

V bazi podatkov GenBank najdemo veliko primerov nedoslednih, pomanjkljivih in napačnih oznak identificiranih genov. Pojav ni izoliran le na gene plazmidov IncF. Tako analiza zaporedja gena *ehxA* za enterohemolizin (priloga J), ki je značilen za črevesno patogeno enterohemoragično *E. coli*, pokaže 99,5 % stopnjo homologije s zaporedjem plazmidov pO104_H7 in pO104_H21 (*E. coli*), ki je anotirano kot gen *hlyA*. A ta oznaka je namenjena

genu za α-hemolizin, ki je značilen za zunajčrevesne patogene *E. coli*. Z nekoliko nižjo stopnjo homologije, vendar vseeno višjo od 98 %, se fenomen ponovi več kot 10 × pri drugih najdenih zaporedjih na različnih plazmidih. Ker gre za dva tipa hemolizina, je določena stopnja homologija pričakovana, vendar Mainil in Daube (2005) v raziskavi opišeta 60 % nukleotidno homologijo med *ehxA* in *hlyA*, zato gre pri omenjenih primerih za napačno oznako gena.

Podobno navzkrižno označene gene je možno opaziti v primeru analize replikona RepFIIA. V zaporedju deponiran z akcesijsko številko HE610901 je gen *repB* 98 % homologen z genom *repA2* v zaporedju AP000342. V zaporedju deponiranem z akcesijsko številko AY509003 je gen *repB* 75 % homologen z genom *repA1* v zaporedju AP000342. V kolikor bi v bazi iskali zaporedje gena *repB*, tovrstne napačne oznake preprečujejo nedvoumen rezultat in zahtevajo dodatne postopke.

Napačne anotacije ribosomske RNA prispevajo k do 90 % napačno pozitivnih rezultatov pri iskanju proteinov (Tripp in sod., 2011).

5.3 GENOMIKA = »Big Data«

Genomika je znanost s produkcijo enormne količine podatkov. Zajem novih genskih podatkov raste z velikansko hitrostjo in se trenutno po obsegu zajetih podatkov podvoji vsakih 7 mesecev. Leta 2015 je NCBI hranila 3,6 peta zlogov surovih genskih podatkov iz približno 32.000 mikrobnih, 5.000 rastlinskih in 230.000 celotnih človeških genomov (Regalado, 2014), vendar to predstavlja le manjši del sekvenciranih genomov, saj jih večina še ni arhivirana. Poročilo Stephens in sod., 2015 ocenjuje, da bo do 2025 sekvenciranih 2,5 milijona organizmov. Znanstveniki iz Združenih držav Amerike, Kitajske, Združenega Kraljestva in drugih držav napovedujejo sekvenciranje do 25 % populacije ljudi, kar ob pričakovani populacijski rasti pomeni do 2 milijarde genomov ljudi.

Projekcije (Stephens in sod., 2015) kažejo, da bodo potrebe po hrambi in obdelavi genskih podatkov do leta 2025 presegle potrebe drugih domen z velikimi obsegi podatkov, kot so astronomija in socialni mediji. Predvideva se, da bo zajem novih genskih podatkov do leta 2025 presegel obseg 1 zeta baznih parov na leto z zahtevanimi kapacitetami hrambe od 2 do 40 novih eksa zlogov letno. Obdelava teh podatkov bo zahtevala 2 bilijona ur centralno procesnih enot in 10.000 bilijonov ur centralno procesnih enot za iskanje poravnav zaporedij. Pretok genskih podatkov med bankami in uporabniki naj bi do 2025 zrasel na 10 tera zlogov na sekundo.

Preglednica 4: Predvidene zahteve za obdelavo in hrambo podatkov štirih domen velikih podatkov leta 2025 (Stephens in sod., 2015)

Faza	Astronomija	Twitter	YouTube	Genomika
Zajem	25 zeta zlogov letno	0,5 – 15 milijard tvitov letno	0,5 – 0,9 milijonov ur letno	1 zeta bp letno
Hramba	1 eksa zlogov letno	1 – 17 peta zlogov letno	1 – 2 eksa zlogov letno	2 – 40 eksa zlogov letno
Analiza	*In-situ* kompresija Obdelava v realnem času Večji obsegi	Analiza konteksta in sentimenta Analiza metapodatkov	Omejene zahteve	Heterogeni podatki in analize Normalizacija podatkov za vnos v baze – 2 bilijona ur centralno procesnih enot Iskanje poravnav – 10.000 bilijonov ur centralno procesnih enot
Posredovanje	Dedicirani sistemi – 600 tera zlogov na sekundo	Minimalne zahteve	Glavnina trenutnega prometa uporabnikov – 10 mega zlogov na sekundo	Več manjših sistemov (10 mega zlogov na sekundo) in manj večjih sistemov (10 tera zlogov na sekundo)

Obdelava tolikšnega obsega podatkov zahteva njihovo normalizacijo, ki v primeru anotacij genov pri mikrobih ni dober primer. Problem raznolike nomenklature istih genov in zaporedij je tipičen problem baz podatkov z ne- ali slabo moderiranim načinom vnosa podatkov ali brez algoritemsko podprtih vnosnih pravil. Normalizacija obstoječe baze zaradi obstoječih referenc na zaporedja in anotirane gene ni mogoča, saj bi spremembe anotacij onemogočile smiselne vpoglede v bazo podatkov iz člankov, ki se navezujejo na posamična vnesena zaporedja in gene.

Rešitev lahko strokovna javnost išče v oblikovanju nove baze podatkov s poprej jasno določenimi pravili pri zajemu in označevanju zaporedij in anotacijah, medtem ko obstoječo bazo ohranijo v obstoječi obliki. Pri oblikovanju take nove baze podatkov bi lahko odločujoče pomagali programi za analizo in detekcijo nomenklaturnih anomalij, kot je to program JuP.

6 SKLEPI

- Rezultat diplomskega dela je spletno analitično orodje JuP, ki omogoča analizo vhodnih genskih zaporedij v smislu iskanja ujemajočih se, že analiziranih in označenih genskih zaporedij iz baze podatkov GenBank in vizualizacija rezultatov, ki ga obstoječa orodja NCBI ne nudijo.
- Grafični in numerični rezultati analize potrdijo mozaično naravo replikonov inkompatibilnostne skupine IncF.
- Replikone je zaradi procesa horizontalnega prenosa možno zaznati tudi v bakterijah različnih družin, ki kažejo podobno stopnjo homologije kot je zaznavna v plazmidih iste družine bakterij.
- Glede na naše rezultate je nadskupina replikonov RepFIIA s predstavniki RepFIC, RepFIIA, RepFIII in RepFVI prevladujoča med replikoni inkompatibilnostne skupine IncF.
- Zaradi nedoločenih ali slabo upoštevanih nomenklaturnih pravil je deponirana zaporedja plazmidov težko računalniško obdelovati, saj so enaki geni večkrat različno anotirani ter različni geni večkrat enako imenovani.

7 POVZETEK

Plazmidi so majhni, krožni, zunaj kromosomski DNA elementi, sposobni avtonomnega od kromosoma neodvisnega podvajanja. V po Gramu negativnih bakterijah imajo pogosto zapise za dejavnike virulence in odpornosti proti različnim protimikrobnim sredstvom. Takšni plazmidi s svojimi protimikrobnimi faktorji ključno vplivajo na spremembo mikrobne populacije, s svojimi virulentnimi lastnostmi pa na patogenost bakterije.

Plazmide so prvič identificirali pri družini *Enterobacteriaceae*, kasneje pa so ugotovili njihovo prisotnost tudi v drugih rodovih in kraljestvih. Pogosteje se pojavljajo pri arhejah in bakterijah, kjer lahko predstavljajo tudi do 25 % skupnega genskega materiala.

Replikacija plazmidov je odvisna od replikacijskih regij, ki jih imenujemo replikoni. Te razvrščamo t.i. inkompatibilnostne skupine. Ena izmed večjih inkompatibilnostnih skupin je skupina IncF, med njimi velja za najbolj mozaično inkompatibilnostno skupino razširjenja družina IncFII. Rezultati analize s programom JuP – navzkrižna primerjava zaporedij plazmidov različnih inkompatibilnostnih skupin pokaže regije visoke homologije in regije slabe homologije oz. njen popoln izostanek. To kaže na mozaično strukturo replikonov, sestavljenih iz različnih virov prek rekombinacijskih dogodkov. Prisotnost zaporedji, podobnih *Chi* zaporedjem na stikih genov iz različnih virov nakazujejo na ključnost *Chi* zaporedij pri rekombinacijskih dogodkih.

Replikoni plazmidov inkompatibilnostne skupine RepFIIA prevladujejo nad replikoni drugih skupin. Rezultati pokažejo 4 × višjo prisotnost zaporedij v genski bazi »nt«, ki kažejo visoko homologijo z replikonom RepFIIA. Razmerje ni absolutno signifikantno, saj nanj vpliva število deponiranih in anotiranih zaporedij replikonov drugih inkompatibilnostnih skupin v bazi podatkov GenBank.

Program JuP pokaže denormaliziranost podatkov replikonov plazmidov v bazi podatkov NCBI. Enaki geni so večkrat anotirani z različnimi imeni in skrajšanimi zaporedji. Pokažejo se tudi primeri, kjer so funkcijsko različni geni z različnimi nehomolognimi zaporedji imenovani z enakimi imeni. Podobno nomenklaturno stanje je prisotno tudi pri drugih zaporedjih in fenomen nekonstitenega anotiranja ni omejen le na replikacijske regije plazmidov, ki so predmet te analize. Če bi želeli ohraniti sposobnost kvalitetne računalniške obdelave, bi bilo glede na pričakovano rast genomskih podatkov smiselno natančneje določiti pravila, ki bi se jih morali deponenti dosledno držati.

8 VIRI

Accogli M., Fortini D., Giufrè M., Graziani C., Dolejska M., Carattoli A. 2013. IncI1 plasmids associated with the spread of CMY-2, CTX-M-1 and SHV-12 in *Escherichia coli* of animal and human origin. Clinical Microbiology and Infection, 19: 238-240

Altschul S. F., Gish W., Miller W., Myers E.W., Lipman D. J. 1990. Basic local alignment search tool. Journal of Molecular Biology, 215, 3: 403-410

Baquero F., Coque T. M., de la Cruz F. 2011. Ecology and evolution as targets: the need for novel eco-evo drugs and strategies to fight antibiotic resistance. Antimicrobial Agents and Chemotherapy, 55: 3649-3660

Bergquist P. L., Lane D., Saadi S., Maas W. K. 1985. Replicon fusion and the origin of the IncF group of plasmids. V: Plasmids in bacteria. Helinski D. R., Cohen S. N., Clewell D. B., Jackson D. A., Hollaender A. (eds.). New York, Plenum Publishing Corporation: 846-846

Bergquist P. L., Saadi S., Maas W. K. 1986. Distribution of basic replicons having homology with RepFIA, RepFIB and RepFIC among IncF group plasmids. Plasmid, 15: 19-34

Blomberg P., Nordstrom K., Wagner E. G. 1992. Replication control of plasmid R1: RepA synthesis is regulated by CopA RNA through inhibition of leader peptide translation. EMBO Journal, 11: 2675-2683

Boot M., Raadsen S., Savelkoul P. H. M, Vandenbroucke-Grauls C. 2013. Rapid plasmid replicon typing by real time PCR melting curve analysis. BMC Microbiology, 13, 83: doi: 10.1186/1471-2180-13-83: 5 str.

Boyd E. F., Hill C. W., Rich S. M., Hartl D. L. 1996. Mosaic structure of plasmids from natural populations of *Escherichia coli*. Genetics, 143: 1091-1100

Campbell I. G., Bergquist P. L., Mee B. J. 1987. Characterization of the maintenance functions of IncFIV plasmid R124. Plasmid, 17: 117-136

Carattoli A. 2009. Resistance plasmid families in *Enterobacteriaceae*. Antimicrobial Agents and Chemotherapy, 53: 2227-2238

Carattoli A., Bertini A., Villa L., Falbo V., Hopkins K. L., Threlfall E. J. 2005. Identification of plasmids by PCR-based replicon typing. Journal of Microbiological Methods, 63: 219-228

Casjens S., Delange M., Ley 3rd H. L., Rosa P., Huang W. M. 1995. Linear chromosomes of Lyme disease agent spirochetes: genetic diversity and conservation of gene order. Journal of Bacteriology, 177: 2769-2780

Chaudhari K. 2014. Microbial genetics. New Delhi, The Energy and Resources Institute, TERI Press: 208-216

Cheah K. C., Skurray R. 1986. The F plasmid carries an IS*3* insertion within *finO*. Journal of General Microbiology, 132: 3269-3275

Cock P. A., Antao T., Chang J. T., Bradman B. A., Cox C. J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., de Hoon M. J. L. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25: 1422-1423

Cohen S. N. 1976. Transposable genetic elements and plasmid evolution. Nature, 263: 731-738

Couturier M., Bex F., Bergquist P. L., Maas W. K. 1988. Identification and classification of bacterial plasmids. Microbiology Reviews, 52: 375-395

Dahmen S., Métayer V., Gay E., Madec J. Y., Haenni M. 2013. Characterization of extended-spectrum β-lactamase (ESBL)-carrying plasmids and clones of *Enterobacteriaceae* causing cattle mastitis in France. Veterinary Microbiology, 162: 793-799

Datta N., Hedges R. W. 1971. Compatibility groups among *fi*R factors. Nature, 234: 222-223

de Been M., Lanza V. F., de Toro M., Scharringa J., Dohmen W., Du Y., Hu J., Lei Y., Li N., Tooming-Klunderud A., Tooming-Klunderud D. J. J., Fluit A. C., Bonten M. J. M., Willems R. J. L., de la Cruz F., van Schaik W. 2014. Dissemination of cephalosporin resistance genes between *Escherichia coli* strains from farm animals and humans by specific plasmid lineages. PLoS Genetics, 10: e1004776, doi: 10.1371/journal.pgen.1004776: 17 str.

DeNap J. C., Hergenrother P. J. 2005. Bacterial death comes full circle: targeting plasmid replication in drug-resistant bacteria. Organic Biomolecular Chemistry, 3: 959-966

Dolejska M., Duskova E., Rybarikova J., Janoszowska D., Roubalova E., Dibdakova K., Maceckova G., Kohoutova L., Literak I., Smola J., Cizek A. 2011. Plasmids carrying *bla*$_{\text{CTX–M–1}}$ and *qnr* genes in *Escherichia coli* isolates from an equine clinic and a horseback riding centre. Journal of Antimicrobial Chemotherapy, 66: 757-764

Eckburg P. B., Bik E. M., Bernstein C. N., Purdom E., Dethlefsen L., Sargent M., Gill S. R., Nelson K. E., Relman D. A. 2005. Diversity of the human intestinal microbial flora. Science, 308, 5728: 1635-1638

F Plasmid – Molecular Biology. 2016. The Crankshaft Publishing: 10 str. http://what-when-how.com/molecular-biology/f-plasmid-molecular-biology/ (julij 2016)

Firth N., Ippen-Ihler K., Skurray R. A. 1996. Structure and function of the F factor and mechanism of conjugation. V: *Escherichia coli* and *Salmonella*: Cellular and molecular biology. 2nd ed. Neidhardt F. C., Curtiss 3rd R., Ingraham J. L., Lin E. C. C., Low K. B., Magasanik B., Reznikoff W. S., Riley M., Schaechter M., Umbarger H. E. (eds.). Washington, D. C., American Society for Microbiology Press: 2377-2401

Fukuhara H. 1995. Linear DNA plasmids of yeasts. FEMS Microbiology Letters, 131: 1-9

Gandy D. 2016. Font Awesome - The iconic font and CSS toolkit. Verzija 4.6.3. Cambridge, CC BY 3.0: programska oprema.
http://fontawesome.io/ (maj 2016)

García-Fernández A., Fortini D., Veldman K., Mevius D., Carattoli A. 2009. Characterization of plasmids harbouring *qnrS1*, *qnrB2* and *qnrB19* genes in *Salmonella*. Journal of Antimicrobial Chemotherapy, 63: 274-281

Garcillán-Barcia M. P., Francia M. V., de la Cruz F. 2009. The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiology Reviews, 33: 657-687

Gibbs M. D., Spiers A. J., Bergquist P. L. 1993. RepFIB: a basic replicon of large plasmids. Plasmid, 29, 3: 165-179

Gruss A., Ehrlich S. D. 1988. Insertion of foreign DNA into plasmids from gram-positive bacteria induces formation of high-molecular-weight plasmid multimers. Journal of Bacteriology, 170: 1183-1190

Gubbins M. J., Will W. R., Frost L. S. 2005. The F plasmid, a paradigm for bacterial conjugation. V: The dynamic bacterial genome. Mullany P. (ed.). Cambridge, Cambridge University Press: 151-206

Guyer M. S. 1978. The γδ sequence of F is an insertion sequence. Journal of Molecular Biology, 126: 347-365

Hall R. M., Vockler C. 1987. The region of the IncN plasmid R46 coding for resistance to β-lactam antibiotics, streptomycin/spectinomycin and sulphonamides is closely related to antibiotic resistance segments found in IncW plasmids and in Tn*21*-like transposons. Nucleic Acids Research, 15: 7491-7501

Hayakawa T., Tanaka T., Sakaguchi K., Otake N., Yonehara H. 1979. A linear plasmid-like DNA in *Streptomyces* sp. producing lankacidin group antibiotics. Journal of General and Applied Microbiology, 25: 255-260

Hayes F. 2003. Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. Science, 301: 1496-1499

Hedges R. W., Datta N. 1971. *fi*R factors giving chloramphenicol resistance. Nature, 234: 220-221

Helinski D. R., Toukdarian A. E., Novick R. P. 1996: Replication control and other stable maintenance mechanisms of plasmids. V: *Escherichia coli* and *Salmonella*: Cellular and molecular biology. 2nd ed. Neidhardt F. C., Curtiss 3rd R., Ingraham J. L., Lin E. C. C., Low K. B., Magasanik B., Reznikoff W. S., Riley M., Schaechter M., Umbarger H. E. (eds.). Washington, D. C., American Society for Microbiology Press: 2295-2324

Ho D. 2016. Notepad++ – free source code editor. Verzija 6.8.8. Paris: programska oprema https://notepad-plus-plus.org/ (maj 2016)

Ho P. L., Chan J., Lo W. U., Law P. Y., Chow K. H. 2013. Plasmid-mediated fosfomycin resistance in *Escherichia coli* isolated from pig. Veterinary Microbiology, 162: 964-967

Holmes M. L., Pfeifer F., Dyall-Smith M. L. 1995. Analysis of the halobacterial plasmid pHK2 minimal replicon. Gene, 153: 117-121

Ingmer H., Cohen S. N. 1993. The pSC101 *par* locus alters protein-DNA interactions in vivo at the plasmid replication origin. Journal of Bacteriology, 175: 6046-6048

Jiang T., Min Y. N., Liu W., Womble D. D., Rownd R. H. 1993. Insertion and deletion mutations in the *repA4* region of the IncFII plasmid NR1 cause unstable inheritance. Journal of Bacteriology, 175: 5350-5358

Johnson T. J., Logue C. M., Johnson J. R., Kuskowski M. A., Sherwood J. S., Barnes H. J., DebRoy C., Wannemuehler Y. M., Obata-Yasuoka M., Spanjaard L., Nolan L. K. 2012. Associations between multidrug resistance, plasmid content, and virulence potential among extraintestinal pathogenic and commensal *Escherichia coli* from humans and poultry. Foodborne Pathogens and Disease, 9, 1: 37-46

Johnson T. J., Wannemuehler Y. M., Johnson S. J., Logue C. M., White D. G., Doetkott C., Nolan L. K. 2007. Plasmid replicon typing of commensal and pathogenic *Escherichia coli* isolates. Applied and Environmental Microbiology, 73, 6: 1976-1983

Johnson T. J. Nolan L. K. 2009. Pathogenomics of the virulence plasmids of *Escherichia coli*. Microbiology and Molecular Biology Reviews, 73: 750-774

Kado C. I. 1998. Origin and evolution of plasmids. Antoine van Leeuwenhoek, 73: 117-126

Kemp K., Kalkur R. 2016. bootstrap-slider. Verzija 7.1.1. Oshkosh: programska oprema https://github.com/seiyria/bootstrap-slider (maj 2016)

Khan S. A. 2000. Plasmid rolling-circle replication: recent developments. Molecular Microbiology, 37: 477-484

Kinashi H., Shimaji M., Sakai A. 1987. Giant linear plasmids in *Streptomyces* which code for antibiotic biosynthesis genes. Nature, 328: 454-456

Kokate C. K., Jalalpure S. S., Hurakadle P. J. 2011. Textbook of pharmaceutical biotechnology. New Delhi, Elsevier Health Sciences: 189-216

Kornberg A., Baker T. A. 1992. DNA replication. 2nd ed. New York, Freeman: 931 str.

Lane D., Gardner R. C. 1979. Second *Eco*RI fragment of F capable of self replication. Journal of Bacteriology, 139: 141-151

Lane H. E. D. 1981. Replication and incompatibility of F and plasmids in the IncFI group. Plasmid, 5: 100-126

Lanza V. F., de Toro M., Garcillán-Barcia M. P., Mora A., Blanco J., Coque T. M., de la Cruz F. 2014. Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. PLoS Genetics, 10: e1004766, doi: 10.1371/journal.pgen.1004766: 21 str.

Li D. X., Zhang S. M., Hu G. Z., Wang Y., Liu H. B., Wu C. M., Shang Y. H., Chen Y. X., Du X. D. 2012. Tn*3*-associated *rmtB* together with *qnrS1*, *aac(6')-Ib-cr* and *bla*$_{\text{CTX-M-15}}$ are co-located on an F49:A-:B- plasmid in an *Escherichia coli* ST10 strain in China. Journal of Antimicrobial Chemotherapy, 67: 236-238

Liao X. P., Liu B. T., Yang Q. E., Sun J., Li L., Fang L. X., Liu Y. H. 2013. Comparison of plasmids coharboring 16s rRNA methylase and extended-spectrum β-lactamase genes among *Escherichia coli* isolates from pets and poultry. Journal of Food Protection, 76: 2018-2023

Lilly J., Camps M. 2015. Mechanisms of theta plasmid replication. Microbiology spectrum, 3, 1:PLAS-0029-2014, doi:10.1128/microbiolspec.PLAS-0029-2014: 11 str.

Liu B. T., Yang Q. E., Li L., Sun J., Liao X. P., Fang L. X., Yang S., Deng H., Liu Y. H. 2013. Dissemination and characterization of plasmids carrying *oqxAB-bla*$_{\text{CTX-M}}$ genes in *Escherichia coli* isolates from food-producing animals. PLoS ONE, 8: e73947, 10.1371/journal.pone.0073947: 9 str.

López J., Rodríguez J. C., Andrés I., Ortiz J. M. 1989. Characterization of the RepFVII replicon of the haemolytic plasmid pSU233: nucleotide sequence of an *incFVII* determinant. Journal of General Microbiology, 135: 1763-1768

López J., Crespo P., Rodríguez J. C., Andrés I., Ortiz J. M. 1989b. Analysis of IncF plasmids evolution: nucleotide sequence of an IncFIII replication region. Gene, 78: 183-187

López J., Delgado D., Andrés I., Ortiz J. M., Rodríguez J. C. 1991. Isolation and evolutionary analysis of a RepFVIB replicon of the plasmid pSU212. Journal of General Microbiology, 137: 1093-1099

Maas R., Wang C. 1997. Role of the RepA1 protein in RepFIC plasmid replication. Journal of Bacteriology, 179: 2163-2168

Maas R. 2001. Change of plasmid DNA structure, hypermethylation, and Lon-proteolysis as steps in a replicative cascade. Cell, 105: 945-955

Madigan M. T., Martinko J. M., Bender K. S., Buckley D. H., Stahl D. A. 2014. Brock biology of microorganisms. 14th ed. London, Prentice-Hall International, Inc.: 1030 str.

Mainil J. G., Daube G. 2005. Verotoxigenic *Escherichia coli* from animals, humans and foods: who's who? Journal of General and Applied Microbiology, 98: 1332-1344

Mathers A. J., Peirano G., Pitout J. D. 2015. The role of epidemic resistance plasimds and international high-risk clones in the spread of multidrug-resistant in *Enterobacteriaceae*. Clinical Microbiology Reviews, 28: 565-591

Miyashita S., Hirochika H., Ikeda J. E., Hashiba T. 1990. Linear plasmid DNAs of the plant pathogenic fungus *Rhizoctonia* solani with unique terminal structures. Molecular Genetics and Genomics, 220: 165-171

Moat A. G., Foster J. W., Spector M. P. 2002. Microbial physiology. 4th ed. New Jersey, John Wiley and Sons, Inc.: 101-167

Moran R. A., Anantham S., Pinyon J. L., Hall R. M. 2015. Plasmids in antibiotic susceptible and antibiotic resistant commensal *Escherichia coli* from healthy Australian adults. Plasmid, 80: 24-31

Murakami Y., Ohmori H., Yura T., Nagata T. 1987. Requirement of the *Escherichia coli dnaA* gene function for *ori*-2-dependent mini-F plasmid replication. Journal of Bacteriology, 169: 1724-1730

Netolitzky D. J., Wu X., Jensen S. E., Roy K. L. 1995. Giant linear plasmids of β-lactam antibiotic producing *Streptomyces*. FEMS Microbiology Letters, 131: 27-34

Novick R. P., Hoppensteadt F. C. 1978. On plasmid incompatibility. Plasmid, 1: 421-434

Novick R. P. 1987. Plasmid incompatibility. Microbiological Reviews, 51: 381-395

Osborn A. M., da Silva Tatley F. M., Steyn L. M., Pickup R. W., Saunders J. R. 2000. Mosaic plasmids and mosaic replicons: evolutionary lessons from the analysis of genetic diversity in IncFII-related replicons. Microbiology, 146: 2267-2275

Pansegrau W., Lanka E., Barth P. T. Figurski D. H., Guiney D. G., Haas D., Helinski D. R., Schwab H., Stanisich V. A., Thomas C. M. 1994. Complete nucleotide sequence of Birmingham IncPa plasmids. Journal of Molecular Biology, 239: 623-663

Python Software Foundation. 2016. Python Language Reference. Verzija 2.7.5. Delaware, Python Software Foundation: programska oprema. https://www.python.org/ (maj 2016)

Perez-Casal J. F., Crosa J. H. 1984. Aerobactin iron uptake sequences in plasmid ColV-K30 are flanked by inverted IS*1*-like elements and replication regions. Journal of Bacteriology, 160: 256-265

Perichon B., Bogaerts P., Lambert T., Frangeul L., Courvalin P., Galimand M. 2008. Sequence of conjugative plasmid pIP1206 mediating resistance to aminoglycosides by 16S rRNA methylation and to hydrophilic fluoroquinolones by efflux. Antimicrobial Agents and Chemotherapy, 52: 2581-2592

Petre S., Aguilar J. 2016. Bootstrap Colorpicker for Twitter Bootstrap. Verzija 2.3.2. Berlin: programska oprema http://mjolnic.com/bootstrap-colorpicker/ (maj 2016)

Picken R. N., Mazaitis A. J., Saadi S., Maas W. K. 1984. Characterisation of the basic replicons of the chimeric R/Ent plasmid pCG86 and the related Ent plasmid P307. Plasmid, 12: 10-18

Ray A., Skurray R. 1983. Cloning and polypeptide analysis of the leading region in F plasmid DNA transfer. Plasmid, 9: 262-272

Regalado A. 2014. EmTech: Illumina says 228,000 human genomes will be sequenced this year. Cambridge, Massachusetts Institute of Technology: 2 str. http://www.technologyreview.com/news/531091/emtech-illumina-says-228000-human-genomes-will-be-sequenced-this-year/ (maj 2016)

Ruiz E., Sáenz Y., Zarazaga M., Rocha-Gracia R., Martínez-Martinez L., Arlet G., Torres C. 2012. *qnr*, *aac(6')-Ib-cr* and *qepA* genes in *Escherichia coli* and *Klebsiella spp*.: genetic environments and plasmid and chromosomal location. Journal of Antimicrobial Chemotherapy, 67: 886-897

Russo T. A., Johnson J. R. 2000. Proposal for a new inclusive designation for extraintestinal pathogenic isolates of *Escherichia coli*: ExPEC. Journal of Infectious Diseases, 181: 1753-1754

Saadi S., Maas W. K., Hill D. F., Bergquist P. L. 1987. Nucleotide sequence analysis of RepFIC, a basic replicon present in IncFI plasmids P307 and F, and its relation to the RepA replicon of IncFII plasmids. Journal of Bacteriology, 169, 5: 1836-1846

Sayers E. W., Barrett T., Benson D. A., Bryant S. H., Canese K., Chetvernin V., Church D. M., DiCuccio M., Edgar R., Federhen S., Feolo M., Geer L. Y., Helmberg W., Kapustin Y., Landsman D., Lipman D. J., Madden T. L., Maglott D. R., Miller V., Mizrachi I., Ostell J., Pruitt K. D., Schuler G. D., Sequeira E., Sherry S. T., Shumway M., Sirotkin

K., Souvorov A., Starchenko G., Tatusova T. A., Wagner L., Yaschenko E., Ye J. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 37: D5-D15. doi: 10.1093/nar/gkn741: 11 str.

Smith G. R., Kunes S. M., Schultz D. W., Taylor A., Triman K. L. 1981. Structure of *Chi* hotspots of generalized recombination. Cell, 24: 429-436

Smith G. R. 1987. Mechanism and control of homologous recombination in *Escherichia coli*. Annual Review of Genetics, 21: 179-201

Solar G. del, Giraldo, R., Ruiz-Echevarria M. J., Espinosa M., Diaz-Orejas R. 1998. Replication and control of circular bacterial plasmids. Microbiology and Molecular Biology Reviews, 62: 434-464

Starčič Erjavec M., Gaastra W., Žgur-Bertok D. 2002. *Tra* region of the natural conjugative *Escherichia coli* plasmid pRK100 is F-like. Acta Biologica Slovenica, 45: 9-15

Starčič Erjavec M., Gaastra W., van Putten J., Žgur-Bertok D. 2003. Identification of the origin of replications and partial characterization of plasmid pRK100. Plasmid, 50: 102-112

Starčič Erjavec M., Žgur-Bertok D. 2006. The RepFIIA replicon of the natural *Escherichia coli* plasmid pRK100. Acta Biologica Slovenica, 49, 2: 3-12

Stephens Z. D., Lee S. Y., Faghri F., Campbell R. H., Zhai C., Efron M. J., Iyer R., Schatz M. C., Sinha S., Robinson G. E. 2015. Big data: astronomical or genomical? PLoS Biology 13, 7: e1002195, doi. 10.1371/journal.pbio.1002195: 11 str.

Szczepanowski R., Braun S., Riedel V., Schneiker S., Krahn I., Puhler A., Schluter A. 2005. The 120.592 bp IncF plasmid pRSB107 isolated from a sewage-treatment plant encodes nine different antibiotic-resistance determinants, two iron-acquisition systems and other putative virulence-associated functions. Microbiology, 151: 1095-1111

Tamang M. D., Seol S. Y., Oh J. Y., Kang H. Y., Lee J. C., Lee Y. C., Cho D. T., Kim J. 2008. Plasmid-mediated quinolone resistance determinants *qnrA*, *qnrB*, and *qnrS* among clinical isolates of *Enterobacteriaceae* in a Korean hospital. Antimicrobial Agents and Chemotherapy, 52: 4159-4162

Tamm J., Polisky B. 1983. Structural analysis of RNA molecules involved in plasmid copy number control. Nucleic Acids Research, 11: 6381-6397

Taylor A. F., Smith G. R. 1995. Strand specifcity of nicking of DNA at *Chi* sites by RecBCD enzyme. Journal of Biological Chemistry, 270: 24459-24467

Taylor D. E., Gibreel A., Lawley T. D., Tracz D. M. 2004. Antibiotic resistance plasmids. V: Plasmid biology. Funnell B., Phillips G. (eds.). Washington, D. C., American Society for Microbiology Press: 473-491

The jQuery Foundation. 2016. jQuery. Verzija 2.2.4. Wallnut, California. The jQuery Foundation: programska oprema https://jquery.com/ (maj 2016)

Thomas C. M., Nielsen K. M. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature Reviews Microbiology, 3: 711-721

Tolun A., Helinski D. R. 1981. Direct repeats of the F plasmid *incC* region express F incompatibility. Cell, 24: 687-694

Toukdarian A. 2004. Plasmid strategies for broad-host-range replication in Gram-negative bacteria. V: Plasmid biology. Funnell B., Phillips G. (eds.). Washington, D. C., American Society for Microbiology Press: 259-270

Tripp H. J., Hewson I., Boyarsky S., Stuart J. M., Zehr J. P. 2011. Misannotations of rRNA can now generate 90 % false positive protein matches in metatranscriptomic studies. Nucleic Acids Research, 39, 20: 8792-8802

Twitter. 2016. Bootstrap, HTML, CSS, and JS framework for developing responsive projects on the web. Verzija 3.3.6. San Francisco. Twitter Inc: programska oprema http://getbootstrap.com/ (maj 2016)

Vanooteghem J. C., Cornelis G. R. 1990. Structural and functional similarities between the replication region of the *Yersinia* virulence plasmid and the RepFIIA replicons. Journal of Bacteriology, 172: 3600-3608

Villa L., García-Fernández A., Fortini D., Carattoli A. 2010. Replicon sequence typing of IncF plasmids carrying virulence and resistance determinants. Journal of Antimicrobial Chemotherapy, 65: 2518-2529

Wilson J. W. 2006. Genetic exchange in bacteria and the modular structure of mobile DNA elements. V: Molecular paradigms of infectious disease: a bacterial perspective. Nickerson C. A., Schurr M. J. (eds.). New York, Springer Science+Business Media: 34-77

Woodford N., Carattoli A., Karisik E., Underwood A., Ellington M. J., Livermore D. M. 2009. Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major *Escherichia coli* lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone. Antimicrobial Agents and Chemotherapy, 53: 4472-4482

Yarmolinsky M. B. 2000. A pot-pourri of plasmid paradoxes: effects of a second copy. Molecular Microbiology, 38: 1-7

Zillig W., Arnold H. P., Holz I., Prangishvili D., Schweier A., Stedman K., She Q., Phan H., Garrett R., Kristjansson J. K. 1998. Genetic elements in the extremely thermophilic archaeon *Sulfolobus*. Extremophiles, 2: 131-140

ZAHVALA

Mentorici prof. dr. Marjanci Starčič Erjavec se iskreno zahvaljujem za mentorstvo, usmeritve in potrpljenje, ki ga je pokazala ob nastajanju te diplomske naloge. Rezultat tega dela je v veliki meri rezultat njene vizije.

Zahvala gre tudi recenzentu, doc. dr. Tomažu Accettu za strokoven in natančen pregled diplomskega dela ter njegove pripombe, ki so delo naredile kvalitetnejše.

Hvala Nacetu Kranjcu, ki mi je pomagal začeti in me uvedel v osnove BioPythona. Hvala Benjaminu Dobravcu za njegove usmeritve in pomoč pri razvoju spletnega vmesnika. Brez vajine pomoči bi bil rezultat gotovo slabši.

Največjo zahvalo dolgujem moji družini, najbolj Nataliji ter mojim staršem, ki so kljub daljši prekinitvi študija verjeli vame in mi nudili moralno podporo pri izdelavi diplomskega dela ob oblici rednih družinskih in delovnih obveznosti.

Hvala vam!

PRILOGE

Priloga A: JSON shema zaledne skripte JuP, po kateri oblikuje izhodne podatke za vizualizacijo


```
{
  "$schema": "http://json-schema.org/draft-04/schema#",
  "type": "object",
  "properties": {
    "sequence-id": {
      "type": "object",
      "properties": {
        "ACC": {
          "type": "string",
          "title": "Accession number",
          "description": "Sequence Accession number"
        },
        "QUERY_LEN": {
          "type": "integer",
          "title": "Query sequence length",
          "description": "Length of query sequence"
        },
        "HSPS": {
          "type": "array",
          "items": {
            "type": "object",
            "properties": {
              "IDENT": {
                "type": "integer",
                "title": "Identities count",
                "description": "Number of positive identities in HSP"
              },
              "QUERY_OFF_END": {
                "type": "integer",
                "title": "HSP End Offset",
                "description": "HSP End Offset on Query sequence"
              },
              "SUBJ_OFF_END": {
                "type": "integer",
                "title": "Subject End Offset",
                "description": "HSP End Offset on Subject sequence"
              },
              "SUBJ_OFF_START_NORM": {
                "type": "integer",
                "title": "Normalized Subject End Offset",
                "description": "Normalized HSP End Offset on Subject sequence"
              },
              "LEN": {
                "type": "integer",
                "title": "HSP Length",
                "description": "HSP Sequence Length"
              },
              "URL_GRAPH": {
                "type": "string",
                "title": "HSP Graphics URL Address",
                "description": "HSP Graphics URL Address at GenBank"
              },
              "URL_GB": {
                "type": "string",
                "title": "HSP URL Address",
                "description": "HSP Record URL Address at GenBank"
              },
              "IDENT_COVER": {
                "type": "number",
                "title": "HSP Coverage percentage",
                "description": "The HSP Coverage percentage with Query Sequence"
              },
              "GAPS": {
                "type": "integer",
                "title": "Number of gaps in HSP",
                "description": "The Number of gaps in HSP"
              },
              "QUERY_STRAND": {
                "type": "string",
                "title": "Query match sequence in HSP",
                "description": "The Query nucleotide seqeunce which matches with subject sequence in this HSP"
              },
              "GAPS_COVER": {
                "type": "number",
                "title": "HSP Gaps Coverage percentage",
                "description": "The HSP Gaps Coverage percentage"
              },
              "CDS": {
                "type": "array",
                "items": {
```


se nadaljuje.

Nadaljevanje priloge A: JSON shema zaledne skripte JuP, po kateri oblikuje izhodne podatke za vizualizacijo


```
"type": "object",
"properties": {
  "ALIGN_MATCHES": {
    "type": "integer",
    "title": "Number of matches with query sequence",
    "description": "The number of matches with query sequence"
  },
  "LOC": {
    "type": "string",
    "title": "CDS location",
    "description": "CDS location on Subject sequence"
  },
  "PRODUCT": {
    "type": "string",
    "title": "CDS product name",
    "description": "The name of the product which CDS encodes"
  },
  "ALIGN_GAPS": {
    "type": "integer",
    "title": "CDS alignment gaps",
    "description": "CDS gaps on Alignment Sequence"
  },
  "NAME": {
    "type": "string",
    "title": "CDS Name",
    "description": "The name of CDS feature"
  },
  "NOTE": {
    "type": "string",
    "title": "CDS Description",
    "description": "The description of CDS feature"
  },
  "ALIGN_OFF_START": {
    "type": "integer",
    "title": "CDS Alignment Start Offset",
    "description": "CDS Start Offset on Alignment sequence"
  },
  "QUERY_OFF_START": {
    "type": "integer",
    "title": "CDS Start Offset",
    "description": "CDS Start Offset on Query sequence"
  },
  "ALIGN_LEN": {
    "type": "integer",
    "title": "CDS Alignment Length",
    "description": "CDS Sequence Alignment Length"
  },
  "LEN": {
    "type": "integer",
    "title": "CDS Length",
    "description": "CDS Sequence Length"
  },
  "URL_GRAPH": {
    "type": "string",
    "title": "CDS Graphics URL Address",
    "description": "CDS Graphics URL Address at GenBank"
  },
  "URL_GB": {
    "type": "string",
    "title": "CDS URL Address",
    "description": "CDS Record URL Address at GenBank"
  },
  "OFF_START": {
    "type": "integer",
    "title": "Subject Start Offset",
    "description": "CDS Start Offset on Subject sequence"
  },
  "ALIGN_OFF_END": {
    "type": "integer",
    "title": "CDS Alignment End Offset",
    "description": "CDS End Offset on Alignment sequence"
  },
  "URL_PRODUCT": {
    "type": "string"
  },
  "QUERY_OFF_END": {
    "type": "integer",
    "title": "CDS End Offset",
    "description": "CDS End Offset on Query sequence"
  },
  "ALIGN_COVER": {
    "type": "number",
    "title": "CDS Coverage percentage",
    "description": "The CDS Coverage percentage with Query Sequence"
  },
  "ALIGN_MATCH": {
    "type": "string",
```


se nadaljuje.

Nadaljevanje priloge A: JSON shema zaledne skripte JuP, po kateri oblikuje izhodne podatke za vizualizacijo


```
                    "title": "CDS Match string",
                    "description": "CDS Alignment match string (sequence of '|' and ' ' characters)"
                  },
                  "OFF_END": {
                    "type": "integer",
                    "title": "Subject End Offset",
                    "description": "CDS End Offset on Subject sequence"
                  },
                  "PRODUCT_ID": {
                    "type": "string",
                    "title": "Gene Product Id",
                    "description": "The idetifier of the gene product protein"
                  }
                }
              }
            },
            "SUBJ_STRAND": {
              "type": "string",
              "title": "Subject match sequence in HSP",
              "description": "The Subject nucleotide seqeunce which matches with subject sequence in this HSP"
            },
            "SUBJ_OFF_START": {
              "type": "integer",
              "title": "Subject Start Offset",
              "description": "HSP Start Offset on Subject sequence"
            },
            "SUBJ_OFF_END_NORM": {
              "type": "integer",
              "title": "Normalized Subject Start Offset",
              "description": "Normalized HSP Start Offset on Subject sequence"
            },
            "SCORE_BITS": {
              "type": "integer",
              "title": "Score in bits",
              "description": "HSP match score in bits"
            },
            "QUERY_OFF_START": {
              "type": "integer",
              "title": "HSP Start Offset",
              "description": "HSP Start Offset on Query sequence"
            },
            "SCORE": {
              "type": "integer",
              "title": "Score",
              "description": "HSP match score"
            },
            "ALIGN_MATCH": {
              "type": "string",
              "title": "HSP Match string",
              "description": "HSP Alignment match string (sequence of '|' and ' ' characters)"
            }
          }
        }
      },
      "URL_GB": {
        "type": "string",
        "title": "Sequence URL Address",
        "description": "Sequence Record URL Address at GenBank"
      },
      "GI": {
        "type": "string",
        "title": "gi identifier",
        "description": "The match sequence gi identifier"
      },
      "DEF": {
        "type": "string",
        "title": "Sequence name",
        "description": "The name of the match sequence"
      }
    }
  }
}
```

Priloga B: Analizirano nukleotidno zaporedje replikona RepFIA v obliki FASTA

```
>gi|8918823:44600-53300 Escherichia coli K-12 plasmid F DNA, complete sequence
CGATCCTCAGGTACATCTTCAACTGCCGCATCAAGCCCTTTCTGGTTCCACAGCTGCCTCAGTATCTTCA
TATTACAGATCTCCGTACAGGCAGCGCAGCCCGGTCAGGGCCTCCGCCAGCGGGCCTTCTTGACACCGCG
TGGCGAGCTCCCGGACCAGCGTCTCAAAGGCCTCCGCCGGCATAAAAATCCGCAGGCCGTCTATCCCGAA
GAAGCCCTCCGGCGTGACGTTTCCGGGCGGGGAGAACATCACTCGCTCCCCCTGGTGACCGTGCGTCAGG
GCTTCCGGATGCCGGGCATACAGCGTGAACAGGCCCAGCAGTTCGGATAACGCTTCCCAGCCGTACGGCA
CCACGAAGTCCTGGCCGGTGATCAGCAGGTGCAGACGGGGTGTGGTGTACCAGGCCCCCGGGCGTTCTCC
GGGGTGCTGGCCGTCAACGCGGATCTCCAGGCGCAGCGTGCCGGCCTCAATGGTTTCCGTGACGGCAGAA
ATAACCGGGCGCAGCTCCTGTGCCAGCGTTCTTGCCTGTTCTGTGTTCCAGTCCAGGCCGCGCAGCATCA
ACGGGAAAACAGCTTTCAGCCGCGGCAGGGTGAAGATTTCTGCCAGCACCGCGTCCGGCACCTCCGCCAG
GCCAAAACAGTCGCTCTCGAGCGGACGCAGGAGGTGTTCGGCATACATCTGATCCACCACGGGGCGCAGC
GCTGCCCGGAAGGCTTCAAACTCTTCCGTCCAGTTTTCCCGGCCAGGCGGAAAATCCCCTTCAGGTAGT
GACCGTCCACAGGGCGATCGTGCTCCACCTGCCAGGCCAGCAGGTTGCCAGTGATGTCCAGCAGCGTGTC
CAGGGCCGGGAGTGTGGCCAGCCTATTATGCCCGCGCAGGAAGGCCTCCCTGACGTGAACCAGGACAAAG
CGCAGTTCGTCGCGGGAGAGCGCTGACGTGCCGAAGGTCTGGCCCAGAATGATATGCCGGTACAGCTGCC
GGACCGTGGCCTCCGGATCGGCAATGAGCTGAAAATACCCGTCACCGGGCGCGACCGTTTTCAGGCCGTC
ATCGAGGAAGCGTTCGGCTAGGGCGTTGACCGAAGTATTTTCCCGACCGGCACGCGTTTTCAGCGCCTCG
ATAAGTTTTTTGTGAAAACGCAGGTTAAGCTGGCTTAGCATCATCGATCCTTTAATGTTTGTTTTGTTCT
TAATGCTAACCTAGCATTTTATTATTAGCTATTAATCAGTTTTTTAGTTTTTGAAAATTGTGATGCTTCT
AAAATTACTAAAATTGAGGATTTTAATGCTACAACAATGCCTGCCTCTTCTTATTTCTCCGGAGATCCGA
AAACCCCAAGTTACGGATCTTCCTCTCCCTCCGCACAGCGTTACATCCCGTCAGCACAGCATGTAGTGCC
TCATACAGTTGCCCATGGCACTATATGTTGTGTTGTATCTCTGGACTGTGATGCGCCGCGCAGGGGCGGA
AAACAGCGATATGATGATTTTCTCAGCGTTGTACACTTCCGGAAAGTCGTTTATTCAAATAAAGTCGGAT
CCATACGAAACGGGAATGCGGTAATTACGCTTTGTTTTTATAAGTCAGATTTTAATTTTTATTGGTTAAC
ATAACGAAAGGTAAAATACATAAGGCTTACTAAAAGCCAGATAACAGTATGCGTATTTGCGCGCTGATTT
TTGCGGTATAAGAATATATACTGATATGTATACCCGAAGTATGTCAAAAAGAGGTGTGCTATGAAGCAGC
GTATTACAGTGACAGTTGACAGCGACAGCTATCAGTTGCTCAAGGCATATGATGTCAATATCTCCGGTCT
GGTAAGCACAACCATGCAGAATGAAGCCCGTCGTCTGCGTGCCGAACGCTGGAAAGCGGAAAATCAGGAA
GGGATGGCTGAGGTCGCCCGGTTTATTGAAATGAACGGCTCTTTTGCTGACGAGAACAGGGACTGGTGAA
ATGCAGTTTAAGGTTTACACCTATAAAAGAGAGAGCCGTTATCGTCTGTTTGTGGATGTACAGAGTGATA
TTATTGACACGCCCGGGCGACGGATGGTGATCCCCCTGGCCAGTGCACGTCTGCTGTCAGATAAAGTCTC
CCGTGAACTTTACCCGGTGGTGCATATCGGGGATGAAAGCTGGCGCATGATGACCACCGATATGGCCAGT
GTGCCGGTCTCCGTTATCGGGGAAGAAGTGGCTGATCTCAGCCACCGCGAAAATGACATCAAAAACGCCA
TTAACCTGATGTTCTGGGGAATATAAATGTCAGGCTCCGTTATACACAGCCAGTCTGCAGTCATGGTACC
GGCAGTGTATTCTGCCGGGCAGCCTGCATCGCTGCCTGTTGCCATTGATTATCCGGCAGCTCTGGCACTT
CGCCAGATGTCGATGGTCCATGATGAACTGCCAAAATATCTGCTGGCTCCGGAAGTGAGTGCCCTGCTCC
ATTACGTCCCGGATCTGCACCGCAAGATGCTGCTGGCCACACTGTGGAACACCGGAGCACGCATTAATGA
AGCACTGGCGCTGACGCGGGGGGATTTTTCGCTTGCGCCTCCGTATCCGTTTGTGCAGCTTGCGACCCTG
AAACAACGGACCGAAAAAGCTGCCAGGACGGCGGGGAGAATGCCCGCCGGTCAGCAGACTCACCGGCTGG
TTCCGCTCTCTGACTCCTGGTACGTCAGCCAGCTGCAGACGATGGTGGCAACACTGAAAATACCCATGGA
GCGGCGTAACCGTCGCACAGGAAGGACAGAGAAAGCGCGGATCTGGGAAGTGACGGACAGAACGGTCAGG
ACCTGGATTGGGGAGGCGGTTGCCGCCGCTGCTGCTGACGGTGTGACGTTCTCTGTTCCGGTCACACCAC
ATACGTTCCGCCATTCCTATGCGATGCACATGCTGTATGCCGGTATACCGCTGAAAGTTCTGCAAAGCCT
GATGGGACATAAGTCCATCAGTTCAACGGAAGTCTACACGAAGGTTTTTGCGCTGGATGTGGCTGCCCGG
CACCGGGTGCAGTTTGCGATGCCGGAGTCTGATGCGGTTGCGATGCTGAAACAATTATCCTGAGAATAAA
TGCCTTGGCCTTTATATGGAAATGTGGAACTGAGTGGATATGCTGTTTTTGTCTGTTAAACAGAGAAGCT
GGCTGTTATCCACTGAGAAGCGAACGAAACAGTCGGGAAAATCTCCCATTATCGTAGAGATCCGCATTAT
TAATCTCAGGAGCCTGTGTAGCGTTTATAGGAAGTAGTGTTCTGTCATGATGCCTGCAAGCGGTAACGAA
AACGATTTGAATATGCCTTCAGGAACAATAGAAATCTTCGTGCGGTGTTACGTTGAAGTGGAGCGGATTA
TGTCAGCAATGGACAGAACAACCTAATGAACACAGAACCATGATGTGGTCTGTCCTTTTACAGCCAGTAG
TGCTCGCCGCAGTCGAGCGACAGGGCGAAGCCCTCGAGTGAGCGAGGAAGCACCAGGGAACAGCACTTAT
ATATTCTGCTTACACACGATGCCTGAAAAAACTTCCCTTGGGGTTATCCACTTATCCACGGGGATATTTT
TATAATTATTTTTTTTATAGTTTTTAGATCTTCTTTTTTAGAGCGCCTTGTAGGCCTTTATCCATGCTGG
TTCTAGAGAAGGTGTTGTGACAAATTGCCCTTTCAGTGTGACAAATCACCCTCAAATGACAGTCCTGTCT
GTGACAAATTGCCCTTAACCCTGTGACAAATTGCCCTCAGAAGAAGCTGTTTTTTCACAAAGTTATCCCT
GCTTATTGACTCTTTTTTATTTAGTGTGACAATCTAAAAACTTGTCACACTTCACATGGATCTGTCATGG
CGGAAACAGCGGTTATCAATCACAAGAAACGTAAAAATAGCCCGCGAATCGTCCAGTCAAACGACCTCAC
TGAGGCGGCATATAGTCTCTCCCGGGATCAAAAACGTATGCTGTATCTGTTCGTTGACCAGATCAGAAAA
TCTGATGGCACCCTACAGGAACATGACGGTATCTGCGAGATCCATGTTGCTAAATATGCTGAAATATTCG
GATTGACCTCTGCGGAAGCCAGTAAGGATATACGGCAGGCATTGAAGAGTTTCGCGGGGAAGGAAGTGGT
TTTTTATCGCCCTGAAGAGGATGCCGGCGATGAAAAAGGCTATGAATCTTTTCCTTGGTTTATCAAACGT
GCGCACAGTCCATCCAGAGGGCTTTACAGTGTACATATCAACCCATATCTCATTCCCTTCTTTATCGGGT
TACAGAACCGGTTTACGCAGTTTCGGCTTAGTGAAACAAAAGAAATCACCAATCCGTATGCCATGCGTTT
ATACGAATCCCTGTGTCAGTATCGTAAGCCGGATGGCTCAGGCATCGTCTCTCTGAAAATCGACTGGATC
ATAGAGCGTTACCAGCTGCCTCAAAGTTACCAGCGTATGCCTGACTTCCGCCGCCGCTTCCTGCAGGTCT
GTGTTAATGAGATCAACAGCAGAACTCCAATGCGCCTCTCATACATTGAGAAAAAGAAAGGCCGCCAGAC
GACTCATATCGTATTTTCCTTCCGCGATATCACTTCCATGACGACAGGATAGTCTGAGGGTTATCTGTCA
CAGATTTGAGGGTGGTTCGTCACATTTGTTCTGACCTACTGAGGGTAATTTGTCACAGTTTTGCTGTTTC
CTTCAGCCTGCATGGATTTTCTCATACTTTTTGAACTGTAATTTTTAAGGAAGCCAAATTTGAGGGCAGT
TTGTCACAGTTGATTTCCTTCTCTTTCCCTTCGTCATGTGACCTGATATCGGGGGTTAGTTCGTCATCAT
TGATGAGGGTTGATTATCACAGTTTATTACTCTGAATTGGCTATCCGCGTGTGTACCTCTACCTGGAGTT
TTTCCCACGGTGGATATTTCTTCTTGCGCTGAGCGTAAGAGCTATCTGACAGAACAGTTCTTCTTTGCTT
CCTCGCCAGTTCGCTCGCTATGCTCGGTTACACGGCTGCGGCGAGCGCTAGTGATAATAAGTGACTGAGG
TATGTGCTCTTCTTATCTCCTTTTGTAGTGTTGCTCTTATTTTAAACAACTTTGCGGTTTTTTGATGACT
TTGCGATTTTGTTGTTGCTTTGCAGTAAATTGCAAGATTTAATAAAAAAACGCAAAGCAATGATTAAAGG
ATGTTCAGAATGAAACTCATGGAAACACTTAACCAGTGCATAAACGCTGGTCATGAAATGACGAAGGCTA
TCGCCATTGCACAGTTTAATGATGACAGCCCGGAAGCGAGGAAAATAACCCGGCGCTGGAGAATAGGTGA
AGCAGCGGATTTAGTTGGGGTTTCTTCTCAGGCTATCAGAGATGCCGAGAAAGCAGGGCGACTACCGCAC
CCGGATATGGAAATTCGAGGACGGGTTGAGCAACGTGTTGGTTATACAATTGAACAAATTAATCATATGC
GTGATGTGTTTGGTACGCGATTGCGACGTGCTGAAGACGTATTTCCACCGGTGATCGGGGTTGCTGCCCA
TAAAGGTGGCGTTTACAAAACCTCAGTTTCTGTTCATCTTGCTCAGGATCTGGCTCTGAAGGGGCTACGT
GTTTTGCTCGTGGAAGGTAACGACCCCCAGGGAACAGCCTCAATGTATCACGGATGGGTACCAGATCTTC
ATATTCATGCAGAAGACACTCTCCTGCCTTTCTATCTTGGGGAAAAGGACGATGTCACTTATGCAATAAA
GCCCACTTGCTGGCCGGGGCTTGACATTATTCCTTCCTGTCTGGCTCTGCACCGTATTGAAACTGAGTTA
ATGGGCAAATTTGATGAAGGTAAACTGCCCACCGATCCACACCTGATGCTCCGACTGGCCATTGAAACTG
TTGCTCATGACTATGATGTCATAGTTATTGACAGCGCGCCTAACCTGGGTATCGGCACGATTAATGTCGT
ATGTGCTGCTGATGTGCTGATTGTTCCCACGCCTGCTGAGTTGTTTGACTACACCTCCGCACTGCAGTTT
TTCGATATGCTTCGTGATCTGCTCAAGAACGTTGATCTTAAAGGGTTCGAGCCTGATGTACGTATTTTGC
TTACCAAATACAGCAATAGTAATGGCTCTCAGTCCCCGTGGATGGAGGAGCAAATTCGGGATGCCTGGGG
AAGCATGGTTCTAAAAAATGTTGTACGTGAAACGGATGAAGTTGGTAAAGGTCAGATCCGGATGAGAACT
```

se nadaljuje.

Nadaljevanje priloge B: Analizirano nukleotidno zaporedje replikona RepFIA v obliki FASTA

```
GTTTTTGAACAGGCCATTGATCAACGCTCTTCAACTGGTGCCTGGAGAAATGCTCTTTCTATTTGGGAAC
CTGTCTGCAATGAAATTTTCGATCGTCTGATTAAACCACGCTGGGAGATTAGATAATGAAGCGTGCGCCT
GTTATTCCAAAACATACGCTCAATACTCAACCGGTTGAAGATACTTCGTTATCGACACCAGCTGCCCCGA
TGGTGGATTCGTTAATTGCGCCGCGTAGGAGTAATGGCTCGCGGTAATGCCATTACTTTGCCTGTATGTGG
TCGGGATGTGAAGTTTACTCTTGAAGTGCTCCGGGGTGATAGTGTTGAGAAGACCTCTCGGGTATGGTCA
GGTAATGAACGTGACCAGGAGCTGCTTACTGAGGACGCACTGGATGATCTCATCCCTTCTTTTCTACTGA
CTGGTCAACAGACACCGGCGTTCGGTCGAAGAGTATCTGGTGTCATAGAAATTGCCGATGGGAGTCGCCG
TCGTAAAGCTGCTGCACTTACCGAAAGTGATTATCGTGTTCTGGTTGGCGAGCTGGATGATGAGCAGATG
GCTGCATTATCCAGATTGGGTAACGATTATCGCCCAACAAGTGCTTATGAACGTGGTCAGCGTTATGCAA
GCCGATTGCAGAATGAATTTGCTGGAAATATTTCTGCGCTGGCTGATGCGGAAAATATTTCACGTAAGAT
TATTACCCGCTGTATCAACACCGCCAAATTGCCTAAATCAGTTGTTGCTCTTTTTTCTCACCCCGGTGAA
CTATCTGCCCGGTCAGGTGATGCACTTCAAAAAGCCTTTACAGATAAAGAGGAATTACTTAAGCAGCAGG
CATCTAACCTTCATGAGCAGAAAAAAGCTGGGGTGATATTTGAAGCTGAAGAAGTTATCACTCTTTTAAC
TTCTGTGCTTAAAACGTCATCTGCATCAAGAACTAGTTTAAGCTCACGACATCAGTTTGCTCCTGGAGCG
ACAGTATTGTATAAGGGCGATAAAATGGTGCTTAACCTGGACAGGTCTCGTGTTCCAACTGAGTGTATAG
AGAAAATTGAGGCCATTCTTAAGGAACTTGAAAAGCCAGCACCCTGATGCGACCACGTTTTAGTCTACGT
TTATCTGTCTTTACTTAATGTCCTTTGTTACAGGCCAGAAAGCATAACTGGCCTGAATATTCTCTCTGGG
CCCACTGTTCCACTTGTATCGTCGGTCTGATAATCAGACTGGGACCACGGTCCCACTCGTATCGTCGGTC
TGATTATTAGTCTGGGACCACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCC
ACTCGTATCGTCGGTCTGATAATCAGACTGGGACCACGGTCCCACTCGTATCGTCGGTCTGATTATTAGT
CTGGGACCATGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCGTATCGT
CGGTCTGATTATTAGTCTGGGACCACGGTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACG
GTCCCACTCGTATCGTCGGTCTGATTATTAGTCTGGGACCACGATCCCACTCGTGTTGTCGGTCTGATTA
TCGGTCTGGGACCACGGTCCCACTTGTATTGTCGATCAGACTATCAGCGTGAGACTACGATTCCATCAAT
GCCTGTCAAGGGCAAGTATTGACATGTCGTCGTAACCTGTAGAACGGAGTAACCTCGGTGTGCGGTTGTA
TGCCTGCTGTGGATTGCTGCTGTGTCCTGCTTATCCACAACATTTTGCGCACGGTTATGTGGACAAAATA
CCTGGTTACCCAGGCCGTGCCGGCACGTTAACCGGGCTGCATCCGATGCAAGTGTGTCGCTGTCGACGGC
CTCCTCACCCGGTCACGTTTCGTCGTTTCTCCTCCACGCGCTGCGGCTTCGGGGCCGCACCTGCATTCGT
ATGCGGTCGCCCGGTTACAGGTGCGGCACGGCCTGATGGAGGCCGCATGTGAGAGGAGAATTCCCATGCC
AAACTGGTGCTCGAATCGTATGTATTTTTCTGGTGAACCGGCACAGATCGCTGAGATTAAACGACTGGCC
AGCGGTGCAGTCACACCGCTTTATCGCCGCGCCACAAATGAAGGTATTCAGCTGTTTCTGGCCGGAAGTG
CCGGACTTCTGCAGACCACTGAAGATGTGCGGTTTGAACCATGCCCCGGACTGACGGCTGCTGGGCGTGG
CGTTGTATCGCCGGAGAATATCGCGTTCACCCGCTGGCTGACACACCTGCAGGACGGTGTACTTCTGGAT
GAGCAAAACTGCCTGATGCTGCATGAACTCTGGCTGCAGAGCGGTACTGGCCGGCGTCGCTGGGAAGAAT
TACCGGATGATGCCAGGGAAAGCATTACCGCTCTTTTCACCCCAAAAAGAGGTGACTGGTGCGACATCTG
GAGTAACGAGGATGTATCGGT
```

Priloga C: Analizirano nukleotidno zaporedje replikona RepFIB v obliki FASTA

```
>gi|8918823:33000-40000 Escherichia coli K-12 plasmid F DNA, complete sequence
TCTGATAATGATGAGCATTATGTAAGACAAACTACATTCCGAAGCAAAGTAATAAACCAGAATTACTATT
CTGTTGCAGTTAATGCCGGTTACTATATTACCCCAGAGGCAAAAGTGTACATCGAGGGTGTATGGAGTCG
TCTCACAAATAAAAAAGGGGATACATCTCTTTACGACCGTAGTGATAATACTTCGGAGCATAATAATAAC
GGGGCTGGAATTGAAAATTACAACTTCATTACGACGGCCGGTCTGAAGTACACGTTTTAACAACTTTAAC
TGTCATCATCCCCGGTGAGAATTATACCGCCGGGGATTCTGTCTGTTGGTCAGGGTAATAATTTACCAAT
GCAGGAGATGCAATGCGTGTGATAATGAAGCCTTTAAGAAGGACTTTAGTATTCTTTATCTTTTCTGTTT
TTCTTTGCGGAACTGTGTCTGCCAGACAGATCGAGTGGCAATCCTGTATGACCAGCCCATACAGTGACTG
GTTTGGTAAAGAAAGCTCTTCACCTGAGTTGTTATGCGGATATCTATCTGTACCGCTAAAATATACCGAC
ACAGGTAAGGATGTGTCAGATGAAAATATACCACTTGTCAGACTGGCGATGACAAAGTTGCCGGCAAAAA
GTAAACGTAAAGGGAGTGTTATTATAATAAGTGGCGGACCTGGATTACCCGGAATTAATCCTTATATCAA
TTTTGACTGGCCGGTCACAAATCTTCGGGAATCATGGGATATTATCGGTTTTGATCCACGAGGTGTCGGT
CAGTCCTTTCCTGCCATAAACTGCCAGCAATCCAATCAAGAAAGATTGGTGAATGTAAGTGAAAAGCAAC
TAATTTTACAGAAAATTAATGCCTGTATCCATAATACAGGAGCTGAAGTCATTCGCCATATTGGATCTCA
TGAGGCTGTATATGATATTGAACGCATCAGACAAGCTCTGGGAGATAAACAACTGACAGCTGTAGCATAT
TCATACGGTACACAGATTGCAGCCTTATATGCCGAACGCTTTCCATCCAGTATCAGATCTATAGTTTTTG
ACGGTGTCGTTGATATTGATGATTTAAATGATAATTTTTCATGGAAACTCAGGCAGGCCCACAGTTATCA
GGAAACGTTTGATCGTTTTGCTGCATGGTGTGCCAGGACAAAAAGTTGTCCACTTTCTTCAGACAGGAAT
CAGGCAATACATCAGTTCCATCAGCTTTTATTAAAATTACATAATTCTCCTTTAACAGACAGCAGGGGGG
AAAGTATATCTTCAGATGATCTCATATCATTAACGACAGAGCTTCTGTTATGGCGCTCATCATGGCCAAC
ACTTGCAACGGCCGTACGTCAGTTCTCTCAGGGTATTGTCAGCAATGAAATTGAAACTGCACTTAATTCT
TCCATAGCATCGGAAAAAGTCAGTGATGCTCTGGGTGTAATATTATGTGTTGATCAGAGTGATGAACAAT
TATCACAGGAACAACGAAAATCGCGAAAAAAAGCCCTTGCTGACGCATTTCCGGCCGTTAATTTTGAGAG
GGAACAGTCCGATTTACCTGAATTTTGTGAATTATGGCCAATTCATCGCGATCTGCAACAGACACGCCTG
AAGAATACTGTTTTACCATCCGGTTTACTGTTTGTGGCACACAAATATGATCCGACAACTCCCTGGATAA
ATGCCCGTAAAATGGCAGATAAATTTTCAGCTCCGTTACTGACTATTAACGGTGACGGGCATACGCTGGC
GCTTGCTGGCACTAATTTATGTGTGGATGAGGCTGTTGTTCGTCACCTCCTGTTTCCGGGGAAATCTGAA
GATATAACGTGTCAGGGAAGTGGTACCGGAGACACTAACTGATATTAAATAAAATATTACTCCCGACATT
AAATAACATCAGTGCTTTGATGTCAGGGCCTGCTAACCCTCTTTTGATATCAGTTGAGCGCAGGAAACGC
CAGTTAAGAGGGGACAGGACTTAGGATAAATAAGAGAGTGGCAATCATGAGTAAGATTCATGTACATGCA
GGTTACGGAAAGAGGATATCAGGCAATGAATAAGGCACTAATATAATATCATATCTCAGACTTGATACAT
TTTAGTTACATATATTTTCTTATTTTATGCGGAAAATGCTATATGGAAATGTAGTAATTATATACATCTT
ATCGAAAGTGATTTTCTGCATAATCATTATTATGATTTATATTAATCATGGGCAGGATTGCATATAAATC
AGTCATTATGCTGAGTCTGCAACCTTCCTGTGTAAAGTCAAACTGTCAGGGGGACGGTGACGTCCCGGGC
CAGAATACACGTGGTGCACCTTCTCCCTTTCTGTTCCCTTTCTCTGTTTTCTCCGATCCCATTCATCCTG
GCATTACCCCAATTCTCCCAAAAACGGCAAACCAGCCCTCAGCCCGCGCCACGACTGGCTTAATACGGGA
TCCCTGCTGCAAAACATACCTTTACGTATTTGGTTTTACCCGCAATAATGCGTGATAAGCAAACAAAAGG
ATCCGCGCGCTGCGATTTATTGTCGTGGAAGGATCCGGGCGTTCAGGTCAGAATACCTGTATCCATAAAC
AGTGCGTGTGCTACGTGAAAAATAACTCATGAACAATGTCATTCCCCTGCAGAATTCACCAGAACGCGTT
TCCCTGTTACCCATTGCGCCGGGGGTGGATTTTGCAACAGCGCTCTCCCTTAGAAGAATGGCCACTTCCA
CGGGGGCCACACCGGCCTACCTGCTGGCCCCGGAAGTGAGTGCCCTTCTCTTCTATATGCCGGATCAGCG
TCACCATATGCTGTTCGCCACCCTCTGGAATACCGGAATGCGTATTGGCGAAGCCCGGATGCTGACGCCG
GAATCATTTGACCTGGATGGAGTAAGACCGTTTGTGCGGATCCTGTCCGAAAAAGTGCGTGCGCGACGCG
GACGCCCGCCAAAAGATGAAGTCCGCCTGGTTCCGCTGACGGATATAAGCTATGTCAGGCAGATGGAAAG
```

se nadaljuje.

Nadaljevanje priloge C: Analizirano nukleotidno zaporedje replikona RepFIB v obliki FASTA

CTGGATGATCACCACCCGGCCCCGTCGTCGTGAACCATTATGGGCCGTGACCGACGAAACCATGCGCAAC
TGGCTGAAGCAGGCTGTCAGACGGGCCGAAGCTGACGGAGTACACTTTTCGATTCCGGTCACACCACACA
CCTTCCGGCACAGCTATATCATGCACATGCTCTATCACCGCCAGCCCCGGAAAGTCATCCAGGCACTGGC
TGGTCACAGGGATCCACGTTCGATGGAGGTTTATACCAGAGTGTTTGCACTGGATATGGCTGCCACGCTG
GCAGTGCCTTTCACAGGTGACGGTCGGGATGCTGCGGAGATCCTGCGTACACTGCCTCCCCTGAAGTGAC
GGCATACGCGCTTCCTTATATAAGCTGTGGTCAGCAGAACAGACATAACATAAGCTGGAGCAGGTAGATA
AGCTGTAGTGAGTAAACCATGTTTTATGAAGGGAGCAATGCCTCAGCATCAGGTTACGGGGTGACTCACG
TAAGGGACAGGCAGATGGCAGCTCAGCCACAGGCAGCACTGCAGGAAACTGAATATAAACAGCAGTGAGC
AGACCACTCACTGCACCTGATAATATAAGCTGTAGTCAGTAAAGGAGCACTCTTCACTGACTACAGTTTA
TGTTCAGCGGGATTTGAAGAGTTTTTCCAGGTCATCCAGTGTGATGCCCAGTTTTTCAAGCAGGGCAAGT
TTCTCCGCCATCTCCGGACTGACTTTTTCCGCAGGTGAAGGTGGCAACGGATTTTCCTTACTCTCATCAT
TCGGCGCTTTTAACCGGGGACGCCGGTAGTGAATACAGAAGAATTTTGTCCGCCCCCGCTGGATCTCCGT
GTAATCAAGATATCCAATCTCGCGCAACTGCTCCATTGCCCGTCTGACCGTCTGGTTCTGGGAAAATACA
GGAGACTTCAGATTGAGGCGTGCACGCAGCCGCGCCAGCGATATCGGTGCCGGATCCCGGGGCAGGCTCT
CTATAAAGGTGTAGAGTGCCTGGGCGGACTCCCGTCGCTTCAGGGCATTAATCGCCTTAAGCTGGAGAAG
GACTTTTCTGTCAAACTGGTACAGTTCAAACAGACGGGGATCAGCCTGTAACTGAACAATATCCCGCTCA
GTATCGTAGTAGGCTGACTGTACCAGATGGGTGATGTATTCCCGGGTGTGCTTCTCATCGGTACGGGAAA
ACGAGATCACGGTACCGGCAATGCGCTTCAGGGAAGGGCTGATGCGCTCACGCAGCCTGCGGGATGACTG
GCTTGAAGGTATACCACACAGTTTTGCAAACTCAACAAAAGGCAGTTCAACTTTGTCACCAATCACGTTA
TGGCGGGCAAAGGAATGAATGATCCCCACCCAGGTCTTGAAATCGTTATCCATATCCAGGCGGGGGCCGG
TGATCTCAACCTTATCGAATCCCTCAGCACGGGCCAGGGAAAGACGAGTCAGCTCTTCCGTGGCATCAGT
ACGTGACAGTGTATTTTTTTTACTGTTCTTCAGTGATTTAAGGGTCGGTACAAAAACGCCCAGACGCATC
AGCGCCACAGGTTGTACGGTGTTGTTATTGTTTGGTGTCAGTGTCACCAGCTCACCGGACGACTTATCCA
CCTCGCCGAACAGTTTTTTGATGTCTGAATTTTCGTTTTCCAGAAAGCCTTCTGCCTGTGGACAGTAAAA
GAATAAACAGAAAATCGTGTGTCAGAACCCCGTCAACTGACTGAAGAAAAACCTGCAGTCTGTATTTAAC
AACTGAGTCTGGCGTACTGACCATAGTTTATGTCATTAACTTACTTCAGTATATGCATTACTGAATACAG
CTTATCATGAACCTGTTACAGTTTATCCATAACATGCCACAGCTTATGTGAACGTGGTCCCGGCCCATGA
GTGACAAGGTTTGTAGCGATATGGGGATCTTTAATGGATCCTTTATTGATCATTCATGGATCTATTTACT
GGATCGGGCCTCTGGATATGTGGATAAATCATTCCGGAAAAACAGCATACACTACCGATTATGAAGGCTT
CAGGGCGTACCTGCTCACTGTGATTACAGATCCATTGCGCTTCCTCGCCGGTTCGCGGGCTTCGCCCGGT
CACCGGTCTGCGGCGAGCACCGTTTTTTCTGAAGTCTGTCTTTACAGGTAATTTTTGACGGAATCTGCCG
TACTTCCTGAGCGTACAGTCATATCCGAAATACACTTCATATGTTTGTTTCAACGCCAGAGGCGCATACA
GACTGGCTCCATGAATATCCTGATGACAGGCACTGCTGCGCAGACACCGGTCATGGATGTAACGGAAATA
CCAGCTGAACCACAGGTCTTCCAGTCGGGGACACTGAGTACAGGGCTGGAGCTTGTGGATACGTATCTCT
GGATCTTCAAACACTTTATGGAAGGAAAGGTGTTAACCAGATCCCTTGCCTGCCTGGTTTATATCAAGCC
CTATAGCGACAGGACGGACAGTGTCACCTTCCAGTCAGTAGCTAAACGATCTAAGGTGTTCCTGGACAAA
CTGCAGGAGCCAACGTCTGAAATGATAAATAAGCGAGGGAATATCTGGTGCTGGAAGAGTTGAGATGACT
GGCACATTGGTATAAATCTTGCCGTCATTCTGATCAGTTTGTAACATTCTGTAATGATCACCATTGGCTG
GCGATTTTTCTGTTCAGTAATGTAATTAACCTTATCTGATGCGCTGGCCACTATTCCATCAGCTGTACTG
ATGGCAGGCTCCCTGTTGAGGTCTGCCGCGAGCTGTCCGCCGGACTCAACGGGCTCTGCCACAATCCGGT
TATGATAGCCTGGAAACTGATCAAGGGTCAGGTCGATATAGTCCCTGGCAGGGGCGGCAATGACGTGGCC
GTATCTGCCAGGACCATTATCCCACTGCATACCGACAATACAGGAGTACAGGTTGCTGCCGTCTTCCGTC
AGATAATCAGTCAGCGTCTGGCTGGCCAGCTCGCTGCAACCGGCCGGGAAATCCTTCAGGTGATCATACC
CGGCGGTGAGAGAAGCCTCACAGGAAGTCAGCTCCAGAGCATGTCGGAAGGCCTGCAGGCGGGTCCAGGT
GTCCGCATTCACTTTTTCACCTTCCCGTCATGACGGGCCAGGGACTGAAGTTTACGGACGTCCTCCCGGG
ATCCTTTCAGCTCAGCTTTCGTATCCAGCAGCTGCGCCGCCAGCGGTCTTTTTCCCGGGAGCCGACTTTC
CTGGCACATATTCTTTTCGCAGTTATCCCCAGAATCTGTGGATAACTGTTGCTCCTCTTTTGAGAGAGAA
GGCTTTACTCACCGCCTGAATGCGGCAGATATGTCTGGACGATATATCTGCATACGCCGGTACTGCTTCG
TCAGATTGCACGGCAGGTGAATTGCCGGGCTGGTACCATACCATATTTGCGCCGAACCCGTCCGGGGCGG
TGCCGATAACAGATGTTAAATGATTGGTCGTTTTTACATCTGTAAAAGCAACAGGGCCCCATGATTATAT
CATGGGGCCCTGTTACTCTGATTCTTCGCAATAGATTGCTCTGAATTAGATGCGCCTTCTCCATTAGACT
AACAGTTCTATATTATTTTACCATCTTTAAATGTGTAAGAGTTTTCTGTTTTTTATTGGTGTTTATCTGT
ATTTTTTGTTTTAGGAACGTTTTTGTAAGAATTATCTTTTATATTTGAGGTGATATCTTATTGTGGTTGT
TGTAATAGTTTAGAGTCGTAATAAATATGATTGATCGCCACTTTAAATGAAGTTAATCTACTGGAGATGA
AGTAACTTCATTAATGCGTCATAAAAATAAAAATTCCTGTTTTGGGAAATATGAATGTGTCTGTCCCCTC
CTTCGGGAGGGTGTTTTTTATCTGGAAACTGGTTTCTGGTTCTGATTTTTCTTTTTGTATTATAGCGAGT
TACGCATATTTACCAATCGATGTCTGGCAAACAGAGTATGGCGATGGTGTGATTTGTGTTTATTGGTAAA
A

Priloga D: Analizirano nukleotidno zaporedje replikona RepFIC v obliki FASTA

```
>gi|1621020|gb|M16167.1|P30REPFIC Escherichia coli Ent plasmid P307 basic replicon REPFIC, copB and repA1 genes, complete
cds
CCCGGGGGGCCGGCTGTTTCTGGCATTTCCAGCTTTTCAGTCATTTCATTATCAAAATCACATTAAACGG
TCGTAATCAGACATGATTTGTGCGCCAACACAGATCATTGTCACAATTCTCAAGTCGCTGATTTCAAAAA
ACTGTAGTATCCTCTGCGAAACGATCCCTGTTTAAGTATTGAGGAGGCGAGATGTCGCAGACAGAAAATG
CAGTGACTTCCTCATCAGGTAACAAGCGTGCATACCGGAAAGGTAACCCTGTTCCGGCCAGAGAGAGACA
AAGGGCTTCTCTAGCTCGCAGAAGCAACACTCATAAGGCTTTTCATGCGGTTATCCAGGCCCGGTTAAAA
GACAGGCTGAGTGAACTGGCAGATGAGGAAGGTATTACCCAGGCGCAGATGCTTGAAAAACTGATTGAAT
CAGAGCTGAAACGTAGAGCGACTTTGTAAATATTCACATTCTTGCTTATCTCAGGCGTGAGTGATAGATT
GCTGATCGTTTAAGGAATTTTGTGGCTGGCCACGCCATAAGGTGGCAGGGAACTGGTTCTGATGTGGATT
TACAGGAGCCAGAAAAGCAAAAAACCCGATAATCTTCATCTAGTTTGGCGACGAGGAGAAGATTACCGGG
GTCCACTTAAACCGTATAGCCAACAATTCAGCTATGCGGGGAGTATAGTTATATGCCCGGAAAAGTTCAA
GACTTCTTTCTGTGCTCACTCCACCTGCGCATTGTAAGTGCAGGATGGTGTGGCTGAAAGATACATCTCA
CAAAGACACTGGAGTCAGCTTCCTCCCGAAGAGCAAATCCGTGTCTGGGAAGACTATGAAGCGGGAAGGG
CGACCACTTTCCTGGTTGAACCGGAAAGGAAGCGCACAAAGCGTCGTCGTGGTGAGCACTCCACTAAACC
CAAATGCGAAAATCCGACCTGGTATCGTCCTGCGCGCTATAAGGCGCTGAGCGGGCAGCTCGGGCACGCC
TATAACCGTCTGGTGAAAAAGGACCCGGTGACCGGCGAACAGAGCCTGCGCATGCACATGTCGCGACATC
CTTTTTACGTGCAGAAACGGACGTTCGCTGGCCGTAAATATGCTTTCCGTCCGGAAAAACAACGCCTCCT
CGATGCTGTCTGGCCGGTTCTGGTCAGTTTTAGTGATGCCGGCACACACACAGTGGGGATGAGTGTTTCC
CGCCTGGCTAAAGAAATCAGCCCGAAAGACAGCAAAGGAAAGGTTATCCCCGAACTCGAGGTGACGGTCT
CCCGGCTTTCCCGTCTGCTGGCAGAGCAGGTGCGTTTTGGTGTGCTGGGTATGTCGGAAGAAACAATGTG
GGACCGTGAAACCCGCCAGCGTCTGCCACGCTATGTCTGGATAACACCGGCAGGGTGGCAGATGCTGGGC
GTCGACATGGTTAAACTTCACGAACAGCAGCAGAAGCGGCTGCGTGAAAGTGAAATCCGCCAGCAGCTCA
TCCGGGAAGGCGTACTGCGTGAGGATGAAGATATCTCCGTACATGCGGCCAGAAAACGCTGGTATCTGCA
GCGCAGCCAGGATGCGCTGAAACACCGTCGGGCGAAAGCTGCAGCCCGCAAGCGTGCTAATCTTCTGAAG
AAGCTACCTGCCGACCAGCAGATTTATGAGATGTCACAGCATATCCTGAAGCGTATGCCGCCGGATGAAG
CCTACTGGTGCACGCCGGAACGCCTGCAGCAACTGGCCATCAGGGAGCTGTATCAGCTTGAACTGACGCT
GGCTGCACCGCCACCGCACTAGACAGCACCATTTCCTCAGCACTGAATCATCGCCAGCCCCTCCGGGGCT
TTCGGCGCTGGTTCCGTTCGACCAGAAACTCCCCGTAACCACCTGAAATATCCTCATCTGGCCATATCTG
GCCACAAAGTCACTCCCCTGTCGTCAGAATGTGGCCACGTCGTTTCAGTTATCCACATAAATCCGCAAAT
AAAGAGTTTTAAGAAGCTGCAAACCAAAAACAGCAAACCTGCAATATAGTCTTACCCCAGTTACTTAATC
CCCTGCGTTGCTTCGCCTCAGGGAAAGTCTTTATCTCTGAAACGCCTATGAACAAGTACAAAGAGGCCTT
CGCTTGCAGGCGGCCAAAGCCGCGCCGCTCAGAATTTAAAAGTACCTCCCACCGCAAGCGGCGGGCCCCG
ACCGGAGCCATTTTAGTTACAACACTCAAATGCGACCACCAAGAAAAACCTAGTCCCGTGCAGAACTGAA
ACCACAAAGCCCCCCCCCTCATAACTTAAAAGCGCCCCCGCCCGGCCCAAAGGGCCGGAACAGAGTCGCT
TTTAATTATGAATGTTGTAACTACATTATCATCGCTGTCAGTCTTCTGGCTGGACGTACTGAGTACACGC
TCGTAAGCGGCCCTGACGGCCCGCTAACGCGGAGATACGCCCCGACTGCGGGTAAACCCTTGTCGGGACC
ACTCCGACCGCGCACAGAAGCTCTGTCATGGCTGAAAGCGGGTATAGCTTAGCAGGACCGGGATGAGTAA
GGTGAAATCTATCAATACGTACCGGCTTACGCCGGGCTTCGGCGGTTTTACTCCGGTATAATATGAAACA
ACAAAGTGCCGCCTTACATGCCGCTGGCGCGGCATATCTTGGTGACAATATCTGAATCGTTATATACTGC
GTATATACGTAGTAATGACGAGGTGATAAATGGCACAGGTTAATATGAGTTTAAGAATCGACGCTGAACC
TGAAGGATGCTTTATGGCTGCTGCAAAAAGCATGGACCGTAATGGCTCTCAGTTAATCCGGGATTTATGC
GCAGACCGTGACGCAGCATATACCGGTCCGTGACCAGGTGCGGCAGGCGCAGCAGCACTCA
```

Priloga E: Analizirano nukleotidno zaporedje replikona RepFIIA v obliki FASTA

```
>gi|5103148:88200-90500 Shigella flexneri 2b plasmid R100 DNA, complete sequence
AAACTGTAGTATCCTCTGCGAAACGATCCCTGTTTGAGTATTGAGGAGGCGAGATGTCGCAGACAGAAAA
TGCAGTGACTTCCTCATCTGGCGCAAAACGAGCATACAGAAAGGGGAATCCGCTTTCTGATGCAGAGAAA
CAAAGATTATCAGTGGCCCGTAAAAGAGCTTCGTTCAAGGAAGTAAAAGTATTTCTTGAACCAAAGTATA
AGGCCATGCTCATGCAAATGTGTCATGAAGATGGTCTGACTCAGGCTGAAGTTCTGACCGCACTGATAAA
AAGTGAAGCGCAAAAACGATGCATGTGATGATGGGCTTACATTCTTGAGTGTTCAGAAGATTAGTGCTAG
ATTACTGATCGTTTAAGGAATTTTGTGGCTGGCCACGCCGTAAGGTGGCAAGGAACTGGTTCTGATGTGG
ATTTACAGGAGCCAGAAAAGCAAAAACCCCGATAATCTTCTTCAACTTTGGCGAGTACGAAAAGATTACC
GGGGCCCACTTAAACCGTATAGCCAACAATTCAGCTATGCGGGGAGTATAGTTATATGCCCGGAAAAGTT
CAAGACTTCTTTCTGTGCTCGCTCCTTCTGCGCATTGTAAGTGCAGGATGGTGTGACTGATCTTCACCAA
ACGTATTACCGCCAGGTAAAGAACCCGAATCCGGTGTTTACACCCCGTGAAGGTGCAGGAACGCTGAAGT
TCTGCGAAAAACTGATGGAAAAGGCGGTGGGCTTCACTTCCCGTTTTGATTTCGCCATTCATGTGGCGCA
TGCCCGTTCGCGTGGTCTGCGTCGACGCATGCCACCAGTGCTGCGTCGACGGGCTATTGATGCGCTCCTG
CAGGGGCTGTGTTTCCACTATGACCCGCTGGCCAACCGCGTCCAGTGCTCCATCACCACGCTGGCCATTG
AGTGCGGACTGGCGACGGAGTCTGCTGCCGGAAAACTCTCCATCACCCGTGCCACCCGGGCCCTGACGTT
CCTGTCAGAGCTGGGACTGATTACCTACCAGACGGAATATGACCCGCTTATCGGGTGCTACATTCCGACC
GATATCACGTTCACATCTGCACTGTTTGCTGCCCTCGATGTATCAGAGGAGGCAGTGGCCGCCGCGCGCC
GCAGCCGTGTGGTATGGGAAAACAAACAACGCAAAAAGCAGGGGCTGGATACCCTGGGCATGGATGAACT
GATAGCGAAAGCCTGGCGTTTTGTTCGTGAGCGTTTTCGCAGTTATCAGACAGAGCTTAAGTCCCGGGGA
ATAAAGCGTGCCCGTGCGCGTCGTGATGCGGACAGGGAACGTCAGGATATTGTCACCCTGGTGAAACGGC
AGCTGACGCGCGAAATCGCGGAAGGGCGCTTCACTGCCAATCGTGAGGCGGTAAAACGCGAAGTTGAGCG
TCGTGTGAAGGAGCGCATGATTCTGTCACGTAACCGTAATTACAGCCGGCTGGCCACAGCTTCCCCCTGA
AAGTGACCTCCTCTGAATAATCCGGCCTGCGCCGGAGGCTTCCGCACGTCTGAAGCCCGACAGCGCACAA
AAAATCAGCACCACATACAAAAAACAACCTCATCATCCAGCTTCTGGTGCATCCGGCCCCCCCTGTTTTC
GATACAAAACACGCCTCACAGACGGGGAATTTTGCTTATCCACATTAAACTGCAAGGGACTTCCCCATAA
GGTTACAACCGTTCATGTCATAAAGCGCCATCCGCCAGCGTTACAGGGTGCAATGTATCTTTTAAACACC
TGTTTATATCTCCTTTAAACTACTTAATTACATTCATTTAAAAAGAAAACCTATTCACTGCCTGTCCTGT
GGACAGACAGATATGCACCTCCCACCGCAAGCGGCGGGCCCCTACCGGAGCCGCTTTAGTTACAACACTC
AGACACAACCACCAGAAAAACCCCGGTCCAGCGCAGAACTGAAACCACAAAGCCCCTCCCTCATAACTGA
AAAGCGGCCCCGCCCCGGCCCGAAGGGCCGGAACAGAGTCGCTTTTAATTATGAATGTTGTAACTACTTC
ATCATCGCTGTCAGTCTTCTCGCTGGAAGTTCTCAGTACACGCTCGTAAGCGGCCCTGACGGCCCGCTAA
CGCGGAGATACGCCCCGACTTCGGGTAAACCCTCGTCGGGACCACTCCGACCGCGCACAGAAGCTCTCTC
ATGGCTGAAAGCGGGTATGGTCTGGCAGGGCTGGGGATGGGTAAGGTGAAATCTATCAATCAGTACCGGC
TTACGCCGGGCTTCGGCGGTTTTACTCCTGTATCATATGAAACAACAGAGTGCCGCCTTCC
```

Priloga F: Analizirano nukleotidno zaporedje replikona RepFIII v obliki FASTA

```
>gi|341551|gb|M26937.1|P36REPA Plasmid pSU316 (from Escherichia coli) replication protein (repA1 and repA2) genes,
complete cds
GATCTTCGTCACAATTCTCAAGTCGCTGATTTCAAAAAACTGTAGTATCCTCTGCGAAACGATCCCTGTT
TGAGTATTGAGGAGGCGAGATGTCGCAGACAGAAAATGCAGTGACTTCCTCATTGAGTCAAAAGCGGTTT
GTGCGCAGAGGTAAGCCTATGACTGACTCTGAGAAACAAATGGCCGCTGTTGCAAGAAAACGTCTTACAC
ACAAAGAGATAAAAGTTTTTGTCAAAAATCCTCTGAAAGATCTCATGGTTGAGTACTGCGAGAGAGAGGG
GATAACACAGGCTCAGTTCGTTGAGAAAATCATCAAAGATGAACTGCAGAGACTGGATATACTAAAGTAA
AGACTTTACTTTGTGGCGTAGCATGCTAGATTACTGATCGTTTAAGGAATTTTATGGCTGGCCACGCCAT
AAGGTGGCAGGGAACTGGTTCTGATGTGGATTTACAGGAGCCAGAAAAGTGAAAACCCCGATAATCTTCT
TTAACTTTGGCGAGTGAGAAAGATTATCGGGGCTAACAAGAAACTGCATAGAAGCGGTTGCTCTATGCGG
GGAGTATAGTTATATGCCCGGAAAAGTTCAAGACTTCTTTCTGTGCTCACTCCTTCTGTGCAACATAAGT
GCAGGATGGTGTGACTGATCTTCAACAAACGTATTACCGCCAGGTAAAGAACCCGAATCCGGTGTTCACT
CCCCGTGAAGGTGCCGGAACGCTGAAGTTCTGCGAAAAACTGATGGAAAAGGCGGTGGGCTTCACCTCCC
GTTTTGATTTCGCCATTCATGTGGCGCATGCCCGTTCCCGTGGTCTGCGTCGGCGCATGCCACCGGTGCT
GCGTCGACGGGCTATTGATGCGCTGCTGCAGGGGCTGTGTTTCCACTATGACCCGCTGGCCAACCGCGTC
CAGTGTTCCATCACCACACTGGCCATTGAGTGCGGACTGGCGACAGAGTCCGGTGCAGGAAAACTCTCCA
TCACCCGTGCCACCCGGGCCCTGACGTTCCTGTCAGAGCTGGGACTGATTACCTACCAGACGGAATATGA
CCCGCTTATCGGGTGCTACATTCCGACCGACATCACGTTCACACCGGCTCTGTTTGCTGCCCTTGATGTG
TCTGAGGATGCAGTGGCAGCTGCGCGCCGCAGTCGTGTTGAATGGGAAAACAAACAGCGCAAAAAGCAGG
GGCTGGATACCCTGGGTATGGATGAGCTGATAGCGAAAGCCTGGCGTTTTGTGCGTGAGCGTTTCCGCTG
TTACCAGACAGAGCTTAAGTCCCGTGGAATAAAACGTGCCCGTGCGCGTCGTGATGCGAACAGGGAACGT
CAGGATATCGTCACCCTGGTGAAACGGCAGCTGACGCGTGAAATCTCGGAAGGGCGCTTTACTGCTAATG
GTGAGACGGTAAAACGCGAAGTGGAGCGTCGTGTGAAGGAGCGCATGATTCTGTCACGTAACCGCAATTA
CAGCCGGCTGGCCACAGCTTCCCCCTGAAACTGACCTCCTCTGAATAATCCGGCCTGCGCCGGAGGCTTC
CGCACGTCTGAAGCCCGACAGCGCACAAAAAATCAGCACCACATACAAAAAACAACCTCATCATCCAGCT
TCTGGTGCATCCGGCCCCCCCTGTTTTCGATACAAAACACGCCTCACAGACGGGGAATTTTGCTTATCCA
CATTAAACTGCAAGGGACTTCCCCATAAGGTTGCAACCGTTCATGTCATAAAGCGCCAGCCGCCAGTCTT
ACAGGGTGCAATGTATCTTTTAAACACCTGTTTATATCTCCTTTAAACTACTTAATTACATTCATTTAAA
AAGAAAACCTATTCATTGCCTGTCCTGTGGACAGACAGATATGCACCTCCCACCGCAAGCGGCGGGCCCC
GACCGGAGCCACTTTAGTTACAACACACAAAAACAACCTCCAGAAAAACCCCGGTCCAGCGCAGAACCGA
AACCACAAAGCCCCTCCCTCATAACTGAAAAGCGGCCCCGCCCCGGCCCTTCGGGCCGGAACAGAGTCGC
TTTTAATTATGAATGTTGTAACTACATCATCATCGCTGCCAGTCTTCTCGCTGGAAGTTCTCAGTACACG
CTCGTAAGCGGCCCTGACGGCCCGCTAACGCGGAGATACGCCCCGACTACGGGTAAACCCTTGTCGGGAC
CACTCCGACCGCGCACAGAAGCTCTATCATGACTGAAAGCGGGTATGCCTTAACAGGGATGGGAATGGGA
TAGGCGAAATCTATCAATCAGTACCGGCTTACGCCGGGCTTCGGCGGTTTTACTCCAGTATCATATGAAA
CAACAGAGTGCCGCCTTCCATGCCGCTGCCGCCGCT
```

Priloga G: Analizirano nukleotidno zaporedje replikona RepFIV v obliki FASTA

```
>gi|545626530:1-1272&43607-43794 Pseudomonas fluorescens R124 plasmid pMP-R124, complete sequence
ATGGAATTGACTTCAAAAGCTATAGGCACGTCCGGTCAGGACGGTAGAGATCGCCGAAATAGCACGTCTG
TAGACAGTTTGTACTCACCCTCCGGGAGATTTTTCGAGGACAACACCGCCCTAAGCCGCGTGTTGCAAGA
AGCTCCCTACCTCGCCCGTTGCAGTGACGACAAAACCGCTACGCGGATCCGTCCTCGCGAATACGCCATC
CGCTATCCCTATATGCAGATCAACCGCCCGGGAATGGTGTCATGGTTGATCTTTGACCTCGATCACACCA
ACTCGCTGATCTGGAATGACGAGGATCTGCCGGCACCGAACCTGGTTGTTCGCAATCGGGCAAACGGTCA
TTCCCACCTCTACTATGCGATCCCCTCTGTTTGCACTACAGAGAATGCGAGGGCCAAGCCGATCCAGTAC
ATGAAGGCCGTCTACGCGGCCTTTGCGGCTCGCCTGAACGCGGACAAGGAATACAACAGCGGCCCTGTGG
CAAAAACACCTGGTCATCCATGGTGGTCGACCACGGAACTACACAACCACGTCTACGAACTCGGCGAACT
AGCGGATTGCGTCGATCTTGCAGTCTCTCCCTGGGGGAAAGGCCCGCAGATCGATGACCTGGGCCACTCT
CGGCACTGCATCCTTTTCGAACAACTGCGGTTCTTCGCCTACAGCATCGTCAACCGGGAACGCGAGCGTG
GAACATTCACCGCATTCACTCGTTCCTTGGAAGCTTTTGCGCACAACAACAACAACTTCCACAAGCACGG
ATTCTCCGAAAACCTGCCGCTCTCCTCGCTCAAGGCCACGGTGAAGTCGGTTGCCAGGTGGACGTGGGAT
CGCTACAGCGGCGATAGCCGGTGTCATCGAGGCGCAATGCAGCTGGACAAGGATTTGCCACTGGTGGAAC
GCCAGAGCCTTGCCGCGAAGCGTACCCATGAGATCCGACACAAGGCGACAGAATCCAAAATCCGCTCCGC
ATGCCGGCAACTGCAGGCACGTGGCGAGAAACTTCGACAAGCTGCGATCGCTGCGATCGCTGGAGTTACC
CGTCAAACCGTAGCGACCTACGCACACGTGCTTAAAGAAGCTGCGCGGCCAGCAGTTGTCGCGGTTCTTA
ACGTCCCTCAAGCGCCTGATCGAGCGCCAATCACCGGCGCGATCGGCCCTCTTCCGGAGGCGTCCGCCGG
CGCCGCTGACGATGTTAATAATGGTGTACATCAGGTACCTGCGGCACCGCGGGGTCTTGGGGAAGATGAA
AACTCTTCATAG
-
CGGAATCGTAGAAACCAAAAAGCCCGGCTGGTAACCGGGCTTTTTGGAAAATCAGAACAGGTCTTTCTCT
TTCGACGGTGAAAACCTGCTCTACACATCAATTGGTGAGCCGATTATGGCGCGCACTGCGGCTTCGTGCA
AGTCGCAGAATGTCAACATAATCAGCTTTCCGCCGTAAGGCGTTGAAA
```

Priloga H: Analizirano nukleotidno zaporedje replikona RepFVI v obliki FASTA

```
>gi|42544|emb|X55895.1| E.coli plasmid pSU212 copB, repA genes and inc determinant
GATCTTCGTCACAATTCTCAAGTCGCTGATTTCAAAAAACTGTAGTATCCTCTGCGAAACGATCCCTGTT
TGAGTATTGAGGAGGCGAGATGTCGCAGACAGAAAATGCAGTGACTTCCTCATCTGGCGCAAAACGAGCA
TACAGAAAGGGGAATCCGCTTTCTGATGCAGAGAAACAAAGATTATCAGTGGCCCGTAAAAGAGCTTCGT
TCAAGGAAGTAAAAGTATTTCTTGAACCAAAGTATAAGGCCATGCTCATGCAAATGTGTCATGAAGATGG
CCTGACTCAGGCTGAAGTTCTGACCGCACTGATAAAAAGTGAAGCGCAAAAACGATGCATGTGATGATGG
GCTTACATTCTTGAGTGTTCAGAAGATTAGTGCTAGATTACTGATCGTTTAAGGAATTTTGTGGCTGGCC
ACGCCGTAAGGTGGCAAGGAACTGGTTCTGATGTGGATGTACAGGAGCCAGAAAAGCGAAAACCCCGATA
ATCTTCTCTAACTTTGGCGAGTGCAGAAAGATTACCGGGGCCATCTAAAAACCGTATAGCCAACAATTCA
GCTATGCGGGGAGTATAGTTATATGCCCGGAAAAGTTCAAGACTTCTTTCTGTGCTCGCTCCTTCTGCGC
ATTGTAAGTGCAGGATGGTGTGACTGATCTTCAACAAACGTATTACCGCCAGGTAAAGAACCCGAATCCG
GTGTTTACACCCCGTAAAGGTGCCGGAACGCTGAAGTTCTGCGAAAAACTGATGGAAAAGGCGGTGGGCT
TCACCTCCCGTTTTGATTTCGCCATTCATGTGGCGCATGCCCGTTCGCGTGGTCTGCGTCGA
```

Priloga I: Analizirano nukleotidno zaporedje inkompatibilnostne determinante *incFVII* v obliki FASTA

```
>gi|341827|gb|M28097.1|P36INC Plasmid pSU316 (from Escherichia coli) incFVII gene
GATCGTTTAAGGAATTTTATGGCTGGCCACGCCATAAGGTGGCAGGGAACTGGTTCTGATGTGGATTTAC
AGGAGCCAGAAAAGTGAAAACCCCGATAATCTTCTTTAACTTTGGCGAGTGAGAAAGATTATCGGGGCTA
ACAAGAAACTGCATAGCAAGCGGTTGCTCTATGCGGGGAGTATAGTTATATGCCCGGAAAAGTTCAAGAC
TTCTTTCTGTGCTCACTCCTTCTGTGCAACATAAGTGCAGGATGGTGTGACT
```

Priloga J: Nukleotidno zaporedje gena *ehxA* za enterohemolizin, deponirano pod identifikatorjem 3654480 v obliki FASTA

```
>gi|73853177:c117132-114136 Escherichia coli EH41 plasmid pO113, complete sequence
ATGACAGTAAATAAAATAAAGAACATTTTCAACAATGCGACATCGACTACAAAATCAGCATTTAATACAG
CATCATCAAGCGTACGTTCCGCTGGAAAAAAACTCATATTATTAATACCTGATAATTATGAAGCTCAGGG
CGTGGGTATTAATGAGTTGGTCAAAGCTGCTGATGAGCTTGGAATAGAAATACACCGTACTGAACGAGAT
GATACAGCGATTGCAAACCAGTTTTTTGGTACAGCAGAAAAAGTTGTAGGATTAACTGAACGTGGTGTTG
CAATATTCGCACCACAACTTGACAAACTTCTGCAGAAGTATCAGAAAGTTGGGAGTAAAATAGGCGGAAC
TGCTGAAAATGTAGGTAATAATCTGGGAAAAGCCGGAACAGTTCTCTCAGCACTACAGAATTTTACGGGG
ATTGCTTTATCAGGCATGGCTCTTGATGAATTGCTGAGAAAACAACGGGAAGGAGAGGATATAAGTCAGA
ATGATATTGCCAAAAGTAGTATTGAACTTATTAATCAGCTTGTAGATACAGTATCAAGTATAAACAGTAC
CGTTGATTCATTTTCTGAGCAGCTTAACCAGCTTGGCTCATTTTTATCCAGTAAACCTCGCTTAAGTTCT
GTTGGTGGAAAATTACAAAATTTACCAGACCTGGGCCCCCTGGGGGATGGGCTGGATGTTGTCTCCGGAA
TTCTTTCTGCTGTATCAGCAAGTTTTATTCTGGGAAACAGTGACGCACATACAGGAACAAAAGCTGCAGC
GGGTATCGAACTGACAACTCAGGTTCTTGGAAATGTTGGTAAAGCTGTTTCGCAATATATTCTGGCTCAG
AGAATGGCACAGGGATTATCGACAACAGCTGCAAGTGCGGGTCTGATCACATCGGCTGTTATGCTGGCTA
TCAGTCCTCTTTCTTTCCTGGCTATTGCAGATAAATTTGAGCGAGCTAAGCAGCTTGAATCATATTCTGA
ACGATTTAAAAAATTCAATTATGAAGGAGATGCTTTACTCGCAGCCTTTCATAAAGAAAGCGGAGCTATA
GATGCAGCCCTGACAACAATAAATACTGTCCTGAGTTCTGTATCTGCGGGAGTTAGTGCAGCCTCCAGTG
CATCCCTCATAGGGGCCCCGATAAGCATGCTGGTGAGTGCATTAACCGGTACGATATCTGGCATTCTGGA
AGCATCAAAACAGGCTATGTTTGAGCACGTTGCAGATAAATTCGCTGCTCGGATCAATGAATGGGAAAAG
GAGCATGGCAAAAATTATTTTGAGAATGGCTATGACGCAAGACATGCTGCGTTTTTAGAAGACTCTCTGT
CTTTGCTTGCTGATTTTTCTCGTCAGCATGCAGTAGAAAGAGCTGTCGCAATAACCCAGCAACATTGGGA
TGAGAAGATCGGTGAACTTGCAGGTATAACCCGTAATGCTGATCGCAGTCAGAGTGGTAAGGCATATATT
AATTATCTGGAGAATGGAGGGCTTTTAGAGGCTCAACCGAAGGAGTTTACACAACAAGTTTTTGATCCTC
AAAAAGGGACTATAGACCTTTCAACAGGTAATGTATCAAGTGTTTTGACATTTGTAACACCAACATTTAC
CCCAGGAGAAGAAGTCAGAGAAAGAAAACAGAGTGGTAAATATGAATATATGACATCTCTTATTGTAAAT
GGTAAGGATACATGGTCTGTAAAAGGCATAAAAAATCATAAAGGTGTATATGATTATTCAAATTTGATTC
AGTTTGTTGAAAAGGATAACAAACACTATCAGGCGAGAATAATTTCTGAGCTCGGAGATAAAGACGATAT
AGTTTATTCTGGGGCAGGCTCATCAGAAGTATTTGCTGGTGAAGGTCATGATACCGTATCTTATAATAAG
ACGGATGTTGGTAAACTAACAATTGATGCAACAGGAGCATCAAAACCTGGTGAATATATAGTTTCAAAAA
ATATGTATGGTGACGTGAAGGTATTGCAGGAAGTCGTTAAGGAACAGGAGGTGTCAGTAGGGAAGAGAAC
AGAGAAAATACAATATCGTGATTTTGAATTCAGAACCGGTGGAATTCCTTATGATGTGATAGATAATCTT
CATTCTGTTGAAGAACTCATTGGCGGAAAACATGATGATGAATTCAAAGGCGGTAAGTTTAATGATATAT
TCCATGGTGCAGATGGGAACGATTATATCGAAGGTAATTATGGTAATGATCGACTATACGGCGATGATGG
GGATGATTATATATCCGGTGGACAGGGAGACGACCAGTTATTTGGTGGTAGTGGAAACGATAAATTGAGT
GGAGGGGATGGTAATAATTATCTGACAGGAGGACGCGGTAATGATGAGCTTCAGGCACACGGAGCTTATA
ATATTCTGTCAGGTGGTACTGGTGATGATAAACTTTATGGTGGTGGTGGTATTGACCTTCTGGATGGAGG
GGAAGGTAATGACTATCTGAATGGTGGTTTTGGTAATGATATTTATGTTTATAGGCAAAACTATGGTCAT
CATACAATTGCAGATGAAGGAGGTAAAGGAGATCGTCTGCACTTATCTGATATTAGCTTTGATGATATCG
CATTTAAGAAAGTTGGAAATGATCTTATCATGAATAAAGCCATTAATGGTGCACTTTCATTTAATGAGTC
AAATGATGTCAATGGGATAACATTTAAAAACTGGTTTGCGAAAGATGCCTCAGGTGAAGATAATCATCTT
GTTGAGGTTATAACAGATAAAGATGGTCGTGAGATAAAAGCTGATAAGATATCTCATAATAATAATGAAC
AGTCAGGTTATATAAAAGCCAGTAATATAGCATCTGAAAAAAACATGGTTAATATCACTAGTGTTGCCAA
TGATATTAACAAGATTATTTCTTCAGTTTCAGGGTTCGATTCAGGTGATGAACGATTAGCATCTTTATAT
AATTTATCCTTACATCAAAATAACACACACTCAACAACTTTAACGACAACTGTCTGA
```